GBS760: BACTERIAL GENETICS & PHYSIOLOGY COURSE PACKET
University of Alabama at Birmingham
Graduate Biomedical Sciences, Microbiology Theme
January / February 2021

Michael J. Gray, M.S., Ph.D.
email: mjgray@uab.edu
twitter: @graymicrolab
website: graymicrolab.com, lab GitHub repository

TABLE OF CONTENTS

LECTURE 1: INTRODUCTION TO MOLECULAR MICROBIOLOGY

INTRODUCTION

The goal of this course packet is to familiarize you with the nomenclature and concepts you will need to participate in each lecture. **Your level of participation in in-class discussions will be a significant determinant of your grade**, and most of the "lecture" time will be dedicated either to small-group problem solving and discussions based on the information and problem sets in each day's reading or to journal club-style discussions of specific scientific papers, so I strongly recommend that you do the reading for each day's class ahead of time. I will begin each lecture with a short question and answer session, so if there's anything in the reading you don't understand or would like clarified, please come prepared to ask. We will then work through the discussion problems during the "lecture" itself. I do not plan to spend very much time actually lecturing at you.

In this first lecture, we will discuss some core principles of scientific literacy, including the basics of the scientific method and using the scientific literature. I will also introduce the basics of molecular microbiology, including the fundamentals of genetic nomenclature in bacteria. This will set the foundation for future lectures, in which we will explore the practical and theoretical implementation of the scientific method for experiments in microbial genetics.

EXPECTATIONS AND LEARNING GOALS

In this course, my goal is for you to learn how to think about and apply the tools of bacterial molecular genetics to solve scientific problems, and then to use that knowledge to build a strong foundation of understanding the molecular mechanisms by which bacteria grow, function, and cause disease. To achieve this goal, we will need to build your skills in two fundamental areas:

- **Scientific literacy**: understanding what is and isn't known, how those facts fit into the larger framework of scientific knowledge, and how to search and read the scientific literature
- **Scientific proficiency**: understanding how to design, carry out, and interpret experiments, knowing what tools you have available and creatively applying those tools to answer specific questions

The product of scientific work is knowledge. I want to give you the tools to effectively access and add to that knowledge.

By the end of this course, I want you to:

- be able to define the steps of the scientific method and develop models and hypotheses based on data
- know where to find information about bacterial genes and proteins
- be able to use and understand the nomenclature of bacterial genetics
- understand the principles of mutagenesis and genetic engineering in microbes
- know how to interpret mutant phenotypes in different kinds of genes and with different kinds of mutations
- have a good working knowledge of bacterial physiology and cell biology
- be able to design rigorous experiments to solve biological problems using bacterial genetics

A glossary of important terms, which are indicated in the text in *italics* the first time they appear, can be found starting on page 198. See page 196 for a concise summary of all of the experimental design principles that we will discuss in the course of these chapters.

Class participation will be evaluated using the following scale and will account for slightly less than half of your grade (76 points out of 156 total):

4 points
Student comes to class prepared; contributes readily to the conversation but doesn't dominate it; makes thoughtful contributions that advance the conversation; shows an interest in and respect for others' contributions; participates actively in all groups.

3 points
Comes to class prepared and makes thoughtful comments when called upon; contributes occasionally without prompting; shows interest in and respect for other's views; participates in small groups.

2 points

Student is poorly prepared or participates in discussion, but in a problematic way: e.g. talks too much, rambles, interrupts instructor and others, or does not acknowledge cues of annoyance from others.

1 point
Has not prepared for class or does not contribute to discussion; displays disrespect towards students and/or faculty.

0 points
Absent without explanation.

There will also be weekly quizzes, which will account for the rest of your grade. These will be made available each Friday morning on the course Canvas page, and are due the following **Monday morning** at **8 AM**. Each quiz is worth 20 points, accounting for 80 points out of 156 total. You will lose 5 points per day that the quiz is late without explanation. Please do not work with other students on the quizzes: they are meant to assess your personal progress. The rubric I will use to grade the quizzes is on page 197.

If you are concerned about your grade or class status at any point during the class, please contact me immediately. I am happy to talk to you outside of class to try to clear up any confusing points. Dr. Yother also has tutors available, if you feel that you need extra help.

It is worth noting, of course, that this course is structured to emphasize the topics and concepts **I** think are important, which are not necessarily the same as another instructor might focus on. Those people are not necessarily wrong, and I am certainly not always right, so I will try not to be too dogmatic in this course. A diversity of opinions and approaches is one of the great strengths of science, and you should draw on all of the resources, mentors, and instructional material available to form your own personal scientific knowledgebase and approach to research.

SCIENTIFIC LITERACY

It has been a very long time since it has been possible for any one person to know everything there is to know about science. What I mean by being scientifically literate has three distinct elements:

- Understanding what scientific knowledge is and is not, and understanding the scientific method.
- Having a good general grasp of the broad state of knowledge across scientific disciplines.
- Having a deep and up-to-date understanding of your own area of specialization.

In this section I will summarize the scientific method, briefly discuss what molecular microbiology is and how it fits into the spectrum and history of science, and describe how to read and understand the scientific literature.

THE SCIENTIFIC METHOD

The goal of science is to learn truths about reality. The *scientific method*, more than anything, is a systematic approach we use to uncover those truths in a reliable way. Understanding and appreciating the scientific method is the core of scientific literacy.

Science begins with a **question**. There is something we don't know that we have reason to look at more closely and which we want to understand more fully. This can be very broad (e.g. what affects the spread of influenza?) or very specific (e.g. what is the role of glutamate 245 in the polyphosphate kinase enzyme of *E. coli*?), but the process always begins by identifying something we don't know.

How do we find an answer to the question? The next step in the scientific method is to develop a *hypothesis*. A hypothesis is a **possible** answer to the question and is informed by whatever else the scientist knows about the subject. The most important feature of a hypothesis is that it must be *falsifiable* or *testable*, which leads directly to the next step in the process.

What distinguishes science from other types of inquiry about the nature of reality is that in science we rigorously test our hypotheses. Whether via *observations* or *experiments*, the scientist puts their ideas to the test, **discarding hypotheses that do not match the facts**. This process of testing hypotheses results in the development of a *model* to explain the **mechanism** underlying the observations the scientist has made, and addition of more observations may strengthen or weaken that model. Models to explain natural phenomena start simple and gain complexity and *predictive power* as more facts are discovered and incorrect hypotheses are discarded. If an observation is made that does not fit with the model, the scientist must change the model to incorporate the new data, and test any new predictions made by those changes. Developing a model is a deeply creative process, drawing on all the knowledge of the scientist, with the fundamental constraint that a valid model must explain **all** of the observations of a system. By reiterating this self-correcting process, scientific knowledge converges on truth.

I really like the way that I heard Nathan Erdmann (from the UAB Department of Medicine) explain his approach to science. To paraphrase, he thinks of the scientist's job as two-fold: 1) to decipher **signal** from **noise** and 2) to ask **meaningful**, **answerable** questions. I think this is really important. How do you make sure that your observations are not just random variation? Has someone already answered the question? Do the tools exist to allow you to ask the question? What impact will knowing the answer to the question have?

In future lectures we will practice developing hypotheses, models, and experiments and go into more detail about what each of those steps entails.

WHAT IS MOLECULAR GENETICS?

Genetics is an approach to understanding biological systems that involves manipulating the genetic material of an organism (its *genotype*) and observing the changes that result from those manipulations (the *phenotype*). It is often contrasted with *biochemistry*, which focuses on the properties of (usually) purified components of cells like particular proteins, nucleic acids, or lipids. Both approaches are essential to understanding how biological systems work. Very often, genetic experiments will provide the first indication of the role of a protein or other cellular component, which will then guide the detailed biochemical analysis of that component. *Molecular genetics* is simply genetics with an understanding of the biochemical nature of DNA, and with tools to directly manipulate that genetic material.

Snyder & Champness's "Molecular Genetics of Bacteria", now in its 5th edition, is an excellent textbook on this topic, if you're interested in more in-depth, detailed discussion of specific topics than I'm aiming to achieve here.

Talking about genetics requires understanding quite a bit of technical terminology, and I'll try to define the essential jargon here as simply as possible, but you will inevitably have to learn the vocabulary. You'll also need to have at least a reasonable grasp of how the basic biological processes of *transcription* and *translation* work. If you need to review the basics, these articles may help:

Transcription (genetics) - from Wikipedia

Translation (biology) - from Wikipedia

(Wikipedia is a generally reliable source for information on biochemistry. There's nothing particularly controversial about, say, the molecular weight of salicylic acid.)

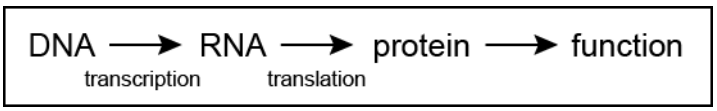
Different organisms are more or less easy to manipulate, and it's important to understand what is technically possible in the species you're studying. An experiment that takes a week in a well-established *model organism* like *Escherichia coli* may take months, years, or be impossible in a slow-growing, poorly characterized, or less well-studied species. Of course, new tools and techniques are constantly being developed to try to accelerate difficult procedures, both in academic labs and by commercial companies. We will talk about modern methods for manipulating DNA molecules in **Lectures 7 - 9**.

Experiments studying the properties of a gene or protein in a living organism, as in genetics, are referred to as *in vivo* studies (Latin for "within the living"). *Ex vivo* ("outside the living") experiments involve the use of cells or tissues removed from a larger organism. *In vitro* ("in glass") experiments, including most biochemistry, involve purified components removed from the cells in which they are normally found. The term *in situ* ("in position") is sometimes used to describe experiments that examine individual cells or organisms in the context of larger systems, without separating them from their natural context. Finally, the term *in silico* (fake Latin for "within silicon") is used to describe experiments performed entirely through computer simulations or calculations. I will note, however, that different fields use these terms differently. See, for an extreme example, this Wikipedia page on the uses of "*in situ*".

MAJOR CLASSES OF BIOMOLECULES AND THEIR FUNCTIONS

Molecular biology is the study of how organisms function at a biochemical level, and requires an understanding of what kinds of molecules make up living cells and what roles those molecules characteristically play. Since all organisms on Earth are descended from a common ancestor, the types of biomolecules are the same in all cells: bacterial, archaeal, or eukaryotic (and in viruses, too).

To review the so-called "central dogma" of molecular biology, genes are encoded as sequences of nucleotide bases on *chromosomes*, which are long molecules of double-stranded helical DNA (deoxyribonucleic acid). These genes are *transcribed* into single-stranded messenger RNA (ribonucleic acid) chains. Messenger RNAs (mRNA) are then *translated* into *proteins* (long polymers of amino acids that fold into complex 3-dimensional structures), which carry out enzymatic or regulatory functions within the cell.



This basic picture is, however, a gross oversimplification of the diversity of biomolecular functions, and you should be aware that, for example, there are many forms of *functional RNA* (ribosomal RNA, transfer RNA, small non-coding RNA, *ribozymes*, *etc.*), that RNA can be *reverse transcribed* into DNA, that some small *peptides* (short proteins) are synthesized without an mRNA template, and that *extracellular DNA* (eDNA) can play a structural role (in bacterial biofilms, for example). We've also entirely left out the roles of lipids and carbohydrates. Nothing in biology is simple!

The goal of research in molecular biology is to understand how the complex interactions of these different molecules fit together to form a functioning living cell. Biochemistry and genetics classes will teach you a lot of detailed theory about what is known so far, and I presume that in order to have gotten this far, you've taken such classes already. In this class, my focus is on giving you the practical and theoretical basis to carry out modern microbial genetics research.

GENES AND GENE PRODUCTS

A *gene* is a nucleotide sequence that encodes a functional *gene product*, which is usually a protein, but could also be an RNA molecule. For historical reasons, the terms *gene* and *locus* are often used interchangeably, although loci can also be functional sequences that are not genes themselves (like *operator sequences* involved in controlling expression of certain genes; see **Lecture 4**) and sometimes the term locus is used to refer to a region containing several related genes. An *open reading frame* (or *ORF*, sometimes also called a *coding domain sequence* or *CDS*) is a gene sequence that encodes a protein, often predicted based entirely on DNA sequence. *Alleles* are versions of a particular gene with different sequences, and sometimes with different functional properties. An *operon* is several genes encoded on the same mRNA, so that their transcriptional expression is linked. In bacteria, operons often (but not always) encode several genes that carry out a single biochemical pathway or otherwise related functions. An mRNA encoding more than one gene is still often called a *polycistronic transcript*, although the use of the term *cistron* as a synonym for gene (coined by Seymour Benzer in 1957) has otherwise almost entirely died out. (An mRNA encoding only one gene might be referred to, similarly, as being *monocistronic*.)

The *genotype* of an organism is a description of what genes and alleles it contains. The *phenotype* describes the measurable properties of that organism. The genotype determines the phenotype, but not all changes in the genotype will result in a measurable phenotypic change. More recently, it has also become clear that *epigenetic* differences in phenotype can exist **without** a corresponding change in the genotype. In bacteria, epigenetics is currently thought to depend mostly on methylation of specific DNA sequences, which changes how genes are expressed.

GENETIC NOMENCLATURE (IN BACTERIA)

For bacteria and archaea, there is a straightforward and consistent system for naming genes and strains that was developed and popularized by Milislav Demerec, a geneticist who was director of the influential Cold Spring Harbor laboratory from 1941 to 1960. The details of this system were published in the journal Genetics in 1966, and spread quickly through the bacterial genetics community. The examples I'll give here are mostly from *Escherichia coli*, the most common laboratory bacterium, but the same rules apply to all prokaryotic organisms.

To illustrate these rules, in the Materials and Methods section of a paper, you might find a table like the following:

Table 1.1. *E. coli* strains used in this study

Strain	Genotype
MG1655	F$^-$, λ$^-$, *rph-1 ilvG*
MJG238	F$^-$, λ$^-$, *rph-1 ilvG* Δ*ppk gloA*::*cat*$^+$

At first glance, of course, you may not get a lot out of that, but the information in this table is actually fairly straightforward, once you know the conventions.

Every strain of bacteria created or used in a lab is given a name, usually the initials of the investigator followed by a number. For example, strain MG1655 was isolated by Mark Guyer in 1981, and was probably the 1,655th strain he stored. MG1655 was one of the first bacterial strains to have its complete *genome* sequenced (in 1997, by Fred Blattner's lab at the University of Wisconsin at Madison), is a very-commonly used lab strain of *E. coli*, and is usually considered to be a "*wild-type*" strain. (Note that "wild-type" can mean anything from "a strain found in nature" to "any strain which doesn't have the mutation I'm interested in", depending on context. MG1655 is itself derived from an *E.*

coli strain called K-12, which was used by Joshua and Esther Lederberg in their foundational studies on bacterial genetic exchange, and is the ancestor of most of the laboratory strains of *E. coli* used today.) MJG238 is a strain I constructed which is derived from MG1655. They are *isogenic strains*, meaning they are identical except for mutations in the genes listed.

The names of some strains of bacteria may include their *serotype*, which describes what antibodies will react with the surface molecules that strain. MG1655 is, for example, a serotype "OR:H48:K-" strain of *E. coli*. Serotypes affect how the animal immune system responds to a bacterium, and for many species of bacteria strains with certain serotypes are more pathogenic than others. (For example, the "Jack-in-the-Box" strain of enterohemorrhagic *E. coli* is famously serotype O157:H7.)

A typical *E. coli* genome contains about 4000 genes in total, which is near the middle of the range for most types of bacteria. *Streptomyces* species can have more than 8000 genes, while simpler lactic acid bacteria often have fewer than 2000 and obligate intracellular pathogens like *Rickettsia* species have under 1000. Genes are given 3 or 4 letter names that are usually a mnemonic reflecting something about their function. For example, the *cbiA*, *cbiB*, and *cbiC* genes of *Salmonella* are three separate genes involved in **c**o**bi**namide biosynthesis, and the *ppk* gene mentioned in the genotype of strain MJG238 encodes **p**oly**p**hosphate **k**inase. In *E. coli* and many other species, *genes of unknown function* (and there are still many hundreds of these) have been given names starting with the letter "y" (*e.g. ydjA* or *yeaG,* or the *yci* genes in Figure 1.1), which indicates the location of that gene on the chromosome. These genes are likely to be renamed once their functions are determined. (Note that genes of unknown function from different species with the same "y gene" symbol may or may not be related in any way. For example, the *yneF* gene of *E. coli* is a putative cytoplasmic diguanylate cyclase, while the gene called *yneF* in *Bacillus subtilis* is an essential membrane protein whose activity is completely unknown. They have no homology to one another.)

Gene names are always written in italics, with the 3 letter first portion in lowercase. The fourth, capitalized letter (not always present, as seen for *ppk*) is used to differentiate separate genes that are all involved in the same pathway or phenotype. The proteins encoded by these genes would normally be capitalized, but not italicized: *e.g.* CbiA, CbiB, and CbiC, although in some cases, especially when the gene has only a 3-letter name, all three letters will be capitalized, as is the case for PPK. Systems of gene and protein naming in eukaryotes and viruses are different, and vary among model organisms.

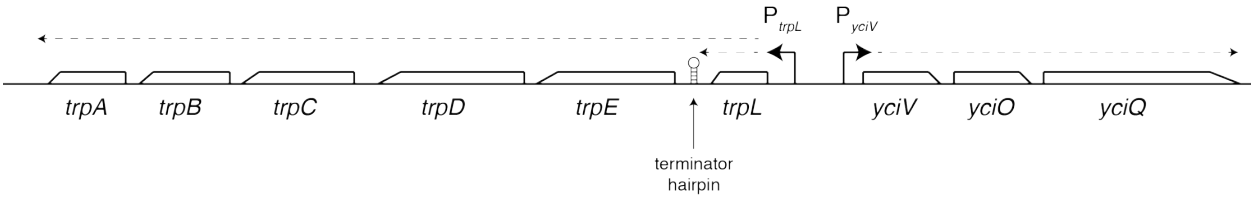


Figure 1.1. The *trp* locus of *E. coli*, to illustrate conventions of drawing genes and operons. The *trpA-E* genes are required for tryptophan synthesis. They and the tryptophan-rich leader peptide encoded by *trpL* are expressed as an operon from the P_{trpL} promoter. Two mRNA transcripts are possible from P_{trpL}: a short *trpL* transcript or, when lack of tryptophan leads to disruption of the terminator hairpin, a full-length 6-gene mRNA (this mechanism of regulation is called "transcriptional attenuation", and we will revisit this later in the course). The *yci* genes are *divergently transcribed* (that is, transcribed in the opposite direction) from the P_{yciV} promoter as a 3-gene operon that has no role in tryptophan synthesis. Genes that are close together and transcribed in the same direction are often, but not always, *cotranscribed* in *operons*. The only way to tell for sure is to directly test whether they are encoded on the same mRNA. Operons often, but again, not always, contain genes involved in related biochemical functions.

When looking at the genotype of a bacterium, the general rule is that any gene **not** mentioned is assumed to have normal, *wild-type* sequence and encodes a functional gene product. Unless otherwise noted, it's assumed that any gene that **is** mentioned has **lost** function. In Table 1.1, "*ilvG*" indicates a *mutation* (or genetic change) destroying the function of the IlvG enzyme (acetolactate synthase, involved in **i**so**l**eucine and **v**aline synthesis). Some mutations, especially in *E. coli*, may also be given *allele numbers* (as in *rph-1*) to indicate that multiple mutations in those genes exist in different strains. You can look up the function of inactivated genes in genomic databases to determine what effect those mutations might have on the phenotype of the strain. For lab strains of *E. coli*, the *Coli* Genetic Stock Center allows you to search for specific mutations by gene or allele number and find information about that mutation and a list of publicly available strains that contain it, although it certainly does not contain every *E. coli* mutation that has ever been made. Other useful databases are discussed below. Google Scholar is particularly useful for tracking down the original references for genes with different allele numbers in the literature.

MJG238 contains two additional mutations, which illustrate additional conventions of genetic nomenclature. The Δ*ppk* allele (that's a delta, for those of you not up on your Greek letters) indicates a *deletion* of the *ppk* gene, in which the DNA sequence encoding *ppk* has been completely removed from the genome. In contrast, the *gloA*::*cat*$^{+}$ allele, while still indicating a loss of function of the *gloA* gene (which encodes **gl**y**o**xalase I), shows (with the double-colon symbol)

that it has been disrupted by *insertion* of additional sequence, in this case the *cat*$^{+}$ gene. The superscript "+" associated with the *cat* gene symbol indicates that it encodes a **functional** gene, in this case encoding **c**hloramphenicol **a**cetyl**t**ransferase, which makes this strain resistant to the antibiotic chloramphenicol. You may see mutations that combine these properties (e.g. **Δ***glpF786*::*kan*$^{+}$), which indicates a deletion of *glpF* with a kanamycin resistance cassette inserted in its place, and this specific mutation has been assigned the allele number 786. Sometimes you will run across people using "Δ" to indicate any mutation destroying gene function, including insertions and point mutations. This is wrong (at least for bacteria). We will discuss types of mutations in much more detail in **Lecture 2**.

The notations "F-" and "λ-" (that's a lambda) are *E. coli*-specific indicators. The F plasmid ("**F**ertility factor") is a large circular DNA element (~ 100 kb) found in natural *E. coli* isolates which is capable of transferring itself (by conjugation, which we will talk about in **Lecture 8**) to other *E. coli* strains. F^{+} strains are sometimes called "male" strains, while F^{-} strains like MG1655 lack the F plasmid and are sometimes called "female". (It's not actually a very good analogy, since only the F plasmid is transferred and the "female" recipient then becomes "male".) "F' strains", rarely encountered today, have additional genes incorporated into the F plasmid. λ is a *lysogenic phage*, a virus which can integrate its genome into the chromosome of *E. coli* as a *prophage*. MG1655 has been cured of this viral genetic element, but many *E. coli* strains contain either wild-type λ or, more commonly, replication-deficient derivatives of λ with important or useful genes inserted into them. Both λ and F and their ability to transfer genes between bacterial strains were discovered by Esther and Joshua Lederberg at the University of Wisconsin around 1950, and were fundamental to the development of bacterial molecular genetics.

Genome sequencing technology has resulted in a tremendous increase in the number of predicted bacterial genes, most of which have no functional information associated with them. To deal with the problem of how to refer consistently to genes from genome sequencing datasets, every predicted gene in a genome is assigned a *locus tag*, which is unique to that gene in that specific strain. There are no established rules for how locus tags are formatted, and they are assigned by whatever research group sequenced and annotated the genome in question. For example, the *ppk* gene has the locus tag b2501 in *E. coli* MG1655, but the locus tag ESCCO14588_5033 in the pathogenic *E. coli* strain O157:H7 TW14588, even though these genes differ in DNA sequence by only 15 nucleotides and encode identical proteins. Locus tags are usually assigned in such a way that genes with consecutive numbers are next to one another in the genome, but even this is not always true, especially for small genes or noncoding RNAs that may have been missed in the initial annotation.

While locus tags are not as easy to understand as classical gene names, they are more specific and should be included whenever the identity of a particular gene needs to be established unambiguously. When publishing a paper, I would recommend mentioning the locus tag of a gene once (perhaps the first time it's referred to in the paper or in the Methods section) and then using an easier-to-remember gene symbol to refer to it in the rest of the paper.

For more details on the rules for writing bacterial genotypes, see the instructions on genetic nomenclature from the Journal of Bacteriology. You can find the genotypes of many lab strains of *E. coli* on the Open WetWare wiki. You'll note that most of them have many more mutations than MG1655. Bacterial strains are available to researchers from a variety of sources, including large *stock centers*. The most comprehensive are the American Type Culture Collection and the German DSMZ collection. These collections maintain stocks of thousands of strains of bacteria and other organisms that have been deposited by researchers around the world. For *E. coli* strains, the *Coli* Genetic Stock Center mentioned above is also a great resource. Many of the most common and useful *E. coli* strains are commercially available from biotechnology companies like Novagen, Agilent, and ThermoFisher. Strains generated by individual labs can be requested directly from those labs. Most researchers are happy to share published strains with their fellow scientists, and in fact, many granting agencies require that they do so. Nevertheless, you may have to do a fair amount of paperwork (called a *material transfer agreement*) to ship bacteria from one university to another.

THE SCIENTIFIC LITERATURE

The *scientific literature* is the summation of all published scientific knowledge. Knowing how to search and interact with it is a critical part of your scientific training. Before embarking on research on a biological system, it's wise to find out what is already known so that you don't waste time repeating experiments someone else has already done. Learning how to find that information is a critical skill. As my Master's thesis advisor once told me, "Six months at the bench can save you an hour at the library."

There are many sources of information about organisms, their genes, and the RNA and protein products of those genes. At the most fundamental level is the *primary literature*: research articles in peer-reviewed scientific journals. These are the basic product of laboratory research, and bacterial genetics papers will often (but not always) be focused on exploring the function of a single gene or protein in a particular organism. Most papers represent a year or more of work from between 2 and 10 scientists. The *first author* listed is generally the person who did most of the experimental

work, while the last author (or *corresponding author*) is usually the head of the lab where the work was done. *Reviews* are articles written by experts, summarizing the current state of knowledge in a particular field and collating information from dozens or hundreds of research articles. They are often the best way of learning about a research topic that is new to you, and are much more detailed and up to date than any textbook. Generally, the more recent the review, the better, at least to start. *Minireviews* are short reviews (a few pages), which usually either give a very brief introduction to or summarize the most recent developments in a specific topic. You can find papers and reviews using specialized search engines, the most useful of which for biomedical research is PubMed, provided by the National Center for Biotechnology Information. PubMed allows you to automate searches of the literature for particular keywords, which is a great way to make sure you don't miss any publications directly relevant to your interests. Google Scholar is also useful, especially since it allows you to search the full text of articles, and not just the title and *abstract* (which is a brief summary of the paper). Both these tools let you set up keyword searches that will automatically send you any new references that fit whatever criteria you define. This is a good way to make sure you don't miss any papers on your specific research area.

Video supplement – Performing a PubMed Search

Video supplement – Gene Searching on PubMed

It is important to note when searching and reading the literature that not all scientific journals are created equal. Some journals have higher quality standards than others. At one end of the spectrum are the *prestige journals*, which only publish what they consider to be the highest quality, most exciting, cutting edge, and influential results. These journals include Nature, Science, Cell, and the New England Journal of Medicine. At the other end are seemingly endless numbers of *predatory journals*, which have very low or no standards for what they will publish and are mostly just scams for separating naïve scientists from their money. In between are most of the journals in which quality research is published. Most scientific societies (like the American Society for Microbiology or the American Society for Biochemistry and Molecular Biology) publish *society journals*, which are not owned by for-profit publishing companies, have rigorous *peer review*, and are generally reliable, trustworthy publications.

Some journals are very specialized (for example, Antimicrobial Agents and Chemotherapy), while others have broader scope (like Molecular Microbiology). There are a number of metrics to measure journal quality, none of which is perfect. The most common is *impact factor*, which is the number of citations received by articles published in that journal during the two preceding years, divided by the total number of articles published in that journal during that time. A higher impact factor indicates that the papers published in that journal have been cited more frequently, but of course, this only takes into account the last two years, and can be skewed by a single highly-cited publication. Journals that publish a lot of reviews tend to have inflated impact factors for this reason. See EIGENFACTOR.org for an alternative, more robust measure of journal quality. While the quality of a paper is not necessarily linked to the quality of the journal in which it is published, higher quality journals will usually have more rigorous peer review and standards for publication, and will tend to publish more reliable work. Read carefully, and exercise good judgment. Do not automatically assume that something that's been published, even in the most prestigious journals, is necessarily correct.

One useful habit that will help you follow the literature outside of your own narrow research area is to subscribe to the *electronic table of contents* of several journals that publish research relevant to your interests. Those journals will then send you regular emails with the tables of contents for each issue, allowing you to quickly scan through the latest papers and keep up with your research community. As a microbiologist, good broad subject matter journals to follow might include mBio, Cell Host and Microbe, the Proceedings of the National Academy of Sciences, Nature Microbiology, the Journal of Bacteriology, the Journal of Biological Chemistry, Molecular Microbiology, and Applied and Environmental Microbiology, but you should subscribe to journals that regularly publish papers you are interested in reading. There are also prestigious journals dedicated solely to publishing reviews, which are tremendously useful. These include the Annual Review of Microbiology, Nature Reviews Microbiology, and Current Opinion in Microbiology, and can help you keep current on the most exciting and active research topics.

DATABASES

A variety of databases exist which compile data from many individual research papers into a single searchable format, and this is usually the best way to find general information (such as sequence and predicted function) about specific genes or proteins. They also typically contain links to the primary literature, which will contain much more detailed information. A common workflow for obtaining information about a gene or protein would be to consult a database to get a general sense for what is known about that gene, and then drill down more specifically into reviews and individual papers to understand the details and biological context of what we know.

The largest of these databases is GenBank, from the National Center for Biotechnology Information (NCBI), which contains all publicly available DNA sequences. A favorite of mine is the Integrated Microbial Genomes system, which

contains all of the information obtained from the full genome sequences of over 86,000 organisms and more than 20,000 *metagenomes* from different environments or bacterial communities (as of October 2020, but that number will rapidly become outdated as more become available).

Additional databases and resources that you may find useful include:

EcoCyc: a very well curated repository of information on the model organism *E. coli*, combining large amounts of manually compiled information from the literature for each gene and pathway in that organism, mostly for the K-12 strain MG1655. Most well-studied model organisms have similar dedicated databases. (*Subti*Wiki for *Bacillus subtilis*, for example.) MetaCyc automatically collates information for all organisms whose genomes have been sequenced, but of course there is generally much less information on genes and pathways in bacteria and archaea that are less well studied than *E. coli*. The BioCyc app (available for iOS) accesses these databases and is a convenient tool for quickly looking up genes of interest.

UniProt: a comprehensive database of **protein** sequence and functional information. This can be very helpful and is a great database, but is not focused on genes and is not an especially good source for nucleotide information. This is most noticeable in the fact that searching UniProt for locus tags will often not give any hits. In this case, you should use GenBank or another database to find the information you need.

PATRIC: a very comprehensive database of bacterial gene information, including genomes, transcriptomes, proteomes, pathways, systems biology, and phenotypic information (including antibiotic resistance), that is intended to be especially useful for those studying pathogenic bacteria. I have not used PATRIC much myself, but students in my lab have and it has a lot of very powerful tools for genome comparison and analysis.

RegulonDB: compiles the known information on how gene expression is controlled in *E. coli*. Much of this information can be found in EcoCyc, as well, but RegulonDB is organized in a different way that you may find helpful. PRODORIC is similar, but contains regulatory network information from a much broader assortment of bacterial species.

BioNumbers: a remarkably useful database that collects biological "trivia" that are very difficult to find elsewhere. Do you want to know something like the volume of a *Bacillus subtilis* cell, the number of cells in a bacterial colony, or the concentration of ATP in *E. coli* grown on glucose? BioNumbers will give you the values and references you need.

KEGG, the Kyoto Encyclopedia of Genes and Genomes: *the* database for pathways in all organisms. KEGG contains a truly vast collection of information on genetics and physiology, with powerful tools for visualizing and comparing pathways in different organisms, although it is less user-friendly than some of the databases listed above, and not **all** of the data in KEGG is freely available.

PDB: the Protein Data Bank contains three dimensional structure information for proteins, mostly determined by X-ray crystallography or nuclear magnetic resonance spectroscopy. To visualize and manipulate the data in this database, you will need a specialized structure-viewing program, such as PyMOL or CCP4.

BRENDA: a comprehensive database of published biochemical information on enzymes. Useful if you want to know things like rate constants for enzymes, cofactor requirements, known inhibitors, and other *in vitro* properties of proteins.

Microbiology Spectrum and EcoSal Plus: good collections of reviews summarizing all aspects of microbiology and of *E. coli* and *Salmonella*, respectively. They are regularly updated with new material, and if your institution has subscriptions (and UAB does) they are well worth consulting.

DISCUSSION PROBLEM SET #1: DATABASES AND LITERATURE SEARCHES

Use the above databases to answer the following questions, and be prepared to discuss your results in class. If you have trouble finding any of the information, that would be a great thing for us to discuss as a group!

** That is, in fact, the purpose of all of the discussion problem sets throughout this packet, so don't stress out if you find yourself stuck on something. The bulk of "lecture" time will be devoted to talking about and working through these problem sets as a class. You are absolutely welcome to work together or discuss the problems before class, if you want to. **

1) What genes are involved in proline synthesis in *E. coli*?

- sketch the pathway of proline synthesis, indicating enzymes and intermediates (no chemical structures necessary)
- draw the operon or operons encoding the genes involved in this pathway
- give a citation for a review article with more information on proline synthesis

2) What is known about the YeaG protein from *E. coli*?

- draw the *yeaG* locus, indicating genes and operons near *yeaG* in the chromosome and their functions (if known)
- summarize briefly what is known about the function or activity of YeaG
- cite two papers from the primary literature that describe research on YeaG

3) What is the function of the gene with locus tag SMc00166?

- what species / strain is this gene found in?
- what is its common name / gene symbol?
- draw the SMc00166 locus, indicating genes and operons near SMc00166 in the chromosome and their functions (if known)
- what is known about the function of this gene? (give 2 citations)

BLAST SEARCHES

Databases allow you to search for genes or proteins by name, function, or by *homology*: how similar they are to other sequences (using a search algorithm called "*BLAST*" (**B**asic **L**ocal **A**lignment **S**earch **T**ool)). Searching by homology is often the most useful, since gene names may not be used consistently and automated genome annotation may not necessarily assign the correct function to a gene (searching by locus tag avoids some of these problems). *Homologs* are genes that share a common ancestor, and may have similar or related functions. *Orthologs* are homologs found in different species, and *paralogs* are homologs found in the genome of a single species. BLAST is the most common algorithm for identifying regions of similarity between sequences, and therefore for inferring homology. It compares nucleotide or protein sequences, identifies sequences that have significant matches to each other, and calculates the statistical significance of those matches. BLAST is commonly used to identify members of gene families or to infer evolutionary or functional relationships between sequences.

The statistical significance of the results of a BLAST search are expressed as an "E-value". Technically speaking, the BLAST E-value is the number of expected hits of similar quality score that could be found by chance in a given database. From a practical standpoint, the smaller the e-value, the more similar the sequences in question are, in much the same way that a smaller P-value in a t-test or ANOVA indicates a more significant difference between samples.

The most common place to do BLAST searches is via the BLAST page at the National Center for Biotechnology Information. This will allow you to search nucleotide or protein sequences against GenBank, NCBI's database of sequence information. GenBank is an extremely large database, and includes essentially all published sequence information. This can be problematic, especially if you are BLASTing a gene from an organism (like *E. coli*) for which there are many very similar or identical matches in the database. You can get around this particular problem by clicking the "Exclude" option in the "Organism" field and excluding *Escherichia* (or whichever genus you don't want to see results from).

For more focused searches of either single genomes or of specific taxa, it is possible to filter your BLAST search by organism, species, or other taxonomic group. Alternatively, you can use the Integrated Microbial Genomes database, which contains only sequences from complete genomes and can filter searches in a variety of ways. The IMG database also has the advantage of providing (in my opinion) more user-friendly information about genes, gene neighborhoods, and pathways. You will need to create a (free) account to access the full capacity of this database (particularly BLAST searching against more than 25 genomes at a time). The "Top IMG Homolog Hits" pulldown menu at the bottom of each gene's page in this database is often exceptionally useful.

It is possible to filter BLAST search results in other useful ways (for example, returning one hit per species or eliminating sequences that are much shorter than your input sequence), but the web-based search platforms do not (at this time) provide for that, and you need to write your own bioinformatics scripts to accomplish these tasks. This is well beyond the scope of this class, and is best addressed by a course in bioinformatics, but I can recommend BioPython as a very accessible and flexible system for writing bioinformatics programs. Many professional bioinformaticians seem to prefer R, a programming language that provides very powerful tools for statistical analysis.

UNDERSTANDING AND ANALYZING BLAST SEARCH OUTPUT

BLAST searches are a key element of almost every project in molecular genetics. The output of a BLAST search will be a list of sequences homologous to your input sequence. The most common format for nucleotide and protein sequences is FASTA format, which looks like this (for the *E. coli* transcription factor RclR):

```
>646312216 NP_414839 transcriptional regulator, AraC family [Escherichia coli str. K-12 substr.
MG1655 chromosome: NC_000913]
MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHALTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDNESTCIVCGTLRLQHS
ARYFLTSLPETLFLAPVNHSVEYNWLREAIPFLQQESRSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGYASESSFHKAFVREFGCTPGEYRERVRQLAP
```

The text on the line after the ">" can be any identifying information for the sequence, from a complex ID like the one above to a simple name or number. The following lines are the amino acid sequence of the protein itself. A FASTA-formatted sequence file can contain any number of sequences in this format. Here, for example, are three *E. coli* genes involved in hydrogenase activity (note that FASTA format can contain either protein or DNA sequences, but not both):

```
>hyfA
ATGAACCGCTTTGTGGTGGCCGAACCACTGTGGTGTACAGGATGTAATACCTGTCTCGCTGCCTGTTCGGACGTGCATAAAACGCAAGGTTTACAGC
AACACCCGCGCCTGGCCCTGGCGAAGACGTCAACAATCACTGCCCCTGTCGTGTGTCATCACTGTGAGGAAGCCCCTTGCCTGCAGGTCTGCCCGGT
CAATGCCATCTCTCAGAGGGATGATGCGATCCAACTCAACGAAAGCCTCTGTATTGGCTGCAAGCTTTGCGCCGTGGTCTGCCCATTTGGCGCAATC
AGCGCTTCAGGAAGCCGTCCGGTGAATGCCCATGCGCAATATGTTTTTCAGGCTGAAGGCTCACTCAAAGACGGCGAAGAAAACGCGCCAACACAAC
ATGCTTTGCTGCGCTGGGAACCTGGTGTCCAGACCGTCGCGGTGAAATGCGACCTGTGTGATTTCTTGCCAGAAGGTCCGGCCTGCGTTCGCGCTTG
CCCGAATCAGGCGTTACGGCTGATCACCGGTGATAGCCTGCAACGTCAGATGAAAGAAAAACAGCGCCTTGCCGCAAGCTGGTTTGCCAATGGCGGG
GAGGATCCCCTTTCCCTCACTCAGGAGCAACGCTAA

>hyfC
ATGAGACAAACTCTTTGCGACGGATATCTGGTCATTTTTGCGTTAGCACAGGCCGTGATTCTGCTGATGCTAACCCCACTTTTTACGGGTATTTCCC
GGCAGATACGCGCGCGTATGCACTCCCGCCGCGGGCCGGGGATCTGGCAGGATTATCGCGATATCCACAAACTGTTTAAACGCCAGGAAGTTGCGCC
GACATCTTCAGGTCTGATGTTCCGCCTGATGCCGTGGGTATTAATCAGCAGCATGCTGGTGCTGGCGATGGCCTTACCACTGTTTATTACCGTTTCC
CCTTTTGCGGGCGGCGGCGATCTGATCACCCTTATCTATCTTCTTGCCCTGTTTCGTTTTTTCTTTGCTCTTTCCGGGCTGGATACCGGAAGTCCGT
TTGCGGGAGTCGGTGCCAGTCGCGAGTTGACGCTCGGCATTCTGGTCGAACCAATGCTTATTCTCTCACTGCTGGTATTGGCGCTGATAGCAGGTTC
CACGCATATCGAGATGATCAGCAATACGCTGGCGATGGGCTGGAACTCGCCGCTAACCACCGTACTGGCGTTACTGGCCTGTGGTTTTGCCTGCTTC
ATTGAGATGGGAAAAATTCCCTTTGATGTTGCTGAAGCAGAACAGGAATTACAGGAAGGCCCGCTGACCGAATATTCCGGTGCCGGGCTGGCGCTAG
CGAAATGGGGGCTGGGGCTGAAACAGGTCGTGATGGCATCACTGTTTGTGGCCCTGTTTCTGCCCTTTGGGCGCGCGCAAGAACTTTCTCTCGCCTG
CCTGCTGACTTCACTTGTCGTTACGCTGCTCAAGGTTTTGCTGATTTTTGTACTGGCCTCAATCGCAGAAAACACGCTGGCACGCGGGCGTTTTTTA
CTCATTCACCATGTGACCTGGCTTGGCTTCAGCCTTGCTGCGCTTGCATGGGTCTTCTGGTTAACCGGTCTGTAA

>hyfE
ATGACCGGTTCTATGATCGTAAATAATCTGGCGGGACTGATGATGCTGACATCGCTGTTTGTGATTAGCGTCAAAAGCTATCGCCTGTCATGCGGAT
TTTACGCCTGCCAGTCACTGGTGCTGGTGTCTATTTTCGCCACTCTCTCGTGCCTGTTCGCCGCAGAGCAACTGCTGATCTGGTCCGCCAGCGCCTT
TATCACCAAAGTGCTGCTGGTACCGTTAATCATGACTTACGCTGCACGAAATATTCCCCAGAACATCCCGGAAAAAGCGTTATTCGGTCCGGCAATG
ATGGCACTGCTCGCGGCGTTAATTGTCCTGCTTTGCGCATTTGTCGTTCAGCCCGTGAAGCTACCGATGGCTACCGGGCTGAAACCGGCGCTGGCGG
TAGCGTTAGGTCATTTTCTGCTTGGCCTGCTGTGCATTGTCAGCCAGCGCAATATCCTGCGGCAAATTTTTGGTTACTGCCTGATGGAAAACGGCTC
CCATCTGGTGCTGGCGCTTCTTGCCTGGCGAGCACCGGAACTGGTGGAAATAGGTATCGCTACCGACGCCATCTTCGCCGTCATTGTGATGGTGTTA
CTGGCAAGAAAAATATGGCGTACCCACGGCACGCTGGACGTGAACAACTTGACCGCGCTGAAGGGATAA
```

Most of the time, after using a BLAST search to identify homologs of your gene of interest, the next step in your analysis will be to generate an *alignment*, which allows you to visualize the regions of homology between the sequences and identify specific positions that are *conserved* between different sequences. Conserved regions are likely to represent the important functional parts of a gene or protein.

I find that for most purposes, amino acid alignments are the most informative, but in specific cases nucleotide alignments are appropriate. These include identifying an unknown DNA sequence and most *phylogeny* experiments, which examine evolutionary relationships among genes and organisms (since there are three nucleotides per amino acid, DNA sequence contains more potential phylogenetic information). *Phylogenetic trees* can be very valuable for exploring alignments and analyzing the evolutionary relationships among genes, but the details of how they are calculated are beyond the scope of this class.

PAIRWISE ALIGNMENT

In some cases, you may be simply interested in calculating the homology between two sequences. This is a "*pairwise alignment*". In this case, I typically use BLAST2seq from NCBI. This program will take two protein or nucleotide sequences and BLAST one (the "query") against the other (the "subject"), giving you a sequence alignment and additional information including an e-value (similar to a p-value, this is a statistical measure of how likely the similarities between two sequences is to have arisen purely by chance), a *percent identity* (how many positions are identical), and a *percent similarity* (for amino acids, how many positions contain residues with similar chemical properties). The output will look something like this, which is an alignment of the *E. coli* RclR protein sequence above with a homologous sequence from *Klebsiella pneumoniae* ("Expect" is the e-value, in this case 2×10^{-36}, which is very significant and indicates that these two sequences are closely related to one another):

```
Length: 284
Score          Expect  Identities     Positives      Gaps
121 bits(303)  2e-36   92/301(31%)    140/301(46%)   27/301(8%)

Query  1    MDSLSHLLALLAPRCEVNLHCRFGGRWQAGHQQMRSGVVPWHVVLRGEGRLNV-GGQTHH  59
            MD+LS LL L AP+  ++ +C  G  WQ  H      V+ WH + +G  +L +  G+
Sbjct  1    MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHALTQGAAKLEMPTGEIFT  60

Query  60   LRAGDVVLLPHGSPHLMESLVEWGQVLPVAHRFNGTVTEMRAGPAEGALEMLCGEFYFGP  119
            LR G+VVLLP  S               AHR +    E           ++CG
Sbjct  61   LRPGNVVLLPQNS----------------AHRLSHVDNESTC--------IVCGTLRLQH  96

Query  120  HVSW-LFSEASTLIHLHTDAREDCPELDALLNILVRESLAQRPGGSAIVRSLGDTLLVLL  178
               + L S   TL     +   +   L   +  L +ES +  PG  A+   +  T   L
Sbjct  97   SARYFLTSLPETLFLAPVNHSVEYNWLREAIPFLQQESRSAMPGVDALCSQICATFFTLA  156

Query  179  LRMLLGEQQPPGGLLRLMSDERLMPAVLAVMATPEQPWTLESMAARAFLSRATFARHFAR  238
            +R  + +      +L L+   RL   +  ++  P   WT+ES+A+ A +SRA+FA+ F
Sbjct  157  VREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAWTVESLASIAHMSRASFAQLFRD  216

Query  239  VYHLTPQAWLSQLRMALAARLLRLERQTNLEVIAERCGFQSLASFSKRFKMRYGVTPGEW  298
            V   TP A L++LR+ +AA++   E    + VIAE  G+ S +SF K F   +G TPGE+
Sbjct  217  VSGTTPLAVLTKLRLQIAAQMFSRE-TLPVVVIAESVGYASESSFHKAFVREFGCTPGEY  275

Query  299  R  299
            R
Sbjct  276  R  276
```

Notice that in this format amino acid residues identical in both proteins (*conserved residues*, "identities" or *percent identity*) are shown with that amino acid letter in between the query and subject sequences and that chemically similar amino acids ("positives" or *percent similarity*; for example, lysine [K] and arginine [R] are both large and positively charged) are indicated with a "+" sign. Dashes indicate regions of sequence in one of the proteins that do not contain matching sequence in the other, so in this case, there are two regions in the query sequence (from *K. pneumoniae*) that are not found in the subject sequence (from *E. coli*). The more residues which are the same in two aligned sequences, the more closely related those sequences are considered to be. Residues that are more highly conserved are generally more likely to have important functions in the final protein product, since mutants lacking amino acids critical for protein function will be selected against by evolution.

DISCUSSION PROBLEM SET #2: BLAST & PAIRWISE ALIGNMENT

Use the tools linked above to answer the following questions, and be prepared to discuss your results in class. (You should probably bring along a laptop so that you can easily share your results with the rest of the class and do additional analysis as necessary.)

For the genes with following locus tags:

- name the species this gene is from
- identify the predicted function of this gene
- align its protein sequence with that of its closest homolog from *E. coli* K-12 MG1655
- report the percent identity and percent similarity between the two proteins

1) aq_2095

2) SFK218_2554

3) USA300HOU_0506

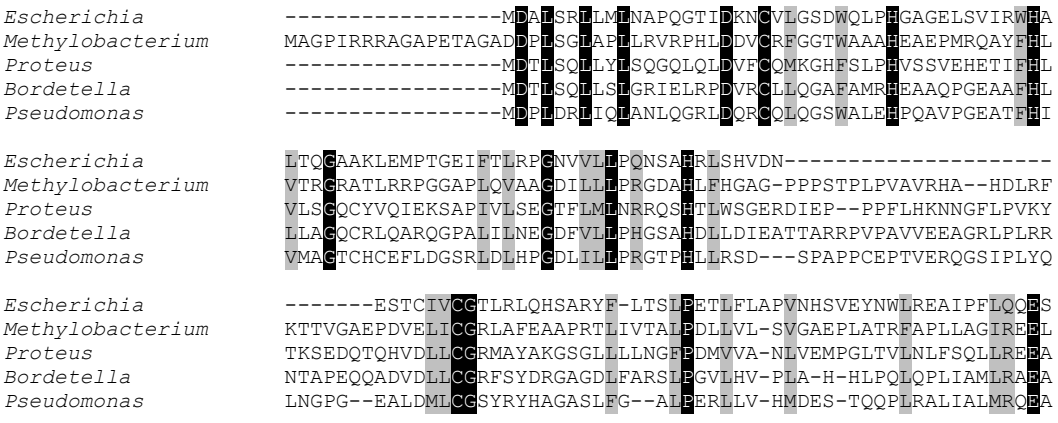
MULTIPLE ALIGNMENT

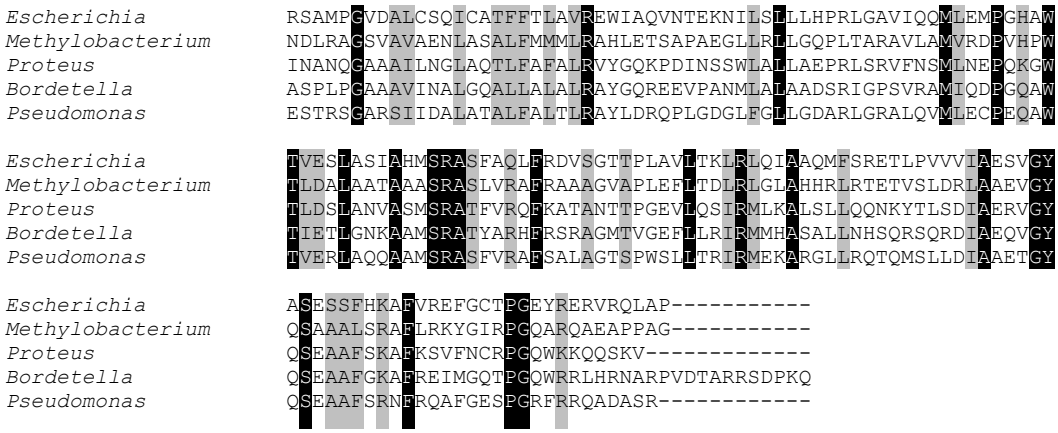
For alignments of more than 2 sequences (*multiple alignments*), there are a variety of tools and algorithms available, many of the best of which can be found at the European Bioinformatics Institute site. I often use MUSCLE for protein alignments, but Clustal Omega is also excellent. Use an alignment program appropriate for your particular samples. A high-quality alignment is important for future analyses (especially for phylogenetic trees). Alignment programs accept

lists of homologous sequences (commonly in FASTA format) and can present the resulting alignments in a variety of formats. One useful one is the human-readable Clustal format:

```
CLUSTAL O(1.2.1) multiple sequence alignment

Escherichia           -----------------MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHA
Methylobacterium      MAGPIRRRAGAPETAGADDPLSGLAPLLRVRPHLDDVCRFGGTWAAAHEAEPMRQAYFHL
Proteus               -----------------MDTLSQLLYLSQGQLQLDVFCQMKGHFSLPHVSSVEHETIFHL
Bordetella            -----------------MDTLSQLLSLGRIELRPDVRCLLQGAFAMRHEAAQPGEAAFHL
Pseudomonas           -----------------MDPLDRLIQLANLQGRLDQRCQLQGSWALEHPQAVPGEATFHI
                                       * *. *  *   .   *  * : . :   *         :*

Escherichia           LTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDN---------------------
Methylobacterium      VTRGRATLRRPGGAPLQVAAGDILLLPRGDAHLFHGAG-PPPSTPLPVAVRHA--HDLRF
Proteus               VLSGQCYVQIEKSAPIVLSEGTFLMLNRRQSHTLWSGERDIEP--PPFLHKNNGFLPVKY
Bordetella            LLAGQCRLQARQGPALILNEGDFVLLPHGSAHDLLDIEATTARRPVPAVVEEAGRLPLRR
Pseudomonas           VMAGTCHCEFLDGSRLDLHPGDLILLPRGTPHLLRSD---SPAPPCEPTVERQGSIPLYQ
                      :  * .  .   .  : :  * .::* :   * :

Escherichia           -------ESTCIVCGTLRLQHSARYF-LTSLPETLFLAPVNHSVEYNWLREAIPFLQQES
Methylobacterium      KTTVGAEPDVELICGRLAFEAAPRTLIVTALPDLLVL-SVGAEPLATRFAPLLAGIREEL
Proteus               TKSEDQTQHVDLLCGRMAYAKGSGLLLLNGFPDMVVA-NLVEMPGLTVLNLFSQLLREEA
Bordetella            NTAPEQQADVDLLCGRFSYDRGAGDLFARSLPGVLHV-PLA-H-HLPQLQPLIAMLRAEA
Pseudomonas           LNGPG--EALDMLCGSYRYHAGASLFG--ALPERLLV-HMDES-TQQPLRALIALMRQEA
                                ::**      .   :   .:*  :    :        :      :: *

Escherichia           RSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
Methylobacterium      NDLRAGSVAVAENLASALFMMMLRAHLETSAPAEGLLRLLGQPLTARAVLAMVRDPVHPW
Proteus               INANQGAAAILNGLAQTLFAFALRVYGQKPDINSSWLALLAEPRLSRVFNSMLNEPQKGW
Bordetella            ASPLPGAAAVINALGQALLALALRAYGQREEVPANMLALAADSRIGPSVRAMIQDPGQAW
Pseudomonas           ESTRSGARSIIDALATALFALTLRAYLDRQPLGDGLFGLLGDARLGRALQVMLECPEQAW
                       .   *  :: . :  ::: : :*            : *  .   .  .  *:. * : *

Escherichia           TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGY
Methylobacterium      TLDALAATAAASRASLVRAFRAAAGVAPLEFLTDLRLGLAHHRLRTETVSLDRLAAEVGY
Proteus               TLDSLANVASMSRATFVRQFKATANTTPGEVLQSIRMLKALSLLQQNKYTLSDIAERVGY
Bordetella            TIETLGNKAAMSRATYARHFRSRAGMTVGEFLLRIRMMHASALLNHSQRSQRDIAEQVGY
Pseudomonas           TVERLAQQAAMSRASFVRAFSALAGTSPWSLLTRIRMEKARGLLRQTQMSLLDIAAETGY
                      *:: *.  *  ***: .: *   :  :   .*  :*:  *   :         :*  .**

Escherichia           ASESSFHKAFVREFGCTPGEYRERVRQLAP-----------
Methylobacterium      QSAAALSRAFLRKYGIRPGQARQAEAPPAG-----------
Proteus               QSEAAFSKAFKSVFNCRPGQWKKQQSKV-------------
Bordetella            QSEAAFGKAFREIMGQTPGQWRRLHRNARPVDTARRSDPKQ
Pseudomonas           QSEAAFSRNFRQAFGESPGRFRRQADASR------------
                       * ::: : *       **. :.
```

In Clustal format, the punctuation under each block indicates conserved positions: "*" indicates completely conserved residues, ":" indicates very similar residues, and "." indicates a lesser degree of conservation. The similarity is based on the chemical properties of the individual amino acids. This format is an excellent way to present alignments of 3 to perhaps as many as 10 sequences. It is often more visually appealing (for publication, for example) to copy the text into a word processing program and replace the punctuation indicating conservation with colored or shaded backgrounds, as shown here:



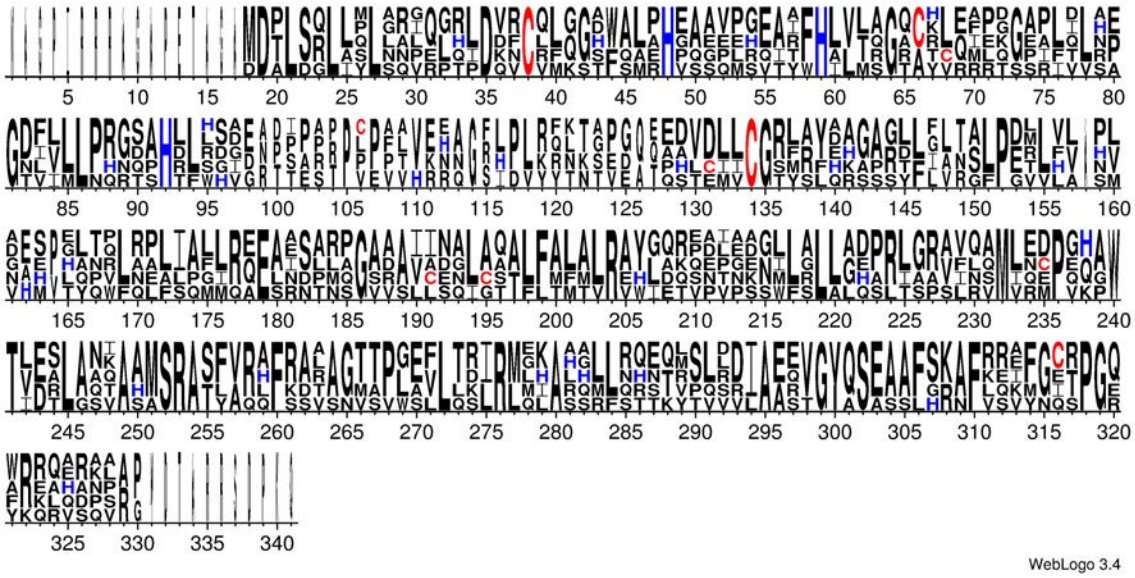


You can also present the same alignment in FASTA format, which is less human-readable, but more convenient for handling larger numbers of sequences:

```
>Escherichia
-----------------MDALSRLLMLNAPQGTIDKNCVLGSDWQLPHGAGELSVIRWHA
LTQGAAKLEMPTGEIFTLRPGNVVLLPQNSAHRLSHVDN--------------------
-------ESTCIVCGTLRLQHSARYF-LTSLPETLFLAPVNHSVEYNWLREAIPFLQQES
RSAMPGVDALCSQICATFFTLAVREWIAQVNTEKNILSLLLHPRLGAVIQQMLEMPGHAW
TVESLASIAHMSRASFAQLFRDVSGTTPLAVLTKLRLQIAAQMFSRETLPVVVIAESVGY
ASESSFHKAFVREFGCTPGEYRERVRQLAP-----------
>Methylobacterium
MAGPIRRRAGAPETAGADDPLSGLAPLLRVRPHLDDVCRFGGTWAAAHEAEPMRQAYFHL
VTRGRATLRRPGGAPLQVAAGDILLLPRGDAHLFHGAG-PPPSTPLPVAVRHA--HDLRF
KTTVGAEPDVELICGRLAFEAAPRTLIVTALPDLLVL-SVGAEPLATRFAPLLAGIREEL
NDLRAGSVAVAENLASALFMMMLRAHLETSAPAEGLLRLLGQPLTARAVLAMVRDPVHPW
TLDALAATAAASRASLVRAFRAAAGVAPLEFLTDLRLGLAHHRLRTETVSLDRLAAEVGY
QSAAALSRAFLRKYGIRPGQARQAEAPPAG-----------
>Proteus
-----------------MDTLSQLLYLSQGQLQLDVFCQMKGHFSLPHVSSVEHETIFHL
VLSGQCYVQIEKSAPIVLSEGTFLMLNRRQSHTLWSGERDIEP--PPFLHKNNGFLPVKY
TKSEDQTQHVDLLCGRMAYAKGSGLLLLNGFPDMVVA-NLVEMPGLTVLNLFSQLLREEA
INANQGAAAILNGLAQTLFAFALRVYGQKPDINSSWLALLAEPRLSRVFNSMLNEPQKGW
TLDSLANVASMSRATFVRQFKATANTTPGEVLQSIRMLKALSLLQQNKYTLSDIAERVGY
QSEAAFSKAFKSVFNCRPGQWKKQQSKV-------------
>Bordetella
-----------------MDTLSQLLSLGRIELRPDVRCLLQGAFAMRHEAAQPGEAAFHL
LLAGQCRLQARQGPALILNEGDFVLLPHGSAHDLLDIEATTARRPVPAVVEEAGRLPLRR
NTAPEQQADVDLLCGRFSYDRGAGDLFARSLPGVLHV-PLA-H-HLPQLQPLIAMLRAEA
ASPLPGAAAVINALGQALLALALRAYGQREEVPANMLALAADSRIGPSVRAMIQDPGQAW
TIETLGNKAAMSRATYARHFRSRAGMTVGEFLLRIRMMHASALLNHSQRSQRDIAEQVGY
QSEAAFGKAFREIMGQTPGQWRRLHRNARPVDTARRSDPKQ
>Pseudomonas
-----------------MDPLDRLIQLANLQGRLDQRCQLQGSWALEHPQAVPGEATFHI
VMAGTCHCEFLDGSRLDLHPGDLILLPRGTPHLLRSD---SPAPPCEPTVERQGSIPLYQ
LNGPG--EALDMLCGSYRYHAGASLFG--ALPERLLV-HMDES-TQQPLRALIALMRQEA
ESTRSGARSIIDALATALFALTLRAYLDRQPLGDGLFGLLGDARLGRALQVMLECPEQAW
TVERLAQQAAMSRASFVRAFSALAGTSPWSLLTRIRMEKARGLLRQTQMSLLDIAAETGY
QSEAAFSRNFRQAFGESPGRFRRQADASR------------
```

In FASTA alignment format, each protein sequence is listed separately, with gaps indicated by "-". As you can see, this does not provide an intuitive way to visualize sequence conservation, and you will need to use a separate alignment-drawing program to present the data. This is a good idea when you are aligning large numbers of sequences, where Clustal format becomes unwieldy.

WebLogo is a convenient online tool for visualizing conservation in large alignments. This program will accept any number of aligned sequences (in FASTA, Clustal, or many other formats), and will generate an image that represents the conserved residues in a very intuitive visual way, called a *sequence logo*. Take the alignment of RclR homologs from above, enter it into the WebLogo interface, and play with the different options to see what the program can do. Here's an example with 80 stacks per line, units of probability, scaled stack widths, no error bars, no y-axis labels, and a custom color scheme highlighting cysteine residues in red and histidine residues in blue:



In a sequence logo, the conservation of residues at each position is indicated by the height of the letters. For example, at position 38, all of the proteins in this alignment have a cysteine ("C") residue, while at position 39, there are approximately equal chances of finding a glutamine ("Q"), leucine ("L"), arginine ("R"), or a valine ("V"). You should adjust the parameters to give the most useful representation of your own data. WebLogo is a very versatile tool.

An alternative to presenting an alignment or logo is to report the *consensus sequence* of a set of sequences. This is a single sequence derived from an alignment by reporting only the most common residue at each point. The consensus sequence for the RclR alignment we've been working with is:

```
>RclR_consensus
--------MDXLSXLLXLXRXQXRLDVRCQLXGXWALPHEAAVPGEAXFHLVXAGQCXLXXPXGAPLXLXXGDFXLLPRGSAHLLXSXE--XPX-
PXPXXVEXAG-LPLRXXTXPGX-EDVDLLCGRLXYXAGAXLLXLTXLPXXLVXPXXESXXLTXLRPLIALLRXEAXSARPGAAAXINALAQALFA
LALRAYXQRXXXXXXLLALLXDPRLGRXVQAMLEDPGXAWTXESLANXAAMSRASFVRAFRAXAGTTPXEXLTRIRMXKAXXLLRQEXXSLXDIA
EEVGYQSEAAFSKAFRRXFGCXPGQWRRQXRXXXX----------
```

Positions where the most common "residue" is no amino acid (gaps, or more accurately, positions where one or two sequences in the alignment have a small insertion) are indicated with "-", and positions where no single amino acid is most abundant are indicated with an "X". As you can see, this is generally less informative than showing an alignment, but it does take up less space, so may be useful in some situations. For nucleotide sequence alignments, "N" is used to indicate a position with no conserved or most abundant nucleotide. There are also single letter codes for combinations of nucleotides (*e.g.* Y = C or T), the complete list of which can be found at this site.

DISCUSSION PROBLEM SET #3: BLAST & MULTIPLE ALIGNMENT

Use the tools linked above to answer the following questions, and be prepared to discuss your results in class.

For the genes with following locus tags:

- name the species this gene is from
- identify the predicted function of this gene
- identify homologs of this gene from species belonging to 5 **different genera**

  (Note that the more distantly related the homologs you choose, the easier it is likely to be to identify highly conserved regions of the protein. Why is that?)
- generate a multiple alignment with all 6 sequences (in whatever format you find most informative)
- based on your alignment, predict domains or specific amino acids that might be important for function of this protein (A *domain* is a structural element of a protein, usually between 50 and 250 amino acids long, that folds independently and may carry out a specific kind of function. Many proteins are constructed of several

domains, and evolution often builds "new" proteins by combining domains. From a sequence-gazing standpoint, a region of high homology could be a conserved domain.)

1) RCAP_rcc03362

2) USA300HOU_0588

3) PGN_1123

LECTURE 2: MUTANTS AND MUTATIONS

INTRODUCTION

In this lecture, we will discuss how bacterial geneticists use mutants and mutations to decipher how biological systems work. We will define different types of mutations and spend considerable time discussing how to interpret mutant phenotypes. We will also begin to explore how observations can lead to models and hypotheses, in the first steps of applying the scientific method to solving biological problems.

SCIENTIFIC PROCESS 1: OBSERVATIONS AND PHENOMENA

Every scientific study begins with an *observation*. The scientist looks at the world around them and sees a *phenomenon* that they think might be important or interesting. The key feature of phenomena is that they can be reliably and objectively measured, and therefore represent some real aspect of the physical world.

Reproducibility is central to the value of scientific observations. If a phenomenon is representative of something real, then it should be observable by different people in different places whenever the appropriate conditions occur. From a practical standpoint as a scientist, detailed record-keeping and recording of your observations is absolutely central. Only then do your observations rise to level of being *data*.

The quality of your observations is, in many ways, directly dependent on the tools and instruments you have available. In the history of microbiology, the invention of ever better microscopes (by van Leeuwenhoek, Hooke, and many others) allowed scientists to directly observe the existence of living things too small to be seen by the naked eye. Robert Koch's invention of solid growth media for bacteria and methods for isolating pure cultures made it possible to distinguish and separate different types of microbes from one another, leading directly to observations of specific bacteria and their relationship with particular diseases or environments. (As an aside, I will note that it was actually Angelina Hesse, an assistant in Koch's lab, who introduced the use of agar to solidify growth media.) Advances in DNA sequencing technology are a more modern example of the same process of technological improvement leading to new kinds of observations.

In this class, our focus is on using genetics, the science of how heritable characteristics are passed from one organism to another, to understand how bacteria function on a molecular level. We will therefore be making observations of how the biochemical and physiological behavior of bacteria is affected by changes in the content and expression of their genes.

When I ask you to describe a set of observations that you plan to make, you should explain:

- What will you be measuring, and how will you measure it?
- When and how often will you measure it?
- Is it a *qualitative* or a *quantitative* measurement?

Quantitative measurements result in numerical data, while qualitative measurements result in categorical or descriptive data. "Growth" or "no growth" is qualitative. "Blue" is qualitative, but "absorbance of 0.87 at a wavelength of 595 nm" is quantitative. Beware of assigning numerical values to categorical measurements and then treating them as quantitative. Occasionally, *binning* quantitative measurements to treat them as categories can be useful: colony diameter is quantitative, "greater than 2 mm" versus "less than 2 mm" is qualitative. Both types of measurement are potentially useful, but they need to be treated differently in statistical analyses.

THE GENETIC TOOLKIT

At a very simple level, molecular genetics techniques do one of two things: move new DNA into a cell or change the genes a cell already has. There are a wide variety of ways to do each of these things, and the methods that allow you to accomplish them in a particular species are referred to as the *genetic toolkit* for that organism. Some species have more fully developed toolkits than others, and this determines what kinds of experiments are possible in each species. In **Lectures 2** and **3**, we'll talk about how and why we can change or remove a cell's genes, and in **Lectures 5 - 9** we'll discuss different ways of moving new DNA into cells, as well as homologous recombination, a mechanism which can incorporate new DNA into a cell's genome.

USEFULNESS OF MUTANTS IN BIOLOGICAL EXPERIMENTS

Any change in the genetic material of an organism is a *mutation*, and the resulting organism is a *mutant*. As noted in the last chapter, mutants are relative to their wild-type *parent strain*, although the definition of "wild-type" is somewhat arbitrary.

Mutations are the geneticist's best and most fundamental tool for understanding biological systems. We isolate mutants to understand what changes in a cell's genotype affect the phenotype we are studying. This allows us to narrow down the tremendous complexity of cells and focus on only the genes, alleles, and loci that directly influence our particular study system. If a mutation affects our phenotype of interest, it tells us something about how that phenotype works. As several senior microbiologists have expressed it to me, "Let the cells tell you what's important." (I've tracked this phrasing back, anecdotally, to Bruce Ames, a pioneer of *Salmonella* genetics.)

An analogy I have found useful for explaining the use of mutants in biology is to imagine that you have no idea how automobiles work, and the only resources you have available to figure it out are a hammer and an infinite supply of Volkswagen Beetles. The geneticist's strategy to solve this problem is to break one thing in each car with the hammer and see what happens. If you break the spark plugs, that car won't run, but the headlights will work (at least for a while). If you break the battery, that car won't run **and** the headlights won't work, telling you that the engine depends on both the spark plugs and the battery, but the headlights only require the battery. The hammer is making "mutations", and by interpreting the "phenotype", we are able to piece together how a complex system functions and how the different components are interrelated.

INTERPRETING MUTANT PHENOTYPES

We extract meaning from mutations by examining the phenotypes that result from genetic changes. If we isolate several different mutants that have mutations in different genes, but have similar phenotypes, we can reasonably conclude that those genes are all involved in that phenotype. We might, for example, identify several different mutations in *Vibrio cholerae* that fail to secrete cholera toxin, and are therefore unable to cause disease. Some of these might be genes encoding the toxin protein itself, while others could be important for transport, processing, or regulation. However, since they all have a "toxin-minus" phenotype, we can conclude that they all must work in concert in the cell to carry out the toxin production process.

DISCUSSION PROBLEM SET #4: BACTERIAL PHENOTYPES

The key feature of a useful mutant is that it has a different phenotype than the wild-type. Mutations can change any of the phenotypes we can measure, and are our primary tool for interrogating biological functions.

What kinds of phenotypes can we measure for bacteria? List as many as you can think of, indicating whether they are quantitative or qualitative.

There are many technical terms that are used to describe phenotypes. An *auxotroph* is a strain that requires a particular nutrient. This is contrasted with a *prototroph*, which does not require that nutrient. A mutant defective in histidine synthesis would be a histidine auxotroph, for example. (This is often written as being "His^-", pronounced "hiss-minus".) A strain that grows slowly in the absence of a particular nutrient is a *bradytroph*, although this is a much less commonly used term.

Phenotypes can be "strong", "weak", or "leaky", terms that are not strictly defined, but generally express how easy the phenotype is to observe. If your mutant dies under conditions where the wild-type grows well, that is a "strong" phenotype. If the difference is a more subtle one in growth rate, that might be referred to as a "weak" phenotype. A complete lack of histidine synthesis would be a "strong" phenotype, while a partial lack, with some histidine still being made, would be a "leaky" phenotype. In this example, you might hypothesize that genes in which mutations result in strong His^- phenotypes might be directly involved in the biochemical pathway for histidine synthesis, while those with leakier phenotypes might play roles in regulating the activity of the pathway or reduce the activity of enzymes without eliminating it completely. (The difference between a "weak" and a "leaky" phenotype is pretty arbitrary, as you can see.)

Mutations that have several apparently unrelated phenotypic effects are said to have a *pleiotropic phenotype*. This often occurs with mutations in genes for *global regulators* (see **Lecture 4**) or in genes with roles in central cellular functions or stress responses (RNA polymerase or protein folding chaperones, for example). See the end of this chapter for more on the use of hypotheses and models in bacterial genetics and how we use mutant phenotypes to develop and test ideas about biological functions.

Does every change in genotype cause a phenotype? I would answer this question with a cautious "no", since many changes in a bacterium's DNA sequence do not cause an obvious change in their appearance or growth. However, this is very much dependent on the growth conditions and on exactly what you are measuring. A mutant defective for uracil synthesis will not appear to have a phenotype until you attempt to grow it on media containing no uracil. A mutant that cannot make flagella forms colonies perfectly well on plates, and only when you look at it through the microscope in liquid culture do you find that it cannot swim. With a mutation that appears to have no phenotype, you may simply have not yet found the appropriate conditions to see the effect, so be cautious in your interpretations. It is also worth remembering that for bacteria we are often limited to relatively crude measures of gene function, like cellular growth rate. Mutations in highly conserved genes which have dramatic effects on multicellular eukaryotes, where developmental problems are very easy to see and can be caused by very subtle biochemical changes, may have no **visible** effect on the growth of a bacterial culture.

KINDS OF MUTANTS

There are many kinds of mutations that differ by exactly what sort of change occurs in an organism's genome sequence. *Point mutations* are changes of a single nucleotide in the DNA (sometimes called a *single nucleotide polymorphism* or SNP). *Transitions* are point mutations in which a purine (A or G) is mutated to the other purine or a pyrimidine (C or T) is mutated to the other pyrimidine. *Transversions* are point mutations from a purine to a pyrimidine or vice versa. *Missense mutations* are point mutations in a protein coding sequence that change the amino acid encoded at that point in the gene to a different amino acid. (See this site for a detailed resource on the consequences this can have.) *Nonsense mutations* are point mutations that change an amino acid-encoding codon to a stop codon (TAA, TAG, or TGA), terminating translation and resulting in a truncated protein product.

Table 2.1. Types of point mutations

transition	purine (AG) to purine, pyrimidine (CT) to pyrimidine
transversion	purine to pyrimidine, pyrimidine to purine
missense	amino acid-encoding codon to different amino acid-encoding codon
nonsense	amino acid-encoding codon to stop codon
silent	amino acid-encoding codon to a different codon encoding the same amino acid

Silent mutations are point mutations that, due to the *degeneracy* of the amino acid code (that is, the fact that more than one codon can encode the same amino acid), do not change the amino acid encoded by that codon. However, because some codons are more efficiently translated than others, "silent" mutations **can** sometimes affect protein expression or even protein folding, since they change the rate of translation. This is because not all tRNAs are present at the same concentration in cells, and all species have some *rare codons* which are infrequently used and therefore are translated by a smaller pool of available tRNA.

Insertions and *deletions* are the addition or subtraction of nucleotides into the chromosome. *Frameshift mutations* are small insertions or deletions of a number of nucleotides not divisible by 3, which disrupts translation of the gene downstream of the frameshift. Frameshifts result in scrambled and often truncated proteins. *Duplications* are mutations in which a region of DNA sequence is duplicated (resulting in 2 or more copies of that region). *Inversions* and *rearrangements* are large-scale changes in the structure of the chromosome, in which substantial regions of DNA are either reversed or moved relative to their position in the wild-type.

It is important to note that *null mutations*, in which the product of a mutated gene loses activity (also called *gene knockouts*), are always more common than *gain-of-function mutations*, where a new activity is generated, but all types of mutation can result in the addition of new functions under certain circumstances. Gain-of-function point mutations are often especially informative when trying to understand how a particular gene works. (There are many ways to break something, but usually only a few ways to make something work **better**.) Large insertions, especially of DNA from a different organism, are the most likely mutations to add new functions, since they may consist of whole new genes. When this happens naturally during evolution it is called *horizontal gene transfer*.

Some mutations will be *lethal*, and will result in a cell that can no longer grow. Like gain- or loss-of-function, this is a property of the phenotype, not the genotype *per se*. You will not be able to isolate lethal mutants in the lab without specialized methods. Lethal mutations could include null mutations of *essential genes* or gain-of-function mutations creating toxic effects. Some mutations result in *conditional phenotypes*, in which phenotypes (often lethality) are only observed under some conditions. A common and useful example of this are *temperature-sensitive* mutants, which result in gene products that are destabilized and do not function at high temperatures.

POLARITY

Because bacterial genes are often found in operons, with more than one protein encoded on a single mRNA molecule, mutations in one gene in an operon can affect expression of *downstream genes* in that operon. This effect is called *polarity*, and can complicate interpretation of mutant phenotypes, since a null mutation in one gene can also prevent expression of several other genes. Large insertions, which can contain entire genes, many stop codons, transcriptional terminators, *etc.* are especially polar, and commonly completely prevent expression of downstream genes in an operon. Some types of point mutations are also polar, and nonsense mutations and frameshifts are much more likely to have polar effects than other types. To understand this, it helps to understand the mechanism by which most polar effects occur.

Normally, when bacterial RNA polymerase is transcribing an operon (shown in the figure below as *orf1 – orf2 – orf3*), the mRNA produced is coated in ribosomes actively translating that mRNA into protein. (Transcription and translation are linked in many bacteria, although recent discoveries suggest that this linkage is not as tight in the Gram-positive model organism *Bacillus subtilis* as it is in *E. coli*.)

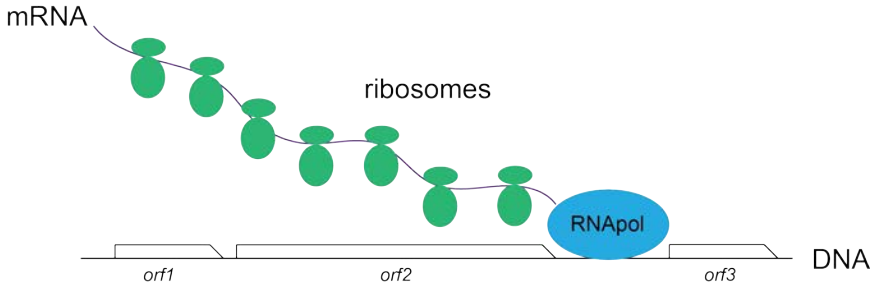


Figure 2.1. Transcription and translation are linked in bacteria. As an mRNA is being synthesized by RNA polymerase (RNApol), it is immediately recognized by ribosomes, which begin translating it into protein before the mRNA is finished being transcribed. Those ribosomes protect the mRNA from degradation and from premature transcript termination.

However, when a mutation prematurely stops translation of a gene in that operon, RNA polymerase continues, producing a stretch of mRNA with no ribosomes on it (until it reaches the ribosome binding site for the next gene – see **Lecture 4** for more on ribosome binding sites):

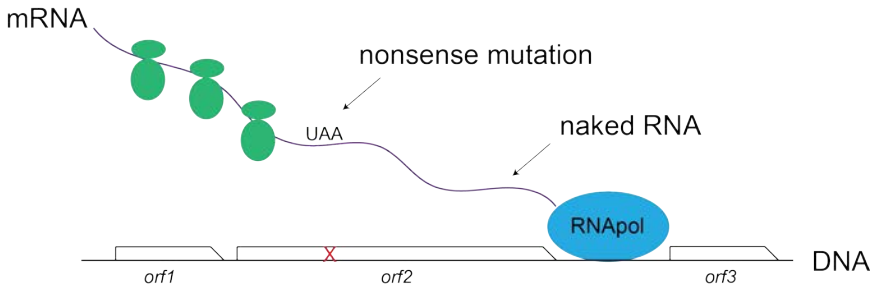


Figure 2.2. Polarity can result from premature stop codons, which result in stretches of "naked RNA" which is susceptible to degradation and premature termination.

Many bacteria contain a homolog of a protein called Rho, recently shown to form a complex with actively transcribing RNA polymerase. Rho is normally involved in transcription termination, and recognizes stretches of untranslated "naked" RNA, catalyzing changes in the structure of polymerase and causing it to disengage with and fall off the DNA. This is normally a mechanism to ensure that the cell doesn't spend a lot of energy transcribing non-coding RNA after stop codons, but in this case, it can cause point mutants to be polar in the same way that insertions can be. Note that *orf3* will never be transcribed, despite the point mutation being in *orf2*.

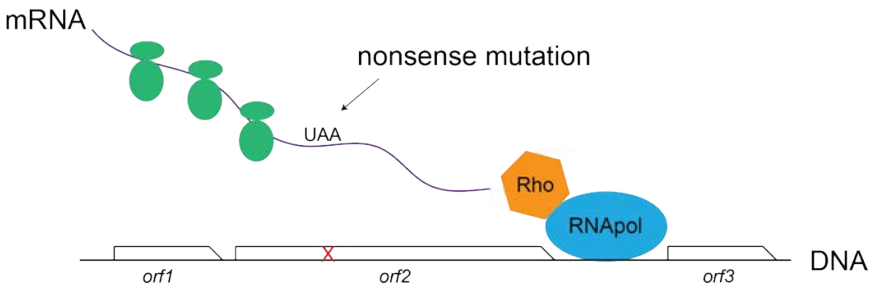


Figure 2.3. Rho recognizes mRNAs that are not covered with ribosomes and interacts with RNApol to terminate transcription.

In cases where polarity may play a role, there are a variety of molecular biology techniques which are used to help clarify which gene is actually responsible for a given phenotype. The two most common are *in-frame deletions*, which we will discuss techniques for generating in **Lecture 8**, and *complementation*, which is a major topic of **Lecture 5**.

MUTAGENESIS

Mutagenesis is the process of making mutations in an organism. There are many ways to do this, and the technique you use will have important effects on what kinds of mutants you can expect to find. *Random mutagenesis* creates mutations at random points within a DNA molecule, and is contrasted with *site-directed* or *targeted mutagenesis*, where you make a specific mutation exactly where you want it. (We will discuss methods for site-directed mutagenesis in **Lectures 7** and **8.**)

In this chapter we will focus on random mutagenesis. The major advantages of random mutagenesis are that you do not need to know in advance what genes or amino acids are important, and in the case of random point mutation, there is the potential to find gain-of-function mutations, which are often extremely informative. Depending on what kind of mutagenesis you want to achieve, you may have a variety of tools available, and those form an important part of the genetic toolkit for your model organism.

Because DNA replication is not perfect, *spontaneous mutagenesis* will occur in any population of growing bacteria, and is, of course, one of the underlying processes behind evolution. This is the simplest way to generate mutants in the lab, but since the various different kinds of mutations occur at low frequencies, it may take a very large number of cells or long period of time to identify the mutations you are looking for. As a rough estimate, in *E. coli*, any given single base pair change will occur in about 1 in every 10^8 cells (a frequency of 10^{-8}), while spontaneous null mutation of any given gene occurs in about 1 in every 10^5 cells, although these numbers can vary widely depending on the region of the chromosome involved.

The frequency at which mutations yielding a particular phenotype arise can be very informative. For example, if the phenotype you are looking for arises spontaneously in 1 in every 10^5 cells, you can reasonably hypothesize that it could be caused by a loss-of-function mutation in a single gene. If it only occurs once in every 10^8 cells, then your working hypothesis might be that it is caused by a specific point mutation.

Table 2.2 lists the rough frequency to be expected for different kinds of spontaneous mutations. We will discuss recombination, plasmids, and transposons in future lectures.

Table 2.2. Approximate Mutation Frequencies (*from Gary Roberts, University of Wisconsin – Madison)*

spontaneous knockout of gene function: 10^{-5}
any particular point mutation: 10^{-8}
reversion of a frameshift, missense, or nonsense mutation: $10^{-6} – 10^{-8}$
spontaneous deletions: $10^{-3} – 10^{-10}$ (*depends on the region to be deleted*)
duplication of a given region: 10^{-3}
loss of tandem duplication: $10^{-1} – 10^{-2}$
loss of various constructed plasmids: $10^{-2} – 10^{-5}$
loss of most natural plasmids: $< 10^{-8}$
precise excision of a transposon: $10^{-6} – 10^{-9}$
site-specific recombination events: $10^{-1} – 10^{-2}$

If obtaining a particular phenotype requires two independent mutations, the frequency of observing that phenotype will be the product of the frequencies of each individual mutation. A phenotype that requires two gene-inactivating knockout mutations would therefore occur spontaneously at a frequency of about $10^{-5} \times 10^{-5} = 10^{-10}$, so once in every 10 billion cells. We will discuss ways of increasing the rate of random mutations in **Lecture 3**.

SCIENTIFIC PROCESS 2: MODELS AND HYPOTHESES

There is more to science than simply recording observations. That's just list making, and a list of, for example, 75 mutations that cause a particular phenotype is not useful in and of itself. You must use that information to try to advance your understanding of how the world functions.

Once you have made a set of observations, you can propose a *model* to explain them. A useful model will not only propose a **mechanism** to explain the observations that have already been made, but even more importantly, will make predictions about what might be observed in the future. A model that can make accurate predictions about the world (has *predictive power*) is both useful and more likely to be correct than a model without such power. Models are always incomplete descriptions of the actual way the world functions (the map is not the territory). Even a model with significant inaccuracies may have some predictive power.

When I ask you to propose a model, it should:

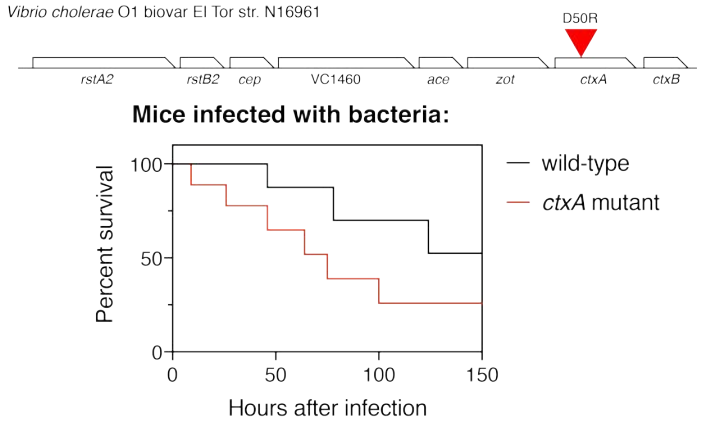
- incorporate all of the available data
- propose a mechanism that explains the behavior of the system
- make testable predictions about the system being studied

DISCUSSION PROBLEM SET #5: PROPOSING MODELS BASED ON DATA

A key skill in science is looking at data and developing models to explain those data. This requires creativity, open-mindedness, and humility (most of your models will end up being wrong, no matter how beautiful or elegant they are), but is the first step in applying the scientific process to solving problems.

Problem #1

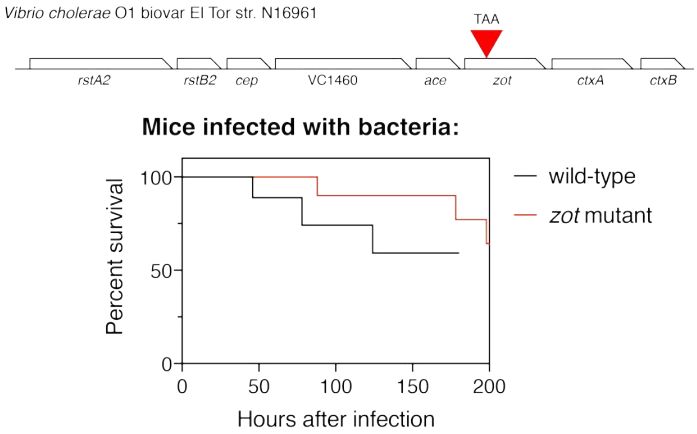
The following figure shows an operon from the cholera pathogen *Vibrio cholerae*. You have isolated a strain with the indicated *ctxA* missense mutation, and compared the survival of mice infected with this mutant and the wild-type strain.



Given these data, propose a model to explain the observed result. (It may be helpful to look up the function of CtxA.) Remember that a model should contain a proposed **mechanism**.

Problem #2

You have isolated a second mutant containing a nonsense mutation in the *rstA2* operon, sequenced it, and tested its phenotype. The results are as follows:



Given these data, propose **two** distinct models that could explain the observed result.

The predictions made by a model are *hypotheses*, and testing whether those predictions are accurate or not is a fundamental part of the scientific process. If a model predicts that you will observe X under a particular set of conditions, but you actually observe Y, then the model is wrong and must be changed to include the new information. To be useful, a hypothesis **must** be *falsifiable*, and therefore testable. "All disease is caused by bacteria," is a valid hypothesis, since it can be disproved by observing even one case of disease which is not caused by bacteria.

When I ask you to propose a hypothesis:

- it should be falsifiable (generally, using methods we have covered in class)
- you should be able to propose a set of observations that can be used to test that hypothesis

These are obviously closely related concepts. In any scientific study, observations lead to models, which lead to hypotheses. Testing hypotheses leads to more observations, the results of which are used to modify the model and improve its predictive power. In this way, science moves ever closer to an understanding of how reality works.

In simple cases, like those we'll be talking about in this class, models and hypotheses are very similar. Models, however, become more and more complex as they incorporate more data and become able to explain more observations. Hypotheses should always be as simple and straightforward as possible.

Hypotheses do not need to test everything about a model, and in fact, generally only test one aspect of it. A good model will lead to many hypotheses.

DISCUSSION PROBLEM SET #6: PROPOSING HYPOTHESES TO TEST MODELS

Problem #1

E. coli colonies **expressing** β-glucuronidase (encoded by the *gusA* gene), are blue on plates containing the indicator compound X-Gluc. Wild-type colonies of *E. coli* MG1655 (whose genome does contain *gusA*) are white. You spread MG1655 on X-Gluc plates, and are able to isolate spontaneous blue colony mutants.

Propose a model and testable hypothesis to explain each of the following possible results:

- blue colonies appear at a frequency of 1 in 10^3
- blue colonies appear at a frequency of 1 in 10^5
- blue colonies appear at a frequency of 1 in 10^8

Problem #2

Sinorhizobium meliloti is a plant symbiont that forms nitrogen-fixing nodules on the roots of alfalfa plants. When studying its metabolism, you find the results in the table below. *Minimal media* are growth media that contain only the compounds that a given organism absolutely requires for growth (as opposed to *rich media*, which contain lots of nutrients). In this case, it indicates media with no amino acids added.

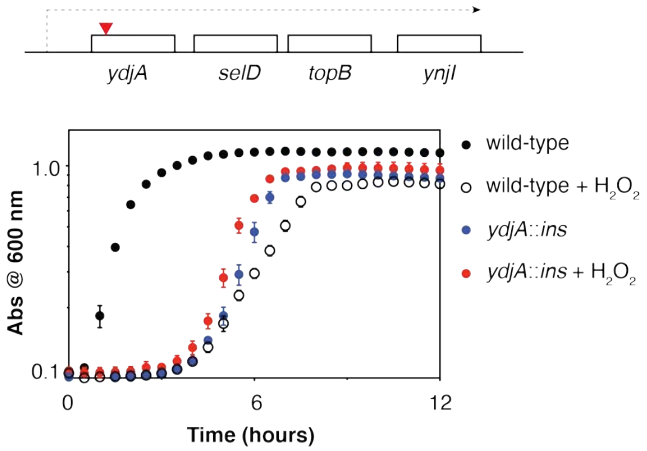
MetE and MetH are two different *isozymes*, non-homologous enzymes that catalyze the same reaction, in this case, the last step of methionine synthesis.

Strain	Ability to form nodules	Ability to grow on minimal media
wild-type	yes	yes
Δ*metE*	yes	yes
Δ*metH*	no	yes
Δ*metE* Δ*metH*	no	no

- Propose a model to explain these data.
- State one testable hypothesis derived from that model.
- Propose one observation you could make to test your hypothesis.

Problem #3

While studying hydrogen peroxide (H_2O_2) resistance in *E. coli*, you isolate a strain with a mutation in the *ydjA* gene. Further analysis reveals that *ydjA* is in an operon with several other genes, that the mutation (indicated with a red triangle in the figure below) is a 30 base pair insertion, and that the mutant has the following growth phenotype:



- Propose a model to explain these data.
- State one testable hypothesis derived from that model.
- Propose one observation you could make to test your hypothesis.

LECTURE 3: MUTANT HUNTS AND EXPERIMENTAL DESIGN

INTRODUCTION

In this lecture, we will go into more depth about the scientific method, discussing the difference between observations and experiments, and exploring the rules and principles of designing good experiments. We will define screens and selections, the two fundamental techniques for finding interesting mutants, and practice devising mutant hunts for different applications. Finally, we will explore different methods for actively mutating bacteria and discuss more advanced types of mutant analysis. We will also talk about alternative approaches and troubleshooting, emphasizing the importance of creativity and rigor for scientific problem solving.

SCIENTIFIC PROCESS 3: EXPERIMENTS, VARIABLES, AND CONTROLS

Observations are basically passive, making measurements of what occurs naturally in a system. To more aggressively test hypotheses, scientists actively manipulate the systems they are studying to see if the effects of those manipulations fit the predictions made by their models. Such a manipulation is called an *experiment*. A well-crafted experiment is a tremendously powerful way to make discoveries about the physical world, but it is important to understand what makes a **good** experiment.

In any experiment, the experimenter changes one or more *independent variable*s (or *treatments*) and observes the effect(s) that these changes have on one or more *dependent variable*s. It is usually best to have only **one** independent variable in an experiment, since this makes interpreting the effects on the dependent variable(s) much simpler. Remember: the independent variable is what you **change**, the dependent variable(s) is what you **measure**.

When designing an experiment to test a hypothesis, you must consider the following:

- **Will it answer the question?** Will the results of the experiment actually test the predictions of your model? Is it possible to learn anything from a result that is different from what you expect? Are there alternate explanations that could lead to the result you predict?
- **Is it possible?** How difficult will it be to carry out your proposed experiment with the resources you have available? What tools will you use to make your manipulations and measurements?
- **Is it elegant?** Some problems can be solved by *brute-force* approaches that simply test all the possible combinations of factors in a system. This can be effective, but is tedious and often expensive. It is often possible and preferable to test hypotheses with simpler, more creative experiments.

The best experiments are those for which **any** possible outcome gives you new information about the system you are studying and lets you improve your model for how it works. This is not always possible, but is definitely something to strive for.

Pilot experiments are preliminary tests, usually done in a relatively quick and inexpensive way, to see whether a new idea or procedure is worth pursuing further. To use an artistic metaphor, they are like sketches done before a real painting. It's especially important to do pilot experiments before embarking on any really labor-intensive or expensive experiment, so that you don't waste a lot of time and energy on something that will not give you meaningful results.

Experiments always need to have *controls*. Controls are experimental treatments with known outcomes, which allow the experimenter to be certain that their experimental setup is working as intended. *Negative controls* are treatments expected to result in **no change** in the dependent variable, while *positive controls* are treatments that **are** expected to result in such a change. Negative controls are important for ensuring that no contamination or other problems are interfering with measurements to give *false positive results* (also called a "type 1 error" in statistical jargon). Positive controls demonstrate that the measurement system is capable of observing the expected changes, and rule out the possibility of *false negative results* (a "type 2 error"). If an experiment's controls don't work as expected, then you cannot interpret the results and need to stop and figure out what has gone wrong with your experimental system.

When I ask you to design an experiment in class, you should explicitly:

- define the dependent and independent variables
- explain what you will measure and how
- describe both positive and negative controls
- describe the possible outcomes of the experiment and what they would mean for your hypothesis

In the previous class, we discussed using mutants to understand biological phenomena. In this class, we will explore this in more depth, beginning to look at the design of experiments in bacterial genetics, and actively manipulating bacterial genomes to test hypotheses.

ARTIFICIAL MUTAGENESIS

If spontaneous mutagenesis does not give you high enough *mutation rates* to isolate the mutants you are interested in, you can treat bacterial cells with *mutagens* that cause DNA damage and increase the rate at which mutations accumulate. *Chemical mutagens* are toxins that react with DNA, and *radiation* (including UV light) delivers energy directly to the DNA. Different mutagens cause different kinds of mutations. For example, UV light primarily causes G:C to A:T transitions, while acridine orange causes frameshifts.

From a practical standpoint, always be extremely careful using mutagens in the lab. You will be adding very toxic chemicals to your bacterial cultures or exposing them to high-energy radiation, which will mutate your DNA just as efficiently as they mutate bacterial DNA, and mutations in your cells may increase the chances of those cells becoming cancerous.

Both mutagens and spontaneous mutagenesis have the disadvantage that it can be difficult to locate where exactly in the genome a mutation causing an interesting phenotype actually is, although this has become quite a bit easier with the advent of inexpensive genome sequencing technology. The main limitation now is that mutagens tend to cause many simultaneous mutations across the genome, and it can be difficult to know which one or ones are causing a particular phenotype.

A common and practical way to make random gene-inactivating null mutations is the use of *transposons* or *insertion elements*. Transposons are parasitic DNA fragments that are able to "hop" or insert themselves into a DNA molecule, and many of them have little or no preference for specific target sequences. Barbara McClintock first discovered transposons in corn during the early 1950's, but her results were not widely accepted until nearly 20 years later, though they did ultimately garner her the Nobel Prize in 1983. Transposon mutants, like other insertions, nearly always destroy the function of the gene they integrate into, and are highly polar, which limits their usefulness for some kinds of mutagenesis experiments. You will, for example, almost never get a gain-of-function mutation or any phenotypes that require more subtle changes in gene function from a transposon screen.

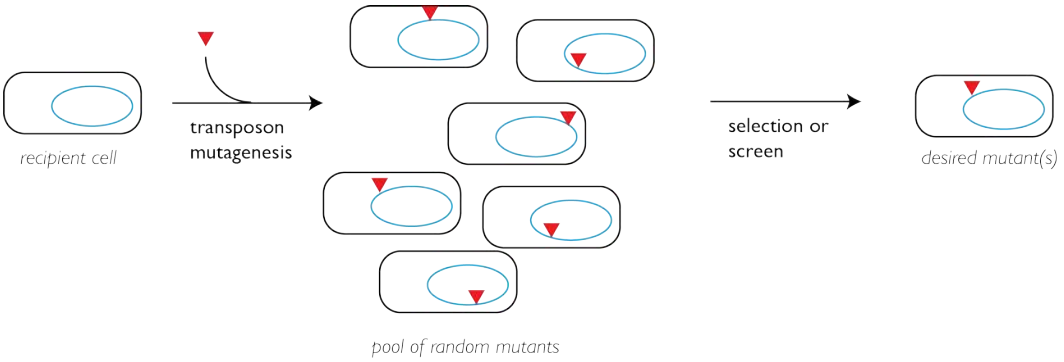


Figure 3.1: Transposon mutagenesis generates a library (or pool) of strains, each containing one randomly located transposon insertion. This library can then be screened to identify insertions that cause phenotypes of interest.

Several different kinds of transposons have been engineered to make them useful for random mutagenesis experiments. Common ones include Tn5 and the Mariner transposon, which are able to insert themselves at essentially any point in a DNA sequence, and have been modified to carry antibiotic resistance genes. This allows you to treat a population of bacteria with the transposon and select for only those cells that have successfully integrated it into their chromosome on media containing the antibiotic. Each individual cell will have only one insertion, but since they occur at random positions, pooling together many cells can give a mixture of mutated cells. This is a *transposon library*, should contain a wide variety of different highly polar insertions, and can be screened or selected for phenotypes as usual (see the next section for more details on screens and selections). Since transposons have a known sequence, it is easy to identify exactly where in the chromosome it has integrated by PCR and Sanger sequencing.

A transposon library should contain enough independent mutants to ensure that at least one insertion is present in every non-essential gene. This is called a "saturated" library. As a general rule of thumb, you need to have at least 4-5 times as many independent mutants in your transposon pool as there are genes in the genome of the organism you're studying. However, the more insertions you have in a transposon library the better, since you will increase the odds of

having at least one transposon even in very small genes. I have seen recent papers using libraries of up to a million mutants, which represent transposons in essentially every possible insertion site in an organism's genome.

Tn-seq (*transposon sequencing*, variations of which are also called INSeq, TraDIS, or HITS) is a very powerful technique that combines transposon mutagenesis with high-throughput DNA sequencing, allowing screening, enriching, or selecting for many transposon mutants in a single experiment without the need to isolate them individually.

DISCUSSION PROBLEM SET #7: LIMITATIONS OF TRANSPOSONS

Problem #1

The integration of transposons into a DNA molecule is carried out by enzymes called transposases. Natural transposons encode transposase, but engineered transposons (like the Tn5 and Mariner-based ones mentioned above) do not, and require expression of transposase from a different source. Why?

Problem #2

Let's suppose you make a transposon library of the cellulose-secreting bacterium *Komagataeibacter xylinus*, with the goal of finding mutants that produce higher than normal amounts of cellulose, which would be useful industrially.

A mat of bacterial cellulose produced by *K. xylinus* (Wikipedia).

However, despite your best efforts (see next section) you are unable to isolate any transposon mutants with increased cellulose production.

Why might this have failed? List as many reasons as you can think of.

FINDING INTERESTING MUTANTS

All kinds of mutations occur spontaneously, but not every mutation is interesting. We use bacterial genetics to ask questions about specific phenomena, which means we need to have methods for identifying mutations that have effects relevant to those phenomena. This kind of experiment is called a *mutant hunt*, and what you're hunting for is mutants that can help answer a specific biological question.

The two broad categories of mutant hunts are *selections* and *screens*.

If there are conditions under which mutants we are interested in will grow but the wild-type will not, then we can *select* for those mutants. Selections are extremely powerful and allow the isolation of very rare mutations. Since as many as 10^8 or 10^9 cells can be spread on a single agar plate, and only mutant cells will survive to form colonies, it is technically very simple to separate mutants from wild-type with a selection. Whenever possible, you should design mutant hunts as selections, since they will give you better results for much less work. However, it is not always possible to design a selection for your desired mutations, and in those cases, you will need to perform a *screen*.

Screens are used to isolate mutants that are different from wild-type in a non-selectable way (color, motility, toxin production, *etc.*) or mutants that die under conditions where the wild-type survives. In either case, the key feature of a screen is that the phenotype of each cell or colony must be examined **individually** to determine if it is an interesting mutant, and even in a best-case scenario no more than about 100-1000 colonies can be screened on a single plate (screens are also commonly now done in liquid media in 96- or 384-well microtiter plates). This means it is rarely practical to screen for mutations that occur at a rate of less than about 1 in 10^5 cells without sophisticated automation. Even with a very expensive robotic setup, screens of more than a few hundred thousand mutants or

conditions are usually impractical, although I have seen some flow cytometry-based screens that can be scaled up very effectively.

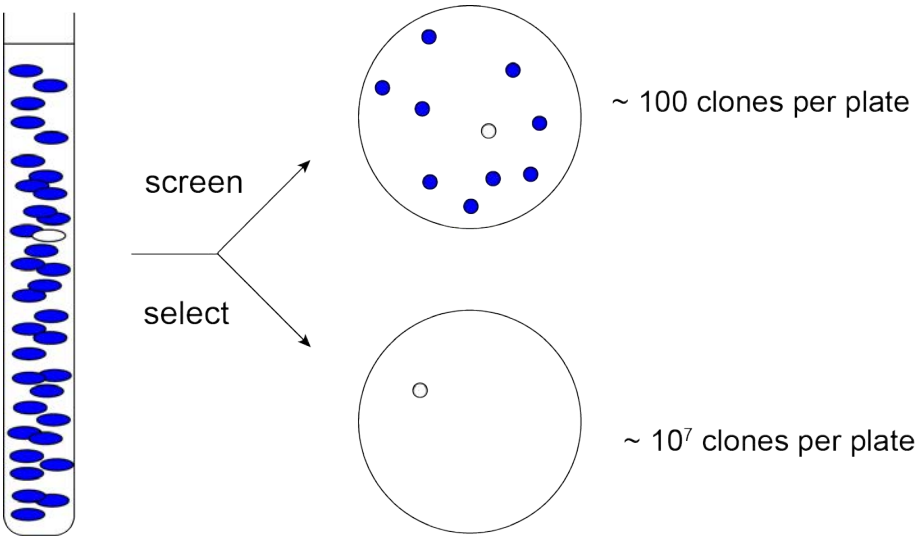


Figure 3.2: Selections allow you to identify much rarer mutants (white cells or colonies in this diagram) than is possible with a screen. A typical bacterial culture contains 10^8 – 10^9 cells per milliliter, and it is possible to spread about 100 microliters of culture on the surface of an agar plate.

An *enrichment* is somewhere between a selection and a screen. For a selection to work, you need conditions where the mutants are alive and the wild-type is dead (or at least does not grow). If you have conditions where the wild-type grows, but the mutants you're interested in grow **faster**, then the mutants will slowly become a larger and larger proportion of the population, thereby "enriching" the population with interesting mutants. You typically follow up an enrichment (or several cycles of enrichment) with a screen to identify individual mutant strains. This can greatly reduce the number of colonies that you need to screen to find mutants of interest.

It can sometimes be challenging to design a mutant hunt that will successfully isolate mutations relevant to a particular biological question. This is where you will need to think creatively about the model you are testing. If your model is correct, what kinds of mutant phenotypes might be possible? Which kind of mutagenesis is most likely to result in interesting and informative changes in the phenotype? We will practice this kind of creative problem solving in class throughout the next several lectures.

I want to end this section by emphasizing one last practical point about looking for mutants: **you get what you select for**. Even if that's not what you think you're selecting for! When you design a mutant hunt to try to identify mutations involved in a particular process, you will have some ideas in mind about what might result in the phenotype you're looking for. Biology is complicated, though, and there may be alternative ways to achieve such a phenotype. Sometimes this is interesting and useful, and leads to discovering unexpected connections between genes, but sometimes it just means you need to think more carefully about your selection conditions.

MUTANT HUNTS AS EXPERIMENTS

It may not be immediately obvious how the principles discussed in the section above on experimental design apply to mutant hunts.

To illustrate, let's look at an example experiment, in which we will use a screening approach to identify mutations in genes involved in sporulation in the opportunistic pathogen *Clostridium difficile*. (We will discuss spore formation in more detail in **Lecture 11**.) Spores are easily visualized through the microscope, and we can design an experiment as follows:

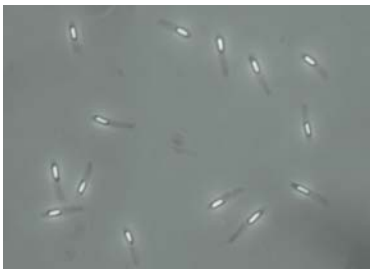


Figure 3.3: Microscopic image of bacterial spores (Wikipedia).

Observation: *C. difficile* can produce ethanol-resistant spores that are easily spread in hospitals.

Hypothesis: There are multiple genes necessary for production of ethanol-resistant spores by *C. difficile*.

Experimental Design:

1) Generate a saturated transposon library of *C. difficile* mutants.

2) Screen mutants microscopically for defects in spore morphology.

3) Sequence the genomes of mutants with spore defects to identify the sites of mutation.

Independent Variable: position of individual transposon insertions

Dependent Variables:

1) spore numbers (a quantitative measurement)

2) spore appearance (a qualitative measurement)

Negative Controls: (eliminate false positive results)

1) Confirm spore production and appearance in the wild-type

Positive Controls: (eliminate false negative results)

1) Confirm that the mutagenesis was successful (transposons confer antibiotic resistance, which is easy to test for).

2) If possible, test a known spore-deficient strain.

Potential Outcomes:

1) Mutations that interfere with spore formation are isolated. This supports the original hypothesis, and will allow us to identify the genes involved.

2) No mutations that interfere with spore formation are isolated. This could result from there not being any such genes (unlikely), from there being redundant genes for each step of spore formation, or from not having screened enough mutants to find (by chance) any with visible differences in spore formation.

BRUTE FORCE AND ELEGANCE IN EXPERIMENTAL DESIGN

"Elegance" is an elusive, but desirable, property of experiments, and nowhere is this more apparent than in the design of mutant hunts. There are multiple ways to solve any experimental problem. The experiment described above will work, but it's a very labor-intensive, brute-force approach to the problem, requiring microscopic examination of tens of thousands of mutants. Can we redesign our experiment to be more elegant?

Here's another possibility (with asterisks indicating steps that have changed):

Observation: *C. difficile* can produce ethanol-resistant spores that are easily spread in hospitals.

Hypothesis: There are multiple genes necessary for production of ethanol-resistant spores by *C. difficile*.

Experimental Design:

1) Generate a saturated transposon library of *C. difficile* mutants.

*2) Screen individual mutants for the ability to grow after ethanol treatment.

*3) Examine mutants that did not survive ethanol treatment microscopically for defects in spore morphology.

4) Sequence the genomes of mutants with spore defects to identify the sites of mutation.

Independent Variable: position of individual transposon insertions

Dependent Variables:

*1) growth after ethanol treatment (a qualitative measurement)

2) spore numbers of ethanol-sensitive strains (a quantitative measurement)

3) spore appearance of ethanol-sensitive strains (a qualitative measurement)

Negative Controls: (eliminate false positive results)

1) Confirm ethanol resistance, spore production, and appearance in the wild-type

*2) If possible, test a known spore-deficient strain. It should not grow after ethanol treatment.

Positive Controls: (eliminate false negative results)

1) Confirm that the mutagenesis was successful (transposons confer antibiotic resistance, which is easy to test for).

*2) Confirm that the mutants do grow **before** ethanol treatment.

*3) Confirm that your ethanol treatment successfully kills vegetative cells of *C. difficile*.

Potential Outcomes:

1) Mutations that interfere with ethanol resistance and spore formation are isolated. This supports the original hypothesis, and will allow us to identify the genes involved. (It will also identify mutants that are ethanol-resistant, but **don't** have visible spore defects.)

2) No mutations that interfere with spore formation are isolated. This could result from there not being any such genes (unlikely), from there being redundant genes for each step of spore formation, or from not having screened enough mutants to find (by chance) any with visible differences in spore formation.

This is still a screen, but it is a much less labor-intensive one, since growth and ethanol treatment of bacteria can be carried out in 96- or 384-well plates and easily scored by measuring absorbance. Only mutants that have demonstrated defects in ethanol resistance will be subjected to time-consuming microscopic examination. It will, however, still be a lot of work screening tens of thousands of individual mutants.

Is there a better way? Here's one more possibility:

Observation: *C. difficile* can produce ethanol-resistant spores that are easily spread in hospitals.

Hypothesis: There are multiple genes necessary for production of ethanol-resistant spores by *C. difficile.*

Experimental Design:

1) Generate a saturated transposon library of *C. difficile* mutants.

*2) Use Tn-seq to identify all of the insertions in the pooled library.

*3) Grow the entire pooled library, then add ethanol to kill vegetative cells.

*4) Regrow the survivors, and use Tn-seq to identify all of the insertions in those surviving cells.

Independent Variable: * ethanol treatment (before and after)

Dependent Variables: * the frequency of each transposon insertion in each pool (a quantitative measurement)

Negative Controls: (eliminate false positive results)

*1) "Before" data will not include insertions in any known essential genes.

*2) Confirm that your ethanol treatment has successfully killed all vegetative cells in your sample.

Positive Controls: (eliminate false negative results)

*1) Confirm that the mutagenesis was successful (the "before treatment" pool contains at least one transposon insertion in every non-essential gene).

Potential Outcomes:

1) You identify transposon insertions that are present in the "before treatment" pool and not present in the "after treatment" pool, and therefore prevent survival of ethanol treatment.

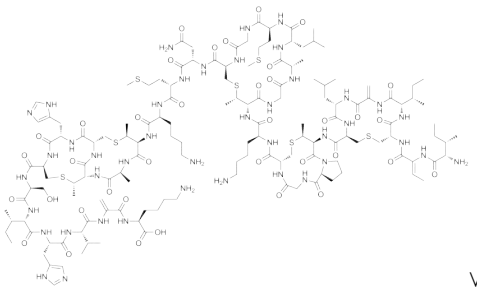
2) The distribution of transposon insertions is the same before and after ethanol treatment. This could result from there not being any such genes (unlikely), from there being redundant genes for each step of spore formation, or from your library being too small and not containing insertions in the genes required.

This experimental design will, in a single step, give you a list of genes in which transposon insertions result in sensitivity to ethanol, which you can plausibly hypothesize will be defective in their ability to form spores. It's still a screen, and you will need to do a secondary experiment to isolate individual mutants and examine them microscopically, but this design elegantly identifies all of the genes in *C. difficile* that are required for ethanol resistance.

DISCUSSION PROBLEM SET #8: SCREENS AND SELECTIONS

Problem #1

Nisin is an antimicrobial bacteriocin produced by some strains of *Lactococcus lactis*. Nisin efficiently inhibits the growth of a wide range of Gram-negative and Gram-positive bacteria. It is commonly used to prevent bacterial growth on the surfaces of food, including hard cheeses.



Wikipedia

L. lactis strains that produce nisin are immune to its effects, as are some strains of *Listeria monocytogenes*. (*L. monocytogenes* does **not** produce nisin.)

1) Design a mutant hunt to identify genes involved in nisin resistance using a **screen**.

2) Design a mutant hunt to identify genes involved in nisin resistance using a **selection**.

For each experimental design, state:

- your hypothesis
- the method of mutagenesis you will use (and why)
- the independent and dependent variables (*what will you change, and what will you measure?*)
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #2

Salmonella enterica can grow on the lipid breakdown product ethanolamine as a sole carbon and nitrogen source. Ethanolamine is abundant in the mammalian gut, especially during inflammation. You hypothesize that *S. enterica* has genes encoding a pathway specifically required for growth on ethanolamine.

Design a mutant hunt that would allow you to identify any such genes, and state:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #3

Mycobacterium tuberculosis strains often become resistant to antibiotics (like rifampicin) over the course of an infection, but do not typically acquire any new genes to do so (probably because they don't encounter any other bacteria when sequestered inside a granuloma). Your lab strain of *M. tuberculosis* is **not** resistant to rifampicin.

Design a mutant hunt that would allow you to identify mutations that lead to a Rif^R phenotype, and state:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?
- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

---

REVERTANTS AND SUPPRESSORS

If a mutation causes a slow growth phenotype or prevents growth without actually killing the cells, you will sometimes observe *secondary mutations* in that strain that restore wild-type-like growth. These arise because in the process of observing the poor growth phenotype, you are also selecting for any mutant that **is** able to grow well under those conditions. Such a mutant is called a *revertant*, because the phenotype has reverted back to "wild-type". In some cases, this may actually be due to a mutation that directly changes the original mutated gene back to the wild-type sequence. This is far more likely with point mutations than with other kinds of mutations. This kind of revertant and other mutations in the same gene as the original mutation are referred to as *intragenic suppressors*.

However, it is often more interesting to identify *intergenic suppressor* mutations, which are mutations in **other** genes that restore the phenotype of your mutant strain. If mutating one gene causes a growth defect, and you identify suppressor mutations in a second gene that restore growth, you have very strong evidence that those two genes are involved in the same biological process.

Multicopy suppressors are genes that, when present in more copies than in the wild-type (see **Lecture 5** on plasmids), suppress the phenotype of a mutation in a different gene.

It is worth noting that revertants may have a wild-type **phenotype**, but nearly always have a mutant **genotype**. It may not be possible to distinguish between the wild-type and a revertant based on the phenotype alone.

Some mutations are **only** ever found with a suppressor elsewhere in the genome, and it can be hard to know when this is the case without whole-genome sequencing. A mutant of *E. coli* lacking the heat shock regulator *rpoH* cannot grow above 18°C, but it is relatively easy to isolate *rpoH* null mutants at 30°C. How does that happen? The strains you isolate turn out to have suppressor mutations that result in an unregulated increase in protein-stabilizing chaperones, but if you didn't know that, you might make the wrong conclusions about the function of *rpoH*.

Figure 3.3: The appearance of spontaneous revertants. Note how faster-growing colonies containing suppressor mutations are arising out of a streak of slower-growing parent cells (which are themselves mutants that do not grow especially well under these conditions).

In a related phenomenon, there are also mutations that have no phenotype on their own, but have measurable phenotypes when they occur in **combination** with another mutation. When either one of a pair of genes can be knocked out, but you cannot delete both of them simultaneously, they are referred to as being *synthetically lethal*. Synthetic lethality is a strong piece of evidence that two genes are involved in related processes, or may in fact be *functionally redundant* genes that encode the **same** essential function.

DISCUSSION PROBLEM SET #9: SUPPRESSORS AND REVERTANTS

In *E. coli*, deletion of the *phoU* gene leads to a massive increase in phosphate uptake, dramatically slowing the growth of the cells and leading to a small-colony phenotype. When you streak out a Δ*phoU* mutant on a plate, though, you very occasionally (about one in every 10,000 colonies) find a colony that grows to an almost normal size.

Propose a hypothesis to explain the phenotype of these suppressors (with a **mechanism** explaining the improved growth), and an observation you could make to test that hypothesis.

MUTATIONS YOU WILL NEVER ISOLATE

There are some kinds of mutations that are very difficult or impossible to obtain, no matter how clever your mutant hunt design might be. The most common example of this is null mutations in *essential genes*, which encode functions that are absolutely required for the cell to survive. These include genes required for key cellular functions like DNA replication, RNA synthesis, and protein translation. Of course, which genes are "essential" depends on what growth conditions you're examining. Certain kinds of gain-of-function mutations may also be difficult or impossible to obtain if they cause toxic effects or consume all of a critical cellular resource in an uncontrolled way.

SCIENTIFIC PROCESS 4: ALTERNATIVE APPROACHES AND TROUBLESHOOTING

There is never only one way to address a scientific question. Testing a hypothesis in multiple independent ways is, in fact, a great way to ensure that any one experiment is not giving you misleading results. Most scientific papers (the good ones, anyway) will use multiple approaches to test and validate their conclusions.

Each approach to a problem has different advantages and disadvantages and gives you different kinds of results, so combining multiple approaches is the most rigorous way to test a hypothesis. When designing experiments for this class, different groups are very likely to come up with different, equally valid approaches to answer each question. This is fine! As we move through the different lectures, the tools you have available will expand and this will make more different kinds of experiments possible.

A related subject is troubleshooting: what do you do when your experiment "doesn't work"?

Be very careful when you say that an experiment has failed. Sometimes equipment breaks or contamination ruins a procedure, and the results of those experiments can be safely ignored while you fix the technical problem. An experiment that just doesn't give you the results you expected is **not a failed experiment**. It is a **discovery**. This is why the "Potential Outcomes" section of an experimental design is so important. You need to think about **all** the possible outcomes of your experiment, and be able to adjust your model to account for the result you actually get, not just the one that fits your preferred hypothesis.

LECTURE 4: PRINCIPLES OF REGULATION

INTRODUCTION

In this lecture, we will discuss regulation in bacteria, with a focus on interpreting the phenotypes of mutations that affect regulation and designing genetic experiments to explore how the expression of bacterial genes are controlled.

GENE EXPRESSION IS NOT CONSTANT

Bacterial genomes typically encode a few thousand different proteins. Not all of these proteins are present at the same concentration or at the same time, and bacteria are able to control the expression and activity of proteins in response to changes in their environments. Recall the basic flow of information from DNA to protein function (the "Central Dogma"), with DNA transcribed to RNA, which is translated into protein, which then has a biological activity:

DNA $\xrightarrow[\text{transcription}]{}$ RNA $\xrightarrow[\text{translation}]{}$ protein $\longrightarrow$ function

The steps in this process that can be regulated include:

1. Transcription initiation
2. Transcription elongation
3. Transcription termination
4. mRNA stability
5. Translation initiation
6. Translation elongation
7. Protein stability
8. Protein activity

Any gene may be regulated at one or more of these steps in response to either internal or external signals. In this chapter, I will summarize what is known about these processes. As with most fundamental biological mechanisms, the details are understood best in the Gram-negative model bacterium *Escherichia coli* and may differ more or less dramatically in other species.

MUTATIONS IN REGULATORS

As geneticists, it is important to understand what kinds of phenotypes arise from mutations in regulators and how we can use and interpret those phenotypes.

At the simplest level, there are two kinds of regulators: *positive* and *negative*. A positive regulator directly **activates** the system being studied in response to a signal. A negative regulator **represses** the system, and that repression is what responds to the signal. When negative regulation is relieved in response to a signal, this is often called *derepression*.

Mutations in positive regulators are often relatively easy to interpret. If a positive regulator is required to activate a particular phenotype, then null mutations in that regulator will have the same phenotype as null mutations in the other genes required for that phenotype. The genes controlled by such a regulator will be *constitutively inactive* (always "off") in the mutant.

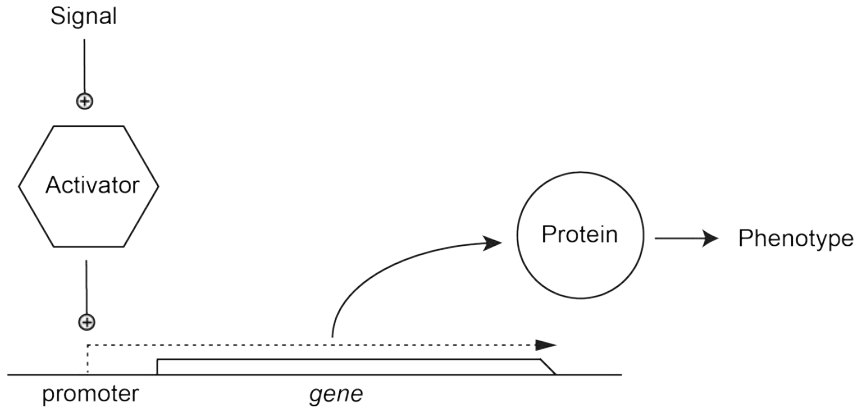


Figure 4.1. An example of a simple regulatory circuit in which gene expression is controlled by a transcriptional activator.

The following table illustrates how null mutations in different components of the circuit shown in Figure 4.1 would be expected to change the observable phenotype, which will only occur when the protein encoded by "*gene*" is produced:

Mutation	Signal	Phenotype
wild-type	absent	**-**
wild-type	present	**+**
activator-	absent	**-**
activator-	present	**-**
gene-	absent	**-**
gene-	present	**-**

Mutations in negative regulators can have less straightforward phenotypes. A very common regulatory circuit in bacteria involves *transcriptional repressors*, proteins that bind to DNA and prevent expression of genes until they detect a signaling molecule or metabolite. When that metabolite is present, the repressor loses its ability to bind DNA, and the repressed genes are then expressed. Other transcriptional repressors may respond to signals by becoming **better** at binding DNA, which will **decrease** gene expression or activity. In either case, the result of a null mutation in a negative regulator is likely to be *constitutive expression* or activity of the genes or proteins being regulated (always "on"), which can have very different effects depending on the genes in question.

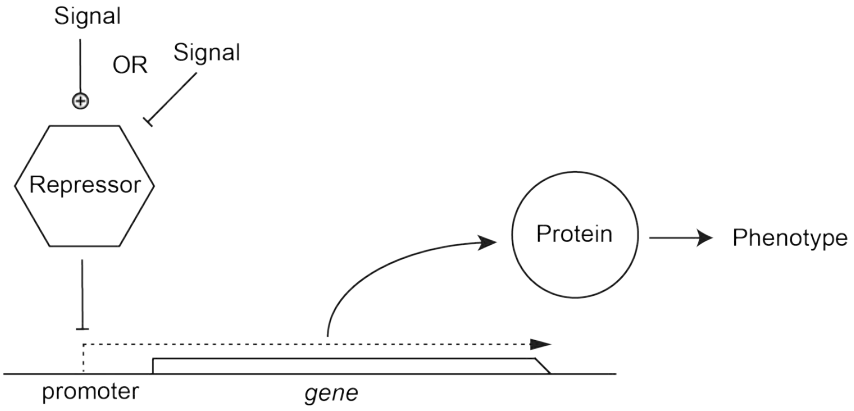


Figure 4.2. An example of a simple regulatory circuit in which gene expression is controlled by a transcriptional repressor.

The following table illustrates how null mutations in different components of the circuit shown in Figure 4.2 would be expected to change the observable phenotype if the repressor responds to a signal that **activates** its repressive functions:

Mutation	Signal	Phenotype
wild-type	absent	**+**
wild-type	present	**-**
repressor-	absent	**+**
repressor-	present	**+**
gene-	absent	**-**
gene-	present	**-**

If the repressor responds to a signal that **inactivates** its repressive functions, null mutations in different components of this circuit would be expected to change the observable phenotype as follows:

Mutation	Signal	Phenotype
wild-type	absent	**-**
wild-type	present	**+**
repressor-	absent	**+**
repressor-	present	**+**
gene-	absent	**-**
gene-	present	**-**

Operator sequences are sites in DNA (usually within the promoters of genes) where proteins or other regulators bind to control gene expression. In eukaryotes these are called "response elements". Mutations in operators can change how well those regulators bind, leading to a variety of phenotypes depending on the nature of the operator mutation and the regulator.

Global regulators control many genes or gene products throughout a cell's genome. They may be positive regulators of some of those genes and negative regulators of others, and mutations of global regulators often have very complex pleiotropic phenotypes. *Local regulators* control only a small set of genes, often in the same locus or operon as the gene encoding the regulator itself. Many genes are regulated by both global and local regulators, which allows them to respond in sophisticated ways to changes in the cell's environment. The classic example of this is the *lac* operon of *E. coli*, which is repressed by the lactose-specific local regulator LacI and activated by the cyclic AMP-sensing global regulator CAP.

It is important to note that not all gene regulation is absolute. In some cases you will have a gene switched entirely on or entirely off, but many regulators only adjust gene expression. This is particularly true for genes with multiple regulators (which is probably most of them, in bacteria).

SURVEY OF REGULATORY MECHANISMS

In the next part of this chapter, I will briefly describe some of the different types of regulation that are known to occur in bacteria without giving many specific examples (which would rapidly become overwhelming). I will also give examples of the different methods available for measuring gene expression in bacteria. My goal here is to give you a broad sense of how complex regulation can be and what tools are available to study it. The experimental problems below focus on how to decipher and understand phenotypes resulting from mutations in regulatory factors, with an emphasis on being able to narrow down the possible mechanisms leading to particular phenotypes.

REGULATION OF mRNA LEVELS

The first step in production of a protein is transcription of the mRNA encoding that protein by RNA polymerase. This involves three steps that can be regulated: *initiation, elongation*, and *termination*. The actual amount of a particular mRNA in a cell is determined by both these factors and by the *stability* of that mRNA.

Regulating transcription initiation is the least wasteful method of regulation, from the cell's point of view, since no nucleotides, amino acids, or energy are wasted producing unwanted gene products. However, it is also the slowest to respond to changes in the environment, since the cell must go through the entire process of transcription and translation to produce a final protein product. The level of an unstable RNA can be changed very rapidly by changes in initiation, elongation, or termination, while it might take several cellular generations to significantly change the levels of a very stable RNA.

1. Transcription initiation. Initiation of transcription takes place at a *promoter* where RNA polymerase binds to the DNA. Promoters vary in sequence, and the sequence of the promoter has a very strong effect on how efficiently a gene is transcribed. The *sigma subunit* (σ or *sigma factor*) of RNA polymerase is a small protein that determines the DNA sequence to which a particular molecule of RNA polymerase will bind. Typically, bacteria encode a *housekeeping sigma factor,* which is the most abundant sigma factor in the cell and recognizes the promoters of genes that need to be transcribed under most growth conditions. In *E. coli*, this is σ^{70}, encoded by the *rpoD* gene, and it recognizes

promoters containing consensus sequences of TTGACA and TAATAT centered at positions 35 nucleotides and 10 nucleotides upstream of the *transcriptional start site*, respectively (the -35 and -10 sites). The more similar the sequence of a promoter is to the consensus sequence for a particular sigma factor, the more strongly it will be bound by that sigma factor, which usually increases the amount of mRNA produced from that promoter.

Alternative sigma factors can replace the housekeeping sigma factor in RNA polymerase, and typically drive the transcription of genes important in responding to particular types of stress (*e.g.* heat shock, stationary phase growth), involved in the construction of complex molecular machines (*e.g.* flagella), or required for processes like virulence or spore formation. They recognize consensus sequences different from those found in promoters transcribed by RNA polymerase containing the housekeeping sigma. The concentration and activity of alternative sigma factors are tightly controlled, often using multiple mechanisms of transcriptional and post-transcriptional regulation, but several sigma factors can be present and active at the same time in a cell, all competing for the pool of core RNA polymerase. Different bacterial species may contain anywhere from one to dozens of sigma factors, depending on the complexity of their environment and developmental pathways. A common form of regulation for alternative sigma factors are *anti-sigma factors*, proteins that bind to sigma factors and prevent them from interacting with RNA polymerase. The stability or activity of the anti-sigma factor can then be regulated to control the activity of the sigma factor itself.

Other features of promoters can also influence the efficiency of transcription initiation. *UP elements* are AT-rich sequences upstream of the -35 site that increase transcription 30 to 70-fold. For some extremely highly active promoters (like those driving transcription of ribosomal RNA), the *initiating nucleotide* (that is, the first nucleotide of the transcribed RNA) can influence initiation in response to the levels of ATP or GTP in the cell, directly linking cellular energy state to gene expression.

Transcription initiation

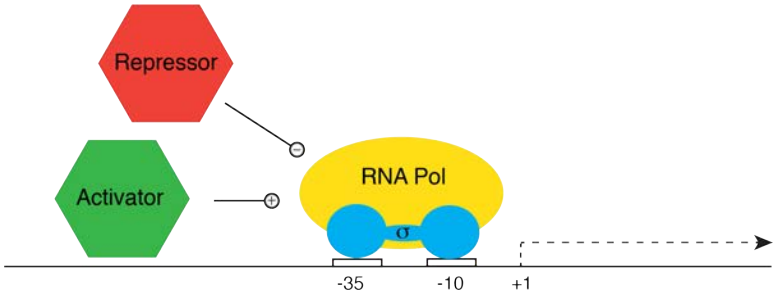


Figure 4.3. Regulators that can affect transcription initiation.

In addition to RNA polymerase itself, there are other proteins that can influence transcription initiation. These are called *transcription factors*. Most of these recognize specific DNA sequences in or near the promoter, but in some cases, they may bind to RNA polymerase without interacting directly with the DNA. *Repressors* are transcription factors that prevent initiation of transcription, often by blocking the -10 or -35 sites or otherwise preventing RNA polymerase from binding to the promoter. *Activators* increase the rate of initiation when bound to a promoter, either by recruiting RNA polymerase to a promoter or by interacting with an RNA polymerase molecule that is already bound to the promoter and stimulating its activity. The same transcription factor can sometimes act as a repressor or as an activator at different promoters, depending on the nature of the protein and the location of the binding site in the promoter, and multiple transcription factors often regulate a single promoter. The DNA-binding or RNA polymerase-influencing activity of transcription factors is often controlled in response to changes in the metabolism or environment of the cell (see Regulation of Protein Activity section below).

2. Transcription elongation. Once RNA polymerase has left the promoter and is producing mRNA, it enters the transcriptional elongation phase. The sequence and structure of the transcribed RNA determines the frequency of *transcriptional pause sites*, where RNA polymerase briefly stops producing mRNA. The number and position of pause sites can affect the speed of mRNA production and how it folds, which can affect both elongation and termination. There are proteins that interact with RNA polymerase to influence elongation speed (*e.g.* NusA or GreA), thereby regulating the amount of transcript produced.

3. Transcription termination. There is considerably more known about regulation of transcriptional termination than of elongation. In *Rho-dependent termination*, the Rho protein, as part of the RNA polymerase elongation complex, recognizes single-stranded RNA with no ribosomes attached and terminates transcription. (As mentioned in **Lecture 2**, this is why nonsense mutations are polar: they result in long stretches of untranslated RNA

in mRNAs.) *Rho-independent termination* also occurs (for about half of transcripts in *E. coli*). In these transcripts, *intrinsic terminators* are encoded in the mRNA itself that lead to the dissociation of RNA polymerase from the transcript. Intrinsic terminators are typically stable, GC-rich stem-loop structures 7 to 20 base pairs long, followed by several uracil residues.

Transcriptional termination can be regulated by *transcriptional attenuation* or by *anti-terminators*. Anti-terminators are proteins that prevent termination at specific termination sites, allowing RNA polymerase to bypass those sites. Attenuation is a mechanism by which an mRNA can take on more than one structural conformation, one of which is an intrinsic terminator. The classic example of this is the tryptophan biosynthesis operon (briefly mentioned in the reading for **Lecture 1**), the first part of which encodes a small Trp-rich *leader peptide*. When this peptide is translated efficiently, the presence of ribosomes on the mRNA causes it to fold into a structure that includes an intrinsic terminator stem-loop. If translation stalls due to a shortage of Trp-charged tRNA, the mRNA folds differently, eliminating the terminator stem-loop and allowing transcription of the entire operon to continue. *Riboswitches* are RNA structures, usually found in the 5' *untranslated region* (UTR) of an mRNA, which bind specific metabolites (*e.g.* amino acids or vitamins) and form structures that can include terminators or anti-terminator loops. UAB Microbiology's own Chuck Turnbough has recently written an excellent review on this topic, if you're interested in more details.

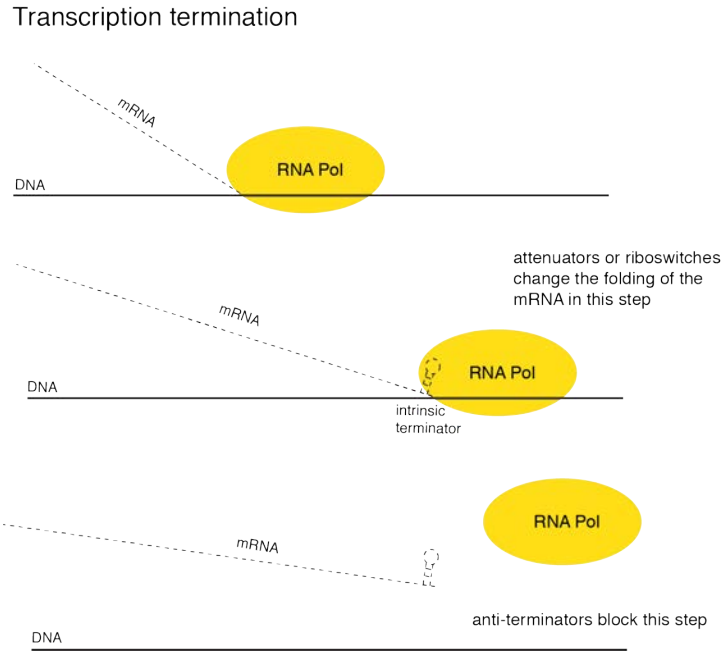


Figure 4.4. Regulators that can affect termination of transcripts with intrinsic terminators.

4. mRNA stability. The final consideration in controlling the amount of a particular mRNA in the cell is *transcript stability*. The half-life of mRNAs varies greatly, ranging in *E. coli* from as little as 40 seconds to longer than 40 minutes. Bacteria contain a variety of *ribonucleases*, which are enzymes that degrade RNA. The stability of a particular mRNA is determined by several factors. An important one is the presence of *endonuclease cleavage sites*, which are more common in some sequences than others. The *translatability* of a particular mRNA (see below) can also affect mRNA stability, since an mRNA that is covered in ribosomes is less susceptible to nucleolytic cleavage.

sRNA regulation of mRNA degradation

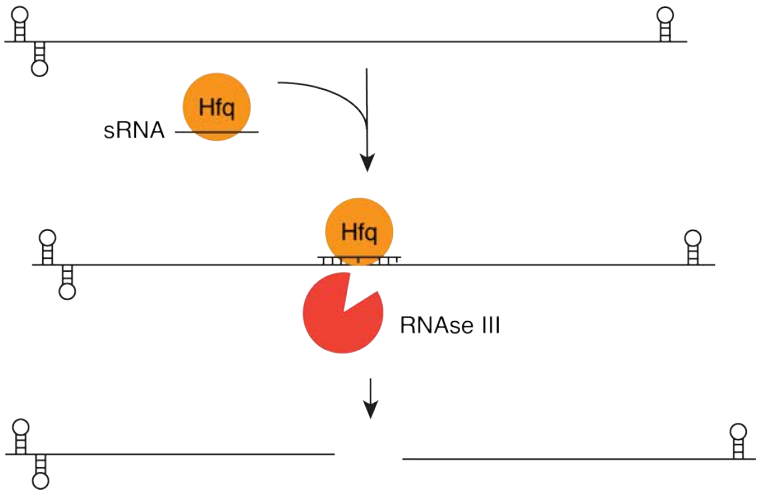


Figure 4.5. sRNA and Hfq-mediated mRNA degradation, by forming a double stranded RNA targeted by RNAse III.

RNA stability can also be regulated in response to environmental factors. The most common mechanism for this involves transcription of small, non-coding RNAs (*sRNAs*) that base-pair with the mRNA to be regulated (often overlapping the *ribosome binding site*, see below). The resulting double-stranded RNA then becomes a target for ribonucleases (specifically, RNAse III, encoded by the *rnc* gene). *Cis*-acting sRNAs are transcribed from the non-coding strand of an open reading frame and are therefore exactly complementary to their target sequences. *Trans*-acting sRNAs are encoded elsewhere in the genome, typically have less exact matches to their target sequences, and can regulate more than one mRNA. They nearly always require the RNA-binding protein Hfq for activity. An *hfq* mutant is therefore defective in all sRNA-mediated regulation, which can be useful for determining whether sRNAs are involved in a regulatory phenotype.

MEASURING GENE EXPRESSION: mRNA

There are a wide variety of techniques available for measuring mRNA levels in a cell. They can be divided into **direct** measurements of RNA and **indirect** measurements, and may be able to measure either expression of a single gene or all of the genes in an entire genome.

	Single Gene	Genome-Wide
Direct	qRT-PCR northern blot	RNA-seq DNA microarray
Indirect	transcriptional reporter fusion	--

Techniques to directly measure the amount of an mRNA in a cell include *northern blotting, quantitative reverse transcriptase PCR (qRT-PCR), DNA microarrays*, and *RNA sequencing (RNA-seq)*. qRT-PCR, in which cDNA is produced from mRNA by reverse transcriptase then PCR amplified in the presence of fluorescent dsDNA reporters and quantified with a specialized instrument, is useful for measuring the levels of individual mRNAs. RNA-seq uses next-generation sequencing technologies to measure the concentrations of mRNAs produced from the entire genome (*transcriptomics*). Northern blotting and microarrays are largely obsolete methods of accomplishing the same things, respectively. RNA-seq is probably currently the best technique for assessing transcript abundance, but can rapidly become prohibitively expensive if a lot of different samples need to be analyzed.

A long-established and common technique to indirectly measure the amount of an mRNA in a cell is by using *transcriptional reporter fusions*. These are plasmids (or occasionally, chromosomal insertions) in which the promoters of genes of interest are cloned upstream of genes encoding products that are easy to measure (*reporter genes*). The level of transcription from that promoter is then inferred from the amount of reporter product produced. Commonly used reporters include fluorescent proteins like GFP or mCherry, enzymes with simple colorimetric assays like β-galactosidase (LacZ) or β-glucuronidase (GUS), or luciferase, which produces light.

Figure 4.6. Transcriptional fusions link a promoter of interest with an easily-measured reporter gene (in this case, *gfp*).

The advantages of using transcriptional reporters are that they tend to be the simplest, cheapest way to measure expression from a given promoter. There are several disadvantages, though. First is that they are intrinsically non-physiological, since they are cloned promoters driving non-physiological products from multi-copy plasmids (see **Lecture 5** for more on plasmids). Secondly, high production of reporter gene products may be toxic (fluorescent proteins) or require large amounts of energy (luciferase). Thirdly, cellular growth conditions can affect reporters in ways that they would not affect the actual gene product. Both GFP and luciferase require aerobic conditions, for example, and both LacZ and GUS can be inactivated by oxidative stress (*e.g.* hydrogen peroxide). Finally, the readout from a reporter fusion is always delayed relative to the actual production of the mRNA due to the time necessary to translate the product and, for fluorescent proteins in particular, the time needed for that product to mature into its active form.

Remember that techniques that directly measure the amount of a particular RNA in a population of cells are measuring the **combined** effect of synthesis and stability. This is not true for indirect assays, since the reporters are generally stable transcripts and proteins. Transcriptional fusions typically only measure **synthesis** rates, since the reporter accumulates over time but does not degrade. This also means, of course, that fusions can only reflect the activity of the promoter, which may or may not accurately describe the regulation of the actual mRNA.

DISCUSSION PROBLEM SET #10: TRANSCRIPTIONAL REGULATION

Problem #1

While studying the actinomycete *Streptomyces griseus*, you identify a mutant that does not produce spores or antibiotics. Sequencing reveals that the mutation is a premature stop codon in a sigma factor homolog.

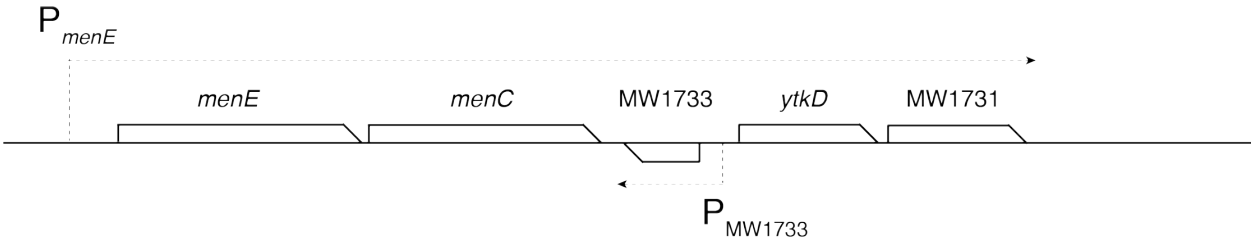
A plate of *S. griseus*, with filamentous colonies typical of *Streptomyces* spp. (GmbH).

Propose a series of experiments to determine which gene or genes in *S. griseus* are regulated by this sigma factor and which ones are required for spore and/or antibiotic production. For each experiment, state:

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #2

An unusual locus called a "non-contiguous operon" has recently been described in *Staphylococcus aureus*, involving 5 genes associated with menaquinone biosynthesis:



The *menE*, *menC*, *ytkD*, and MW1731 genes are all encoded on a single polycistronic mRNA. The MW1733 gene, located between *menC* and *ytkD* on the opposite strand, has its own promoter and is encoded on its own monocistronic mRNA. Both MW1731 and MW1733 encode conserved hypothetical proteins with no known functions.

Under conditions where the MW1733 mRNA is expressed, the amount of *menE-menC-ytkD*-MW1731 mRNA decreases. Consistent with this, replacing P_{MW1733} with a strong constitutive promoter (increasing transcription of MW1733) dramatically reduces the amount of *menE-menC-ytkD*-MW1731 mRNA.

1) Propose a model and testable hypothesis to explain the regulation of the *menE-menC-ytkD*-MW1731 operon by MW1733.

2) Propose a genetic experiment that will test your hypothesis. All standard genetic tools are available to manipulate *S. aureus*. State:

   - the independent and dependent variables
   - both positive and negative controls
   - potential outcomes of your experiment, and how you will interpret them

---

REGULATION OF PROTEIN LEVELS

Similarly to mRNA, protein levels in a cell are controlled at the level of both production and degradation. Translation can be regulated at the *initiation* or *elongation* stages, and protein *stability* is controlled by the activity of protein-degrading enzymes called *proteases*.

Similar considerations must be taken into account when considering protein regulation as for mRNA. Regulation of translation allows the cell to maintain a pool of mRNA that it does not need to transcribe before producing protein, speeding regulatory response. Cellular levels of unstable proteins can be changed much more quickly than stable ones can, and regulated proteolysis is a fast and irreversible way to stop a particular protein from carrying out its function in the cell.

5. Translation initiation. The first step in translation is binding of the 16S ribosomal subunit to the Shine-Dalgarno (S.D.) sequence (also known as a *ribosome binding site* or *RBS*) upstream of the start codon in an mRNA. The sequence of the 3' end of the 16S rRNA (the *anti-Shine-Dalgarno sequence*) of *E. coli* is 5' **ACCUCCU**UA 3', and therefore the consensus sequence for S.D. sites in *E. coli* is 5' AGGAGGU 3', which base pairs with the bolded region of the 16S rRNA. The more similar a gene's RBS is to the consensus, the more efficiently ribosomes will bind to that site, and the more efficiently translation will be initiated. Each gene in a polycistronic mRNA typically has its own RBS, meaning that different genes encoded by the same RNA can be translated at different rates.

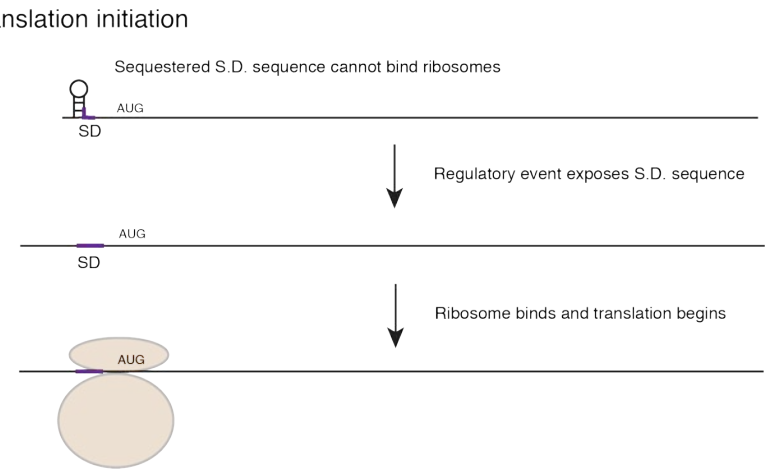


Figure 4.7. An example of how RBS accessibility can regulate translation initiation.

(Note that every organism's 16S rRNA sequence is different, and therefore the consensus RBS in each species is also different, although bacterial anti-S.D. sequences are typically A/C rich.)

Several types of regulation work by changing the *accessibility* of the RBS. There are proteins that compete with ribosomes for binding to mRNAs, and a variety of factors that can change the structure of the mRNA to make it more or less accessible to ribosome binding. These include riboswitches which fold to expose or hide the RBS when bound to metabolites, sRNAs which base pair with the RBS or change the folding of the 5' UTR, and structural features of the mRNA itself which can conceal or expose the RBS in response to changing conditions. A straightforward example of this are thermosensors in which the RBS forms part of a stem-loop structure at low temperature but unfolds at higher temperatures (found, for example in the virulence-associated *prfA* transcript from *Listeria monocytogenes*).

The identity of the start codon also has a strong effect on translation initiation. Most protein-coding gene sequences begin with an AUG codon, but some begin with GUG or UUG and are therefore less efficiently translated.

6. Translation elongation. The rate of elongation by ribosomes is determined by a number of factors, but the most important one for regulating the relative amounts of protein produced from different transcripts is *codon usage*. While most organisms contain tRNAs capable of translating all of the possible amino acid-encoding codons, different tRNAs are not all present in the same concentrations. A gene with many *rare codons* will not be translated efficiently, since the ribosome will need to pause frequently to wait to encounter an appropriate charged tRNA. Different species have different codon usage patterns, but for example, in *E. coli* the arginine codons AGG and AGA are very rare, and an mRNA with these codons will not be well translated (and is likely to be prone to termination or degradation, as described above). There is also some evidence that certain combinations of adjacent codons are particularly poorly translated, possibly due to steric clashes between tRNAs in the A and P sites of the ribosome, but the rules determining what combinations those are have not yet been well defined.

7. Protein stability. Protein stability is determined by cytoplasmic *proteases*, which themselves are tightly regulated to prevent uncontrolled degradation of cellular proteins. They are typically large multi-protein complexes with barrel-like structures. The active sites are inside the barrel, inaccessible to most protein substrates. In *E. coli*, the primary ATP-dependent proteases are ClpP (in complex with either ClpA or ClpX), Lon, HslUV, and FtsH. These are widely conserved, but some other bacteria have different protease complexes, such as the "bacterial proteasome" found in mycobacteria. Each protease has different specific substrates, although they often overlap extensively. Proteases recognize specific signal sequences (*degrons*) in their target proteins, and the presence of degrons in a protein will determine which proteases degrade it. Lon, for example, recognizes aromatic amino acids that are normally buried in the hydrophobic core of proteins, and is therefore an important protease for degrading unfolded or damaged proteins. The ClpA and ClpX adaptor proteins recognize different degrons and target them for degradation by ClpP.

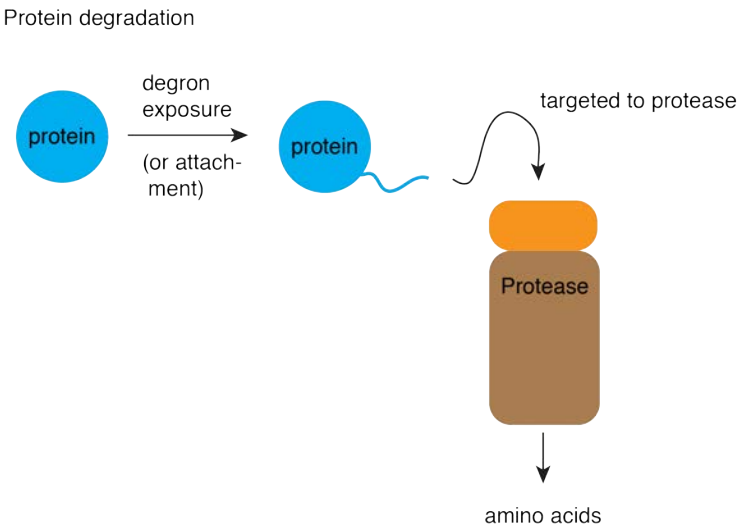


Figure 4.8. Targeting and degradation of proteins by protease complexes.

The N- and C-terminal ends of proteins often contain degron sequences that determine their stability. The "N-end rule" describes a phenomenon in which the N-terminal amino acid(s) of a protein have a dramatic effect on that protein's degradation. Proteins which still have their N-terminal formyl-methionine (fMet) residue are degraded more quickly (probably by FtsH) than those in which fMet has been removed. Proteins with N-terminal leucine, tyrosine, tryptophan, or phenylalanine residues are recognized by ClpS and then degraded by ClpAP. Proteins with N-terminal arginine, lysine, or methionine residues can have N-terminal phenylalanine residues added by the L/F-tRNA-protein transferase, targeting them to the same system. *Endopeptidases* that cut within proteins can generate previously unexposed degrons.

In eukaryotes, proteins destined for degradation by the proteasome are post-translationally modified by addition of ubiquitin. Bacteria do not contain ubiquitin, but actinobacteria (including *Mycobacterium* spp.) have a similar system in which they conjugate the small, intrinsically disordered Pup protein (**p**rokaryotic **u**biquitin-like **p**rotein) to lysine residues in proteins that are then targeted to a protease complex known as the bacterial proteasome. This *pupylation* system is only found in actinobacteria.

MEASURING GENE EXPRESSION: PROTEIN ABUNDANCE

There are multiple techniques available for measuring protein abundance in a cell. As for measurements of transcripts, they can be divided into **direct** measurements of protein and **indirect** measurements, and may be able to measure either expression of a single protein or a large fraction of the *proteome*.

	Single Protein	Proteome-Wide
Direct	western blotting ELISA	mass spectrometry 2-D gels
Indirect	translational reporter fusion	ribosome profiling

Techniques to directly measure the amount of a protein in a cell include *western blotting, mass spectrometry,* and *2-dimensional gel electrophoresis.*

Western blotting (also called *immunoblotting*) relies on antibodies specific to a particular protein to detect and quantify that protein, and is by far the most common method for measuring protein abundance in cells. Whole cell protein extracts can be spotted directly onto membranes or run on polyacrylamide gels and then transferred to membranes before detection by western and quantification by comparison to a standard curve of purified protein. There are numerous variations on using antibodies for protein detection, notably including *ELISA* (**e**nzyme-**l**inked **i**mmuno**s**orbent **a**ssay), which allows high-throughput quantitation of particular proteins in complex biological samples. ELISA kits are available for many mammalian proteins of interest, but are not as commonly used or available for bacterial systems. To

quantify a protein by immunoblotting, you must have a high-quality antibody, which is to say, one that is both sensitive and specific to the protein you want to detect. Generating such antibodies can be very challenging.

To get around this issue, there are a number of *epitope tags* that can be engineered into proteins to allow them to be detected with commercially available high-quality antibodies. Common examples include the HA-tag, derived from a fragment of the influenza virus hemagglutinin protein, the *myc*-tag, derived from human *c-myc* protein, and the FLAG-tag, an entirely artificial antigen with the amino acid sequence DYKDDDDK. Adding such a tag to a protein makes it far easier to detect, but careful controls must be done to make sure that the tag itself does not have an effect on the abundance or activity of the protein. It is also easier to add epitope tags to proteins encoded on plasmids than to chromosomal genes, and effects of plasmid copy number, *etc.*, must be taken into account in such experiments (see **Lecture 5** for more on plasmids).

Proteomics studies attempt to quantify the abundance of all (or a large subset) of the proteins in a cell simultaneously. There are a very wide range of sophisticated methods to do this, but nearly all of them rely on *mass spectrometry* to identify proteins by their molecular weight. The details of how this works are well beyond the scope of this course, but in general, one weakness of this kind of approach is that proteomics is not able to detect low-abundance proteins.

2-D gels are an older "proteomics" method that you will sometimes run across in the literature which uses gel electrophoresis to separate proteins (often from cells fed radioactively-labeled amino acids) on large acrylamide gels in two stages. Proteins are first separated by size, and then by isoelectric point. This technique is very technically challenging and has been almost entirely supplanted by more modern techniques which give higher-quality data for less work, but does allow separation, visualization, and quantification of hundreds of separate proteins simultaneously.

It is possible to indirectly measure the amount of a protein in a cell by using *translational reporter fusions*, which are closely related to the transcriptional fusions discussed earlier in this chapter. The difference is that instead of only including the promoter of the gene of interest, the entire upstream region of that gene, including the RBS and often several codons of the gene itself, is fused to the reporter gene. This makes expression of the reporter dependent on both the transcriptional and translational control signals associated with the gene of interest. All of the same caveats listed for transcriptional fusions apply to translational fusions, and it's important to remember that any translational signals (pause sites, rare codons, *etc.*) found within the coding sequence of the gene will not be present in the fusion.

Figure 4.9. Translational fusions link the promoter, RBS, and first few codons of a gene interest with an easily-measured reporter gene.

A relatively recent development is *ribosome profiling*, an indirect method to infer whole-genome protein translation. In this method, cells are treated with a chemical that reversibly crosslinks ribosomes to mRNA. The ribosome-mRNA complex is purified, treated with RNAse to degrade any RNA that is not protected by ribosome binding, and then the crosslinking is reversed to obtain the pool of protected mRNA fragments. Next-generation sequencing is used to compare the ribosome-bound RNA fragments to the total mRNA pool, which quantifies the proportion of any given mRNA that is ribosome-bound at the moment of measurement. The presence of ribosomes on an mRNA is interpreted as a measure of the amount of translation of that mRNA, and therefore a readout for the amount of that protein being produced. This is a powerful technique that has a lot of potential uses, but is currently expensive and requires considerable technical expertise.

Protein stability can be difficult to measure independently from synthesis. One common approach is to add a translation inhibitor (such as the antibiotic chloramphenicol) to cells and then measure the abundance of a particular protein over time. This has the disadvantage, of course, of having serious effects on cellular physiology in general. A second approach is a *pulse-chase experiment*, which, in its original form, involves adding radioactively labeled amino acids to a cell for a short period of time, then replacing them with unlabeled amino acids and tracking how long the radioactively labeled proteins produced during that "pulse" are maintained in the cell. More sophisticated labeling or immunodetection techniques can be used to focus pulse-chase experiments on a single protein or set of proteins.

REGULATION OF PROTEIN ACTIVITY

8. Protein activity. The amount of a particular protein in a cell does not necessarily determine the level of activity of that protein. Many proteins' activities vary depending on the concentration of metabolites within the cell or can be

regulated by covalent modifications or by physical interactions with other cellular components. These regulatory events can be much more difficult to measure than changes in the amount of mRNA or protein.

Regulation of protein activity is the fastest, most agile mode of regulation available to the cell, since all of the components needed are already present. However, it can be quite wasteful, since producing inactive proteins requires the same resource expenditure as producing active ones does.

The rest of this section is more biochemistry than genetics, but I think it's important to understand how all of these layers work together in living cells. It is certainly possible to isolate mutations in proteins that effect the regulation of protein activity, and we will discuss the interpretation of such mutations in class.

The nature of enzyme kinetics means that the activity of many enzymes varies depending on the concentrations of their substrates and products. The reaction rate of a reversible enzyme operating close to thermodynamic equilibrium can change dramatically or even reverse in response to modest changes in the ratio of substrates and products. Large changes in the activity of an enzyme can result from quite small changes in substrate concentration for enzymes whose K_m (*Michaelis constant*; the concentration of substrate at which reaction rate V is half of V_{max}) is close to the concentration of substrate found in the cell. Many of the enzymes of central metabolism have these properties, and flux through these pathways therefore rapidly responds to changes in conditions without any changes in gene expression. This mode of regulation is, however, somewhat wasteful, since an enzyme operating near its K_m cannot, by definition, be working at its maximum efficiency, and enzymes operating near thermodynamic equilibrium will spend most of their time catalyzing exchange reactions between substrates and products with no net flux in one direction or the other.

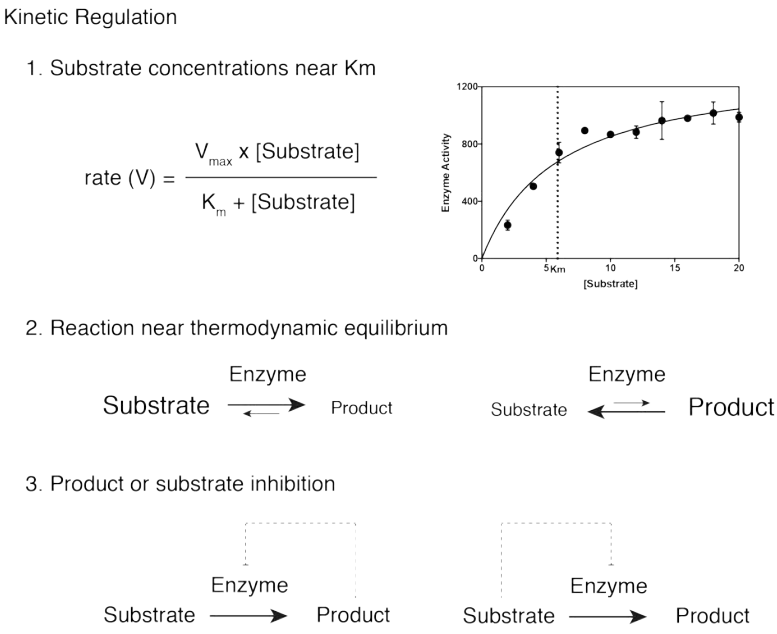


Figure 4.10. A variety of ways enzymes can be regulated by their biochemical properties.

Many enzymes are competitively inhibited by their products and some are inhibited by their substrates, providing additional layers of kinetic control that can affect enzyme activities. This kind of inhibition usually occurs by competition for binding in the active site of the enzyme.

Allostery is a regulatory mechanism by which a molecule controls protein activity by non-covalently binding to a site that is **not** the active site of that protein. *Allosteric effectors* can activate or inhibit protein activity, and are generally thought to function by causing changes in the structure of the protein. Allostery is particularly common in metabolic enzymes. For example, the first enzyme of a complex biosynthetic pathway is often allosterically inhibited by the final product of that pathway, ensuring that the pathway will be inactive when enough of the product is present in the cell.

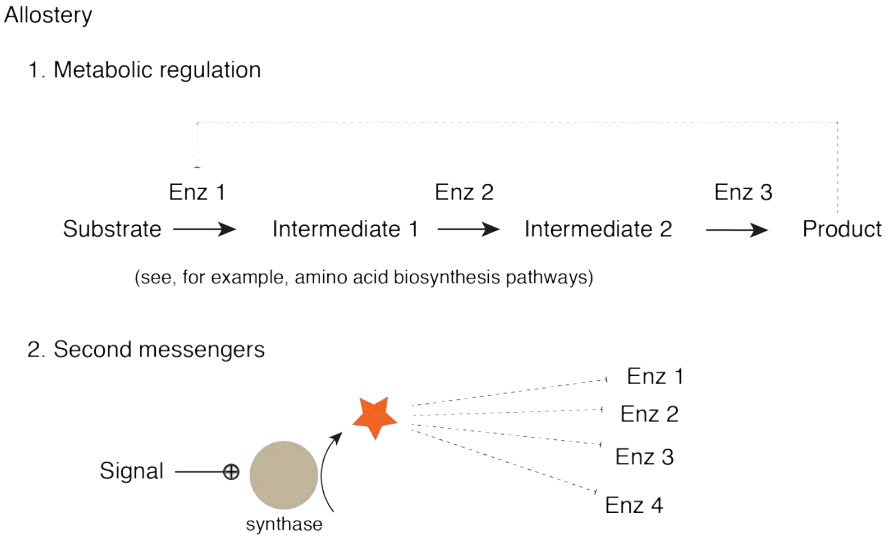


Figure 4.11. Examples of allosteric regulation of protein activity by small molecules.

There are many examples of allosteric regulation by *second messengers*, which are small molecules produced under certain conditions which affect the activity of proteins throughout the cell. Many of these are derived from nucleotides, and important examples include cyclic AMP (cAMP), cyclic di-GMP (c-di-GMP), and guanosine tetraphosphate (ppGpp). Second messengers regulate complex processes in cellular stress response and development, and typically have multiple enzymes controlling their synthesis and degradation.

Allostery is also important for many transcription factors, whose DNA-binding activity or interactions with RNA polymerase are changed when they bind to the specific small molecules they sense. Riboswitches are an example of allosterically-controlled regulators which are not proteins.

Covalent modifications (often called *post-translational modifications* or *PTMs*) can also affect protein activity.

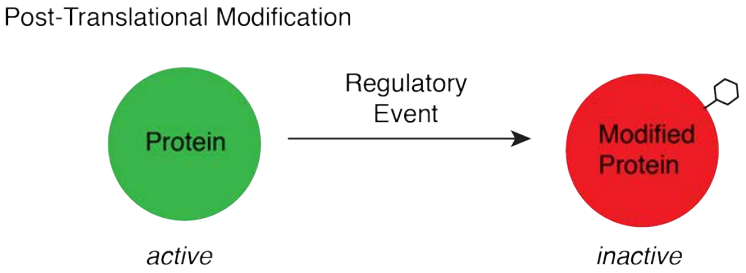
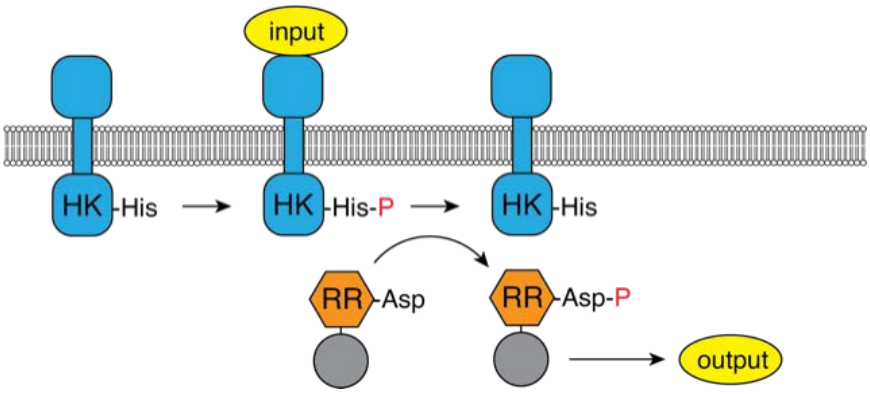


Figure 4.12. Protein activity can be changed by covalent modification of the protein. These modifications are usually the result of the activity of other enzymes, which are typically regulated in turn by one or more of the mechanisms discussed in this chapter.

Serine, threonine, tyrosine, histidine, aspartate, arginine, and (very rarely) cysteine residues can be *phosphorylated*, reversibly adding an ATP- or GTP-derived large negatively charged phosphate group that can dramatically affect protein structure and activity. These are controlled by specific *kinases* and *phosphatases*, which are enzymes that add or remove phosphate groups, respectively, and which often function in signaling pathways.

DISCUSSION PROBLEM SET #11: TWO COMPONENT REGULATORS

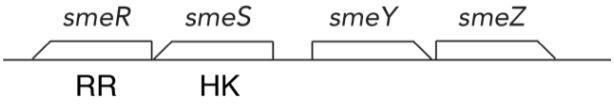
A very common family of regulators in bacteria that use post-translational modifications to regulate protein activity are the so-called **t**wo-**c**omponent **s**ystems (or TCS). An archetypical TCS consists of two proteins, one with a sensor domain fused to a **h**istidine **k**inase (HK) domain and a second protein with an output domain fused with a phosphate-accepting **r**esponse **r**egulator (RR) domain. Many HKs, but not all, are integral membrane proteins with their sensor domains on the outside of the cell, allowing them to detect changes in the exterior environment.



When the sensor domain of the HK detects its cognate input signal (whatever that may be), the HK autophosphorylates itself on a conserved histidine residue, then transfers the phosphate to a conserved aspartate residue in the RR. Phosphorylation affects the activity of the RR, which is often (but not always) a DNA-binding transcription factor. Many "TCS" actually consist of *phosphorelays* with multiple HK and RR domain containing proteins.

TCS are involved in regulating responses to a wide variety of environmental signals. The number of TCS systems encoded by the genome of any particular species depends on the complexity of that species' environment: obligately intracellular pathogens like *Anaplasma* may have none at all, while free-living bacteria with complex developmental programs (like *Myxococcus* or *Nostoc*; see **Lecture 11**) can have close to 200.

Stenotrophomonas maltophilia is ubiquitous in the environment, and occasionally causes opportunistic infections in humans. This is problematic, because *S. maltophilia* is naturally resistant to nearly all antibiotics. While studying this organism, you identify a locus containing homologs of a TCS (encoding a predicted RR called SmeR and a predicted HK called SmeS) divergently transcribed from a predicted small-molecule export pump (SmeYZ).



A **Δ***smeYZ* mutant is **more sensitive** to aminoglycoside antibiotics than the wild-type, so you make mutations in *smeR* and *smeS* and use qRT-PCR to measure the expression of *smeY*, to determine whether SmeRS is involved in regulating antibiotic resistance. You get the following results:

Mutation	*smeY* mRNA
wild-type	+
Δ*smeR*	+
Δ*smeS*	-
Δ*smeRS*	-

Based on these results, propose a model of how SmeRS regulates the *smeYZ* operon. In your model, is SmeR an activator or a repressor? How do you predict that phosphorylation by SmeS affects the activity of SmeR?

N-acylation is the conjugation of acyl groups (acetyl-, propionyl-, succinyl-, *etc.*) to lysine residues by acyltransferases, and plays an important role in controlling metabolic enzymes. To give one illustrative example, in *Salmonella enterica*, acetyl-CoA synthase (Acs) is inactivated by acetylation (by the Pat acetyltransferase, which uses acetyl-CoA as a substrate) when acetyl-CoA levels rise in the cell. When acetyl-CoA levels drop, acetylation of Acs is reversed by the activity of the CobB sirtuin deacetylase, reactivating it for acetyl-CoA synthesis. Proteins also can be reversibly *methylated*, which plays a notable role in controlling the activity of proteins involved in chemotaxis (**Lecture 15**). Note that, of course, the activity of the modification and demodification enzymes for each of these mechanisms must themselves be regulated.

Oxidative modifications of cysteine or methionine residues are common regulators of protein activity in response to changes in redox conditions. Cysteine is normally found in a reduced thiol state (-SH), and can be reversibly oxidized to sulfenic acid (-SOH) or, if two cysteines are in close proximity to each other, to a disulfide bond (-S-S-), either of which can dramatically affect the structure and activity of a protein. Cysteine residues can also be covalently modified by electrophilic compounds. In other proteins, oxidation of methionine to methionine sulfoxide regulates activity, and can be reversed by the activity of methionine sulfoxide reductases.

Most PTMs are reversible, but some regulatory events are irreversible. Cysteine can be oxidized irreversibly to sulfinic ($-SO_2H$) or sulfonic ($-SO_3H$) acid, and the *Bacillus subtilis* transcription factor PerR responds to peroxide stress via the irreversible oxidation of a histidine residue. Presumably, the resulting inactive proteins are subsequently degraded by proteases. Arguably, any modification that leads to proteolysis is an irreversible PTM.

Finally, a protein's activity can be controlled by physical interactions between the protein and other components of the cell, including proteins, ribosomes, DNA, or the cell membrane. This is a kind of allostery, since no covalent modifications of the proteins are involved, and the interaction surface is often not the active site. Some proteins are only active when they are in complex with other proteins, and the formation of these complexes can be regulated by the mechanisms described above. In other cases, proteins can be sequestered in an inactive state by interactions with other cell components, and become active only when they are released from these interactions.

MEASURING GENE EXPRESSION: PROTEIN ACTIVITY

Measuring protein activity is a very direct way to assess the function of a gene product, but can be technically challenging. The techniques required depend on the function of the gene product in question, and very often differ for every protein. There are, however, some general categories of assays which are commonly used, and which I will describe below. One key consideration for protein activity assays is whether they can be performed *in vivo* or if they require the *in vitro* analysis of purified proteins or cell lysates. *In vivo* activity measurements are affected by both how active a given protein is and how abundant it is in the cell, while *in vitro* assays typically allow much simpler normalization for protein abundance.

Enzyme activity assays are the most direct way to assess whether an enzyme is active in a cell or not, but how easy this is to measure depends entirely on the particular enzyme in question. Some enzymes are very simple to assay. Many are not. This is not a biochemistry class, so we won't go into tremendous detail here, but when you're thinking about measuring the activity of an enzyme, consider the following:

1) Is the enzyme cytoplasmic, periplasmic, secreted, or membrane-bound?

Alkaline phosphatase (PhoA) in *E. coli* is a surface-exposed enzyme for which a colorimetric substrate is available, so cells can simply be resuspended in buffer for measurements of PhoA-dependent accumulation of a yellow product. Cytoplasmic enzymes (like LacZ) may need cell permeabilization to allow substrate access. Membrane proteins might or might not retain activity when solubilized with detergents. A secreted protein might need to be concentrated from the spent growth medium of the culture.

2) What are the substrate(s) of the enzyme, and how can you measure them?

How can you measure the conversion of substrate into product? Are they different colors? Do they have different absorbance or fluorescence properties? Can they be separated by chromatography? Are there *substrate analogs* available that are easier to measure than the physiological substrate? (This is what the commonly used indicator substrate X-Gal is; a colorimetric analog of lactose that turns blue when cleaved by LacZ.)

3) Are there other enzymes in the cell that act on the same substrate(s)?

Many cellular enzymes act on common substrates, like ATP or NADH. Trying to measure the activity of this kind of enzyme *in vivo* or in a complex mixture of proteins is not possible due to interference from other enzymes. You will need to purify the protein and study it *in vitro*.

4) How fast does the enzyme act? Do the products accumulate *in vivo*? Are they stable *in vitro*?

Some enzymes catalyze very slow reactions, others turn over in milliseconds. Both situations make it difficult to measure the activity of the enzyme accurately. Some enzyme products are immediately consumed in cells by the next enzyme in the pathway, making it impossible to measure the synthesis of product *in vivo*. If the product of an enzyme is chemically unstable, it will also be difficult to measure *in vitro* unless it can be trapped or stabilized somehow.

5) How stable is the enzyme?

Some purified enzymes are highly stable. Others lose activity rapidly *in vitro*. It's generally a good idea to keep enzymes cold, but some will lose activity when frozen. Reducing agents and metal chelators are often added to enzyme storage buffers to prevent oxidation and inhibit contaminating proteases, respectively, but can be problematic for enzymes with metal cofactors. *In vivo*, you have much less control over protein stability, although of course, the cell itself has more control, which can be a form of regulation in and of itself, as discussed above.

6) How easy is the enzyme to purify?

When purifying proteins for *in vitro* studies, there are potential problems at each of several steps. Can the protein be overexpressed without toxicity? Is the protein soluble? Membrane proteins are never soluble, so there are different detergents available that can be used to try to keep them in solution. Will the protein tolerate having an affinity chromatography tag (e.g. 6xHis or GST) fused to it? If so, does the tag need to be removed after purification in order for the protein to be active? If not, how can you separate the protein from other cellular proteins without a tag?

Allosteric and kinetic regulation is typically easiest to measure for proteins that can be purified and assayed *in vitro*, but are generally challenging, especially if you don't know what the small molecule regulators might be.

The activity of DNA- or RNA-binding proteins (like transcription factors) can be measured by a variety of methods, some of which take advantage of modern high-throughput sequencing technology. Purified DNA binding proteins can easily be mixed with different DNA fragments to see if they interact *in vitro*. The most common method uses gel electrophoresis to separate unbound DNA from protein-bound DNA, which migrates more slowly. This is called an ***e****lectrophoretic* ***m****obility* ***s****hift* ***a****ssay* (EMSA). *ChIP-seq* (**ch**romatin **i**mmuno**p**recipitation **seq**uencing) is an *in vivo* technique to identify all of the genomic binding sites of a DNA binding protein. In a ChIP-seq experiment, cells are treated with a chemical to crosslink proteins and DNA, the DNA is fragmented, then an antibody to a particular DNA-binding protein of interest is used to pull down only those fragments of DNA bound to that protein. The resulting pool of DNA fragments is sequenced with next-generation sequencing and compared to the entire genome sequence.

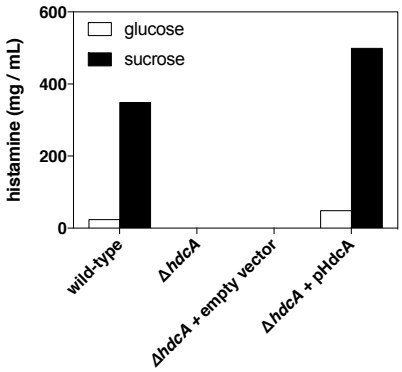
Most PTMs are detectable by mass spectrometry, although some can be detected by other means (antibodies, radioactive tracers, *etc.*). This is, of course, simplest with purified proteins, but can often be done in a high-throughput way in the course of a mass spectrometric proteomics experiment.

Protein-protein interactions can be measured both *in vivo* and *in vitro*, although *in vitro* techniques with purified proteins are much more likely to give quantitative measurements of binding affinity. *Two-hybrid assays* are clever *in vivo* screens (or sometimes selections) that link protein-protein interactions to easily measured phenotypes. They typically involve generating plasmids with fusions between the proteins of interest and two halves of a protein that has a measurable activity when brought within close proximity to each other. This could be an enzyme or, in the case of the most common yeast two-hybrid system, a transcription factor. Libraries of different proteins fused to these kinds of reporters have been used to generate maps of all of the two-way protein-protein interactions in various kinds of cells.

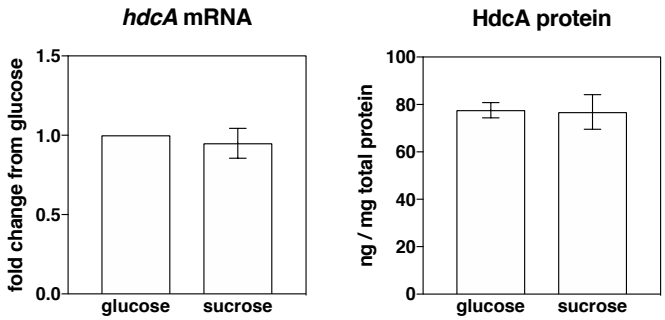
Metabolomics uses mass spectrometry to measure the concentration of molecules in cells that are not proteins or nucleic acids (*metabolites* or "small molecules"). This can be especially useful to assess how much *metabolic flux* is passing through different pathways, by quantifying the amount of each substrate, intermediate, and product that accumulates under different conditions. As protein levels and activity change, this flux will shift, reflecting changes in cellular metabolism.

DISCUSSION PROBLEM SET #12: POST-TRANSCRIPTIONAL REGULATION

You isolate a *Lactobacillus* strain from the microbiome of a mouse and discover that it synthesizes the anti-inflammatory compound histamine when grown in media containing sucrose, but not in media containing glucose. This is dependent on the presence of the *hdcA* gene, which encodes histidine decarboxylase, an enzyme that converts the amino acid histidine to histamine. (Note that in the figure below, "pHdcA" indicates a plasmid or "vector" encoding the *hdcA* gene. We will discuss plasmids in more detail in **Lecture 5**.)



You would like to understand how histamine synthesis is regulated. You do qRT-PCR to measure *hdcA* mRNA levels and quantitative Western blots to measure HdcA protein levels, with the following results:



Based on these data and your knowledge of regulation, propose a hypothesis to explain the regulation of histamine synthesis in response to sucrose. Describe an experiment or series of experiments that would allow you to test your hypothesis. State:

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

CONCLUSIONS

The take-home message from this section is that regulation of bacterial systems can be very complex and that bacteria can regulate multiple steps between gene expression and protein function. Even whole-cell measurements of mRNA levels, protein levels, or enzyme flux only tell part of the story, which is very important to remember when designing and interpreting experiments.

LECTURE 5: PLASMIDS

INTRODUCTION

In this lecture, we will discuss correlation and causation, and how to design experiments that establish causal relationships. Because these kinds of experiments in bacterial systems very often use plasmids, we will also spend considerable time discussing what plasmids are and how they are used in different experimental applications.

SCIENTIFIC PROCESS 5: CORRELATION AND CAUSATION

It is extremely important to distinguish between phenomena that are *correlated* with each other and phenomena that *cause* other phenomena. How can we distinguish between these experimentally?

In the earliest days of microbiology, there was a very serious debate about whether the microbes found in diseased humans and animals were the **cause** of disease or a **symptom** of disease. A great many observations were made and bitter arguments were had over the course of decades, until Robert Koch was finally able to settle the issue with a series of experiments based on what have come to be known as Koch's Postulates:

1. A specific microbe must be found in abundance in all host organisms suffering from the disease, but should not be found in healthy hosts.
2. The microbe must be isolated from a diseased organism and grown in pure culture.
3. The cultured microbe should cause the same disease symptoms when introduced into a healthy host.
4. The microbe isolated from inoculated host must be identical to the originally isolated microbe.

Koch used these postulates to prove that *Bacillus anthracis* was the causative agent of anthrax in 1884, and he and his coworkers spent much of the next 30 years following essentially this process to identify and isolate the bacterial pathogens that cause various diseases (including cholera, diphtheria, tetanus, typhoid fever, pneumonia, and bubonic plague, among others).

The key aspect of Koch's Postulates that allows the scientist to establish *causality* is the careful addition and subtraction of a single independent variable, in this case a specific microbe. In step 1, a correlation between microbe and disease is established, and then steps 2 - 4 demonstrate that adding **only** that microbe to a healthy host organism leads to development of the same disease. Similar principles can be applied to a wide variety of scientific questions.

In 1988, Stanley Falkow proposed a set of "Molecular Koch's Postulates" which he applied to the problem of figuring out whether particular genes contribute to the pathogenesis of disease-causing microbes, and which are more directly relevant to this course. Falkow's Postulates (from Falkow 1988 Rev Infect Dis 10 supp 2: S274-6) are:

1. The phenotype or property under investigation should be associated with pathogenic members of a genus or pathogenic strains of a species.
2. Specific inactivation of the gene(s) associated with the suspected virulence trait should lead to a measurable loss in pathogenicity or virulence.
3. Reversion or allelic replacement of the mutated gene should lead to restoration of pathogenicity.

He also included the alternative steps:

2A. The gene(s) associated with the supposed virulence trait should be isolated by molecular methods. Specific inactivation or deletion of the gene(s) should lead to loss of function in the clone.

3A. The replacement of the modified gene(s) for its allelic counterpart in the strain of origin should lead to loss of function and loss of pathogenicity or virulence. Restoration of pathogenicity should accompany the reintroduction of the wild-type gene(s).

Falkow's argument was that observing a phenotype that went away when a particular gene was deleted and which came back when that gene was reintroduced is strong evidence that the gene in question **causes** the phenotype. Nearly every molecular genetics experiment follows this logic, and Falkow's postulates are still the gold standard for demonstrating genetic causality. (I would argue, of course, that virulence is not the only interesting bacterial phenotype.)

When designing experiments for this class, think carefully about whether the observations and manipulations you are making test correlation or causation, and interpret the results accordingly. Correlations can be very valuable information. Most of the time, however, an experiment that tests causality is superior to one that tests correlation.

DISCUSSION PROBLEM SET #13: CORRELATION AND CAUSATION

Problem #1

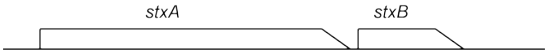
People (and mice) with inflammatory diseases of the gut have different proportions of bacteria in their gut microbiomes than do healthy people. This typically includes higher populations of *E. coli* and lower populations of *Faecalibacterium* species.

Propose an experiment to determine whether inflammation causes changes in bacterial populations in the intestine or *vice versa*. State:

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #2

The lysogenic ϕStx prophage of *Shigella dysenteriae* carries the *stxAB* operon:



Nonsense mutations isolated in either *stxA* or *stxB* prevent *S. dysenteriae* from causing disease, but expressing a wild-type copy of *stxA* in an *stxA* null mutant does **not** restore virulence.

Is *stxA* a virulence factor?

Propose a model to explain these results, and an experiment (based on Falkow's postulates) to test your model. State:

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

---

The rest of this chapter discusses the use of plasmids, one of the main tools for genetic manipulation of microbes. **Lecture 7** will address the technical aspects of constructing and manipulating plasmids.

PLASMIDS

Plasmids are genetic elements that replicate independently from the chromosome. The term "plasmid" was first proposed by Joshua Lederberg, and settled on more or less its current meaning around 1968. It replaced François Jacob and Élie Wollman's term *episome*, which is no longer much used in bacterial genetics, but is still used to describe some autonomously replicating DNA molecules in eukaryotes. You will also often hear plasmids referred to as *vectors*, because they are used to transfer genes from one cell to another. (The analogy is to disease vectors, like ticks or mosquitos. We will discuss gene transfer in more detail in **Lecture 8**.)

Naturally occurring plasmids vary widely in their size and properties. They may be present in a *copy number* anywhere from one per cell up to hundreds. They may carry a wide variety of genes, some of which are involved in maintaining their own copy number or encode conjugation machinery to transfer themselves to other cells (see **Lecture 8**), and some which may provide evolutionary advantages to their host cell. The classic example of this is antibiotic resistance, but there are many other examples. They may be less than 1 kb in size or as large as several Mbp (*megabase pairs* = 1,000,000 bp), at which point it becomes difficult to distinguish clearly between a plasmid and a chromosome. Generally speaking, in such cases, if an essential gene is encoded on the "plasmid", and it has a copy number of one, it is likely to be considered a chromosome. The purple photosynthetic bacterium *Rhodobacter sphaeroides*, for example, has two chromosomes: one of 3.1 Mbp and one of 0.9 Mbp. Like bacterial chromosomes, plasmids are usually, but not always, circular. Linear plasmids have been studied in spirochetes and in *Streptomyces* species, but you are unlikely to encounter them in most labs.

The plasmids we use most often in the lab have generally been engineered to make them easy to work with. They are typically quite small (2 – 8 kb), and usually have high copy numbers, which makes them easy to purify and manipulate. They also nearly always encode at least one antibiotic resistance gene, which makes it possible to select for their

presence in transformed cells. See below for a reasonably comprehensive list of plasmid features and **Lecture 7** for details on the methods by which plasmids can be engineered and manipulated.

COMPLEMENTATION ANALYSIS

In genetic experiments, plasmids are arguably most important for *complementation analysis*. This is an experimental design that uses plasmid-encoded genes to ensure that the interpretation of mutant phenotypes is correct by fulfilling Falkow's postulates. In a complementation experiment, you replace a mutated gene by expressing the wild-type gene from a plasmid, testing to see if this restores the wild-type phenotype. This is an excellent way to test whether a mutant phenotype is due to polar effects on downstream genes.

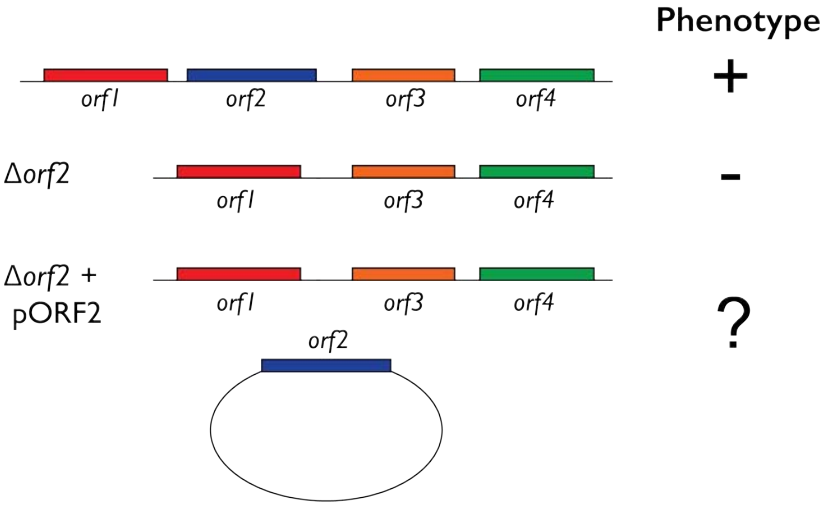


Figure 6.1. Illustration of complementation analysis testing whether deletion of *orf2* is responsible for the "-" phenotype.

As a real-world example (drawn from my dissertation research), let's consider the *bluB* gene of the photosynthetic bacterium *Rhodospirillum rubrum*. I constructed a Δ*bluB* mutant and observed that it grew poorly in the absence of the vitamin B_{12} precursor dimethylbenzimidazole (DMB). This was intriguing, and suggested that BluB might be necessary for DMB synthesis, but how could I be sure that the phenotype I observed was actually due to the deletion of *bluB* and not to polar effects or to an unrelated mutation elsewhere on the chromosome? It took me most of a year to construct the Δ*bluB* mutant (site-directed mutagenesis of *R. rubrum* is not trivial!), so there was certainly a chance that other mutations could have arisen.

The following figure (from Gray & Escalante 2007 PNAS 104:2921), shows how I was able to demonstrate this using a plasmid encoding the *bluB* gene (here indicated as "p*bluB*$^+$"):

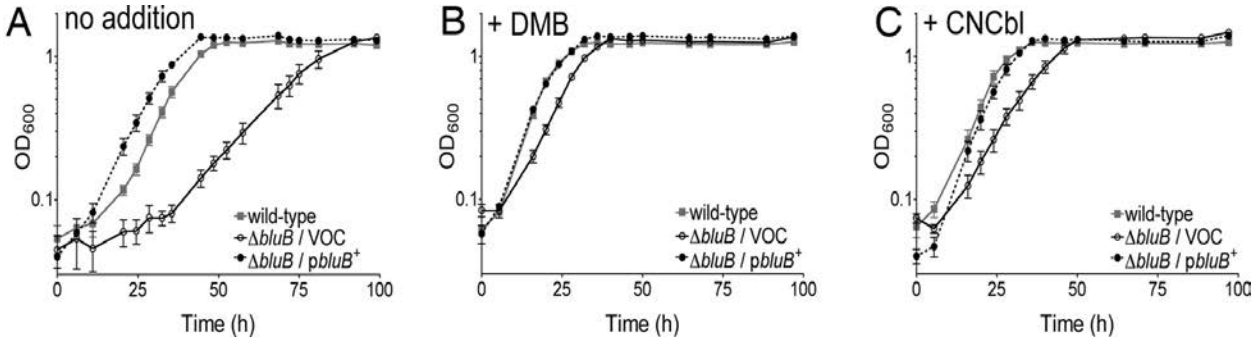


Figure 6.2. BluB is necessary for DMB synthesis. *R. rubrum* wild-type and Δ*bluB* mutant cultures containing the indicated plasmids were grown photosynthetically in the presence of DMB or cyanocobalamin (CNCbl), as indicated.

"CNCbl" is cyanocobalamin, the chemical name for vitamin B_{12}. Observe that the wild-type (grey squares) grows well under all three conditions, but the Δ*bluB* mutant containing an empty plasmid (*vector-only control* or *VOC*, open circles) has a significant growth defect in the absence of DMB or B_{12}. Critically, complementing the mutant with the p*bluB*$^+$ plasmid (black circles) restored growth in the absence of DMB or B_{12} (actually allowing it to grow better than wild-type), demonstrating that it was **only** the lack of *bluB* that was responsible for the observed growth defect phenotype.

Whenever possible, you should complement any mutants you make to confirm that the mutation you have made is actually causing the phenotype you observe, and plasmids are by far the simplest way to do this. This is an especially useful technique when examining mutations that you suspect may have polar effects, since it allows you to distinguish which gene or genes in an operon are responsible for a particular phenotype.

OTHER USES FOR PLASMIDS IN EXPERIMENTS

Far and away the most common use for plasmids in microbiology is as *cloning vectors*. Putting a gene into a plasmid is called *cloning* because it generates many identical copies of the gene in question. The resulting plasmid can then be used for complementation (as with p*bluB*$^+$ in the example above) or for a variety of other purposes.

A gene on a plasmid is far easier to manipulate than a gene on the chromosome (see **Lecture 7** for details). Expression of genes on plasmids can be tightly controlled, depending on the promoter present in the plasmid (see **Lecture 4** for more about promoters and gene expressions), so you can tune the amount of its encoded RNA or protein that is produced. You can do this by replacing the promoter entirely or by using an *inducible promoter* whose activity can be modified by addition of a chemical (often a sugar, in practice).

Protein products encoded on plasmids can be fused to GFP or other proteins for detection or purification. Plasmids can easily be mutated, either *in vivo* or *in vitro*, to rapidly test the effect of specific mutations on gene activity. Randomly mutating a plasmid (for example, by propagating it in a *mutator strain* that lacks DNA repair genes) allows *localized mutagenesis* of a single gene, rather than the entire genome. *Site-directed mutagenesis* is much easier on a plasmid than in the chromosome, and allows very precise experimental designs.

Another common and very useful technique is the construction of *plasmid libraries*, which are pools of plasmids containing many different cloned inserts. A *genomic library* contains random fragments of the entire genome of an organism, and a *cDNA library* contains DNA reverse transcribed from an mRNA preparation. cDNA libraries are useful not only for enriching for genes which are being actively expressed, but also (for libraries derived from eukaryotic organisms) will lack introns. Libraries can be screened to rapidly identify genes encoding specific functions. A *bacterial artificial chromosome* or *BAC* is a plasmid derived from the F plasmid that can be used to clone very large inserts (up to 350 kb). BACs are commonly used in the construction of genomic libraries from eukaryotic organisms whose genes are much larger than those in bacteria (due to the present of introns).

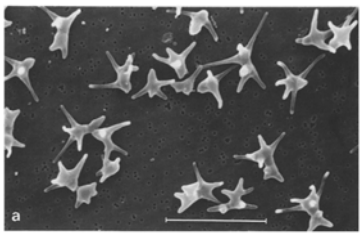
Genomic or cDNA libraries are useful for mutant hunts when you don't know which gene encodes a particular function. They allow you to select or screen for phenotypes that require the **addition** of a gene. In this way, they are kind of the opposite of transposon libraries (**Lecture 3**).

For more information on plasmids, as well as a place you can obtain many useful vectors, the nonprofit plasmid repository Addgene is an excellent source. Their Plasmid Guide is particularly valuable.

DISCUSSION PROBLEM SET #14: USING PLASMIDS IN GENETIC EXPERIMENTS

Problem #1

You are interested in identifying genes involved in determining the cell shape of the structurally complex bacterium *Ancalomicrobium adetum*.



MicroBestiary

Design a genetic experiment using plasmids to identify and confirm which genes are required for cell shape determination in *A. adetum*.

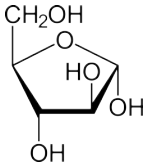
State:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?

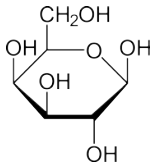
- what are you using plasmids for in your experiment?
- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #2

Arabinose and galactose are dietary sugars that can affect the levels and proportions of bacteria in the gut. *E. coli* can grow on both of these sugars, although the pathways utilized to break them down are very different (we will discuss metabolic pathways in **Lecture 17**).



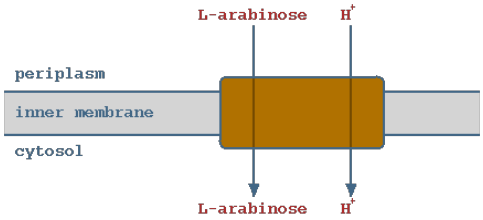
arabinose



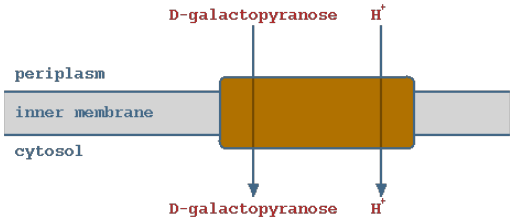
galactose

Surprisingly, the transporters for importing arabinose and galactose (AraE and GalP, respectively) in *E. coli* are 65% identical to each other at the amino acid level. They are both > 470 amino acid proteins. AraE cannot transport galactose and GalP cannot transport arabinose (images below from EcoCyc).

AraE



GalP



Design a genetic experiment using plasmids to identify amino acids involved in substrate specificity in AraE and/or GalP. (Note that neither *araE* nor *galP* are in operons.) State:

- the method of mutagenesis you will use (and why)
- are you using a screen, a selection, or an enrichment to identify relevant mutants?
- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

Problem #3

Bordetella bronchiseptica is a pathogen of pigs whose genome encodes genes for a type 3 secretion system (T3SS). In *Salmonella* and several other pathogens, T3SS systems function to transfer bacterial proteins into host cells, where they contribute to pathogenesis. (We will discuss T3SS and other protein secretion systems in detail in **Lecture 13**.)

It is unfortunately not possible to predict which proteins are T3SS substrates by examining gene or protein sequences *in silico*. However, proteins fused to T3SS substrates can often be secreted, either into the supernatant or into the cytoplasm of mammalian cells.

Design an experiment using plasmids to identify T3SS substrates in *B. bronchiseptica*. State:

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

NAMING CONVENTIONS FOR PLASMIDS

There are no rules set in stone for the naming of plasmids, but some guidelines may be helpful. The names of plasmids nearly always start with a lowercase "p". This is followed by a short name consisting, usually, of capital letters and numbers:

pBR322
pUC18
pET-21b
pBAD18

Like strain identifiers, these letters are often the initials of the researcher(s) who first built or isolated the plasmid. The B and R in pBR322 (one of the original cloning vectors developed for use in *E. coli*) stand for **B**olivar and **R**odriguez, who were postdocs in Herbert Boyer's lab. However, this is not at all a universal rule. The "UC" in pUC18 stands for "**U**niversity of **C**alifornia", the "ET" in pET-21b (one of a very large family of "pET vectors") stands for "**e**xpression by **T**7 RNA polymerase", and the "BAD" in pBAD18 refers to the presence of the arabinose-inducible P_{ara}***BAD*** promoter in that plasmid. The numbers typically refer to the order in which the plasmid was constructed.

This is all very well, but it becomes somewhat more confusing once a researcher begins to manipulate plasmids for their own work. In a very simple case, a scientist may insert a single gene (say, for example, the metabolic gene *mgsA*) into a common plasmid, such as pUC18. In that case, what should the resulting plasmid be called? There are many possibilities, none of which are really wrong, but I do have my own preferences:

Many people will simply append the name of the gene onto the end of the name of the plasmid and call it a day:

pUC18-*mgsA*

This is OK in a simple case, but rapidly becomes unwieldy with more complex constructs. Say, for example, that you were constructing a vector with a GFP fusion to a mutant form of the enzyme MgsA. You might end up with something like the following:

pUC18-GFP-*mgsA*(A745T, G746C, C747T)

While informative, this system is a real nuisance to write and work with, and I personally find it inelegant.

On the other end of the naming spectrum, you might decide that all of this is too complicated and you will just put your initials on every plasmid you build and number them consecutively. In that case, the two plasmids above would just be:

pMJG01
pMJG02

Super simple and concise, and a very common system, but also not very informative, especially since you are likely to be building plasmids for multiple different projects in multiple labs over the course of your research career.

Personally, I like to give plasmids names based on the gene or genes that they encode. I find this strikes a good balance between the two systems above:

pMGSA1
pMGSA2

Concise, but also reasonably informative, in that it's easy to see that these two plasmids encode alleles of the *mgsA* gene. A table in the publication, along with a detailed description of how each plasmid was constructed in the Methods section, is the appropriate place to describe in detail exactly what alleles and constructs of *mgsA* each plasmid encodes. Your PI will probably have their own preferences, but you will certainly see all three of these methods (and more!) used in the literature.

FEATURES AND TYPES OF PLASMIDS

The following list is not a comprehensive list of everything that might be found on a plasmid, but covers the most common and useful features of plasmids you are likely to encounter, along with some practical considerations for their use.

origin of replication (*ori* or *oriC*): Every plasmid will have an origin of replication, which controls the ability of the plasmid to replicate within the cell. There are many different types, each of which is associated with a characteristic *copy number* (the number of plasmids per cell) and *host range* (the species in which the plasmid will replicate). Origins of replication that work in Gram-negative bacteria will often not work in Gram-positive bacteria, for example. *Shuttle*

vectors will replicate in more than one species, and sometimes have separate origins of replication for each species. A *suicide vector* is a plasmid that can be introduced into a cell but does **not** have a functioning origin of replication for that species (useful, for example, in allelic exchange mutagenesis procedures – see **Lecture 8**). This can be accomplished either by using a plasmid with an origin that does not function in the recipient species or by using a vector with a *temperature-sensitive origin*, which will allow replication at a *permissive temperature* (often 30°C) but not at a *restrictive temperature* (often 42°C).

If you want to put more than one plasmid into a single strain of bacteria, you need to ensure that they have **different** origins of replication. Two plasmids with the same origin will be *incompatible*, so origins of replication are sometimes called *compatibility groups*. A single plasmid **cannot** contain two origins of replication that function in the same organism.

selectable marker: A gene encoding a product which allows you to select for cells containing the plasmid. This is most often an antibiotic resistance gene, in which case only bacteria with the plasmid will survive in media containing that antibiotic. Such a plasmid can only be used in a strain that is otherwise sensitive to that antibiotic, and if you want to have more than one plasmid in a strain, they must have **different** selectable markers. Plasmids may carry more than one selectable marker. Most plasmids we use in the laboratory are unstable and are lost fairly quickly in the absence of selection (see Table 2.2), and therefore you should always include the appropriate antibiotics in media used to grow strains containing plasmids. (Natural plasmids are typically much more stably maintained.)

The phenotypes conferred by antibiotic resistance markers are abbreviated in the form "Ab^R", as opposed to cells that are sensitive to that antibiotic, which are sometimes indicated as "Ab^S". The abbreviations for some common laboratory antibiotics are listed below.

ampicillin = Ap, Amp
chloramphenicol = Cm, Cam
kanamycin = Kn, Kan
tetracycline = Tc, Tet
streptomycin = Sm, Str
spectinomycin = Sp
nalidixic acid = Nx
gentamycin = Gm
rifampicin = Rif
erythromycin = Em, Erm

(A practical note that may save you some headaches: when making antibiotic stock solutions, be sure to look up what concentration and solvent are appropriate. Not all antibiotics are water-soluble. Cm, for example, must be dissolved in 100% ethanol, and Tc will only dissolve in 70% ethanol.)

counter-selectable marker: A gene encoding a product which allows you to select for cells that **don't** contain the plasmid. These typically encode conditionally toxic gene products, and the most common is the *sacB* gene, which confers sucrose sensitivity on many Gram-negative bacteria. This is useful for allelic exchange procedures, for example, which we will discuss in **Lecture 8**.

A potentially useful side note is that it is possible (at least in *Salmonella* and *E. coli*) to select **against** tetracycline resistance, allowing Tc^R to serve as both a selectable **and** a counter-selectable marker. The method was developed by Barry Bochner, and works because the *tetA* gene makes cells resistant to tetracycline by changing the properties of the cell membrane. Those same changes make the cells **sensitive** to killing by fusaric acid. Maloy and Nunn (1981) J Bacteriol 145(2):1110-2 and Li *et al.* (2013) Nucl Acids Res 41(22):e204 expand on this method and describe media for using it in *E. coli*, should that ever happen to be useful to you.

Multiple Cloning Site (MCS): A small region of the plasmid with several closely spaced restriction sites to allow simplified insertion of cloned genes (see **Lecture 7**).

promoter: A DNA sequence which allows expression of genes on the plasmid. While every gene on the plasmid must have a promoter to be expressed, in most cloning vectors there is a separate promoter directed at the MCS, so that inserted genes will be expressed from that promoter.

Constitutive promoters express genes at a constant level, while *inducible promoters* can be turned on and off or have their level of expression tuned by addition of *inducers* to the growth medium. Common inducible promoters used in plasmids include *lac* operon-derived promoters that respond to lactose or unnatural lactose analogs like *IPTG* (**i**so**p**ropyl β-D-1-**t**hio**g**alactopyranoside) and promoters controlled by other sugars, like arabinose or xylose. Many *overexpression vectors* used to produce proteins for purification contain a very strong promoter from the T7

bacteriophage which, when provided with T7 RNA polymerase (usually on the chromosome of specialized *overexpression strains* and itself controlled by a *lac* promoter), drives **extremely** high levels of gene expression. It is not uncommon for a protein expressed from a T7 promoter to make up 50% of the total protein within a cell.

Like origins of replication, not all promoters work equally well in all species, and you must use a promoter compatible with the organism you are working with. In lactic acid bacteria, for example, the most common inducible promoter in use is activated by the polycyclic peptide nisin.

ribosome binding site (RBS): also called the *Shine-Dalgarno sequence* after John Shine and Lynn Dalgarno, the Australian scientists who identified it, this is a short AG-rich sequence required for ribosomes to interact with mRNA. In order for a protein to be translated, there must be an RBS between the promoter and the start codon, and different RBS sequences may lead to more or less efficient translation. Some plasmids include an RBS, and some do not. In the latter case, you must include an RBS in gene sequences you clone in order for them to be translated.

terminator: a sequence which stops transcription, often included on the opposite side of the MCS from a promoter to prevent *read-through transcription* of other genes on the plasmid from that promoter. These usually consist of stable DNA secondary structures (*hairpins*) that block the progress of RNA polymerase.

fusion proteins / tags: Some plasmids are designed to allow inserted gene sequences to be linked to sequences already encoded on the plasmid. This results in a *chimeric protein* or *protein fusion* with sequence derived from both your inserted gene and another protein. These can be used for purification of the fused protein (as with the 6xHistidine or GST tags), changing the physical properties of the protein (fusion with the maltose binding protein MBP increases the solubility of a protein, and fusion with a *signal sequence* can target a protein for secretion out of the cell), or for easier detection of the expressed protein either *in vivo* or *in vitro* (as with green fluorescent protein, the easily assayed enzyme β-galactosidase, or the FLAG tag, which is detectable with commercially-available antibodies). To use this kind of plasmid, you must make sure that your gene of interest is *in-frame* with the fusion protein (that is, forms a single continuous open reading frame) and that you do not include a stop codon in between your cloned sequence and the fusion protein (if it is a C-terminal protein fusion).

origin of transfer (*oriT*): A DNA sequence allowing the plasmid to be *mobilized* by conjugation. The **tra**nsfer genes (or *tra* functions) necessary for mobilization may be encoded on the plasmid with the *oriT*, on a separate plasmid, or on the chromosome. See **Lecture 8** for more on conjugation.

f1 origin: Many older plasmids will contain an origin of replication derived from the filamentous phage f1, and are referred to as *phagemids*. This is a site that allows the plasmid to be packaged as long repeating single-stranded DNA molecules when the host bacterium is infected with f1. This was useful when DNA sequencing technologies required large amounts of single-stranded DNA, but is now largely obsolete.

***cos* sites**: Like the f1 origin, *cos* sites are sequences that allow plasmids to be packaged into phage particles, in this case those of λ phage. Plasmids containing *cos* sites are called *cosmids*, and can contain much larger DNA sequences than is practical in normal plasmids, limited only by the size of the DNA molecule which will fit in a λ phage capsid (up to about 45 kb). Like f1 phagemids, cosmids are much less commonly used now than they used to be.

Typically, when working with a plasmid, you will have a *plasmid map*, which is a drawing showing the location of the various features of that plasmid. You will probably also have a sequence file with the exact DNA sequence of the vector. If you are unable to find a map for the plasmid you're working with, you should go ahead and draw it yourself, based on the sequence. There is specialized software intended for just this kind of thing. My lab uses SnapGene, but there are many options, including a free web-based tool provided by the company Genscript.

The following figure illustrates what some plasmid maps might look like:

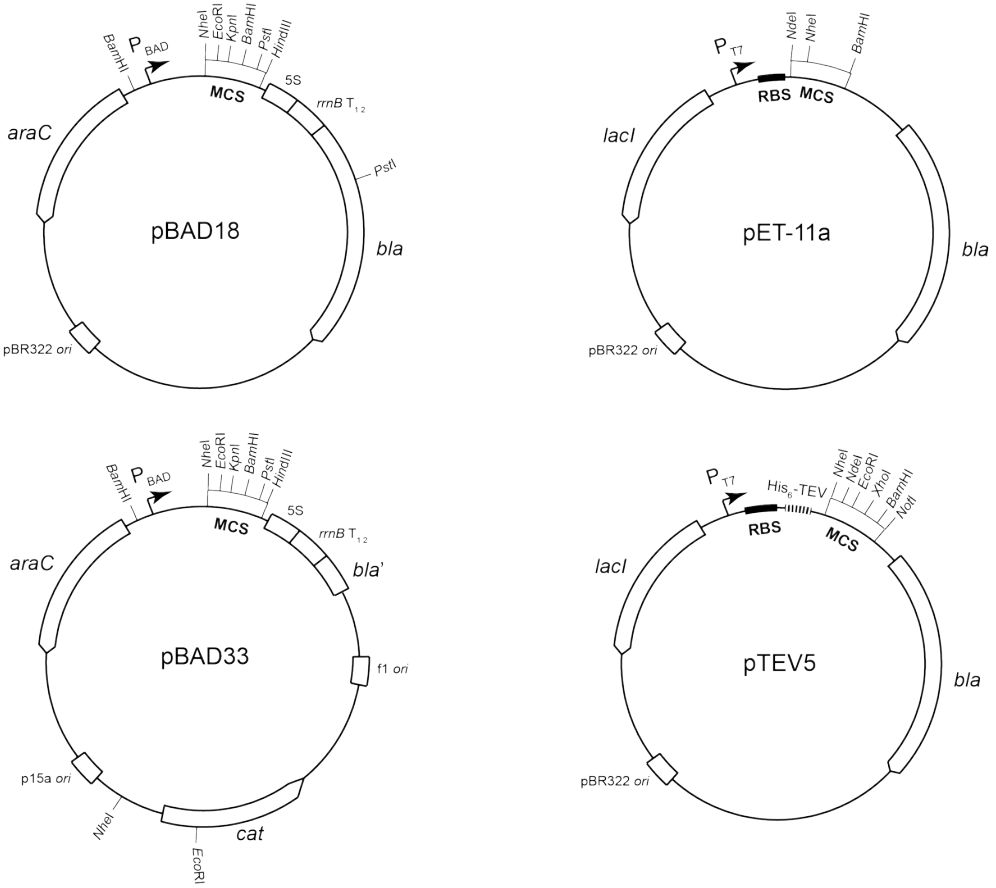


Figure 6.3. Sample plasmid maps.

• pBAD18 and pBAD33, from Guzman *et al.* (1995) J Bacteriol 177(14):4121, are common vectors used for expressing genes in *E. coli* and other Gram-negative bacteria. They have the P_{BAD} promoter and the *araC* gene encoding the arabinose-sensitive regulator of that promoter, so that expression of genes inserted into the MCS can be controlled by addition of arabinose to the medium. They also have the "5S *rrnB* $T_{1\,2}$" transcriptional terminator immediately after the MCS, to prevent read-through transcription.

pBAD18 carries the *bla* gene, encoding **β-la**ctamase, which breaks down the antibiotic ampicillin, and the high copy number pBR322 origin of replication. In contrast, pBAD33 carries the *cat* gene, encoding a protein that protects the cell against chloramphenicol, and a lower-copy number p15a origin of replication. It also has an f1 phage origin, making this technically a phagemid. Because pBAD18 and pBAD33 have different origins of replication **and** different antibiotic resistance markers, both of these plasmids can coexist in a single bacterial cell.

• pET-11a (available from EMD Millipore) has a much more limited set of restriction sites in its MCS, but includes an RBS, which the pBAD vectors do not. It also encodes the IPTG-sensing transcription factor LacI and has a powerful T7 promoter to drive very high-level expression of cloned genes. This only works, of course, in strains containing the T7 RNA polymerase (such as the overexpression strain BL21[DE3]).

• pTEV5, from Rocco *et al.* (2008) Plasmid 59(3): 231-239, is a similar protein overexpression vector, but has a much improved MCS and incorporates an N-terminal, TEV protease-cleavable 6xHis purification tag into proteins produced from this plasmid. This allows easy purification of the tagged protein, and then removal of the tag from that protein by addition of the sequence-specific TEV protease.

Note that restriction sites found in the MCS may cut elsewhere in the plasmid (as indicated in the pBAD vectors), so you should take care that any sites you plan to use for cloning (see **Lecture 7**) only cut once. Not all possible restriction sites are included in most maps, which is where having the complete sequence becomes helpful. (RestrictionMapper is an online tool that searches DNA sequences for restriction sites.)

LECTURE 6: CRITICAL READING (MUTAGENESIS AND MUTANT HUNTS)

INTRODUCTION & EXPECTATIONS

In today's class, we will discuss a scientific paper from the recent literature in detail, to see how the principles of bacterial genetics we've discussed have been applied to an actual scientific problem. This kind of deep dive into a paper is very valuable for thinking about experimental design and rigor, as well as keeping on top of the current literature. It's also good practice for peer reviewing manuscripts. You will probably participate in *journal clubs* that function more or less this way throughout your career.

To prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
    - What is the central **question** of this paper?
    - Is the experimental design clear and appropriate to address that question?
    - Do you understand the methods used?
    - Are the data clearly presented, with appropriate statistics?
    - Do you agree with the conclusions the authors came to based on their data?
    - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

> Eickhoff & Bassler (2020) "*Vibrio fischeri* siderophore production drives competitive exclusion during dual species growth" Mol Microbiol 114:244-261.

As we discussed in **Lecture 1**, you can retrieve this paper from a number of databases. Either PubMed or Google Scholar is probably the simplest option, although since this is a recent paper in a non-open access journal, you will probably need to be logged into a UAB network to access the full-length document.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including the Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

LECTURE 7: PRINCIPLES OF GENETIC ENGINEERING

INTRODUCTION

Constructing a DNA molecule with sequences from two or more different organisms is called *recombinant DNA technology*, is the basis of all modern biotechnology, and most frequently involves the use of plasmids. As we saw in the last chapter, plasmids are a key tool for molecular genetics of bacteria. This chapter is a discussion of the principles and techniques used to engineer plasmids.

PRINCIPLES OF GENETIC ENGINEERING FOR MOLECULAR BIOLOGY

Biotechnology and molecular genetics depend on being able to manipulate the genetic material of cells. From a practical standpoint, this means that your success as a molecular biologist hinges on understanding the technology for constructing and changing the sequence of DNA molecules. At first glance this seems like a very daunting proposition. There are hundreds of different protocols for manipulating DNA, some of which have many steps and seem very complicated or specialized, and new techniques are being invented all the time. However, all of these techniques are built up from a framework of only a few different fundamental procedures. The goal of this chapter is to provide a practical resource that will explain what those building blocks are and how they can be combined to build a DNA molecule of almost any desired sequence.

First, I will describe the six fundamental procedures that make up all molecular genetics protocols and the current technology for carrying out these procedures both on purified DNA *in vitro* and on living cells *in vivo*. Then I will show how these procedures are combined to construct and modify DNA molecules, using common lab techniques as specific examples. I will include notes with links to resources describing specific technologies in detail for readers who want to explore them in more depth. I will also try to highlight common mistakes and points of confusion.

By the end of this section, you should be able to understand any molecular biology protocol by breaking it down to its basic building blocks. I will focus here on molecular genetics in bacterial systems, but the fundamental concepts apply to all molecular biology, and in fact almost all DNA molecular construction is done in *E. coli*, where the most highly developed tools are available. The resulting DNA products are then transferred to other species of interest.

THE SIX FUNDAMENTAL PROCEDURES OF MOLECULAR BIOLOGY

All of molecular biology is based on carrying out combinations of six different procedures on DNA: *reading*, *writing*, *copying*, *cutting*, *pasting*, and *swapping* sequences, either *in vitro* or *in vivo*. The following table lists these procedures, along with the method(s) we use to accomplish them (some of which we will not discuss until **Lecture 8**).

	In vitro	***In vivo***
Read	DNA sequencing	--
Write	Oligonucleotide synthesis	--
Copy	PCR	Replication
Cut	Nucleases	CRISPR
Paste	Ligase	DNA nick repair
Swap	--	Homologous recombination

In the text and figures below, appropriate icons will be used to indicate each type of procedure.

Read

The technology for determining the nucleotide sequence of DNA molecules continues to advance rapidly, and it is now straightforward and relatively inexpensive to sequence DNA up to and far beyond the length of an organism's genome.

For routine sequencing of short sections of DNA molecules (< 1000 bp), *Sanger sequencing* is the most common and cheapest method. A variety of so-called "*next-generation sequencing*" (NGS) and emerging "third generation" sequencing technologies exist that allow us to sequence whole genomes and complex mixtures of DNA from many organisms (*metagenomes*), usually by computationally aligning millions of very short sequence reads. Practically speaking, in most labs you will not do your own DNA sequencing, but will outsource it to a company or university core facility.

The Microbial Genome Sequencing Center in Pittsburgh provides extremely affordable whole-genome sequencing (about $100 for a bacterial genome and associated bioinformatic analysis), and my lab has had a very good experience working with them.

Every molecular biology protocol ends with a DNA sequencing *read* step to confirm that the correct DNA molecule has been constructed. DNA is always extracted from the organism before sequencing, so the read step always happens *in vitro*.

Write

It is possible to chemically synthesize DNA molecules with a desired sequence *in vitro*, but current methods typically only allow accurate synthesis of single-stranded DNA chains up to about 100 nucleotides long. These are called *oligonucleotides* (or "oligos"), and are relatively inexpensive (as little as 20 cents per nucleotide, if ordered in bulk).

Since even a single gene is usually hundreds or thousands of nucleotides long, it is not typically practical to make large and complex DNA molecules from scratch (although see "Gene Synthesis" below, for a common, commercially available, and not **too** expensive way to do so, when needed). There is no equivalent technique for generating entirely new DNA sequences from scratch *in vivo*.

Almost all of the protocols we'll discuss below begin with an oligonucleotide synthesis *write* step. Very few labs have the specialized equipment to synthesize their own oligos, and you will typically order them from a company.

There is lots of software available to help you design primers for different applications, often provided by the companies that want to sell you the oligos. Here are a few free web-based options you may find useful:

- Primer-BLAST
- IDT Primer Quest
- NEBuilder
- WebPrimer (*this is what I use, mostly*)

Copy

In order to sequence or manipulate DNA, we typically need to make many copies of the specific DNA molecule of interest. We can do this either *in vivo* or *in vitro*. In either case, DNA polymerase is the essential enzyme for copying DNA sequences, and uses an existing DNA molecule as a template to synthesize new DNA.

If the DNA molecule in question has an origin of replication, it is simple to grow large amounts cells containing that DNA and allow normal cellular replication and reproduction to generate copies *in vivo*. This is useful for generating large amounts of chromosomal DNA and plasmids.

One of the key technologies of molecular biology is *PCR* (the **p**olymerase **c**hain **r**eaction), a technique for copying DNA sequences *in vitro*. (See this Wikipedia page for a detailed explanation.) For PCR, short oligos (15 – 30 nucleotides) called *primers* are designed which are complementary to the sequence of a double-stranded DNA template molecule. These are annealed to the template and then extended with purified DNA polymerase. If two primers are used which are directed towards each other on the same template and multiple cycles of annealing and extension are repeated, the result is exponential copying of the sequence between the two primers. PCR works best on relatively short sequences of a few hundred to a couple of thousand base pairs, but can be used to amplify linear DNA molecules up to about 10,000 base pairs long. Many different thermostable DNA polymerases are available for use in PCR, some of which are especially good at amplifying long templates or are engineered to make fewer errors during amplification.

Since we can only *write* very short DNA sequences, constructing complex DNA molecules always involves at least one *copy* step. In order to get enough DNA to *read* the resulting sequence, it is also essentially always necessary to *copy* the product of your protocol *in vivo*.

Cut

A key step in many genetic engineering protocols is *cutting* DNA molecules into smaller fragments. *Nucleases* are enzymes that cleave DNA molecules by breaking the bonds between nucleotides. The most common and useful nucleases for molecular biology are those that cleave DNA only at specific sequences, but some protocols use nucleases with less specificity for particular purposes. This is important in Gibson Assembly, for example, a protocol you will examine later in this chapter.

Purified nucleases are used to cut DNA molecules *in vitro*. *Restriction enzymes*, the most commonly used type, are nucleases that recognize specific short DNA sequences (usually 4 – 8 base pairs long, called "*restriction sites*") and introduce a double-strand break in the DNA at or near that recognition sequence. Hundreds of different restriction enzymes with different recognition sequence specificities are commercially available. See, for example, the lists of enzymes available from ThermoFisher or New England Biolabs. RestrictionMapper is extremely helpful for determining which restriction enzymes will cut a given DNA sequence.

Purified restriction enzymes are used in many protocols to cut DNA molecules into defined fragments. The names of restriction enzymes are based on the species they were originally isolated from. *Eco*RI and *Eco*RV are the first and fifth restriction enzymes isolated from *E. coli* strain R, for example, and *Hind*III was isolated from *Haemophilus influenza* strain Rd.

Since we are able to *read* DNA sequences, we can reliably predict where a restriction enzyme will cut any given DNA molecule. Some restriction enzymes break both DNA strands at the same base pair, generating a "blunt ended" cut. Others, which are typically more useful for molecular biology, break the two strands in a staggered way, generating "sticky ends" with short single-stranded overhangs at the end of the cleaved DNA molecule, as illustrated below:

DNA fragment 1:

CATATGTTTAAAAAATCTGTTTTATTTGCAACACTATTATCTGGCGTTATGGCATTTTCCACCAATGCAGATGATAAAATAATTCTGATAA**GGATCC**
GTATACAAATTTTTTAGACAAAATAAACGTTGTGATAATAGACCGCAATACCGTAAAAGGTGGTTACGTCTACTATTTTATTAAGACTATT**CCTAGG**

DNA fragment 2 (part of plasmid pET-11a):

…GAAGGAGATATA**CATATG**GCTAGCATGACTGGTGGACAGCAAATGGGTCGC**GGATCC**GGCTGCTAACAAA…
…CTTCCTCTATAT**GTATAC**CGATCGTACTGACCACCTGTCGTTTACCCAGCG**CCTAGG**CCGACGATTGTTT…

DNA fragment 1, digested with *Nde*I and *Bam*HI:

TATGTTTAAAAAATCTGTTTTATTTGCAACACTATTATCTGGCGTTATGGCATTTTCCACCAATGCAGATGATAAAATAATTCTGATAA**G**
ACAAATTTTTTAGACAAAATAAACGTTGTGATAATAGACCGCAATACCGTAAAAGGTGGTTACGTCTACTATTTTATTAAGACTATT**CCTAG**

DNA fragment 2 (part of plasmid pET-11a), digested with *Nde*I and *Bam*HI:

…GAAGGAGATATA**CA** **TATG**GCTAGCATGACTGGTGGACAGCAAATGGGTCGC**G** **GATCC**GGCTGCTAACAAA…
…CTTCCTCTATAT**GTAT** **AC**CGATCGTACTGACCACCTGTCGTTTACCCAGCG**CCTAG** **G**CCGACGATTGTTT…

The most recent major addition to the molecular biology tool kit is a technology for cutting DNA *in vivo*. *CRISPR* (which stands for "**c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats", referring to the genomic context in which the relevant genes were discovered) takes advantage of a nuclease called Cas9 that can be targeted to a specific DNA sequence *in vivo* by a short guide RNA. This confers great specificity to Cas9 and allows it to introduce double-stranded DNA breaks at very precise locations in the chromosomes of living cells. Applications of CRISPR are in very active development, and are allowing previously impossible genetic engineering procedures in a wide range of species. The biotech company Genscript has a very nice summary of the history and uses of CRISPR available for download here, and we will discuss applications of CRISPR in bacterial genetics in **Lecture 8**.

Both restriction enzymes and CRISPR are derived from naturally occurring systems bacteria use to defend themselves against infection by viruses. Since restriction enzyme recognition sites are short and occur commonly in the genomes of the bacteria encoding those enzymes, each restriction enzyme is paired *in vivo* with a *DNA methylase* that methylates the recognition site and protects the host cell's DNA against restriction. Unmethylated DNA, such as the genome of an invading virus, is therefore cut by the restriction enzyme, preventing infection. (Infection by phage is what is "restricted" by restriction enzymes.) Practically speaking, this means that we can protect a DNA molecule from digestion by a particular restriction enzyme by treating it with the corresponding methylase *in vitro* or by copying it *in vivo* in a strain expressing that methylase. PCR products are always unmethylated, which is important to remember, and is useful in some cloning and mutagenesis procedures.

CRISPR targets longer (~ 20 bp), less common DNA sequences, and bacteria defend themselves against their own CRISPR systems by simply not encoding the target sequences anywhere in their genomes, or occasionally by encoding CRISPR-inhibiting proteins.

Paste

Recombinant DNA technology depends on being able to *paste* two or more DNA molecules together into a single molecule. This reaction is catalyzed by an enzyme called *DNA ligase*.

Ligase forms a phosphodiester bond between the 5' phosphate of one linear single DNA strand and the 3' hydroxyl of another. *In vivo*, this is part of a cell's DNA repair mechanism, and repairs "nicks" or breaks in a single strand of a double-stranded DNA molecule. If a molecular biology protocol results in a DNA molecule with a single nick or a few

widely-spaced nicks, this will be repaired when that molecule replicates *in vivo*. Bacteria do not typically ligate double strand breaks *in vivo*, although this does happen in eukaryotic cells (where it is called "non-homologous end-joining").

Generating recombinant DNA *in vitro* with purified ligase is more versatile. The most common enzyme used is the ligase from the bacteriophage T4. At high enzyme concentrations, T4 ligase will join blunt-ended linear double stranded DNA fragments into linear or circular products. Sticky-ended DNA fragments allow more precision, since fragments with complementary sticky ends will anneal to each other, in essence creating loosely fused DNA molecules with two nearby nicks, one on each strand. Ligase efficiently forms phosphodiester bonds to repair these nicks, allowing construction of composite DNA molecules with their components joined in a particular orientation and order.

Insert, digested with *Nde*I and *Bam*HI:

```
TATGTTTAAAAAATCTGTTTTATTTGCAACACTATTATCTGGCGTTATGGCATTTTCCACCAATGCAGATGATAAAATAATTCTGATAAG
  ACAAATTTTTTAGACAAAATAAACGTTGTGATAATAGACCGCAATACCGTAAAAGGTGGTTACGTCTACTATTTTATTAAGACTATTCCTAG
```

Vector, digested with *Nde*I and *Bam*HI:

```
…GAAGGAGATATACA            GATCCGGCTGCTAACAAA…
…CTTCCTCTATATGTAT              GCCGACGATTGTTT…
```

Ligated product:

```
…GAAGGAGATATACATATGTTTAAAAAATCTGTTTTATTTGCAA…CACCAATGCAGATGATAAAATAATTCTGATAAGGATCCGGCTGCTAACAAA…
…CTTCCTCTATATGTATACAAATTTTTTAGACAAAATAAACGTT…GTGGTTACGTCTACTATTTTATTAAGACTATTCCTAGGCCGACGATTGTTT…
```

Essentially all genetic engineering protocols involve a *paste* step, although for some procedures that step is relatively invisible, since it happens *in vivo* at the same time as the final *copy* step before sequencing.

Swap

Finally, in some protocols you will take advantage of the ability of cells to *swap* sequences from one DNA molecule to another. This is dependent on another DNA repair mechanism called *homologous recombination*, and only occurs *in vivo*. We will discuss the mechanism and use of recombination in **Lecture 8**.

Beware of confusing terminology here: "recombinant DNA" and "DNA recombination" are not the same thing!

For an alternative discussion of the principles of molecular biology, from a different perspective and with some more technical details, see this very nice article from Addgene.

EXAMPLES OF COMMON MOLECULAR BIOLOGY PROTOCOLS

In this section, I will break down a series of protocols into their component steps, both in outline and graphical form. I will proceed from fairly simple procedures to more complex ones. Notice, however, that many of the steps are the same for all or almost all of the protocols. For example, essentially every protocol ends with an *in vivo copy* step and an *in vitro read* step to confirm the sequence of your engineered DNA product.

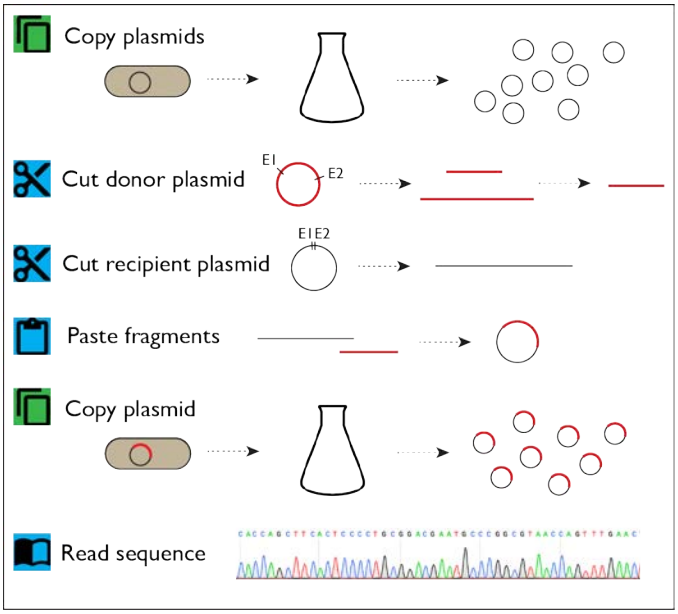
This is the key principle I want you to take away from this chapter: **complicated protocols are just combinations of simple procedures**.

Molecular biology is essentially a creative endeavor. Like any artist, you are using the tools at your disposal to solve problems in a creative way. This is your toolbox.

Icons on a **green** background indicate *in vivo* steps, while those on a **blue** background indicate *in vitro* steps.

SUBCLONING

Subcloning is a protocol in which a DNA fragment from one plasmid is moved into another plasmid. Most plasmids contain arrays of defined restriction enzyme recognition sites called multiple cloning sites (see Figure 6.3) to make this kind of procedure straightforward.



Protocol:

1. *Copy – in vivo*
- grow cells containing donor and recipient plasmids to make large amounts of each

2. *Cut – in vitro*
- digest the donor plasmid with restriction enzymes that cut on either side of the gene, ideally two different enzymes that each create a different sticky end
- optionally, separate the resulting fragments on a gel and purify the fragment you wish to subclone

3. *Cut – in vitro*
- digest the recipient plasmid with the same restriction enzyme(s)
- optionally, treat with a phosphatase (*e.g.* shrimp alkaline phosphatase) to remove the 5' phosphate from the DNA; this prevents ligase from rejoining the fragments of the recipient plasmid to each other

4. *Paste – in vitro*
- mix the digested recipient plasmid and donor gene and treat with ligase to covalently join the sticky ends

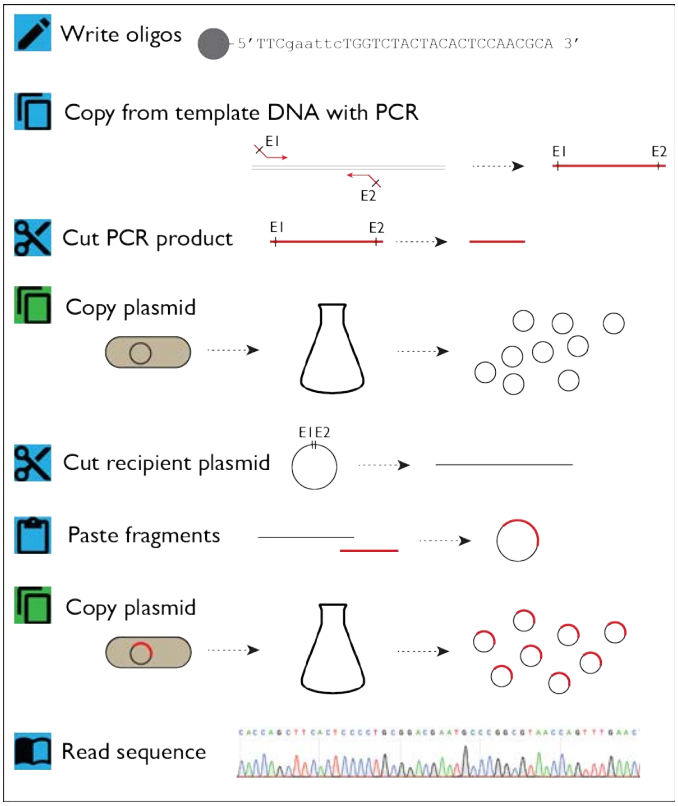
5. *Copy – in vivo*
- transform the ligation mixture into a fresh bacterial strain and grow the culture under conditions that select for the desired plasmid to make a large amount of recombinant plasmid product

6. *Read – in vitro*
- sequence the recombinant plasmid to confirm that it has the desired sequence

Exactly the same procedure can be done with completely or partially restriction-digested genomic DNA instead of a donor plasmid, which results in a pool of plasmids containing a variety of different inserts. This is a *genomic library* and is useful for many kinds of mutant hunts, as we discussed in **Lecture 5**.

CLONING

The protocol most frequently referred to as "cloning" in a modern molecular biology lab involves generating a gene sequence by PCR and then inserting it into a plasmid's multiple cloning site. Because PCR primers can be synthesized directly, this allows you to place any restriction site you like at the ends of the DNA to be inserted and means you do not have to depend on whatever restriction sites are naturally present in the original source of that DNA.



Protocol:

1. *Write – in vitro*
- design PCR primers that amplify your DNA of interest
- add desired restriction site sequences to the 5' end of the primers (with a few extra nucleotides, since many restriction enzymes don't cut well at the very end of a DNA fragment)

2. *Copy – in vitro*
- PCR amplify the DNA of interest from a template (for example, genomic DNA) using the primers designed in step 1

3. *Cut – in vitro*
- digest the PCR-amplified DNA with the restriction enzymes whose sites you added to the primers

4. *Copy – in vivo*
- grow cells containing recipient plasmid and make a large amount of it

5. *Cut – in vitro*
- digest the recipient plasmid with the same restriction enzyme(s) used in step 3
- optionally, treat with a phosphatase to remove the 5' phosphate from the DNA

6. *Paste – in vitro*
- mix the digested recipient plasmid and donor gene and treat with ligase to covalently join the sticky ends

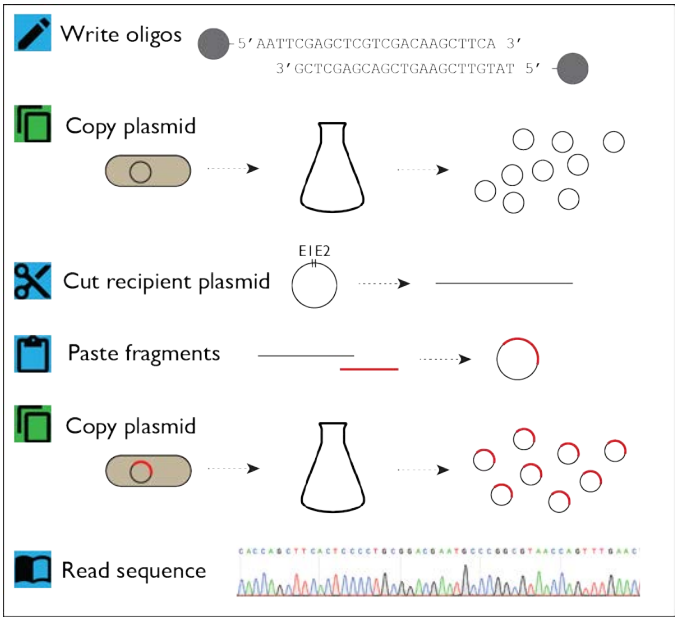
7. *Copy – in vivo*
- transform the ligation mixture into a fresh bacterial strain to make a large amount of the recombinant product

8. *Read – in vitro*
- sequence the recombinant plasmid to confirm that it has the desired sequence

Since PCR is generally limited to amplifying no more than 10 kb of DNA, it is difficult to obtain enough PCR product for cloning larger fragments than this, but most plasmids are not very good at maintaining inserts this large anyway. For large inserts, specialized vectors (*e.g.* BACs) are a better choice.

CLONING SMALL FRAGMENTS

It is often useful to clone very short DNA sequences (< 100 bp). Many older plasmids have simple multiple cloning sites with only a few restriction sites, and you might want to add more. You might have a gene in a plasmid that you would like to add a promoter to, or a short amino acid tag for protein purification. In cases like this, you can *write* the sequence to be cloned directly. The main complication for this kind of cloning is screening for small inserts, which are often too small to be seen easily on a gel. It's therefore often a good idea to include a unique restriction site within the insert that will allow you to rapidly distinguish between your desired product and the original plasmid without having to sequence every possible candidate.



Protocol:

1. *Write – in vitro*
- design PCR primers that contain your sequence of interest and are complementary to each other, with sticky ends for cloning
- anneal the primers to each other by mixing them, heating to 95°C, then cooling slowly to room temperature
- optionally, use T4 polynucleotide kinase to phosphorylate the 5' end of the annealed DNA fragment (oligos are not normally synthesized with a 5' phosphate, and that phosphate is necessary for ligase activity)

2. *Copy – in vivo*
- grow cells containing recipient plasmid and make a large amount of it

3. *Cut – in vitro*
- digest the recipient plasmid with restriction enzyme(s) that match the sticky ends you designed into your primers
- treat with a phosphatase to remove the 5' phosphate from the recipient DNA (if you have phosphorylated the insert)

4. *Paste – in vitro*
- mix the digested recipient plasmid and donor DNA and treat with ligase to covalently join the sticky ends

5. *Copy – in vivo*
- transform the ligation mixture into a fresh bacterial strain to make a large amount of the recombinant product

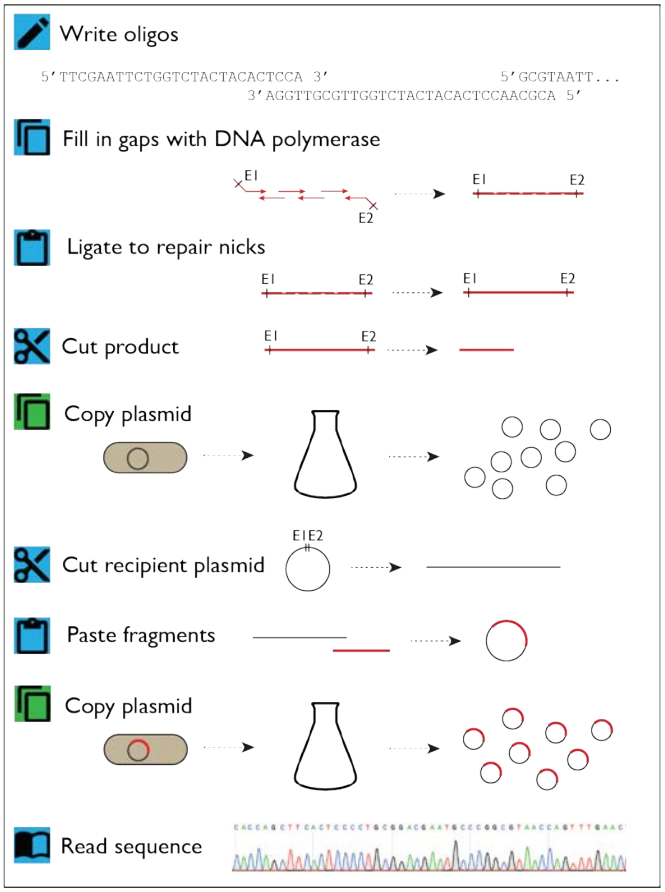
6. *Read – in vitro*
- sequence the recombinant plasmid to confirm that it has the desired sequence

All three of the previous techniques fall into the broad category of "molecular cloning", and if you want a different, more detailed description of this category of protocols, once again Addgene has an excellent webpage describing the various techniques.

GENE SYNTHESIS

PCR isn't the only way to generate a large DNA fragment to be cloned into a plasmid. It is possible to build up DNA molecules of any sequence and, in theory, any length by synthesizing a series of overlapping oligonucleotides and stitching them together in a process called "overlap extension". The resulting DNA can then be cloned as usual. Many companies will synthesize DNA for you in this way fairly inexpensively (about 35 or 40 cents per base pair). You could also do it yourself, although it requires careful primer design. In practice, synthesizing sequences longer than a single gene is usually not worth it, but the Craig Venter Institute used this method to (very expensively!) synthesize an entire bacterial genome in 2008. Cost is the most important limitation to gene synthesis in general lab use.

One common reason to have a gene synthesized rather than cloning it directly from genomic DNA is to optimize the gene's codon usage for expression in your target organism. Different species translate the various codons for specific amino acids at different efficiencies, and this can strongly affect how much protein is produced. For example, *E. coli* very rarely uses the AGG codon for arginine, and has low levels of the tRNA for that codon. Another useful application is to obtain genes from organisms which are difficult to grow or for which genomic DNA is not readily available. It may also be a simple way to obtain complex protein fusions that would be labor-intensive to construct by cloning.



Protocol:

1. *Write – in vitro*
- design 40-50 nucleotide oligos that overlap at their ends and together encode your desired sequence, with restriction sites as desired at the ends of the final product
- anneal the oligos to each other

2. *Copy – in vitro*
- add DNA polymerase to fill in the gaps in the annealed oligo chain

3. *Paste – in vitro*
- treat with ligase to repair nicks and form a single double stranded linear DNA product

4. *Cut – in vitro*
- digest the synthesized DNA with the restriction enzymes whose sites you added to the primers

5. *Copy – in vivo*
• grow cells containing recipient plasmid and make a large amount of it

6. *Cut – in vitro*
• digest the recipient plasmid with the same restriction enzyme(s) used in step 4
• optionally, treat with a phosphatase to remove the 5' phosphate from the DNA

7. *Paste – in vitro*
• mix the digested recipient plasmid and donor gene and treat with ligase to covalently join the sticky ends

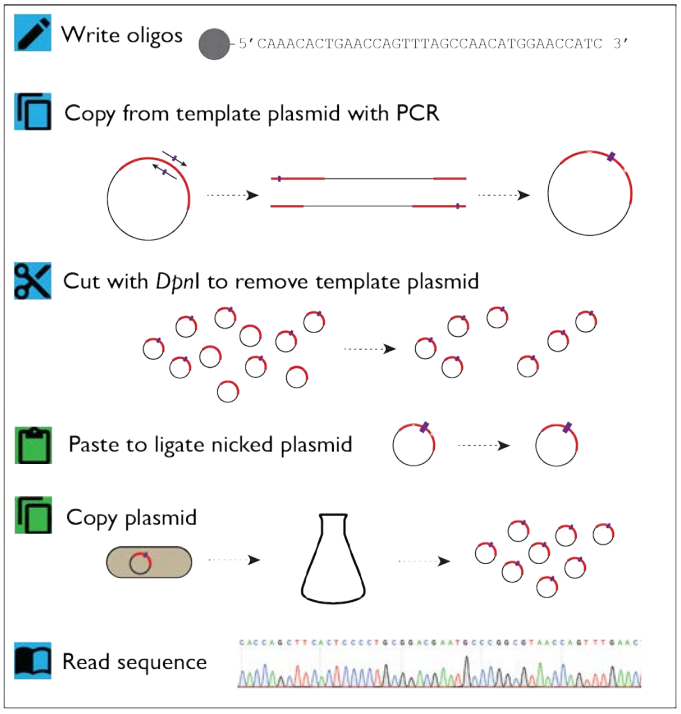
8. *Copy – in vivo*
• transform the ligation mixture into a fresh bacterial strain to make a large amount of the recombinant product

9. *Read – in vitro*
• sequence the recombinant plasmid to confirm that it has the desired sequence

SITE-DIRECTED MUTAGENESIS OF PLASMIDS

The methods I've described so far focus on constructing plasmids from large component parts, which is a very common molecular biology procedure. However, you will often want to make more subtle changes to a DNA molecule, including changing single base pairs or codons. There are a variety of ways to accomplish this. Here is one of the most common. It works well for small mutations of all kinds.

This is the basis of the "QuikChange" mutagenesis kit, available from Agilent. You don't need to buy their reagents to do this kind of mutagenesis, but they do have a nice description of how it works in the manual available here.



Protocol:

1. *Write – in vitro*
• design a pair of oligos that are complementary to your plasmid, with the desired mutation centered in the oligo sequences (the PrimerX tool is very useful for this)

2. *Copy – in vitro*
• using the oligos designed in step 1 as primers, use PCR to amplify the entire plasmid; this will require using a high-fidelity DNA polymerase with high enough processivity to generate a full-sized plasmid product
• the resulting single stranded products will anneal into a nicked, double stranded circular DNA molecule

3. *Cut – in vitro*

- treat with the restriction enzyme *Dpn*I, which cuts methylated DNA at the commonly-occurring sequence GATC; this eliminates the original vector, methylated by the natural Dam methylase of *E. coli*, while leaving the unmethylated PCR-synthesized DNA intact

 4. *Paste – in vivo*

- transform the resulting nicked circular DNA product into a fresh bacterial strain; the DNA repair system of the recipient strain will repair the nicks in the plasmid

 5. *Copy – in vivo*

- grow up the transformed strain to make a large amount of the recombinant product

 6. *Read – in vitro*

- sequence the recombinant plasmid to confirm that it has the desired sequence

Surprisingly, it is not actually necessary to generate a double-stranded DNA product for this type of site-directed mutagenesis to work. The procedure above works very well with only a single primer. This generates a single-stranded, linear mutated DNA product, which *E. coli* is able to repair into a circular double-stranded DNA, probably by first synthesizing the second strand and then circularizing the resulting DNA by recombination (*in vivo copy* and *swap* steps).

Note that, because the site-directed mutagenesis procedure happens almost entirely *in vitro*, there is no opportunity for DNA repair mechanisms to repair the mutation. This is why it's so much easier to make precise mutations of plasmids than of the chromosome.

DISCUSSION PROBLEM SET #15: UNDERSTANDING NEW PROTOCOLS

There are a lot of different techniques available in the literature that people have used to construct particular kinds of recombinant DNA molecules. It is useful to be able to break them down into their component steps to make sure that you understand how a particular protocol is done, and how it can be used.

Here, for example, are links describing two protocols commonly used to construct plasmids with complex inserts:

SOEing PCR

Gibson Assembly

Based on these links (and any other resources you can find), break down each of these procedures into a series of "read", "write", "copy", "cut", and "paste" steps, in the same way that protocols were broken down in the previous section. Make sure to include **all** necessary steps!

DISCUSSION PROBLEM SET #16: DESIGNING CONSTRUCTS FOR GENETIC EXPERIMENTS

Problem #1

Describe a detailed protocol for generating a plasmid which will allow inducible expression of a protein fusion between the MreB cytoskeletal protein of *E. coli* (which we will discuss in detail in **Lecture 11**) and the red fluorescent protein mCherry. (This is one of the most commonly-used of the "mFruit" family of fluorescent proteins: see Shaner *et al.* [2004] Nat Biotechnol 22:1567.)

As raw materials, you have wild-type *E. coli* MG1655 genomic DNA and the plasmids linked below (and whatever standard genetic tools you care to use):

pBAD30

pmCherry

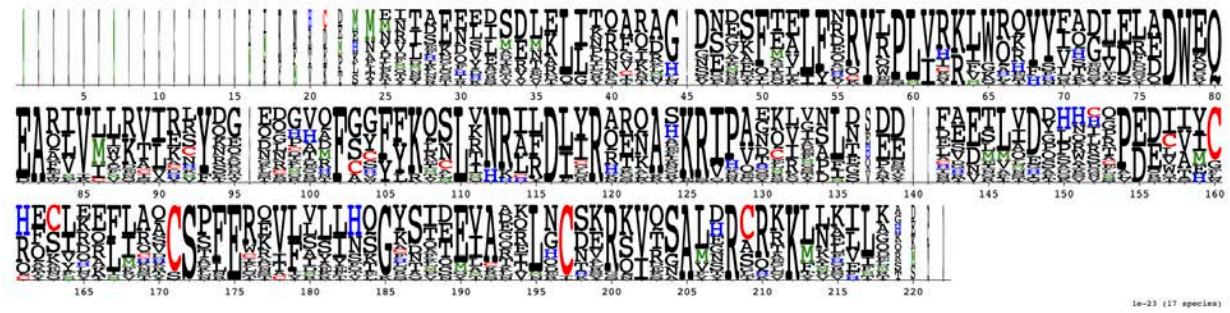
You can find the sequence of the mCherry gene on GenBank.

Be sure to include **all** of the necessary steps (but do not worry about exact primer sequences and such), and draw a map of the resulting plasmid product you intend to construct.

What phenotype do you expect to observe when you express the MreB-mCherry fusion protein you have constructed in *E. coli*?

Problem #2

The gut-inhabiting lactic acid bacterium *Lactobacillus reuteri* has only one alternative sigma factor: SigH, which you suspect controls gene expression in response to changes in oxygen levels. SigH homologs are found in many lactobacilli, so you generate an alignment of SigH sequences from 17 different species:



Based on this alignment, and knowing that the redox state of cysteine residues often regulates protein activity (see **Lecture 4**) you hypothesize that Cys171 and Cys197 are required for oxygen sensing by SigH. Propose an experiment using plasmids to test this hypothesis. State:

- a **detailed** description of how you will construct the necessary plasmids

  (Note that there are a number of useful *Lactobacillus-E. coli* shuttle vectors available. For the purposes of this experiment, use pTRKH2.)

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

---

LECTURE 8: GENE TRANSFER AND RECOMBINATION

INTRODUCTION

In this lecture, we will discuss how genetic material can be transferred between bacterial cells, both naturally and in the lab. This will lead to a discussion of homologous recombination and techniques for genetic engineering that depend on gene transfer and recombination. We will design experiments using these techniques and discuss the benefits and disadvantages of such approaches.

GENE TRANSFER IN BACTERIA

There are three ways by which a bacterial cell can take up new DNA: *transformation*, *conjugation*, and *transduction*. All of these occur in nature, and are mechanisms by which bacteria can acquire new genetic material from other, distantly related organisms (*horizontal gene transfer*).

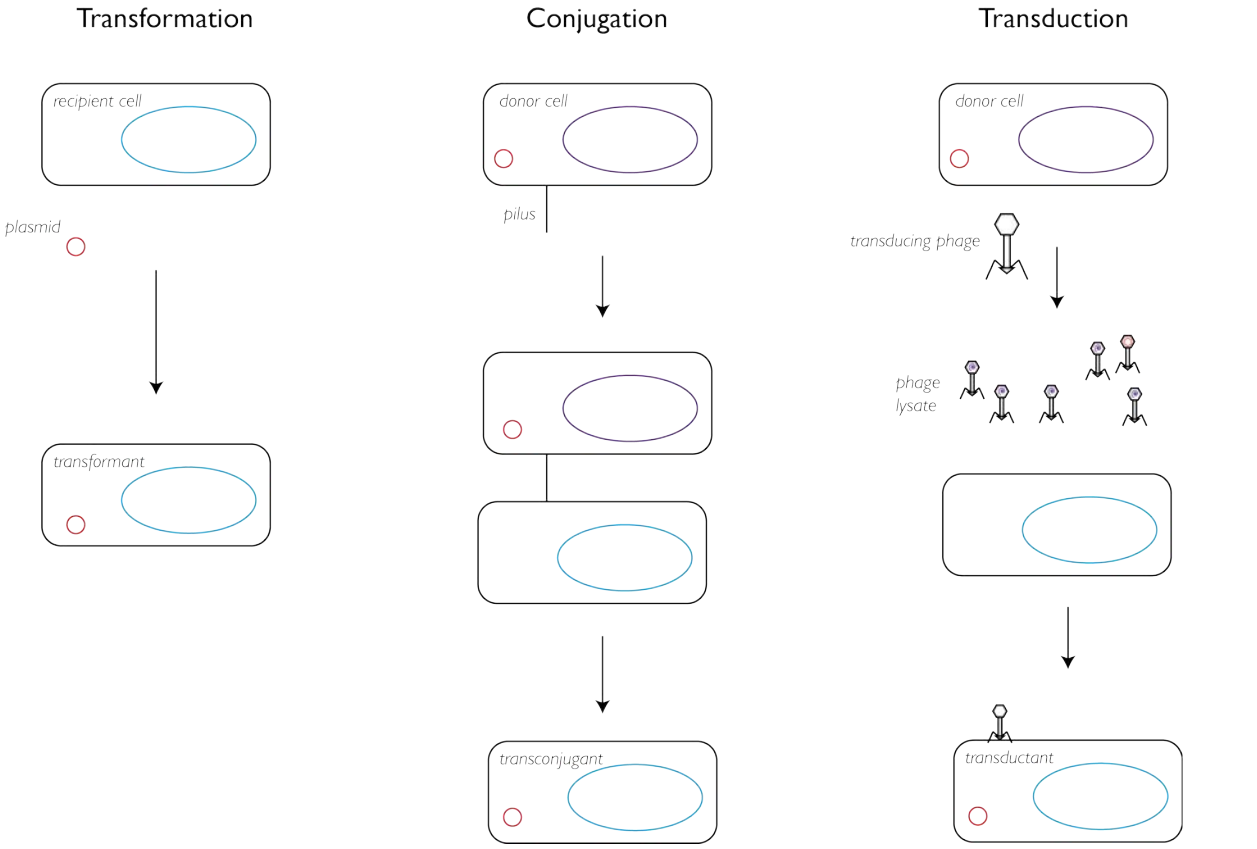


Figure 8.1. Moving a plasmid (red circle) into a recipient cell (blue chromosome) by three different methods. Note that transduction also results in phage particles containing fragments of chromosomal DNA from the donor cell (purple chromosome), which may then be transferred into the recipient cell. See the "Common Protocols" section of this chapter for more details.

Transformation is a process in which cells directly take up DNA from their environment and incorporate it into their genetic material. Cells that can do this are called *competent cells*. It's called "transformation" because the uptake of new genes can transform the phenotype of a strain. (In fact, Oswald Avery's 1944 experiments showing that adding very pure DNA could change the colony morphology phenotype of the pneumonia-causing pathogen *Streptococcus pneumoniae* were among the first pieces of evidence that DNA is the genetic material of cells.)

Some species are *naturally competent* (e.g. *S. pneumoniae*, *Bacillus subtilis*, *Neisseria gonorrhoeae*) and will take up DNA from their environment on their own, but most species require special treatment to allow transformation. *E. coli* and some other Gram-negative bacteria can be made *chemically competent* by resuspending them in very cold $CaCl_2$ solutions and then briefly heat shocking them at 42°C. Many types of cells can be transformed by *electroporation*, in which cells are mixed with DNA in a cold, low ionic-strength solution then subjected to an electric shock. These methods are thought to work by disrupting the cell membrane enough to allow DNA through. (Note that for eukaryotic cells, direct uptake of DNA is called "transfection" instead of transformation.)

Conjugation is a process in which bacterial cells form hair-like tubular structures (called *pili*, which is the plural of *pilus*) on their surfaces and transfer DNA through those pili into other cells. The genetic elements that allow specific DNA molecules to be conjugated are called *tra* factors (short for **tra**nsfer). Only DNA molecules containing an *origin of transfer* (*oriT*) for the particular *tra* system in a donor bacterium can be conjugated. Note that conjugation is not species-specific, and is in fact a common method used in the lab for transferring DNA from easy-to-work-with species (like *E. coli*) into species that are more challenging to transform, which do not even necessarily have to be bacteria. In nature, the plant pathogen *Agrobacterium tumefaciens* (which causes crown gall disease) conjugates a genetic element called T-DNA into plant cells, causing formation of tumors in the host plant. Conjugation was discovered in *E. coli* in 1947 by Joshua Lederberg and Edward Tatum in the first demonstration that bacteria can mate and exchange genes, a discovery which really made bacterial genetics (and, eventually, molecular biology) possible. Tatum, Lederberg, and George Beadle later won the 1958 Nobel Prize for this and other contributions to molecular genetics.

Transduction is the use of *bacteriophage* (recall that these are viruses that infect bacteria, often just called "phage") to transfer DNA from one bacterial strain to another. *Generalized transducing phage* are phage which are able to package plasmids or random fragments of DNA from the chromosome of their host cell into virus particles. These particles can then attach to and inject that DNA into another bacterial cell, where it can potentially be incorporated into the host chromosome by homologous recombination (see below). This is contrasted with *specialized transducing phage*, which are less useful and are only able to package host genes directly adjacent to the single site where the phage integrates into the host chromosome. Most wild-type phage normally only package viral DNA, of course, but many lab strains of transducing phage have been selected to package host DNA at higher frequency. (About 1 in 30 phage particles produced during a P1*vir* infection of *E. coli* contains host DNA instead of phage DNA, for example.)

Like other viruses, phage are typically extremely species- or even strain-specific. The P1 phage will only work for transductions in certain strains of *E. coli*, for example, while the P22 phage is specific for *Salmonella enterica*. Generalized transduction (by phage P22) was discovered by Joshua Lederberg and Norton Zinder in 1951, and specialized transduction (by phage λ) was discovered by Esther Lederberg in 1956.

DISCUSSION PROBLEM SET #17: GENE TRANSFER

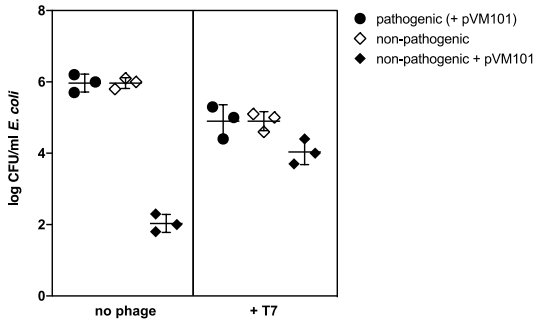
Problem #1

While studying antibiotic resistance in *S. aureus*, you discover that incubating an erythromycin-resistant strain together with a chloramphenicol-resistant strain in media **without** antibiotics results in the appearance of some cells that are resistant to **both** antibiotics.

Propose an experiment to determine the mechanism by which this genetic exchange occurs.

Problem #2

During intestinal infections, the number of phage particles in the gut increase dramatically. Pathogenic strains of *E. coli* often contain plasmids which encode toxins, siderophores, and other virulence factors. One of these, called pVM101, is more than 150 kbp in size. You infect mice with a combination of pathogenic (pVM101-containing) and non-pathogenic (plasmid-free) *E. coli* and measure the rate of transfer of pVM101 to the non-pathogenic strain during the infection. You find that adding the *E. coli*-infecting phage T7 greatly increases this transfer.



T7 is an obligately lytic phage with a 40 kbp genome. Propose a model to explain your observations, and an experiment to test that model.

PRACTICAL CONSIDERATIONS

When working in the lab, there are some practical considerations you should take into account when attempting to move a particular piece of DNA into a bacterial strain:

1. Is the gene you want to move on the chromosome or on a smaller DNA element like a plasmid? Highly competent cells with efficient recombination systems (see below) may be able to take up and incorporate genomic DNA, but this is likely to result in incorporation of a lot of genetic material from the donor. Transduction can move smaller fragments of chromosomal DNA, but generally only between fairly closely related strains. It is important to remember that plasmids **can** be moved from one cell to another by transduction, it is just usually less convenient than the other two methods.
2. Are generalized transducing phage or conjugative plasmid systems available for your model organism? While these tools exist for many species, they have not been developed for all bacteria. It has become particularly unfashionable to identify generalized transducing phage for new model organisms. (It's a lot of work to do, and not that many labs are equipped to do it anymore.)
3. Can you easily make the bacteria you are working with competent? If so, transformation is likely to be the most convenient method to move a plasmid into those cells.

Moving DNA between species can present a particular challenge. Most bacteria possess defense mechanisms that will attempt to break down any foreign DNA molecules that enter their cells. These include restriction enzymes, which we have discussed as molecular tools, and which recognize and cut specific DNA sequences. As mentioned in **Lecture 7**, in nature these function to protect bacteria against attack by phage, and what they "restrict" is the ability of particular phage to infect that strain. The bacterium protects its own DNA from restriction digestion with a sequence-specific *restriction methylase* that adds a methyl group to the DNA sequence recognized by its cognate restriction enzyme, preventing them from being cut. If you are trying to move DNA into a cell with a restriction enzyme system from a cell without the appropriate methylase, the transformation efficiency will be very low. Daisy Dussoix, a graduate student in Werner Arber's lab, was the first to recognize the existence of restriction-modification enzyme systems and their effects on DNA transfer around 1960.

HOMOLOGOUS RECOMBINATION

When a bacterial cell takes up a plasmid with an appropriate origin of replication, the plasmid is able to replicate and be maintained in that cell and its descendants. DNA molecules without their own origin of replication have to be incorporated somehow into the host chromosome in order to be passed down to the next generation. One very common mechanism by which this can occur is known as *homologous recombination*. (For many more details and links, this Wikipedia article is a great place to start.)

The main physiological function of homologous recombination in cells is in DNA damage repair, and the complex details of its mechanism are beyond the scope of this course. However, it is important to have a general sense of how it works, since many genetic engineering procedures depend on it.

As shown in Figure 8.2, when there are two pieces of DNA in a cell with similar sequences, the RecA single-stranded DNA binding protein and RecBCD *recombinase* proteins can recognize single- or double-strand breaks in those DNA molecules, bind to them, and create stretches of hybrid base-paired DNA with crossover points called *Holliday junctions* (after Robin Holliday, who first proposed their existence in 1964). This requires stretches of DNA with very similar sequences at the crossover points, which is why this is called **homologous** recombination. The higher the homology between the two DNA molecules, the more likely RecABCD is to be able to generate Holliday junctions.

The Holliday junctions are resolved into replication forks by the RuvABC complex, and subsequent DNA replication results in two intact chromosomes, each of which incorporates the new DNA on one strand, where it may be inherited by some of the cell's progeny or repaired by other DNA repair mechanisms (*e.g.* the mismatch repair system). As you might expect, mutants lacking any of the *rec* or *ruv* genes are unable to carry out homologous recombination and are extremely sensitive to DNA-damaging chemicals and radiation.

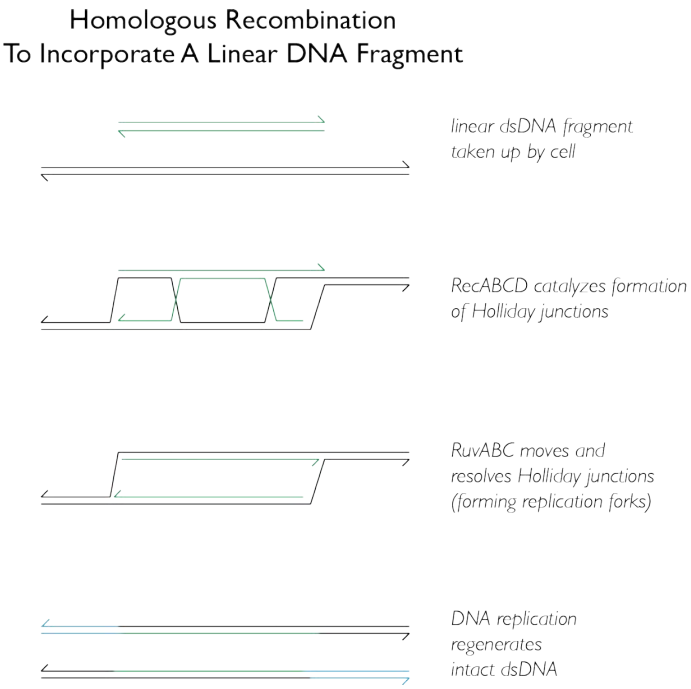


Figure 8.2. A very simplified diagram illustrating incorporation of a linear DNA fragment into a bacterial chromosome by homologous recombination.

Many *cloning strains* of *E. coli*, which contain mutations to make them more competent and easier to work with in the lab (common examples are the strains DH5α and JM109), are *recA* mutants. This is to prevent the possibility of recombination between genes on plasmids and genes on the chromosome, and improves plasmid stability. It does, however, mean that no genetic engineering protocols that rely on homologous recombination will work in these strains.

Lysogenic bacteriophage encode their own recombinases, which they use to integrate themselves into the host chromosome at a specific site (the *att* site, which is different for each phage). These often require much shorter regions of homology than RecABCD to stimulate recombination, which has made them useful tools for molecular genetics, as we will discuss below.

USING HOMOLOGOUS RECOMBINATION FOR GENETIC ENGINEERING

Swap

As mentioned in the last chapter, one of the six fundamental procedures in molecular biology is recombination, which I abbreviated as "swap", since it results in swapping or exchanging sequences from one DNA molecule to another. In the next section of this chapter, I will describe a variety of genetic engineering protocols that depend on recombination, to illustrate what is possible. Swap steps always happen *in vivo*. The recombination machinery is very complex, and reconstituting it *in vitro* is impractical for general purposes.

The protocols described in the previous chapter were useful for engineering plasmids, which are relatively easy to manipulate. Recombination allows us to expand our toolkit and generate site-directed mutations in the bacterial chromosome itself.

It is important to distinguish between *single-crossover* and *double-crossover* recombination events, which are distinguished by requiring either one or two independent homologous recombination events. Single-crossover occurs much more frequently, and combines two DNA molecules into a single product, but single-crossover recombination between one linear and one circular DNA molecule results in a linear product, which, if the circular DNA was the chromosome, constitutes a lethal double strand break.

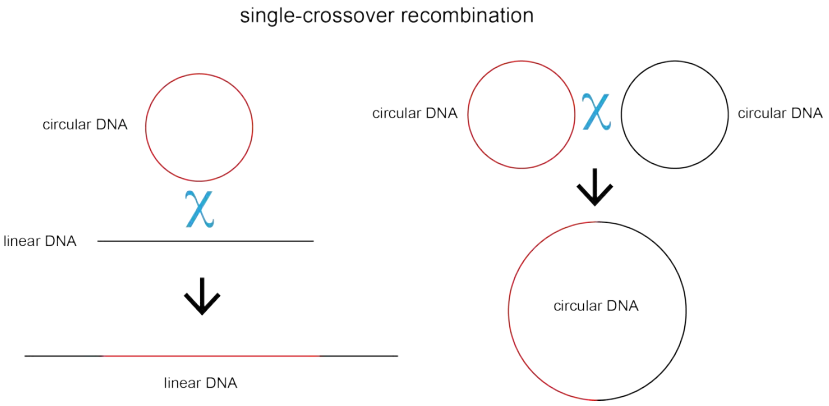


Figure 8.3. The results of single-crossover homologous recombination events between a circular and a linear DNA molecule (on the left) or between two circular DNA molecules (on the right).

The protocols described in this chapter almost exclusively rely on double-crossover recombination, which generates 2 products, but no DNA strand breaks in the circular molecule.

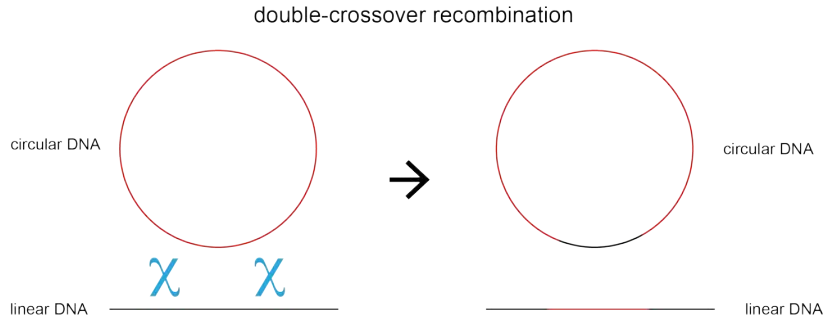


Figure 8.4. The results of a double-crossover homologous recombination event between a circular and a linear DNA molecule. Double crossover recombination between two circular DNA molecules results in two circular DNA molecules.

In either case, but especially when demanding a double-crossover product, recombination is a rare event, so having a strong selection for strains that contain the desired final product is essential. This is usually accomplished by including an antibiotic resistance gene in the DNA that you want to incorporate into the chromosome.

EXAMPLES OF COMMON MOLECULAR BIOLOGY PROTOCOLS

In the next section, I will break down a series of protocols into their component steps, both in outline and graphical form. The protocols in this section all depend on recombination, although many of them **also** require some amount of plasmid engineering, which can be done using the methods described in the previous chapter.

TRANSDUCTION

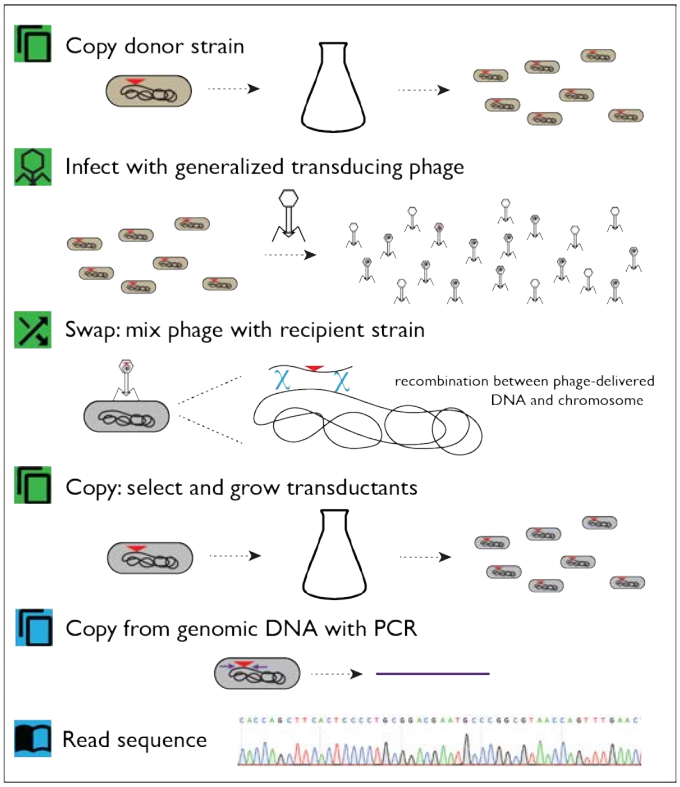
As described above, generalized transduction uses transducing phage to transfer selectable markers between bacterial strains, which are then incorporated into the chromosome of the recipient cell by homologous recombination.

Protocol:

1. *Copy – in vivo*
- grow donor cells containing the selectable marker you plan to transduce

2. *Infect – in vivo* (not really one of the "six steps"…)
- infect donor cells with generalized transducing phage and harvest phage particles, some of which will contain DNA from the donor cell chromosome

3. *Swap – in vivo*
- add phage containing selectable marker to recipient cells, and allow time for DNA injection and recombination to occur
- you will need to include a step to **stop** the phage infection, since most of the phage particles you added will be virulent; this is commonly done by chelating away calcium, which many transducing phage require for attachment



4. *Copy – in vivo*
- select for recombinants and grow them

5. *Copy – in vitro*
- use PCR to amplify the region of the chromosome containing the desired mutation

6. *Read – in vitro*
- sequence the PCR product to confirm that the selected transductant has the desired sequence derived from the donor strain

Since successful transfer and incorporation is a relatively low-frequency event, a selection is required to identify successful *transductants*. When the mutation you want to move is itself selectable, this is straightforward. However, since transducing phage package large fragments of host chromosomal DNA (100 kb in the case of the *E. coli* transducing phage P1, for example, or 40 kb for the *Salmonella* phage P22), just having a selectable marker **near** your mutation of interest (a *linked marker*) is sufficient. This is commonly a transposon insertion in a nearby gene or intergenic region. The closer two mutations are on the chromosome, the more frequently they will be *cotransduced*. (This can also be used to calculate the distance between two mutations on the chromosome, if you don't already know that information. This is called *linkage mapping* and has been made almost entirely obsolete by inexpensive genome sequencing.)

Note that step 5, using PCR to amplify the genomic region containing the putative mutation for sequencing, can be done with purified genomic DNA or, for many species, simply by suspending some cells in the PCR reaction mix. This is "colony PCR", and works because the 95-98°C melting step of the PCR cycle lyses some of the bacteria, releasing their DNA into solution.

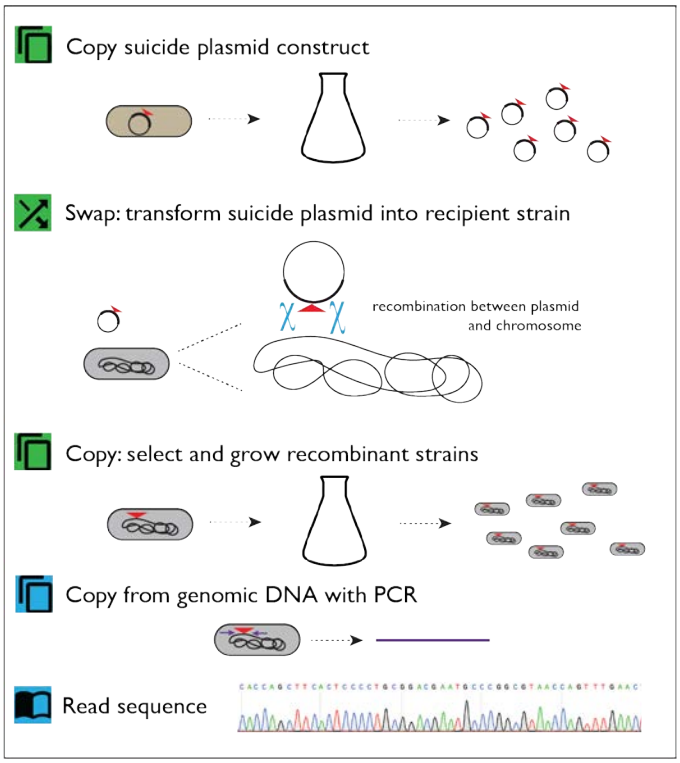
ALLELIC EXCHANGE

Allelic exchange procedures involve the construction of plasmids containing the desired mutant allele, which are then recombined into the chromosome of the recipient strains using the native RecA-dependent recombinase activity of that strain. RecA usually requires very long regions of homology for recombination to occur (500 to 1000 bp).

Normally, allelic exchange templates will consist of a suicide vector containing an *antibiotic resistance cassette* (a gene encoding a product that confers antibiotic resistance, along with all of the additional sequences needed to ensure its expression) flanked by sequences homologous to the target region in the host chromosome. This makes it

straightforward to select for *recombinants* on plates containing the relevant antibiotic. Any of the plasmid construction methods described in the previous chapter can be used to construct this vector.

Use of a suicide vector that cannot replicate in the recipient strain is essential. Otherwise selecting for antibiotic resistance will only yield transformants (cells containing replicating plasmids), not recombinants. Transformation is orders of magnitude more efficient than recombination. See **Lecture 5** for more details on suicide vectors.



Protocol:

1. *Copy – in vivo*
• grow cells containing suicide plasmid and make a large amount of it

2. *Swap – in vivo*
• transform the suicide plasmid into recipient cells and allow time for recombination to occur

3. *Copy – in vivo*
• select for recombinants and grow them

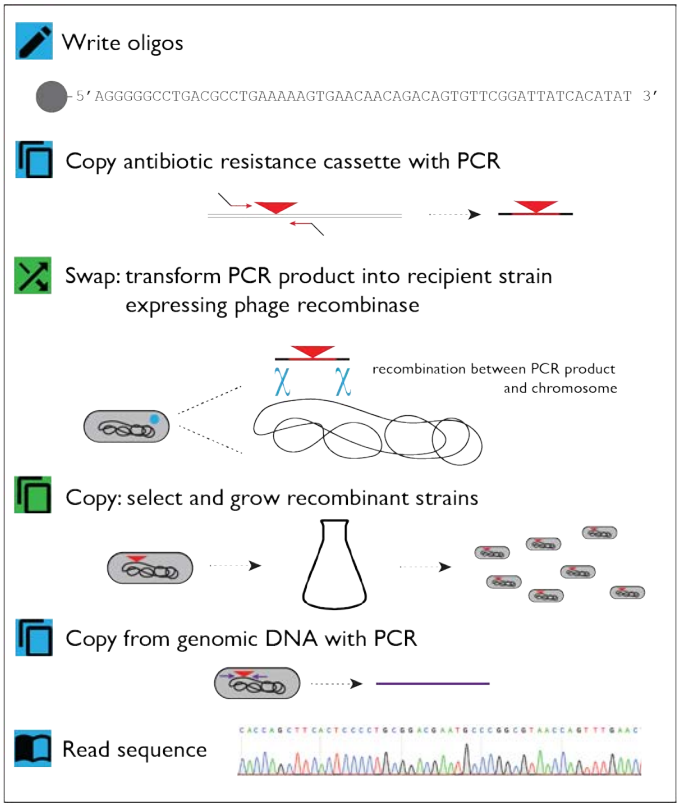
4. *Copy – in vitro*
• use PCR to amplify the region of the chromosome containing the desired mutation

5. *Read – in vitro*
• sequence the PCR product to confirm that the selected strain has the desired mutation

While the protocol above illustrates allelic exchange with double-crossover recombination, do note that this is the only protocol in this chapter which can use single-crossover recombination, since the suicide vector is a circular DNA molecule. In this case, the "swap" step might involve one recombination step to integrate the plasmid into the chromosome, followed by a second single-crossover recombination step to "loop" the integrated plasmid out, which will (about 50% of the time) result in the chromosome containing the allele that was originally in the vector. You might need to do this if the double-crossover recombination event is too rare, since single-crossover recombination happens much more frequently.

RECOMBINEERING

Recombineering uses double-stranded linear DNA fragments (typically PCR products) as templates for recombination in cells expressing highly active *phage recombinases* that can integrate DNA fragments with as little as 40 to 50 bp of sequence homologous to the host chromosome. The PCR products used for recombineering almost always contain an antibiotic resistance gene to allow selection of recombinants.



Protocol:

1. *Write – in vitro*
- design PCR primers that amplify an antibiotic resistance cassette
- add sequences homologous to the desired insertion site in the chromosome to the 5' end of the primers

2. *Copy – in vitro*
- PCR amplify an antibiotic resistance cassette using the primers designed in step 1

3. *Swap – in vivo*
- transform the PCR product into recipient cells expressing a phage recombinase and allow time for recombination to occur

4. *Copy – in vivo*
- select for recombinants

5. *Copy – in vitro*
- use PCR to amplify the region of the chromosome containing the desired mutation

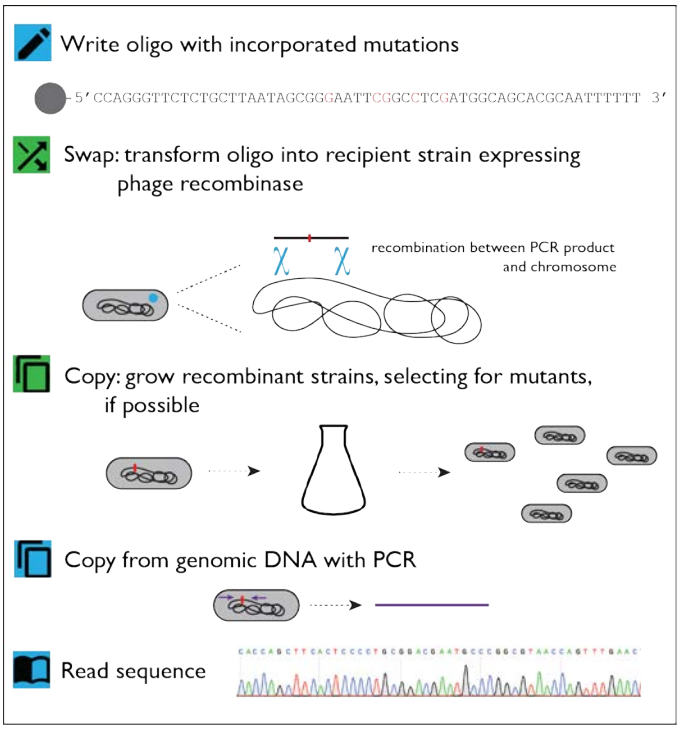
6. *Read – in vitro*
- sequence the PCR product to confirm that the selected strain has the desired mutation

The phage recombinase (very commonly the Red recombinase from phage **λ**, especially in Gram-negative bacteria) must be expressed from an inducible promoter on a plasmid, which can be constructed using any of the plasmid construction methods described in **Lecture 7**. It is not generally healthy for bacteria to constitutively express recombinases, which can lead to unwanted chromosome rearrangements. Recombinase expression plasmids for recombineering often have temperature-sensitive origins of replication to make curing the plasmid easy after the desired chromosomal mutation(s) have been made.

OLIGO-DIRECTED RECOMBINEERING

The PCR products used as recombination templates in the previous protocol are double stranded DNA molecule. The phage recombinases used for recombineering also work with single-stranded DNA templates, like oligos. It is only possible to order oligos up to about 100 bp long, so oligo-directed recombineering cannot be used to insert large sequences (like antibiotic resistance genes), but if a point mutation you're interested in has a selectable or easily screenable phenotype, this approach can work very well.

When recombineering primers are carefully designed to avoid triggering the host cell's DNA repair mechanisms it is sometimes possible to generate non-selectable alleles, including point mutations, using this method. Of course, in this case, you need to screen the resulting colonies to determine which ones contain your desired mutation, usually by sequencing the affected gene. Efficiency may be quite low (< 1%), however, making this a labor-intensive approach. See below for how we can use CRISPR to greatly increase the efficiency of oligo-directed recombineering.



Protocol:

1. *Write – in vitro*
- design an oligo homologous to the bacterial chromosome with the desired mutation near its center
- adjust the sequence of the oligo to avoid mismatch repair and increase recombination efficiency (see below)

2. *Swap – in vivo*
- transform the mutagenic oligo into recipient cells expressing a phage recombinase and allow time for recombination to occur

3. *Copy – in vivo*
- select for recombinants, if possible, or dilute and plate for individual colonies to screen

4. *Copy – in vitro*
- use PCR to amplify the region of the chromosome containing the desired mutation

5. *Read – in vitro*
- sequence the PCR product to confirm that the selected strain has the desired mutation

The bacterial mismatch repair system does not work well on mutations that change 5 or more sequential nucleotides or on several closely spaced point mutations (as shown in the figure above). Making 3 or 4 silent mutations directly adjacent to the mutation of interest can improve efficiency greatly, as can synthesizing the oligo with more-stable phosphorothioate linkages at the 5' or 3' ends. Mutagenic oligos are also more efficiently incorporated into the

chromosome when they are complementary to the lagging strand during DNA replication, possibly because the cell mistakes them for Okazaki fragments.

CRISPR

The most recent addition to the molecular genetics toolkit is CRISPR, which stands for **c**lustered **r**egularly **i**nterspaced **p**alindromic **r**epeats. The name is derived from the fact that *CRISPR arrays* of short, repetitive DNA sequences were observed in many bacteria and archaea long before their function was known. In the early 2000's, Philippe Horvath and Rodolphe Barrangou, working for the food company Danisco, realized that CRISPR was involved in protecting the yogurt-fermenting bacterium *Streptococcus thermophilus* from bacteriophage, and in fact, functioned as a kind of adaptive immune system for those bacteria. The CRISPR array contains short pieces of DNA derived from parasitic phage or plasmids, and the *CRISPR-associated* (Cas) *proteins* are then able to recognize and bind to the matching sequences in those parasites and cause double strand breaks in their DNA, protecting the bacterium from infection.

Extraordinary work from many labs (including those of Jennifer Doudna, Emmanuelle Charpentier, and Feng Zhang) has turned this bacterial defense system into a bioengineering tool that can efficiently introduce double- or single-strand breaks in **any** targeted DNA sequence. In the framework I have laid out for genetic engineering processes, this is an ***in vivo*** "cut" step. The nuclease most often used in genetic engineering protocols is called Cas9, and Cas9 is directed to target a roughly 20 nucleotide target sequence by a *guide RNA* (gRNA) that base pairs to the target.

There are other variations on CRISPR, which we will not discuss in detail here. For example, using CRISPR with inactive Cas proteins that bind DNA but do not cut it (dCas9) allows precise targeting of proteins fused to the inactive Cas protein to specific DNA sequences. This has been used with fluorescent proteins to visualize where particular DNA elements are found in cells, and has also been used to either activate or (more commonly) repress gene expression (resulting in a *gene knock-down* instead of a gene knock**out**).

CRISPR is a tremendously versatile and powerful tool. It works in both bacteria and in eukaryotic organisms, and is far simpler and faster than other techniques for genetic manipulation of eukaryotes (although this may not be true for bacterial systems, where many different genetic engineering technologies exist). The development of CRISPR has stimulated an active debate in the scientific community about the ethics of genetic engineering in higher organisms.

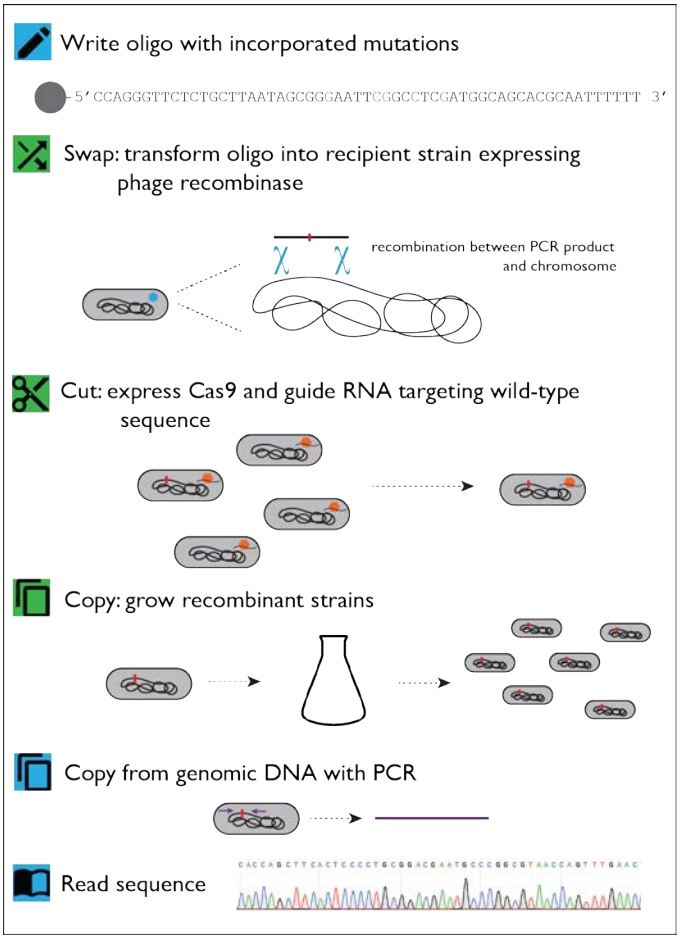
CRISPR-ASSISTED RECOMBINEERING

The most common use of CRISPR in bacterial genetics is in combination with oligo-directed recombineering, where it is used as a selection after the recombination step. Many recombineering procedures have low efficiency, especially when they are used to generate point mutations. Combining recombineering with a CRISPR system that targets the wild-type sequence for double strand breaks efficiently kills any cells that are not mutated. In most of the bacterial systems I have seen, the recombinase and Cas9 are expressed from one plasmid, while the guide RNA is expressed from another.

One additional cloning step is required compared to the previous protocol: constructing a plasmid that will express the guide RNA to target Cas9. Any of the cloning methods from the previous chapter can be used, but since guide RNAs are very small (about 20 nt), "Cloning Small Fragments" is the most common approach.

As with recombinase expression plasmids, it is important that the plasmid(s) expressing Cas9 and the guide RNA be easily curable so that the final strain construct does not contain unnecessary plasmids. This is often accomplished with temperature-sensitive origins of replication.

It is well worth searching the literature (and the plasmid repository Addgene) to see if anyone has developed a CRISPR-based mutagenesis system for your organism of interest, although since it is so new, many species do not yet have such a system. Depending on what you need to do, in that case it might be a good idea to make one yourself.



Protocol:

1. *Write – in vitro*
- design an oligo homologous to the bacterial chromosome with the desired mutation near its center
- adjust the sequence of the oligo to avoid mismatch repair and increase recombination efficiency

2. *Swap – in vivo*
- transform the mutagenic oligo into recipient cells expressing a phage recombinase and allow time for recombination to occur

3. *Cut – in vivo*
- simultaneously, express Cas9 and a guide RNA targeting the wild-type sequence

4. *Copy – in vivo*
- plate recombinant colonies, most of which will have the desired mutation

5. *Copy – in vitro*
- use PCR to amplify the region of the chromosome containing the desired mutation

6. *Read – in vitro*
- sequence the PCR product to confirm that the selected strain has the desired mutation

SINGLE-COPY INSERTION ELEMENTS

The last type of "site-directed" mutagenesis I want to mention briefly is somewhat old-fashioned and not much used any more, but you may encounter examples of it in older papers or strains. Some transposons and most lysogenic bacteriophage do not insert into the bacterial genome randomly, but always insert at the same *attachment site*, which is usually between genes or within a conserved non-essential gene. Cloning genes into such insertion elements can be useful when you want to integrate a single copy of a gene or operon into a strain in a very stable way. Plasmids have

higher and sometimes variable copy numbers and are less stable than a chromosomal insertion. Single-copy insertions are a very clean method to use for complementation experiments.

For phage-based systems, it is important that the inserted sequence not include the genes required for production of live phage particles, since cultures with active infections behave very differently from uninfected cells. The resulting irreversible insertion is called a *defective prophage* or *stable lysogen*. Perhaps the most common stable lysogen encountered in molecular biology is the defective λ phage DE3, which carries a *lac* promoter driving expression of the powerful RNA polymerase from phage T7. This is found in the protein overexpression *E. coli* strain BL21(DE3), for example, which is used in many protein purification procedures.

The transposon Tn7 is a widely-used system for making single-copy insertions. It can be relatively easily engineered to carry a sequence of interest and integrates into a wide variety of bacteria at the end of the highly conserved *glmS* gene, so is less species-specific than phage integrants.

For most purposes, it is now simpler to use recombineering to construct single-copy chromosomal gene insertions. Recombineering also has the advantage that genes can be inserted into the chromosome anywhere you desire, instead of just at the specific attachment site for a given transposon or prophage.

DISCUSSION PROBLEM SET #18: EXPERIMENTAL DESIGN WITH RECOMBINATION

Problem #1

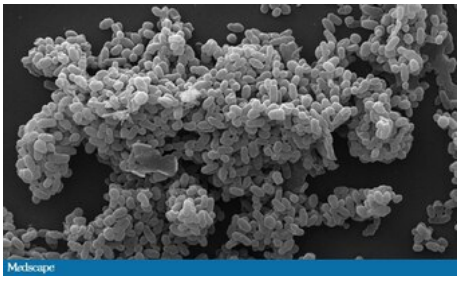
For many model organisms, *knockout collections* have been generated that consist of thousands of individual mutants, generally one in each non-essential gene for that organism. These may be generated by isolating individual transposon insertion mutants (like the Nebraska Transposon Mutant Library for *Staphylococcus aureus* USA300_FPR3757) or by recombineering (like the Keio collection for *E. coli* BW25113), but regardless of how they are constructed, are a tremendously useful resource.

While studying responses to starvation stress in *E. coli*, you find that Tn10(*tet*$^+$) transposon mutants of either *rpoS* or *rpoN* reduce the ability of *E. coli* to grow in minimal media. These genes encode alternative sigma factors, and you hypothesize that they each drive the expression of different sets of genes needed under those growth conditions. As part of a series of experiments to test this hypothesis, you decide to construct a double mutant lacking both *rpoS* and *rpoN*.

Describe a detailed protocol to build an *E. coli rpoS rpoN* mutant. You have access to the Keio collection (in which each mutation is a replacement of the gene in question with a kanamycin resistance gene) and all standard genetic tools.

Problem #2

Akkermansia muciniphilia is a mucus-degrading bacterium found in the mammalian intestine that is associated with reductions in obesity. Human derived strains of *A. muciniphilia* do not colonize mice efficiently, and *vice versa*. You use UV mutagenesis to mutagenize a human strain and select for mutants that colonize mice well.



MedScape

All of the mutants you isolate have multiple point mutations throughout their chromosomes, but you isolate several strains with a particular C to T point mutation in *waaL*, a gene involved in **l**ipo**p**oly**s**accharide (LPS) biosynthesis (see **Lectures 10** and **14** for more on LPS). You hypothesize that this mutation changes the surface properties of *A. muciniphilia*, contributing to host-specific colonization.

For the purposes of this discussion problem, assume that all standard genetic tools are available for *A. mucinophilia*.

Propose an experiment using recombination to test this hypothesis. State:

- a **detailed** description of how you will construct the necessary strains

- the independent and dependent variables
- both positive and negative controls
- potential outcomes of your experiment, and how you will interpret them

---

LECTURE 9: CRITICAL READING (GENETIC ENGINEERING)

EXPECTATIONS

As a reminder, to prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
    - What is the central **question** of this paper?
    - Is the experimental design clear and appropriate to address that question?
    - Do you understand the methods used?
    - Are the data clearly presented, with appropriate statistics?
    - Do you agree with the conclusions the authors came to based on their data?
    - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Nozaki & Niki (2019) "Exonuclease III (XthA) Enforces *In Vivo* DNA Cloning of *Escherichia coli* to Create Cohesive Ends." J Bacteriol 201:e00660-18.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

You may also find the following minireview / methods paper interesting or relevant, although we will not be discussing it in detail in class:

Watson & García-Nafría (2019) "*In vivo* DNA assembly using common laboratory bacteria: A re-emerging tool to simplify molecular cloning." J Biol Chem 294(42):15271-15281.

LECTURE 10: BACTERIAL CELL ENVELOPES

INTRODUCTION

The second half of this course will focus on bacterial *physiology*: the structure, metabolism, energetics, and development of bacteria as living organisms. In each of the following chapters, we will review the fundamentals of one aspect of bacterial physiology, and design molecular biology experiments to answer questions about those topics. You will have plenty of opportunities to practice using the experimental design principles we have discussed (summarized on page 196), the methods for exploring gene and protein expression and activity we explored in **Lecture 4**, and the genetic engineering techniques described in **Lectures 7** and **8**, as well as learning more about some more specialized techniques and applications.

My goals for the rest of this class are really two-fold:

1. I want you to be familiar with the basic principles of bacterial physiology, so that you will be prepared to take on more advanced treatment of those topics in future classes (for example, the upcoming Bacterial Pathogenesis module directed by Drs. Scoffield and Swords) and your own research career.
2. I want you to get lots of practice thinking about and developing ideas for experiments on a variety of different aspects of microbial biology. The problem sets are going to become more complicated and (I hope) realistic from here on forward.

I will only really be able to include the basics for each physiological topic in each upcoming chapter, any of which could easily be expanded into an entire course on their own, and there are certainly whole important topics I will miss entirely due to time constraints. If you're interested in a comprehensive and detailed reference on bacterial physiology, I'd recommend the following book:

"The Physiology and Biochemistry of Prokaryotes" by David White *et al.* (now in its 4th edition)

I will also note again that the examples and topics I have chosen are inevitably biased towards those that I personally think are the most broadly important or interesting (or happen to have run across recently!). This is an idiosyncratic selection process, and it's important to remember that other scientists would certainly make different choices. My approach is not the only right way, and my real goal is to give you the tools to build your own scientific literacy and proficiency. In other classes, you will encounter teachers with different approaches and emphases, and that's great! Learn as much as you can from as many people as you can.

In this chapter in particular, we will discuss the major components of bacterial *cell envelopes*, including *lipid bilayer membranes* and *cell walls*, that separate the interior of bacterial cells from their external environments. Different groups of bacteria organize their cell envelopes differently, and we will discuss the structures of four major groups of bacteria in some detail.

For a **much** more in-depth treatment of microbial cell structure, the Atlas of Bacterial and Archaeal Cell Structure by Catherine Oikonomou and Grant Jensen is a phenomenal resource.

BACTERIAL MEMBRANES

All cellular organisms have a lipid membrane that encloses their *cytoplasm* and defines the "inside" and "outside" of the cell. The lipids that make up these membranes vary considerably in structure, but have key features that allow them to assemble into bilayer membranes. First, they are *amphipathic*, meaning they have a polar or charged hydrophilic group as well as hydrophobic *fatty acid* chains. This means that, in aqueous media, the polar headgroups form outer layers around an inner hydrophobic layer.

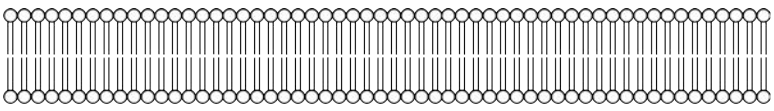


Figure 10.1. A cartoon of a lipid bilayer membrane, made up of amphipathic lipids with polar headgroups (circles) and hydrophobic tails (lines).

Some examples of *E. coli* membrane lipids are shown in Figure 10.2. The *phospholipids* phosphatidylethanolamine (PE) and phosphatidylglycerol (PG) are the most abundant lipids in the *E. coli* membrane (roughly 70% and 20% of the inner membrane, respectively), with cardiolipin (CL) making up an additional 5-10% of the total. Note that since the headgroups of these lipids contain phosphate, the aqueous surfaces of typical bacterial bilayer membranes are negatively charged at roughly neutral pH.

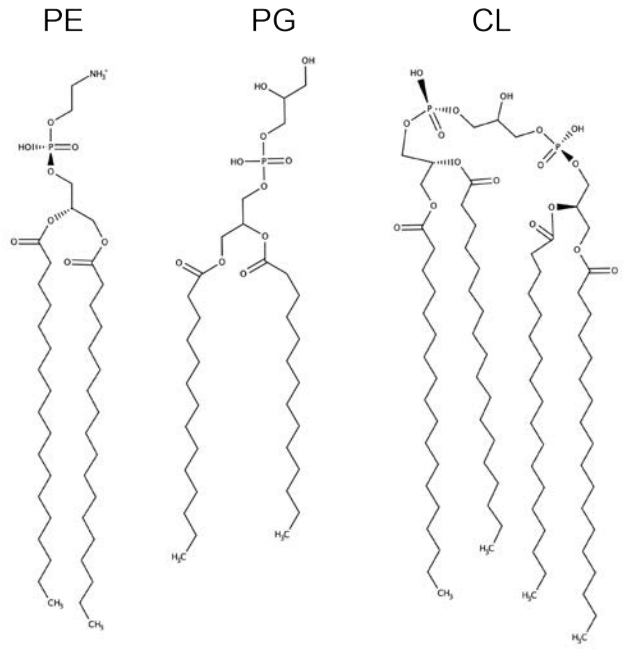


Figure 10.2. The chemical structures of phosphatidylethanolamine (PE), phosphatidylglycerol (PG), and cardiolipin (CL), common lipids in the membrane of *E. coli*. The fatty acids shown are all unsaturated, but bacterial membrane lipids often contain a single double bond or, occasionally, a cyclopropane group. Bacterial lipids do not typically contain more than one double bond.

As in eukaryotes, bacterial membrane lipids consist of fatty acids linked by ester bonds to glycerol. The length of the fatty acid chains can vary, as can their *saturation* (the number and position of double bonds in the fatty acid chain). Different species have different repertoires of membrane lipids, and the membrane composition of a single species can change dramatically according to environmental conditions.

A key property for membrane function is *fluidity*. The cell membrane must be fluid enough to allow free movement of lipids and proteins within the membrane in order for the cell to function. Think of how cooking fats can be solid (butter, shortening) or liquid (vegetable oils), and how easily changes in temperature can soften or harden those fats. The same holds true for membrane lipids. The fluidity or melting temperature of a particular fat depends on the mixture of fatty acids in it: double bonds tend to decrease the melting temperature (butter has more saturated fatty acids, and therefore fewer double bonds, and vegetable oil has more unsaturated ones). Bacteria adjust the mixture of fatty acids they produce to keep their membranes at the correct fluidity.

Bacterial membranes have lots of proteins associated with them, including proteins involved in transporting hydrophilic compounds and proteins from one side of the membrane to the other (see **Lectures 13**, **14**, and **16**), adhesion to surfaces and to other cells (**Lecture 15**), generating energy (**Lecture 16**), and a wide variety of other functions.

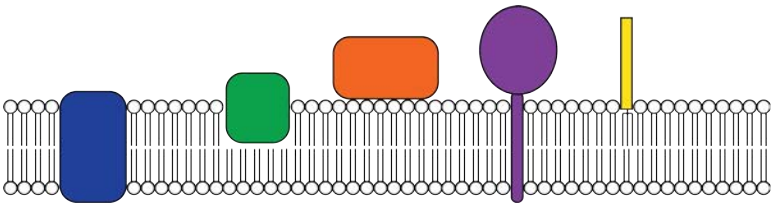


Figure 10.3. Cartoon showing the different ways in which proteins can be associated with membranes.

As illustrated in Figure 10.3, proteins can associate with membranes in a variety of ways. *Integral membrane proteins* have hydrophobic domains that insert into or span the membranes completely, and these may be as little as a single α-helix or as much as a complete hydrophobic β-barrel. *Peripheral membrane proteins* are more loosely attached to membranes, and may interact largely via charge-charge interactions with the polar headgroups of membrane lipids. *Lipoproteins* have covalently attached lipid groups that insert into the membrane and anchor the protein there. Membrane-associated proteins may also, of course, associate with each other to form complexes or arrays in the membrane.

CELL WALLS

Most bacteria have rigid cell walls that define their cell shape. These are composed of crosslinked polymers of sugars and amino acids called *peptidoglycan* or *murein*. The sugar polymers are long chains of alternating *N*-acetylmuramic acid (MurNAc) and *N*-acetylglucosamine (GlcNAc) amino sugars. In *E. coli*, as shown in Figure 10.4, each MurNAc sugar has a pentapeptide of *L*-alanine, *D*-glutamate, *meso*-diaminopimelate (*m*DAP, a lysine derivative), and two *D*-alanines attached. These are cross-linked to other pentapeptides by the formation of peptide bonds between *m*DAP and *D*-alanine to form the mesh-like cell wall. The nature and abundance of the cross-linked peptide component of the cell wall, as well as the extent of cross-linking, varies considerably among different bacteria. In *S. aureus*, for example, MurNAc has an *L*-alanine, *D*-glutamine, *L*-lysine, *D*-alanine tetrapeptide, and the peptidoglycan includes a 5-glycine peptide bridge between linked lysine and *D*-alanine residues.

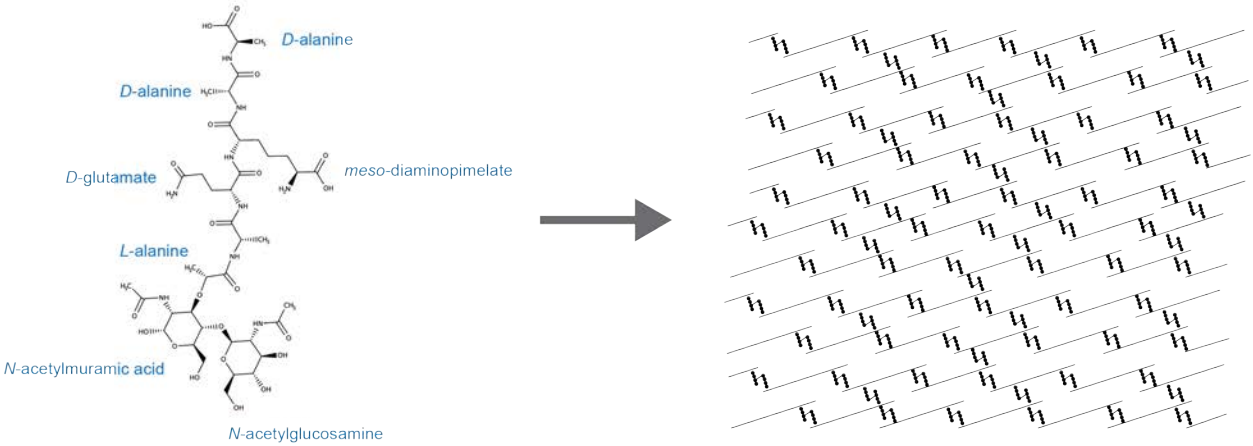


Figure 10.4. The chemical structure of an *E. coli* peptidoglycan monomer, which polymerize and crosslink to form the peptidoglycan sacculus of the bacterial cell wall. The amino acids and type of crosslinking varies greatly among different bacterial species.

The entire crosslinked shell of peptidoglycan is called a *sacculus*, and cells from which the cell wall has been removed by enzymatic treatment are called *sphaeroplasts*. As you might expect, such cells lose their shape and become extremely sensitive to lysis by *osmotic stress* (changes in the concentration of solutes in the environment).

Different bacteria also contain different amounts of peptidoglycan. This can range from a 7-nm thick single layer making up about 10% of the dry weight of Gram-negative bacteria (see below) to a 30- to 100-nm thick shell making up 20-25% of the dry weight of Gram-positive bacteria. In either case, the peptidoglycan mesh is generally permeable to molecules less than 2 nm in diameter, or proteins of less than 25-50 kDa. Note the presence of abundant *D*-amino acids in peptidoglycan, in contrast to the *L*-amino acids found in proteins.

The monomers of peptidoglycan are synthesized, of course, in the cytoplasm. They are exported across the membrane by a process we'll explore in some more detail in **Lecture 14** and added to growing glycan chains by *glycosyltransferases*. These are cross-linked by *transpeptidases* (of both the *DD*- and *LD*- variety), which belong to a family of enzymes also called ***p**enicillin-**b**inding **p**roteins* (PBPs). *Endopeptidases*, *amidases*, and *lytic transglycosylases* are all enzymes that break different bonds in peptidoglycan, which is necessary to allow new material to be incorporated into the sacculus as the cell grows and divides.

Many bacteria (including *Clostridium difficile, Bacillus anthracis, Deinococcus radiodurans*, and *Caulobacter crescentus*) also have an outermost *S-layer* composed of a shell of identical proteins or glycoproteins that cover the entire **s**urface of the cell (hence the name). These can be important for protection against environmental stresses, attachment, structural stabilization, antigenic variation, or a variety of other functions, and can vary dramatically among even closely related species.

GRAM-NEGATIVE BACTERIA

In 1884, Danish scientist Hans Christian Gram published a method for differentially staining bacteria with the dyes crystal violet and safranin, which results in some kinds of bacteria taking on a dark purple color (*Gram-positive*) and others turning pink (*Gram-negative*) when viewed through a microscope. The *Gram stain* remains a useful tool in diagnostic microbiology. The differences in cell structure that distinguish Gram-negative and Gram-positive bacteria were not discovered until 80 years later, in 1964, when Howard Bladen and Stephen Mergenhagen used thin-section electron microscopy to show that Gram-negative bacteria have two membranes, separated by a narrow *periplasm*

(also called the *periplasmic space*). Because of this "two-skinned" structure, Gram-negative bacteria are also called *diderms* or *diderm bacteria*.

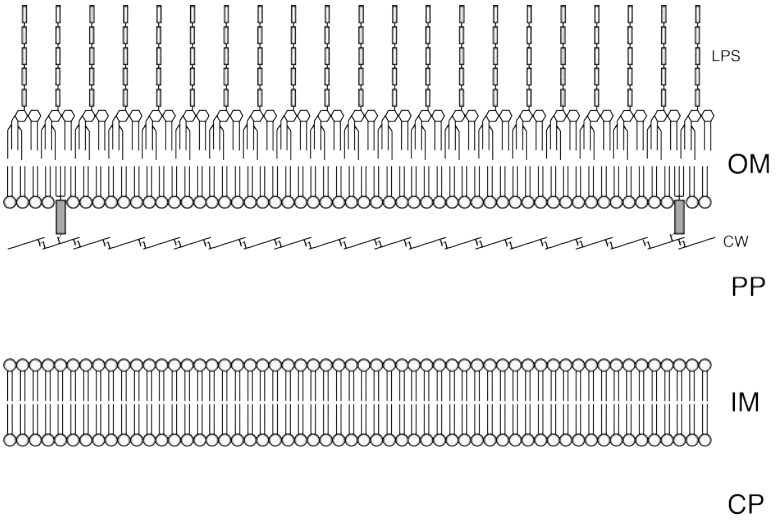


Figure 10.5. A schematic view of the cell envelope of a Gram-negative bacterium. Abbreviations: CP, cytoplasm; IM, inner membrane; PP, periplasm; CW, peptidoglycan cell wall; OM, outer membrane; LPS, lipopolysaccharide. Grey rectangles represent Braun's lipoprotein (Lpp).

The periplasm is about 20 nm across, and it estimated to constitute between 5 and 20% of the total volume of a Gram-negative cell. It contains a distinct set of proteins, including many involved in nutrient acquisition and stress resistance. In *E. coli*, roughly 15% of the proteome is targeted to the periplasm (see **Lecture 13**). However, it does **not** contain ATP or other small, physiologically important cytoplasmic molecules like NADH, $FADH_2$, or glutathione.

The inner and outer membranes are quite different in both lipid and protein content. Major **o**uter **m**embrane **p**roteins include the *porins* OmpC, OmpF, OmpD, *etc*., which are β-barrel proteins that form channels through the outer membrane and allow diffusion of solutes and small molecules into the periplasm. The outer membrane is therefore much more permeable than the inner membrane. This means that, for example, while the pH of the cytoplasm is maintained at neutrality, the pH of the periplasm is the same as that of the outside environment. The exact permeability is regulated by the relative expression of different porins under different environmental conditions. The most abundant protein in *E. coli* by number (about 500,000 copies per cell) is the lipoprotein Lpp (also called Braun's lipoprotein), which is attached to the inner surface of the outer membrane by a covalently-attached lipid group. About 1 in 3 Lpp proteins are also covalently bound to peptidoglycan (by an *LD*-transpeptidase), securely attaching the outer membrane to the cell wall. In many Gram-negative bacteria, the cell wall is covalently bound to outer membrane β-barrel proteins, which has the same general functional role.

Notably, the outer membrane is asymmetric with respect to its lipid content. The *inner leaflet* (facing the periplasm) is similar in composition to the inner membrane, but the *outer leaflet* is composed almost entirely of *lipopolysaccharide* (LPS), a complex molecule with both lipid and carbohydrate components. LPS is also sometimes called *endotoxin*, because the human immune system can mount an extremely aggressive immune response to LPS, leading to septic shock. (This term was coined by Richard Pfeiffer while working in Robert Koch's lab in the late 1880's, to distinguish it from *exotoxins*, which are secreted away from the bacterial cell.)

The innermost component of LPS, called *Lipid A*, consists of two phosphorylated glucosamine sugars with four to six ester-linked fatty acids as the lipid component that forms the hydrophobic interior of the membrane. Attached to one of the glucosamines is the "core region", a branched and phosphorylated polysaccharide that is divided into inner and outer cores. The core region of *E. coli* contains two keto-deoxyoctulosonate, four *L*-glycero-*D*-manno-heptose, three glucose, and one galactose sugar in a characteristic arrangement, but this varies widely from one species to another. The rest of LPS is known as the *O-antigen*, and is extremely variable, even within a single species. It is typically a repetitive polysaccharide composed of up to 40 three- to five-sugar repeat units (see **Lecture 14**). At least 20 sugars are known to occur in O-antigens in different bacteria, including some that are very rare in other contexts. More than 160 antigenically-distinct O-antigens are known for *E. coli* alone, and are a key tool for distinguishing pathogenic strains in a clinical context (think of enterohemorrhagic *E. coli* O157:H7, where "O157" is the serotype of that strain's O-antigen).

As the outermost surface of the Gram-negative cell, LPS is important for interactions between bacteria and host microbes (both pathogenic and symbiotic), ranging from inhibition of phagocytosis in *Salmonella*, to mimicry of human

Lewis blood group antigens by *Helicobacter pylori*, to a role in establishment of symbiotic interactions with legume plants by nitrogen-fixing *Rhizobium* bacteria.

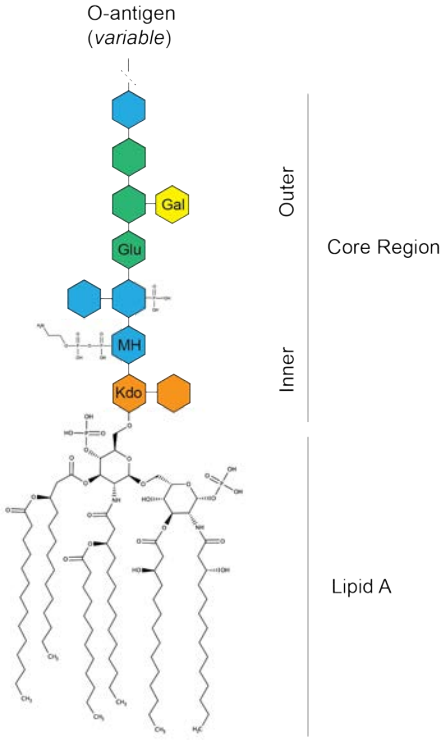


Figure 10.6. The structure of lipopolysaccharide from *E. coli*. Kdo = keto-deoxyoctulosonate, MH = *L*-glycero-*D*-manno-heptose, Glu = glucose, Gal = galactose. The O-antigen is a polymer of hexose and hexosamine repeats and is highly variable from strain to strain.

Many bacteria important in human, animal, and plant health and the environment have what we now know to be a typical Gram-negative bacterial cell envelope. These include human commensals like *Escherichia coli* and *Bacteroides thetaiotaomicron*, important pathogens like *Salmonella enterica, Pseudomonas aeruginosa*, *Vibrio cholerae*, and *Yersinia pestis*, plant pathogens like *Pectobacterium* (formerly *Erwinia*) *carotovera*, and vast numbers of others, including the green photosynthetic cyanobacteria and the predatory bacterium *Bdellovibrio bacteriovorus*. Basically, in any environment where bacteria are found, you will find Gram-negative species.

GRAM-POSITIVE BACTERIA

The Gram-positive bacteria, members of the phyla *Firmicutes* and *Actinobacteria*, have cell envelopes quite different from those of Gram-negative bacteria. They are *monoderms*, and have only a single cell membrane with a (typically) very thick peptidoglycan cell wall outside of it.

In addition to peptidoglycan, Gram-positive cell walls contain large amounts of *teichoic acids*, which are polymers of repeating sugars and phosphate groups that are either anchored in the cell membrane by lipid groups (***l****ipo****t****eichoic* ***a****cids* or LTAs) or covalently attached to peptidoglycan (***w****all* ***t****eichoic* ***a****cids* or WTAs). Teichoic acids can be highly charged, with both positively charged amino groups and negatively charged phosphate groups. There are also anionic and uncharged teichoic acids in some species. Teichoic acids in general are important for cell wall rigidity, maintenance of cell shape, attachment of S-layer proteins, chelating cations like sodium and magnesium, and also play roles analogous to those of LPS in interactions with the host immune system and other kinds of host-microbe interactions.

Lacking an outer membrane to keep extracellular proteins from diffusing away, many secreted proteins in Gram-positive bacteria are covalently bound to peptidoglycan (by enzymes called *sortases*) or bound to teichoic acids. There is some data that argues that the innermost volume of the cell wall space in Gram-positives, between the outer surface of the cell membrane and the bottom of the peptidoglycan matrix, is different enough in composition from the environment to constitute a "Gram-positive periplasm", but this is somewhat controversial.

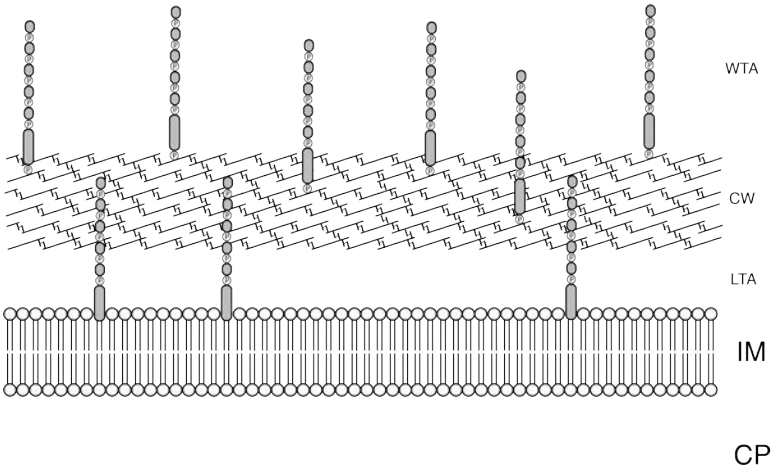


Figure 10.7. A schematic view of the cell envelope of a Gram-positive bacterium. Abbreviations: CP, cytoplasm; IM, inner membrane; LTA, lipoteichoic acids; CW, peptidoglycan cell wall; WTA, wall teichoic acids.

Gram-positive bacteria are also ubiquitous in humans, animals, and the environment. They include human pathogens like *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Clostridium difficile*, and *Bacillus anthracis*, as well as commensals like *Bifidobacterium bifidum* and *Faecalibacterium prausnitzii*. The Gram-positive lactic acid bacteria (commonly found on plants, especially fruit) are important in food fermentation and preservation (e.g. *Lactococcus lactis, Streptococcus thermophilus, Staphylococcus carnosus*) as well as as health-promoting probiotics (e.g. *Lactobacillus reuteri* or *Lactobacillus rhamnosus*). In the environment, the filamentous Gram-positive *Streptomyces* species and their relatives (the actinobacteria) are very abundant and are notable for producing many complex chemical compounds, including most of the known kinds of natural antibiotics (see **Lecture 18**) and petrichor, the characteristic aroma of damp soil. Many Gram-positive bacteria are able to differentiate into extremely stress-tolerant *spores* (see **Lecture 11**). Before the advent of more sophisticated taxonomic methods, aerobic spore-forming rods were given the genus name "*Bacillus*" while anaerobic spore-formers were called "*Clostridium*". Neither of those phenotypically-defined "genera" turned out to be *monophyletic*, so each has now been divided into many more phylogenetically meaningful genera.

I will note here that not every microbe that stains Gram-positive with crystal violet is a monoderm bacterium (the diderm *Deinococcus radiodurans* stains Gram-positive, as do eukaryotic yeast cells), and some monoderm bacteria do not stain strongly Gram-positive (late stationary phase cultures of *Clostridium* species are notorious for appearing Gram-negative). Modern phylogenetic methods do not rely on staining, and the current *phylogenetic tree* of the bacteria, a rough version of which is illustrated below, is based on the sequences of highly conserved genes (most commonly 16S ribosomal RNA, although more comprehensive methods incorporate other genes as well).

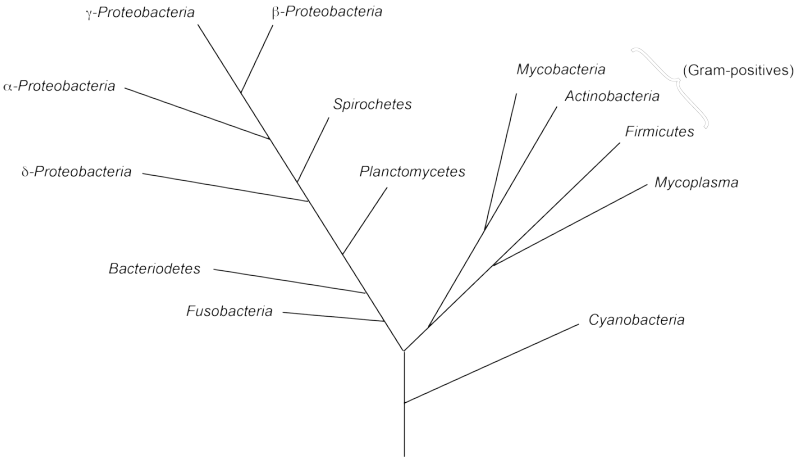


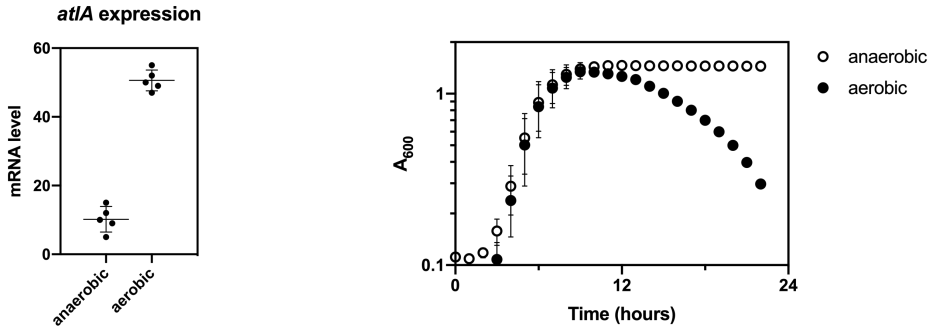
Figure 10.8. The approximate phylogenetic relationships between major groups of bacteria, as determined by the sequence of their 16S rRNA and other highly conserved genes.

DISCUSSION PROBLEM SET #19: REGULATION OF AUTOLYSIS IN *STREPTOCOCCUS GORDONII*

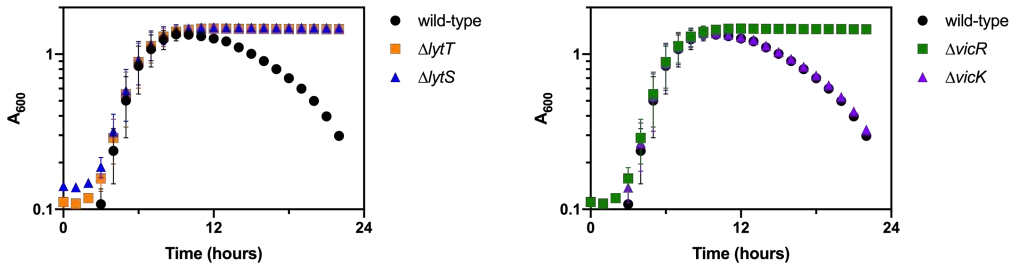
Streptococcus gordonii is a Gram-positive lactic acid bacterium that is an early colonizer of the oral cavity and member of the healthy oral microbiome. It is able to antagonize the growth of the oral pathogen *S. mutans*.

The genome of *S. gordonii*, like that of many streptococci, encodes an *autolysin*, an enzyme with peptidoglycan-degrading *N*-acetyl-muramidase activity that can degrade the cell wall and cause cell lysis. The *S. gordonii* autolysin is encoded by the *atlA* gene, which is in an operon with genes encoding a two-component system histidine kinase (*lytS*) and response regulator (*lytT*) (recall TCS regulators from **Lecture 4**). The signal that LytS responds to is unknown.

Expression of *atlA* increases in the presence of oxygen, and aerobic cultures lyse more rapidly than anaerobic ones:



There is a known TCS in *S. gordonii* that responds to oxygen, composed of the VicK histidine kinase and VicR response regulator. Mutants lacking each of the TCS regulators have the following **aerobic** growth phenotypes:



Propose a model to explain the regulation of AtlA expression in response to oxygen. Design an experiment to test this model. State:

- your hypothesis, and how your experiment will test that hypothesis
- the independent and dependent variables
- both positive and negative controls
- a description of how you will construct any necessary strains
- potential outcomes of your experiments, and how you will interpret them

For discussion in class: What advantage(s) might bacteria gain from expressing a protein whose sole purpose seems to be to break down the cell's own cell wall? (Or from bacterial *programmed cell death* in general.)

MYCOBACTERIA

Mycobacteria are members of the genus *Mycobacterium*, and include both free-living environmental species and obligate pathogens. The most important of these in human medicine are *Mycobacterium tuberculosis*, the cause of tuberculosis, and *Mycobacterium leprae*, the causative agent of leprosy. Phylogenetically, mycobacteria are members of the Gram-positive actinobacteria, but they have a unique cell wall structure that sets them apart from other members of that clade, and in fact, they cannot be stained by the Gram stain procedure at all. They do not take up the crystal

violet stain very well, and have long been known as "acid-fast bacilli", due to their resistance to destaining by acids, both of which properties are due to their unusual cell envelope structure.

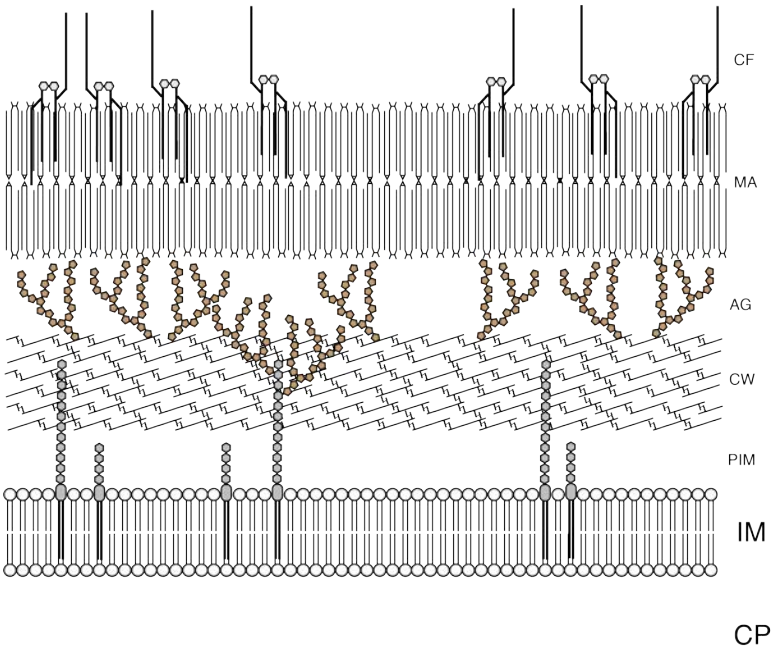


Figure 10.9. A simplified schematic view of the cell envelope of a mycobacterium. Abbreviations: CP, cytoplasm; IM, inner membrane; PIM, phosphatidylinositol mannosides; CW, peptidoglycan cell wall; AG, arabinogalactan; MA, mycolic acids; CF, cord factor (trehalose dimycolate).

As in other bacteria, the inner membrane of mycobacteria is composed mostly of phospholipids, including phosphatidylethanolamine, phosphatidylinositol, and cardiolipin, but with an very high abundance of phosphatidylinositol mannosides (PIMS) that are not found in other groups of bacteria. These glycolipids span the space between the inner membrane and the peptidoglycan cell wall, and both the PIMs and the peptidoglycan are decorated with highly branched polymers of arabinogalactan that form an additional layer outside of the cell wall. Attached to the arabinogalactan layer is another lipid bilayer that is chemically very different from that of the inner membrane or of the outer membrane of Gram-negative bacteria. It is composed of *mycolic acids*, which are long-chain (70 – 90 carbon atoms), extremely hydrophobic lipids unique to the mycobacteria.

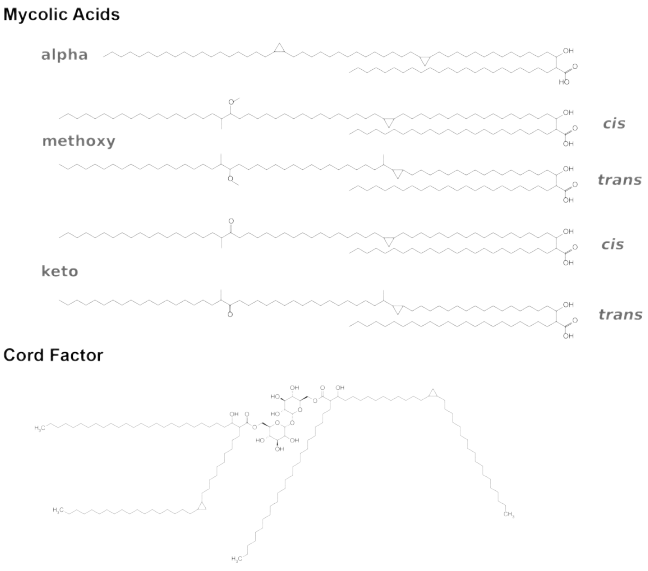


Figure 10.10. The chemical structures of mycolic acid-containing lipids from *Mycobacterium tuberculosis*. (Images from Wikimedia Commons.)

Mycolic acids make up at least 50% of the dry weight of a mycobacterial cell, and their waxy, hydrophobic properties are largely responsible for the unique nature of the mycobacterial cell envelope. Cord factor (trehalose-6,6'-dimycolate) is an abundant mycolic acid-containing glycolipid associated with virulence and resistance to antimicrobial compounds in *M. tuberculosis*. There are a complex variety of other lipids in the outer membrane of mycobacteria, including other glycolipids, phthiocerol dimycocerosate, and sulfolipids, all of which contribute to the stability and overall impermeability of mycobacterial cells.

Compared to other groups of bacteria, mycobacteria are generally very resistant to toxins, stressful environmental conditions, and antibiotics. This is largely due to their cell envelopes, and the fact that very few compounds can penetrate the many layers to reach the interior of the cell. This is particularly important when considering how pathogenic mycobacteria resist attack by the host immune system.

Mycobacteria face an interesting physiological challenge in transporting compounds across their impermeable cell envelopes. In order to survive, they must import nutrients and export waste products, surface-associated proteins, and (as has recently been discovered by the Niederweis lab at UAB) toxins. The mechanisms by which they accomplish these transport functions are not completely understood, and are an area of active research. See, however, the discussion of type VII secretion in **Lecture 13** for one mechanism.

MYCOPLASMAS

Mycoplasmas are another group of bacteria that are phylogenetically related to the Gram-positive bacteria, but which have distinct cell envelope structures. In the case of the mycoplasmas, they lack cell walls entirely. They are also extremely small, both physically (less than half a micron in diameter in many cases) and genomically. The genome of *Mycoplasma genitalium*, a pathogen that causes urethritis, contains fewer than 600 kilobases of DNA, and that of *Mycoplasma pneumoniae*, the causative agent of "walking pneumonia", is just over 800. This makes the mycoplasmas the smallest and simplest free-living cellular organisms currently known.

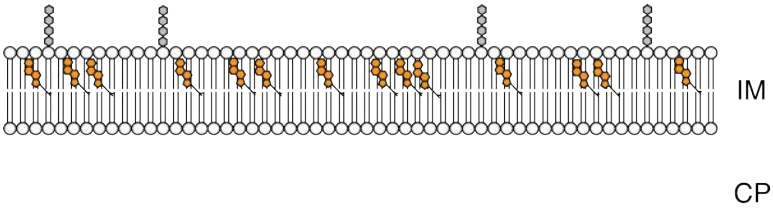


Figure 10.11. A schematic view of the cell envelope of a mycoplasma. Abbreviations: CP, cytoplasm; IM, inner membrane. Cholesterol is indicated in orange, and cell surface glycolipids are indicated in grey.

Unlike other bacterial membranes, most mycoplasma membranes contain large amounts (25-30% by weight) of the sterol lipid cholesterol, which they cannot synthesize, but must obtain from eukaryotic cells. The function of cholesterol is to help maintain membrane stability and fluidity. The phospholipids in the outer leaflet of the mycobacterial membrane are heavily glycosylated, which is also important for maintaining membrane integrity. Since this is the outer surface of the cell, these glycolipids are also highly immunogenic, and are a major point of interaction between pathogenic mycoplasmas and the host immune system.

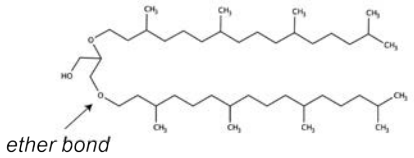
Because they are small and flexible, mycoplasma cells pass easily through filters designed to sterilize media. The pores of these filters are usually 0.2 or 0.45 μm in diameter, large enough to block passage of most bacteria, but not mycoplasmas. Mycoplasmas are also immune to antibiotics that target cell wall synthesis, like ampicillin. This is a serious problem in laboratories that use animal or human cell culture systems, which can easily become contaminated by mycoplasma species, and since mycoplasma cells are so small, this can be difficult to detect by normal microscopic examination.

OTHER PROKARYOTIC CELL ENVELOPES

The four examples described above, which include most of the important groups of human bacterial pathogens, do not encompass the entire diversity of bacterial cell envelopes, the evolution of which appears to have been quite complicated. If you're interested in more details, I'd recommend this review article, which argues convincingly that the common ancestor of all bacteria was probably a diderm, and that modern monoderm bacteria are examples of diderms that have lost their outer membranes. The existence of bacteria like the oral commensal *Veillonella parvula*, a deep-branching member of the *Firmicutes* clade with an outer membrane and periplasmic space, is one example that supports this hypothesis.

The cell envelopes of archaea are dramatically different from those of bacteria or eukaryotes, and are well worth reviewing briefly here, although most of you are unlikely to study archaea intensively during your microbiology careers. Unlike the ester-linked lipids found in bacteria and eukaryotes, archaeal membranes contain **ether**-linked lipids (Figure 10.12). This, along with the occurrence in some species of tetraether lipids that form a lipid monolayer (rather than the more typical lipid bilayer), means that archaeal membranes have very distinctive chemical properties, including the potential for much higher heat tolerance than bacterial membranes. This is probably why the most heat-resistant living organisms are all archaea. (The current record-holder, to the best of my knowledge, is the deep-sea volcanic vent archaeon *Methanopyrus kandleri*, which can grow at 122°C.)

archaeol



caldarchaeol

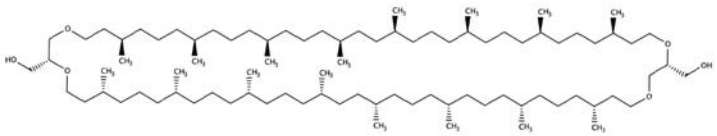


Figure 10.12. Structures of two representative archaeal membrane lipids, the biether lipid archaeol from *Halobacterium* spp. and the membrane-spanning tetraether lipid caldarchaeol from *Pyrococcus* spp. Note the presence of ether bonds rather than the ester bonds characteristic of bacterial and eukaryotic membrane lipids.

Nearly all archaea have S-layers, which we discussed above, and in species that lack cell walls, the S-layer plays an key structural role in maintaining cell shape. Other archaea **do** have cell walls, but they are not made up of peptidoglycan (or of the chitin or cellulose predominant in fungal and plant cell walls, respectively). In some species of methanogens, cell walls are made of *pseudomurein*, a peptide-linked polysaccharide similar in general concept to peptidoglycan, but composed of different sugars and lacking *D*-amino acids. Other methanogens have a cell wall composed of a polysaccharide called methanochondroitin, while *Ignicoccus hospitalis* has an outer lipid bilayer membrane surrounding a "periplasmic space" as much as 0.3 µm across. This is much wider than the periplasm of Gram-negative bacteria, and some data suggests that *I. hospitalis* is able to synthesize ATP in its "periplasmic space", making it functionally extremely different from a bacterial periplasm.

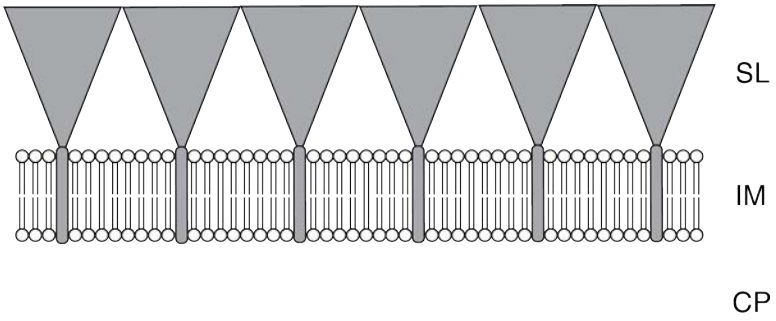


Figure 10.13. A cartoon of a representative archaeal cell envelope, showing the inner membrane composed of both bi- and tetraether lipids and an S-layer (SL), a crystal-like array of glycosylated surface proteins. This arrangement is found in *Sulfolobus* spp., for example, but many archaea have more complicated cell envelopes, which may contain additional protein, carbohydrate, or lipid layers.

Archaea are weird and diverse, and their biology is fascinating, but because there are no known archaeal pathogens, studies of archaea are not funded as intensively as those of bacteria. Some recent hints that methanogenic archaea may play a role in the health and function of the gut microbiome may help to change this in the future, though.

DISCUSSION PROBLEM SET #20: NOT ALL GENETIC SYSTEMS ARE CREATED EQUAL

Up to this point in the class, we've been assuming that all of the tools of molecular biology are available for any organism you might be interested in working on. This is 100% **not** the case, and you will often be limited in what techniques and tools you have available to you in any particular species.

For future problems, I will list what genetic tools are available for the species in question (to the best of my knowledge), and notable limitations for working with that species in the lab, which may limit the kinds of experimental approaches you can take. This is, however, a great opportunity for you to be creative in thinking about ways around the limitations of a particular study organism. If you want to propose **developing** a tool to help you solve a particular problem, that's great, but I will want to hear exactly how you plan to do that.

Here is an incomplete list of techniques and tools which may or may not be available for a particular species:

Essentially available for every species:
- can extract total DNA, RNA, and proteins
- has a complete genome sequence
- is susceptible to chemical or radiation mutagenesis

Available in many or most bacteria:
- can grow in pure culture and/or outside of host cells
- is or can be made competent
- can take up DNA by conjugation
- efficiently carries out homologous recombination with foreign DNA
- has suitable cloning, shuttle, or suicide vectors available
  - has usable selectable and/or counter-selectable markers
  - has known inducible promoters
- has a compatible transposon mutagenesis system

Not available for very many bacteria:
- can carry out oligo-directed recombination
- has a CRISPR-based mutagenesis system (and variations thereof)
- has a genome-scale knockout collection
- has a generalized transducing phage

For discussion in class: which of these do you think is **most important** for being able to do genetic experiments in a bacterial species? In what order would you prioritize developing these tools for a new study organism?

DISCUSSION PROBLEM SET #21: ANTIBIOTIC RESISTANCE IN *FAECALIBACTERIUM PRAUSNITZII*

Faecalibacterium prausnitzii is a strictly anaerobic, non-sporulating, Gram-positive commensal bacterium that is very abundant in the human gut microbiome (making up as much as 5% of the bacteria in the large intestine). It digests dietary fiber and produces butyrate and other short-chain fatty acids which have significant effects on the metabolism of intestinal epithelial cells. Low levels of *F. prausnitzii* are correlated with some diseases, including Crohn's disease, asthma, and major depressive disorder.

The β-lactam antibiotics kill bacteria by inhibiting the DD-transpeptidases (PBPs, or **p**enicillin-**b**inding **p**roteins) necessary for peptidoglycan synthesis. Surprisingly, some isolates of *F. prausnitzii* are sensitive to the first-generation β-lactam ampicillin but resistant to the more advanced β-lactams of the cephalosporin and cefoxitin classes. This suggests that *F. prausnitzii* may be a source of clinically important antibiotic resistance genes that could be transferred to pathogens by horizontal gene transfer. The genome of *F. prausnitzii* encodes only 3 PBPs. (Compare this to the model Gram-positive organism *B. subtilis*, which encodes 11, and is typically sensitive to all β-lactam antibiotics.)

The following limited set of methods are available for *F. prausnitzii*, which is extremely resistant to transformation (for reasons that are not entirely clear, but may have to do with high activity from either restriction enzymes or a native CRISPR system):

growth in pure culture (strictly anaerobic)	✓
can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓

Design an experiment to identify gene(s) or gene features required for either ampicillin sensitivity or cephalosporin resistance in *F. prausnitzii*. State:

- your hypothesis, and how your experiment will test that hypothesis
- the independent and dependent variables
- both positive and negative controls
- a description of how you will construct any necessary strains
- potential outcomes of your experiments, and how you will interpret them

---

LECTURE 11: BACTERIAL CYTOSKELETON AND DEVELOPMENT

INTRODUCTION

This chapter is concerned with the broad topic of bacterial *cell biology*, the study of the molecular mechanisms that determine the shape, size, division, differentiation, and development of bacterial cells. This includes the *cytoskeleton*, composed of different protein filaments with a variety of functions, as well as the complex machinery of cell division. We will also discuss development and differentiation in different bacterial species, and conclude with a discussion of the use of fluorescent protein fusions, which have been invaluable tools in the study of these systems.

THE BACTERIAL CYTOSKELETON

Bacteria contain homologs of all of the main cytoskeletal proteins found in eukaryotes, with a few extras for good measure. These typically have very little primary sequence similarity to their eukaryotic equivalents, but the homology becomes clear when their three-dimensional structures are compared, and they carry out many of the same kinds of functions in maintaining cell shape, organizing organelles within the cell, segregating DNA molecules, and driving cell division. All of the types cytoskeletal proteins are characterized by an ability to form filaments (at least *in vitro*), although the physical properties and dynamics of the resulting filaments differ.

Most rod-shaped bacteria contain one or more actin homologs, usually called "MreB" after the *E. coli* protein (encoded by the second gene in the *mre* **m**u**r**ein formation gene cluster **E** operon). Gram-negative rods typically encode one MreB homolog, while many Gram-positive rods encode multiple copies. Mutations in *mreB* affect the shape of the bacterial cell, with single amino-acid changes in MreB able to change the width, evenness, or curvature of the cells, and even lead to branching morphologies. MreB is essential in most bacteria that encode it, but inhibition or depletion of MreB leads to rounding of cells, loss of rod shape, and eventually (after several rounds of cell division) lysis.

In vitro, MreB assembles into filaments structurally very similar to those of actin. This is affected by the presence of magnesium, other salts, and in some species, ATP or GTP concentrations. *In vivo*, MreB filaments are of varying length, associated with the inside of the cell membrane, and **move** around the circumference of the cell. In *B. subtilis*, the three MreB homologs are colocalized and distributed along the length of the cell throughout growth. In the α-proteobacterium *Rhodobacter sphaeroides*, MreB is found primarily as a ring in the middle of the cell. In *E. coli* and *Caulobacter crescentus*, these two patterns interconvert, depending on the stage of cell division.

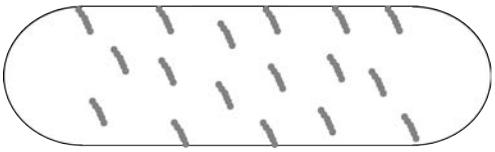


Figure 11.1. MreB filaments distributed around the length of an elongating rod-shaped bacterial cell. The MreB filaments move around the circumference of the cell, and are colocalized with the peptidoglycan synthesis machinery. Early microscopy data suggested that MreB formed a continuous helix, but higher-resolution images clarified that MreB filaments are relatively short, but highly mobile.

MreB is the central organizer of peptidoglycan-synthesis enzyme complexes. MreB colocalizes with the sites of new peptidoglycan synthesis, and peptidoglycan synthesis is necessary to drive the motion of MreB filaments. The protein complex involved is called the *elongasome*, and links the cytoplasmic and periplasmic components of the cell wall synthesis machinery (**Lecture 10**). The diameter of each species' cells correlate with the degree of curvature of that species' MreB filaments, and MreB localization depends on the local geometry of the cell surface. By directing peptidoglycan synthesis to regions of appropriate curvature, MreB creates a feedback loop that maintains the desired cell shape.

MreB and MreB homologs can also have other functions, which vary from species to species, but include roles in chromosome segregation, motility, development, and the positioning of different cellular components.

Many bacteria are not simple spheres or rods, and one of the simpler variations in the rod shape is the presence of varying degrees of curvature or helicity (which are actually the same thing, depending on the degree of curvature relative to the length of the cells).

In *C. crescentus*, cell curvature depends on a protein called *crescentin* (or CreS) that is homologous to eukaryotic intermediate filaments. Crescentin forms non-mobile, stable filaments along the inside of the inner membrane along the short axis of the gently-curved cells and maintains curvature by a combination of mechanical force and local effects on peptidoglycan synthesis. The localization of crescentin depends on MreB activity, but the interactions between these proteins must be in well-conserved regions of MreB, since expressing CreS in *E. coli* is sufficient to curve *E. coli* cells.

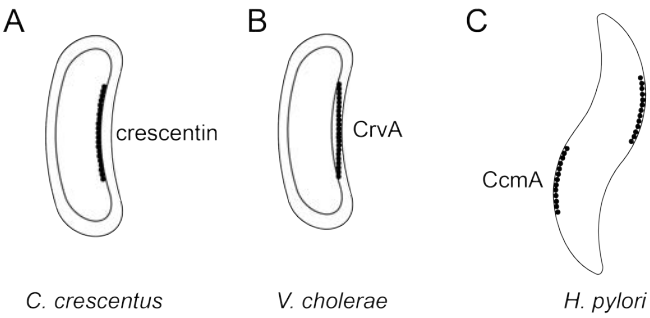


Figure 11.2. The localization of cytoskeletal proteins involved in cell curvature in *Caulobacter crescentus*, *Vibrio cholerae*, and *Helicobacter pylori*.

In *Vibrio cholerae* (and presumably other *Vibrio* spp.), cell curvature is maintained by an unrelated filament-forming protein called CrvA. CrvA is not related to any known eukaryotic cytoskeletal proteins, and, unlike crescentin, assembles into a stable filament in the **periplasm**. CrvA localizes to the inner curve of the cell, where it biases the relative rates of peptidoglycan synthesis and turnover, so that the cell wall grows faster on the other side of the cell.

Another cytoskeletal protein family common in bacteria but absent from eukaryotes are the *bactofilins*. These have a variety of functions, but as one example, in the tightly coiled helical bacterium *Helicobacter pylori*, the bactofilin CcmA localizes to the major axis of the helix and **enhances** pegptidoglycan synthesis (by a mechanism that is still unclear). This is essentially the opposite of the negative effects that CreS and CrvA have on cell wall synthesis, but has the same effect of adjusting cell shape by biasing peptidoglycan synthesis to one side of the cell.

Finally, the most widely conserved cytoskeletal element in bacteria is the tubulin homolog FtsZ (**f**ilamentous **t**emperature-**s**ensitive mutant **Z**), which plays a central role in cell division, a topic that certainly deserves its own section heading:

BACTERIAL CELL DIVISION

Cell division is an exceptionally complicated process, in which DNA replication, chromosome segregation, growth, *cytokinesis* (the separation of the cytoplasm of daughter cells from one another), cell wall synthesis, and daughter cell separation all must be carefully coordinated. I'm not going to go into great detail here, since that would take a textbook all on its own, but the role and function of FtsZ is critical and is an important topic to cover in any discussion of the bacterial cytoskeleton.

With only a few known exceptions, all bacteria contain FtsZ, and the *ftsZ* gene is essential. Like eukaryotic tubulin, FtsZ is a GTPase whose assembly into filaments is dynamic and regulated by GTP binding, which stimulates assembly, and hydrolysis, which stimulates disassembly. FtsZ is the key component of the *Z-ring* that forms at the septum between dividing bacteria cells, recruiting a cascade of proteins to direct cell division (the *divisome*) and providing at least part of the contractile force that shrinks the Z-ring and separates the two daughter cells from one another.

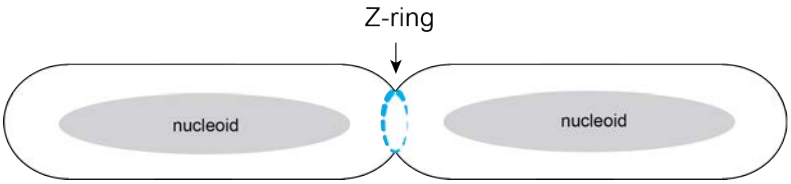


Figure 11.3. FtsZ is the central organizer of the Z-ring and divisome, the protein complex that carries out the process of cell division.

FtsZ in the Z-ring is not contiguous, but forms highly dynamic short filaments that move around the circumference of the cell by *treadmilling*, rapidly polymerizing at one end while simultaneously depolymerizing at the other end. FtsZ filaments are anchored to the cell membrane by proteins called FtsA (an actin homolog) and, in some species (including *E. coli*), an additional protein called ZipA. FtsA, in turn, recruits a complex of peptidoglycan synthesis enzymes that synthesize the septum dividing the two daughter cells. The divisome contains many proteins that are recruited in a very specific order, the details of which are beyond the scope of what we want to discuss here. This review is a good place to start for more information, if you're interested.

Z-rings only form at the site of cell division (mid-cell in bacteria that divide by binary fission), and there are several mechanisms by which this localization is enforced. Here, I will discuss the three partially redundant systems known in *E. coli* as well as a separate system from *C. crescentus*.

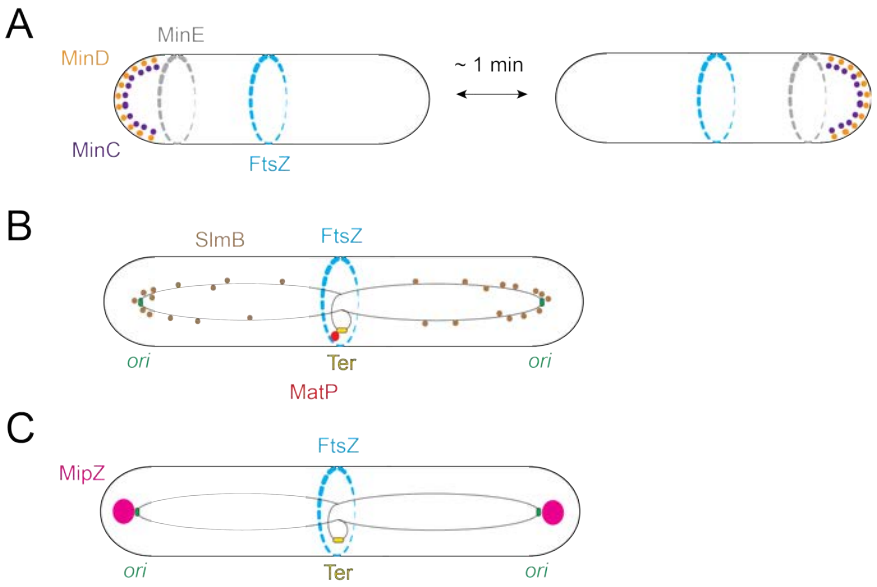


Figure 11.4. Mechanisms of Z-ring positioning. (A) In *E. coli*, MinC (purple) is an inhibitor of FtsZ (blue) polymerization. MinD (orange) localizes MinC to the membrane near the poles of the cell and, due to the activity of MinE (grey) oscillates from one pole of the cell to the other about once every 60 seconds, keeping the concentration of MinC at the center of the cell low. (B) Also in *E. coli*, the FtsZ inhibitor SlmB (brown) binds throughout the genome, but is enriched near *ori* and not present near Ter, while the FtsZ binding protein MatP (red) tethers FtsZ directly to Ter. (C) In *C. crescentus*, the FtsZ inhibitor MipZ (pink) binds to *ori*.

The *E. coli min* (**min**i-cell) mutants were discovered in 1967, and are characterized by the abundant production of tiny, DNA-free "cells" that cannot replicate, but do remain metabolically active for many hours. Ultimately this phenotype was linked to the *minCDE* operon, disruption of which causes the Z-ring to be mislocalized in many dividing cells.

MinC is an inhibitor of FtsZ polymerization. MinD binds to MinC, and tethers it to the cell membrane. MinE is a small (88 amino acid) protein that has multiple functions: it forms filaments and rings, it binds to both the cell membrane and MinD, and it stimulates the ATPase activity of MinD, resulting in MinD monomerization, detachment from the membrane, and the release of MinC. MinD immediately rebinds to ATP, allowing it to reform the complex that tethers MinC to the membrane, but it tends to reassemble as far as possible from the MinE ring. The ultimate result of these dynamic interactions is that the MinCD proteins **oscillate** back and forth from one pole of *E. coli* to the other with a regular period of about 1 minute. This means that, on average, the concentration of MinC is highest at the poles of the cell, and Z-rings tend to form at mid-cell, where MinC is, on average, at its lowest concentration.

B. subtilis lacks MinE, and the MinCD proteins are tethered to the cell poles by proteins called MinJ and DivIVA. There is therefore no oscillation in the Min system of *B. subtilis*, but the principle of minimizing MinC concentration at mid-cell to allow Z-ring formation is the same.

E. coli has two additional known systems for Z-ring positioning, both of which link Z-ring formation with chromosome replication. Most bacteria have circular chromosomes with a single origin of replication (*ori*) and, on the opposite side of the circle, a somewhat larger region called the *Ter macrodomain*, which is where the two replication forks meet and replication terminates. The DNA *replisome*, containing DNA polymerase and both replication forks, tends to be found near the center of the cell, and as the two daughter chromosomes are produced, they are pushed outward, towards the poles. This means that, as the DNA is being replicated, *ori* (which was replicated first) will tend to be found near the poles and Ter (replicated last) will be in mid-cell.

MatP (**ma**crodomain **T**er **p**rotein) binds to the chromosome at *matS* sites in the Ter macrodomain and also forms a protein-protein complex with ZapAB (**Z** ring-**a**ssociated **p**roteins), which are FtsZ-binding proteins. This directly links FtsZ and the Z-ring to Ter, providing a positive localization signal.

SlmA (**s**ynthetically **l**ethal with a defective **m**in system) is a negative regulator of FtsZ polymerization that binds to specific DNA sequences that are found throughout the chromosome, but are enriched close to *ori* and not present in Ter. The activity of SlmA is the basis of a phenomenon called *nucleoid occlusion*, the observation that Z-rings in *E. coli* cannot form around regions of the cytoplasm containing large amounts of DNA. Nucleoid occlusion in *B. subtilis* is mediated by the Noc protein, which is unrelated to SlmA, but functions similarly.

C. cresentus lacks homologs of both the Min and nucleoid occlusion systems, but uses a protein called MipZ (**mi**d-cell **p**ositioning of Fts**Z**) to achieve the same goal. MipZ binds to *ori* and interacts with proteins involved in chromosome segregation, and is therefore localized to the poles. It is also an FtsZ inhibitor. MipZ is therefore more or less the functional opposite of MatP, and provides an *ori*-specific Z-ring inhibition signal.

Other Z-ring localizing systems have been discovered in other bacteria, including filamentous and spherical species, but the mechanisms by which these function are much less well understood.

PLASMID PARTITIONING

As a final example of bacterial cytoskeletal elements, I want to discuss two distinct systems that are involved in the segregation of low copy number plasmids into daughter cells during cell division. High copy plasmids do not typically have dedicated partitioning systems, since the odds of a daughter cell not containing any plasmids is relatively low.

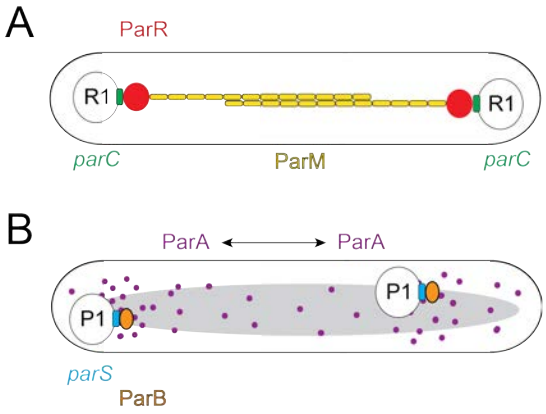


Figure 11.5. Examples of plasmid partitioning mechanisms. (A) For plasmid R1, ParR (red) binds the *parC* site on the plasmid (green). ParM (yellow) polymerization is stabilized by binding to ParR and by interactions with other ParM filaments, pushing the plasmids to the poles of the cell. (B) For plasmid / phage P1, ParB (orange) binds the *parS* site on the plasmid (blue). The MinD homolog ParA (purple) oscillates along the nucleoid (grey), creating concentration gradients along the length of the cell. ParB binds transiently to ParA, and moves P1 to regions of higher ParA concentration.

The R1 plasmid, which confers multi-drug antibiotic resistance to *Salmonella* and other *Enterobacteriaceae*, depends on the actin-like protein ParM to partition replicated plasmids between daughter cells. ParM very dynamically polymerizes and depolymerizes at about equal rates. However, when ParM interacts with ParR, a protein that specifically binds the *parC* site, a DNA sequence near the origin of replication of R1 which is functionally similar to the centromere of eukaryotic chromosomes (which is what the "C" in *parC* stands for), the ParM filament is partially stabilized. If two ParM filaments encounter each other, they can interact to form an antiparallel complex that is much more prone to assemble longer filaments than to disassemble. The end result is that the ParM filaments extend, pushing the attached ParR protein and attached plasmids as far apart as possible within the cell (*i.e.* to the poles).

Bacteriophage P1 exists as a single-copy plasmid in lysogenic *E. coli*, and has the best studied example of a very common plasmid segregation system which consists of the DNA-binding protein ParB and the MinD homolog ParA. Like ParR, ParB binds to a specific DNA sequence on the plasmid, in this case called *parS*. ParA interacts both with ParB and, in a non-specific, ATP-dependent manner, with the DNA of the nucleoid. ParA oscillates along the nucleoid, creating concentration gradients along the length of the cell, and the ParB / plasmid complexes are dragged along these gradients, ultimately settling in positions as far apart from one another as possible. The exact mechanism by which this occurs is an area of active research.

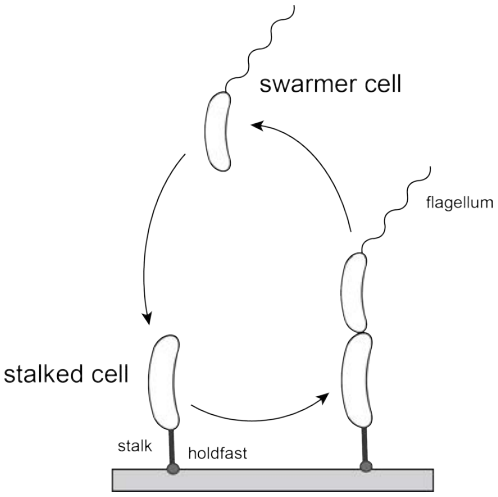
These are by no means the only mechanisms for DNA segregation in bacteria, which are extremely diverse. Some species bacteria use Par systems similar to the plasmid systems described above to segregate their chromosomes. The *C. crescentus* chromosome, for example, has a ParAB system, with *parS* near *ori*, that is required for proper chromosome partitioning between daughter cells.

DISCUSSION PROBLEM SET #22: BACTOFILINS FROM *CAULOBACTER CRESCENTUS*

As you can perhaps tell from the mention of crescentin, MipZ, and ParAB above, *Caulobacter crescentus* is one of the flagship model organisms of bacterial cell biology, largely due to the pioneering work of Lucy Shapiro and many of her

trainees. See this review for a great summary of this fascinating and important organism, whose cell biology differs substantially from that of the more "usual" model organisms *E. coli* and *B. subtilis*.

C. crescentus is an α-proteobacterium found in freshwater streams and other dilute environments under flow. To thrive in these environments, it has a relatively complex cell cycle, alternating between motile "swarmer cells" and surface-attached "stalked cells", as shown in the figure below.



Only the stalked cells divide, budding swarmer cells off from the non-stalked pole. This asymmetric cell division from one pole is typical of α-proteobacteria, and distinct from the binary fission found in other model organisms like *E. coli* and *B. subtilis*. Swarmer cells swim until they encounter a nutrient-rich environment, then differentiate into stalked cells, growing a stalk (or *prostheca*), which is a membrane-enclosed, cytoplasm-containing appendage tipped with an adhesive *holdfast*. The holdfast allows them to stick firmly to surfaces, anchoring them in place in the new location.

The genome of *C. crescentus* encodes two bactofilins: BacA, encoded by the *bacA* gene (locus tag CC1873) and BacB, encoded by *bacB* (locus tag CC3022). Purified BacA and BacB can form filaments *in vitro*. In *C. crescentus*, these proteins form flat sheet-like structures and localize to the stalked pole by an unknown mechanism. Mutants lacking either bactofilin have defects in stalk formation, and there seems to be some interaction between BacAB and MreB.

In order to study the *in vivo* function and assembly of BacA and BacB in isolation, you decide to express these proteins in *E. coli*, which has no native bactofilin homologs. While codon usage is similar in these two organisms, promoters are quite different, and you cannot count on expression signals from one species functioning properly in the other.

All molecular methods are available for *E. coli*, and the following methods are available for *C. crescentus*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓
plasmids can be introduced by conjugation	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
inducible promoters available	✓
compatible transposon	✓

generalized transducing phage	✓

Describe a genetic engineering strategy to express BacA and BacB in *E. coli*. Explain the necessary features of any plasmids you intend to use (**Lecture 5**), the molecular methods you will employ (**Lectures 7 & 8**), and the rationale behind your choices.

Describe an experimental design using your BacAB expression construct(s) to determine the effect(s) of BacAB expression in *E. coli*. State:

- what observations you plan to make
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- potential outcomes of your experiments, and how you will interpret them

---

DEVELOPMENT

Development in bacteria can be defined in a number of different ways, including *asymmetric cell division,* like that of *Caulobacter*, where each daughter cell is genetically identical but has a different phenotype, or *differentiation*, where a proportion of cells in a population change their phenotypes (and sometimes genotypes, in cases of irreversible *terminal differentiation*) according to a defined program.

By its nature, development of either variety involves sophisticated regulation and coordination of gene expression within and between cells, and we will examine just a few fairly well-characterized examples here.

SPORULATION

We will begin by discussing *sporulation* in Gram-positive bacteria, a form of asymmetric cell division where one daughter cell is a metabolically-inactive, highly stress-tolerant *spore*. The spores (technically, *endospores*, since they are formed inside of the *mother cell*) are exceptionally resistant to killing. To kill *Clostridium botulinum* spores and prevent the possibility of food-borne botulism, for example, food canning protocols must ensure that every part of the product spends at least 10 minutes at 121°C. This is why we sterilize bacteriological growth media in an autoclave. Boiling is not sufficient to sterilize any solution that contains spores, which *germinate* to produce growing *vegetative cells* when conditions are permissive for growth. Bacterial spores can remain viable for centuries.

(Before autoclaves were available, English physicist, mountaineer, inventor, and occasional microbiologist John Tyndall discovered that briefly boiling a liquid on several consecutive days [*Tyndallization*], will eventually sterilize it. He hypothesized, correctly, that this is because all of the heat-resistant spores present will have germinated, and the resulting vegetative cells are killed by the boiling. The growth of bacteria in boiled liquids was a real sticking point in mid-nineteenth century debates about spontaneous generation.)

The process of sporulation has been studied in the model organism *Bacillus subtilis* for many decades, and the terminology used here is from that organism. Sporulation in other species differs slightly, but the basic process is conserved. The species that can form endospores have classically been divided into the genera *Bacillus*, for aerobes, and *Clostridium*, for anaerobes, but both of those genera have been subdivided extensively based on modern phylogenetic methods.

Sporulation has classically been divided into a 7-step process: Stage 0 is regular cell division. Stage I is characterized by the formation of a Z-ring off-center in the cell, and the beginning of constriction of the peptidoglycan *septum* in between the *mother cell* and the *forespore*, the cell which will eventually become the spore. In stage II, the DNA is completely segregated between the cells and the cytoplasm of the mother cell and forespore are separated. Next, the mother cell engulfs the forespore, leading to a double-membrane around the forespore in stage III. In stage IV, the peptidoglycan *cortex* is synthesized in between the two membranes of the forespore. In stage V, the outer layers of the spore (the *spore coat*) are added, which consist of a thick modified peptidoglycan layer and several different protein layers. In stage VI the spore is dehydrated and matures into its stress-tolerant final state, and in stage VII the mother cell lyses to release the spore.

The nomenclature of genes involved in sporulation is tied to the 7-stage process described above. Genes which, when mutated, cause the cell to be stuck in stage 0 are called "*spo0*", so, for example *spo0A, spo0B,* and *spo0F* are all genes

that are necessary for the initiation of the sporulation cascade. Similarly, *spoIID* and *spoIIM* mutants have engulfment defects and never reach stage III. There are over 100 genes absolutely required for sporulation in *B. subtilis* and many more that influence the process, which illustrates the complexity of even this "simple" developmental program.

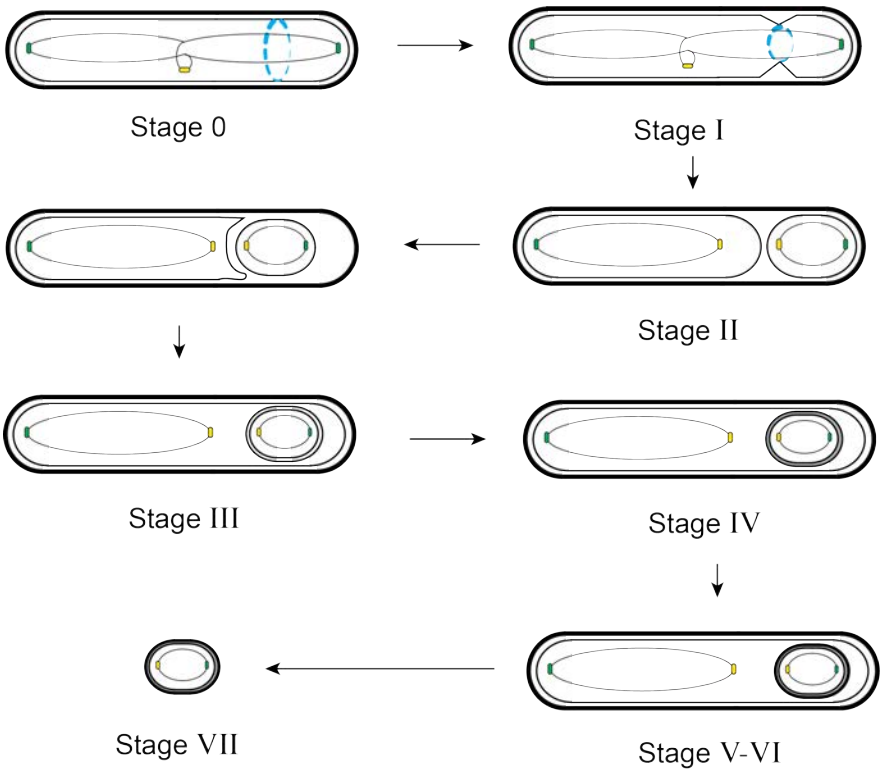


Figure 11.6. The stages of sporulation in *B. subtiis*. Stage 0 is vegetative cell growth, and sporulation begins with the formation of an off-center Z-ring (blue). Stage I begins the separation of the mother cell and forespore, with about 1/3 of the replicated chromosome trapped in the forespore. At stage II, septation is complete and the spore chromosome is completely translocated into the forespore. The mother cell engulfs the forespore at stage III, and a peptidoglycan cortex is assembled between the double membranes of the forespore in stage IV. Stages V and VI involve the assembly of the outer coat layers and final maturation of the spore, and stage VII is the death and lysis of the mother cell, releasing a mature endospore.

The signal that triggers sporulation is nutrient limitation, and in *B. subtilis*, starvation in stationary phase causes about half of the bacterial population to sporulate. Many of the remaining cells lyse, about 10-20% become competent, and some increase their mutation rate, all of which is presumably a bet-hedging strategy to ensure that the population as a whole has the best chance of surviving starvation. We will only consider the regulation of the sporulation cascade here, but many of the same pathways do feed into the regulation of this broader *phenotypic heterogeneity*.

The master regulator of sporulation is the response regulator Spo0A, which is the endpoint of a *phosphorelay* consisting of 5 histidine kinases (KinA-E), two *phosphotransferases* (Spo0B and Spo0F) and several phosphatases and kinase inhibitors (Figure 11.7). Note that phosphorelays in general are composed of more than two of the same histidine kinase and response regulator domains that make up "two-component" regulatory systems (**Lecture 4**). Phosphorylated Spo0A (Spo0A-P) regulates the expression of about 120 genes, and is the trigger that begins the sporulation process.

Among the five histidine kinases, KinA and KinE are cytoplasmic and KinB, KinC, and KinD are transmembrane proteins. They coordinate to sense environmental conditions, but the mechanism by which they do so and what signal(s) they sense is not fully known. KinA and KinB are the most important activators of sporulation, and some evidence suggests that they may be responding directly somehow to slow growth.

KinA and KinB phosphorylate Spo0F, which, in turn, transfers that phosphate to Spo0B, from where it can be transferred to Spo0A. KinC and KinE can directly phosphorylate Spo0A, and KinD acts to reduce Spo0A phosphorylation, although it is not known whether this is direct or indirect.

Acting against the kinases that phosphorylate Spo0A are phosphatases, including Spo0E, which dephosphorylates Spo0A-P, and RapA, RapB, RapE, and RapH, which dephosphorylate Spo0F-P. There are also kinase inhibitors like KipI and Sda that inhibit KinA and/or KinB, and, in turn, negative regulators of those inhibitors (KipA and ClpXP, respectively).

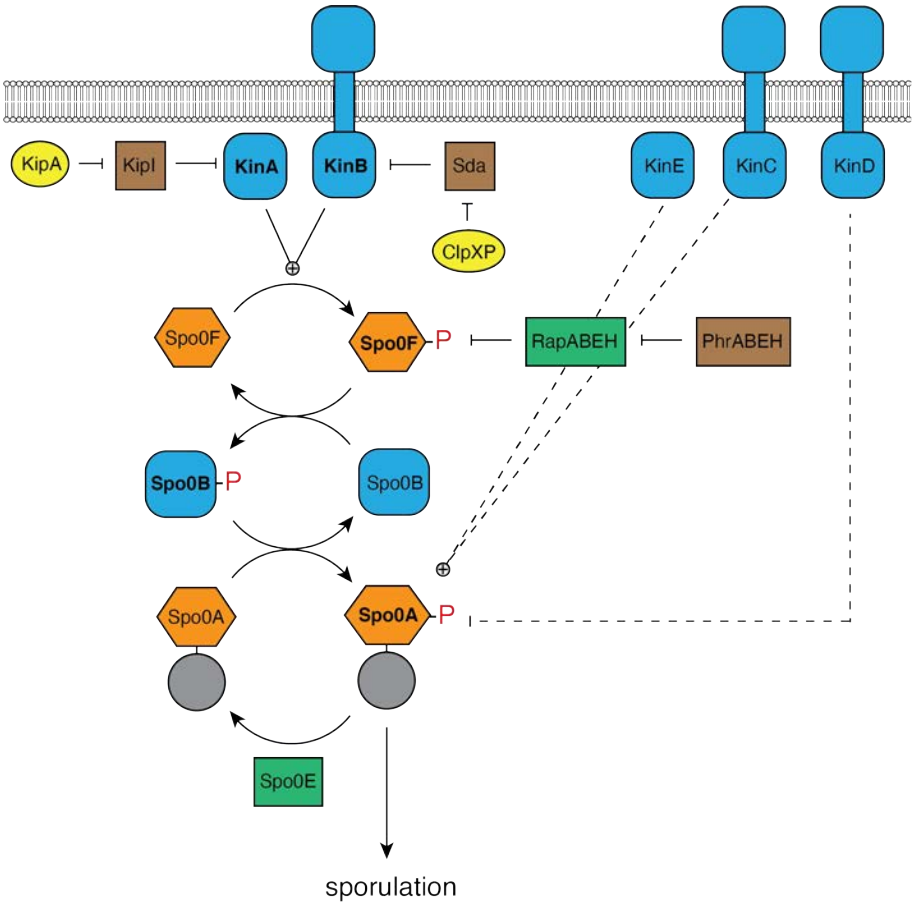


Figure 11.7. The phosphorelay that regulates initiation of sporulation in *B. subtilis*, as described in the text. This is, by necessity, an oversimplification of the actual phosphorelay, which incorporates many additional signals and proteins.

What is the advantage of this intricate phosphorelay? The general idea is that a more complex "input" system allows many different signals and regulators to be integrated into a single "output": in this case, the phosphorylation state of Spo0A. Sda, for example, inhibits sporulation during periods of active chromosome replication, while the Rap phosphatases are inhibited by secreted *quorum sensing* peptides derived from the PhrA, B, E, and H proteins, therefore ensuring that sporulation only occurs at high cell density.

(Like other quorum sensing regulatory signals or *autoinducers*, the Phr peptides are secreted by bacterial cells and only exert their regulatory effects when they pass a threshold concentration in the extracellular milieu, therefore providing a chemical signal of culture density. This is a common mechanism for regulating population-level behaviors in bacteria, although the chemical signals used vary widely. Gram-positive quorum sensing systems often use peptide autoinducers. See **Lecture 14** for more on quorum sensing.)

Once Spo0A reaches a threshold level of phosphorylation and sporulation begins, the developmental progression is driven by a series of alternative sigma factors (**Lecture 4**) that progressively direct different transcriptional programs in the mother cell and the spore. The first two of these are σ^E and σ^F, expression of both of which is activated by Spo0A-P. Both of these sigma factors are normally kept in an inactive state by their respective anti-sigma factors.

At stage I of sporulation, the septum between the mother cell and forespore is almost complete, separating the cytoplasm of the two cells, but only about a third of the replicated chromosome is present in the forespore. The gene for σ^F itself is located near *ori*, but the gene for SpoIIAB (the σ^F anti-sigma, which is a proteolytically unstable protein) is located near Ter. This means that SpoIIAB levels drop in the forespore while σ^F levels rise, leading to expression of the σ^F regulon (about 50 genes) **only** in the forespore. Among the genes activated by σ^F is *spoIIR*, also close to *ori* and

encoding the SpoIIR protein, which is secreted from the forespore into the mother cell. SpoIIR activates a protease (SpoIIGA) that degrades the σ^E anti-sigma factor, leading to expression of the σ^E regulon (about 300 genes) **only** in the mother cell. This is the beginning of the gene expression asymmetry that underlies differentiation.

As development continues, expression of the σ^E-dependent genes in the mother cell generates a signal that leads to activation of the sigma factor σ^G in the forespore at stage III, which in turn, causes expression and secretion of a signal that activates the σ^K regulon in the mother cell at stage V. This "criss-cross" mechanism of compartment-specific sigma factor activation is well-conserved among endospore-forming bacteria, even when many of the details of structural sporulation genes differ.

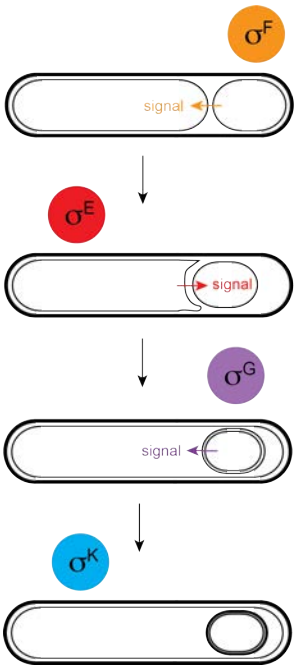


Figure 11.8. Alternating sigma factor activity in the forespore and mother cell that controls the progress of sporulation in *B. subtilis*. Each sigma factor activates expression of a signal that is transmitted to the other cell compartment, leading to activation of the next sigma factor in the cascade.

DEVELOPMENT IN MYXOBACTERIA

The myxobacteria are Gram-negative δ-proteobacteria that exhibit a variety of group or multicellular behaviors, including a multicellular developmental process that lead to the production of stress-tolerant *myxospores*. Myxospores are resistant to UV light, desiccation, and freezing, but are not especially heat-resistant, unlike Gram-positive endospores. As we will discuss below, their development is also quite different.

Myxobacteria are *predatory bacteria* that get their nutrients by killing other microbes. They do not do this as individual cells, but rather as a *swarm* of millions of individual cells that move collectively across surfaces and within the soil environment where they are found. We will discuss the mechanisms of myxobacterial motility in **Lecture 15**. The swarm secretes antibiotics and hydrolytic enzymes to kill, lyse, and digest prey bacteria. Before their bacterial nature was discovered, the myxobacteria were thought to be slime molds, multicellular eukaryotic soil organisms with a similar macroscopic spreading slime appearance and behavior.

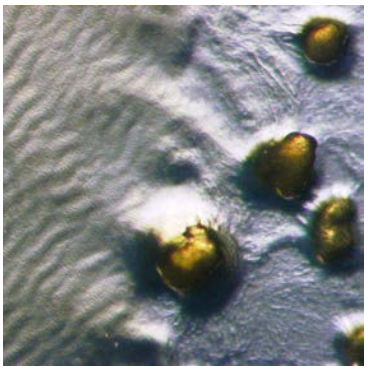


Figure 11.9. Low-magnification image of *M. xanthus*. When prey becomes limiting, fruiting bodies (yellow) form. Zalman Vaksman and Heidi Kaplan, University of Texas Medical School. CC BY 4.0. Obtained from Wikipedia.

Myxospore formation, like that of endospores, is triggered by starvation, which, in the case of the myxobacteria, means that the density of potential prey bacteria has dropped below the threshold necessary to maintain growth of the myxobacterial swarm. Under those conditions, rather than spreading, the individual members of the swarm will move together, clumping together to form *fruiting body* structures. Within the fruiting body, 10% of the cells (usually about 100,000 per fruiting body) differentiate into myxospores, which are stuck together by a dense extracellular polysaccharide (EPS) matrix (see **Lecture 14** for more on EPS). This ensures that when environmental conditions improve and the myxospores germinate, there will be a large enough population of bacteria for efficient predation and growth. Note that, instead of forming within a mother cell during cell division, vegetative cells of myxobacteria differentiate directly into myxospores. Myxospores are metabolically inactive, dehydrated, spherical, and have thicker cell envelope structures than vegetative cells, all of which protect them from adverse environmental conditions.

Myxococcus xanthus is the best-studied model species of myxobacterium, and the regulation and developmental process is best understood in that species, although there are still many unknowns. It involves several interacting regulatory pathways, including two second messengers (**Lecture 4**), the alternative sigma factor σ^{54}, and multiple intercellular communication signals.

There are at least 5 signaling systems that are used by *M. xanthus* for intracellular communication during development, called the A-, B-, C-, D,- and E-signals. While genes required for the production of each signal have been identified (the *asg, bsg, csg, dsg,* and *esg* genes, respectively), only the chemical identity of the A-, C-, and E-signals have been identified, and the functions of only the A- and C-signals are reasonably well-understood. The E-signal is the branched lipid iso-15:0 O-alkylglycerol, but its role in development and regulation is currently unclear.

Starvation signals are sensed by multiple regulators in *M. xanthus*, but one of the key factors required for development is the second messenger guanosine tetraphosphate (ppGpp), which is a conserved stress-responsive signaling molecule across the bacterial domain. This signaling pathway, commonly called the *stringent response*, relies on ppGpp synthases and hydrolases (called RSH enzymes) that control growth rate and metabolism via the interaction of ppGpp with a wide range of proteins, including nucleotide synthases, translation factors, and, notably, RNA polymerase. The transcriptional signature of the stringent response typically includes upregulation of, for example, amino acid biosynthesis genes and downregulation of ribosomal RNAs, and has been extremely well-studied in *E. coli*, *B. subtilis*, and other model organisms.

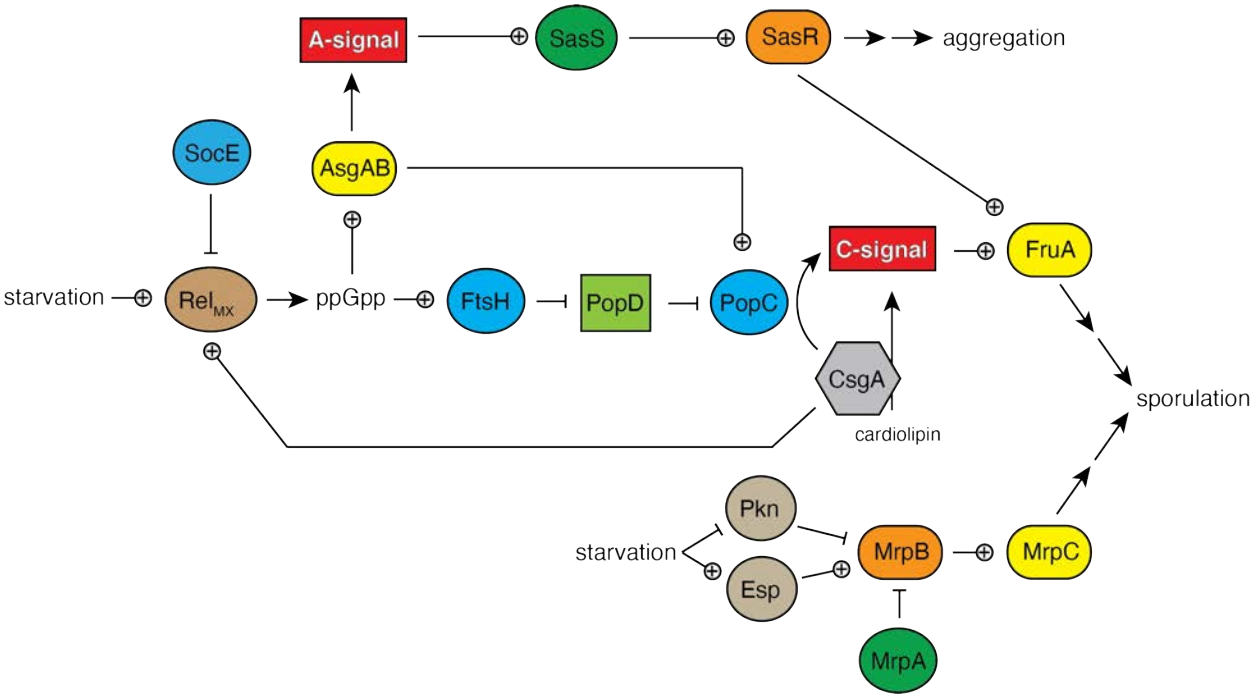


Figure 11.10. Stress-sensing pathways involved in the production of A- and C-signal in response to starvation stresses, leading to sequential expression of genes required for aggregation of fruiting bodies and those necessary for myxospore formation.

In *M xanthus*, ppGpp is synthesized by the RSH protein Rel_{MX}, which is associated with the ribosome and senses amino acid starvation by detecting uncharged tRNAs in the A-site of the ribosome. When ppGpp accumulates, the AsgAB transcriptional activator responds and upregulates genes involved in the production of the A-signal, which accumulates early in development. The A-signal is a mixture of a specific set of amino acids (primarily Y, P, F, W, L, and I) produced by proteolytic cleavage of cell surface proteins. The presence of these amino acids at low concentration (about 10 μM) acts as a quorum sensing signal, detected by SasS, which, in turn, activates SasR, an activator of genes necessary for the aggregation of swarming cells into a fruiting body. The A-signal is thought to ensure that development can only begin when there is high cell density, although, since the A-signal is composed of amino acids, additional regulators (including SocE and CsgA) are needed to maintain ppGpp production by Rel_{MX} during development.

The production of the C-signal, which is important for later stages of development and differentiation into spores, is also dependent on ppGpp. The C-signal is dependent on the cleaved form of the CsgA protein, which localizes to the poles of the *M. xanthus* cell and has cardiolipin cleavage activity. It is unclear whether cleaved CsgA itself or the products of cardiolipin cleavage are the C-signal, but these possibilities are not mutually exclusive. Physical interaction between the C-signal-expressing poles of adjacent cells in a fruiting body are necessary for successful development. CsgA is cleaved by the protease PopC, whose expression is activated by the ppGpp-activated regulators AsgAB, and whose activity is inhibited by the PopD protein. PopD is degraded by the protease FtsH in the presence of ppGpp, so ultimately ppGpp affects C-signal production via two different pathways.

The C-signal is detected by the transcription factor FruA, which activates expression of the genes for sporulation and development. Many of those genes **also** require activation by the MrpC transcription factor, which is activated by ppGpp-independent starvation signals from the Pkn and Esp sensing systems, by way of the MrpB activator of *mrpC* expression.

A more recently-discovered pathway, required for synthesis of the EPS matrix, depends on production of a different second messenger, cyclic di-GMP (often abbreviated c-di-GMP):

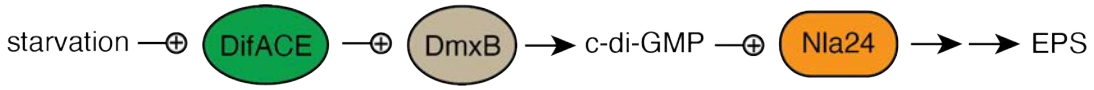


Figure 11.11. The c-di-GMP-dependent pathway leading to expression of genes needed for extracellular polysaccharide synthesis in the fruiting body.

In this pathway, starvation signals are detected by the DifACE sensor system, which leads to activation of the DmxB enzyme. DmxB is a diguanylate cyclase that converts GTP into c-di-GMP. Accumulation of c-di-GMP leads to activation of the transcriptional activator Nla24, which is required for production of EPS. The genes regulated by Nla24 have not been exhaustively identified.

Like ppGpp, c-di-GMP is nearly universally conserved among bacteria, and plays diverse roles in controlling bacterial development, cell cycle progression, virulence, and biofilm formation. We will return to c-di-GMP in **Lecture 14**, when we will be discussing biofilm formation in depth.

Finally, layered on top of the regulatory pathways described above is a cascade of alternative sigma factor-dependent regulons. However, unlike *B. subtilis*, which employs alternative sigma factors homologous to the housekeeping sigma factor σ^{70}, *M. xanthus* depends on its homolog of the alternative sigma factor σ^{54}, which is **not** homologous to σ^{70}. The key feature of σ^{54}-dependent transcription is that σ^{54} on its own cannot activate transcription by RNA polymerase. That means that σ^{54}-dependent promoters require additional activators, which are transcription factors called **b**acterial **e**nhancer **b**inding **p**roteins or bEBPs. *M. xanthus* contains a very high number of bEBPs (53, compared to 11 in *E. coli*), meaning that it is able to very precisely control the activity of multiple σ^{54}-dependent regulons. bEBPs are often activated by phosphorylation, so both their expression and activity can be regulated. *M. xanthus* is somewhat unusual in that the gene encoding σ^{54} is essential.

Three of the regulators we've already discussed are bEBPs: SasR, MrpB, and Nla24, all of which are required for successful sporulation. However, there are at least 7 more that regulate and are required for different aspects of the developmental process. These form what is called the "bEBP cascade", since they are activated sequentially to drive a developmental program. This is analogous to the sigma factor cascade involved in *B. subtilis* sporulation, where each regulon includes activators of the next regulon in the sequence. In the case of the bEBP cascade, each bEBP's regulon includes the gene encoding the bEBP for the next step in the developmental sequence. However, the exact role(s) of the other genes controlled by each bEBP are much less well understood.

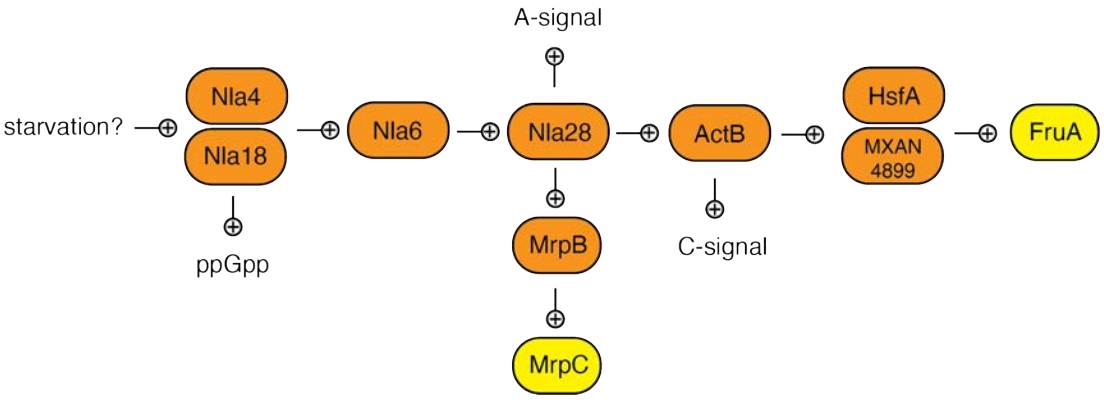


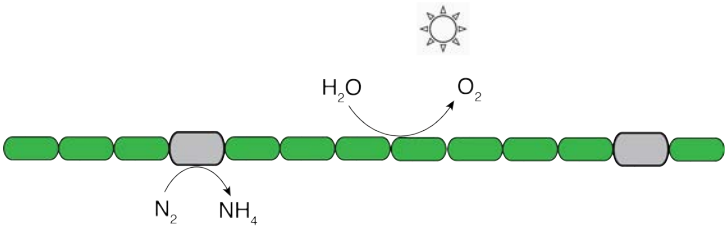
Figure 11.12. The bacterial enhancer binding protein cascade, consisting of sequentially activated σ^{54}-dependent regulons, all of which are required for successful development.

By an unknown mechanism, starvation leads to expression and activation of the bEBPs Nla4 and Nla18. Mutants lacking Nla4 and Nla18 do not accumulate ppGpp, although the mechanism responsible is unknown. Nla4 and Nla18 drive expression of Nla6, which is necessary for expression and activation of Nla28. Nla28 is required for production of both the A-signal and the bEBP MrpB (and therefore of MrpC). Nla28 is also required for expression of the bEBP ActB. ActB, in turn, is necessary for production of the C-signal and the final bEBPs in the cascade: HsfA and MXAN4899 (which really ought to be given a real name). HsfA and MXAN4899 are required for FruA expression. As you might expect, each bEBP drives expression of quite a few genes, some of which have known functions and many of which do not. However, nearly a quarter of the genes in the *M. xanthus* genome are differentially regulated over the course of development.

There remain many unknowns in the field of myxobacterial development, and the regulators of differentiation overlap considerably with the regulators of multicellular predatory behavior and motility. This is an area of active research, although, since myxobacteria are not human or animal pathogens, it is also an area in which progress is relatively slow and limited to a small number of labs worldwide.

DISCUSSION PROBLEM SET #23: CYANOBACTERIAL HETEROCYSTS

Not all bacterial developmental processes lead to the formation of metabolically inactive resting cells. Cyanobacteria are the Gram-negative oxygen-generating photosynthetic bacteria formerly known as the "blue-green algae", and share a common ancestor with the chloroplasts of plants. In filamentous cyanobacteria belonging to the order Nostocales, some cells in a filament are able to differentiate into *heterocysts*, specialized non-photosynthetic cells that fix atmospheric nitrogen (N_2) into bioavailable ammonia (NH_4).



The problem that nitrogen-fixing cyanobacteria face is that nitrogenase, the enzyme that fixes nitrogen, is **very** sensitive to inactivation by oxygen, and oxygenic photosynthesis produces a **lot** of O_2. Under nitrogen starvation conditions, therefore, some individual cells in each filament differentiate, losing the ability to photosynthesize, upregulating O_2-consuming enzymes, and developing a thick, gas-impermeable cell envelope. This is an irreversible process, and heterocysts cannot revert back into photosynthetic cells. The photosynthetic cells provide carbon, in the form of sucrose, to the heterocysts, and the heterocysts, in turn, export fixed nitrogen in the form of glutamine.

The master transcription factor controlling heterocyst formation in the model cyanobacterium *Anabaena* sp. strain PCC 7120 is called HetR, and activation of HetR is required to initiate differentiation. In wild-type cells, the spacing of hetrocysts in a filament is extremely regular. This is regulated by the HetR-dependent expression of a diffusible peptide (PatS) that inhibits heterocyst formation. There is some debate about whether PatS diffuses through channels connecting the cytoplasm of adjacent cells or in the contiguous periplasm shared by all of the cells in a filament.

There are 14 diguanylate cyclases (c-di-GMP synthases) in *Anabaena* sp. strain PCC 7120, but only All2874 is required for heterocyst formation. An *all2874* null does not form heterocysts or express *patS*. However, over-expressing HetR in the *all2874* mutant restores normal heterocyst formation. This indicates that c-di-GMP is in some way acting **upstream** of HetR in the differentiation signaling pathway.

The following methods are available for *Anabaena* sp. strain PCC 7120:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
plasmids can be introduced by conjugation	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
compatible transposon	√

Describe an experiment to identify regulators involved in c-di-GMP-dependent heterocyst formation in *Anabaena*. State:

- your hypothesis
- the independent and dependent variables
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids

- potential outcomes of your experiment, and how you will interpret them

---

FUSION PROTEINS AND BACTERIAL CELL BIOLOGY

In **Lecture 5**, I mentioned the construction of fusion proteins that combine domains from different proteins into a single polypeptide chain. This technique has been very important in the study of the internal structures of bacteria, especially using fusions with fluorescent proteins (most often derivatives of GFP, the **g**reen **f**luorescent **p**rotein from the jellyfish *Aequorea victoria* or of DsRed, a red fluorescent protein from the cnidarian coral *Discosoma*). Note that the term *fluorophore* is used to refer to fluorescent molecules, including both proteins and small molecule dyes, which are commonly used together in microscopy experiments.

In this section, I will discuss how fluorescent protein fusions are constructed in more detail, some considerations about their use, and caveats that must be taken into account when designing experiments with protein fusions.

There are a wide variety of fluorescent proteins available with different properties that make them useful for different kinds of applications. The FPbase database is an excellent resource that describes the sequences and characteristics of fluorescent proteins. Key properties include color (both of the light needed to excite the proteins and of the fluorescent emission), brightness, and maturation time (fluorescent proteins take time to fold into fully active fluorescent forms, and this can vary from a few minutes up to several hours). Most, but not all, fluorescent proteins also require oxygen to mature into their active form, which limits the usefulness of fluorescent protein fusions in anaerobic bacteria (see this review for a discussion of the current state of the art in this area).

Fusions between bacterial proteins and fluorescent reporters (Figure 11.13) can be constructed using the techniques described in **Lectures 7** and **8** for expression in bacterial cells, either from a plasmid or from the chromosome. Fluorescent proteins can be fused to either the N- or C-terminus of a protein or into a surface-exposed loop within the protein. Normally, flexible linkers made up of relatively inert amino acids (serine, alanine, *etc.*) are included between the proteins to minimize possible effects of the fluorescent protein (typically large, stable β-barrel domains) on the structure and function of the protein of interest. However, it is absolutely essential to confirm that the fusion protein retains the function of the wild-type protein.

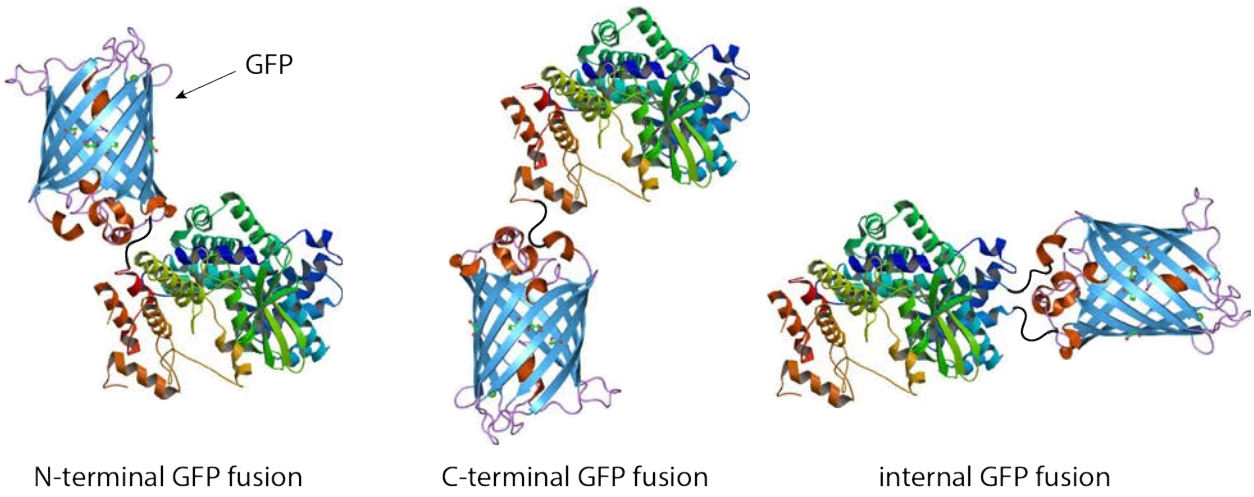
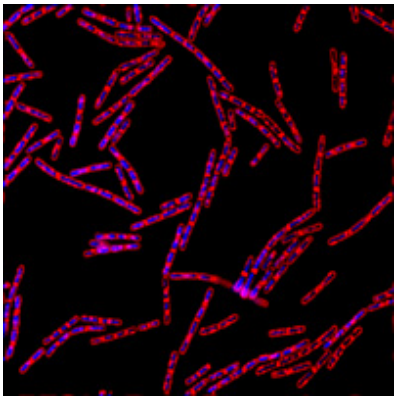


Figure 11.13. Fluorescent protein fusions. Fluorescent proteins can be fused to either terminus of the protein of interest or within a surface-exposed loop, as long as the presence of the fluorescent domain does not influence the folding or activity of the protein of interest. Linkers are represented by black lines. Ribbon structures from Wikimedia Commons.

The expression level of fusion proteins is also a key variable in this kind of experiment. Especially when using plasmids, it is important to not express fusion proteins at much higher levels than the wild-type protein. Not only could this lead to artifacts in visualizing the presence and localization of the protein of interest, high level expression of fluorescent proteins is often toxic to bacterial cells (due to increased production of reactive oxygen species). Careful selection of promoters is important, and Western blotting (see **Lecture 4**) to confirm that protein levels are in the expected range is a common control. It is also a good idea to confirm that the protein of interest remains stably fused with the fluorescent protein *in vivo*, since proteases may degrade either the poorly-structured linkers or the protein of interest itself. Fluorescent protein domains are typically extremely stable and protease-resistant.

If the correct conditions are met, fluorescence microscopy can be used to visualize where protein fusions and, by extension, presumably also where the wild-type proteins localize within the bacterial cell. Since bacteria are so small, this is less precise than with eukaryotic cells, and larger bacterial species (*e.g. B. subtilis*, which are about twice as long and slightly wider than *E. coli*) can be used to make this easier.



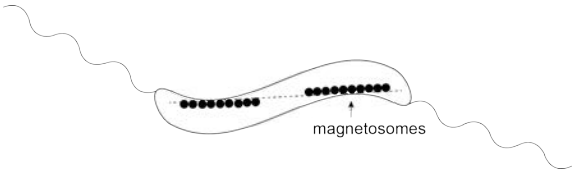
UC Berkeley

Figure 11.14. A fluorescent micrograph of *Bacillus subtilis* cells, illustrating how cellular structures in bacteria can be visualized with fluorophores of different colors.

Confocal microscopy takes successive images of the same bacteria at different planes, allowing the reconstruction of 3D images of fluorescence for more precise localization of fluorescent proteins within a cell. At the absolute cutting edge of fluorescence microscopy, various methods for *super-resolution microscopy* have been developed that allow the detection of single molecules of fluorescent protein within cells (see this Wikipedia article for a nice summary). These require extremely specialized equipment and sophisticated data analysis, but are allowing genuinely astonishing advances in our understanding of how living cells are structured.

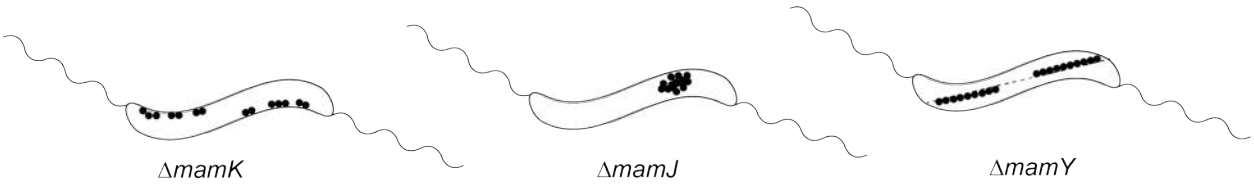
For an up-to-date review on the use of fluorescence microscopy in bacterial cell biology, see this paper.

DISCUSSION PROBLEM SET #24: MAGNETOSOME POSITIONING



Magnetospirillum magneticum is an α-proteobacterium that orients itself along magnetic field lines (so that it can swim up or down in the water column to its preferred oxygen concentration without having to randomly explore 3D space – see **Lecture 15**). Like other *magnetotactic bacteria*, it contains multiple *magnetosomes*, which are membrane-bound organelles containing magnetite (Fe_3O_4) crystals about 50 nm in diameter. The magnetosome membrane is derived from the inner membrane. The dotted line in the figure above indicates the axis along which linear arrays of magnetosomes are arranged in wild-type cells, which is tangent to the inner curvature of the membrane.

By using allelic exchange to make targeted deletions of genes of unknown function in an operon with genes known to be necessary for magnetosome formation, you identify three mutants that have distinct disruptions in the positioning of the magnetosomes:



MamK is an MreB/actin homolog, and MamY has a predicted transmembrane helix.

The following methods are available for *M. magneticum*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓
plasmids can be introduced by conjugation	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposon	✓

Propose a model for the function of the MamK, MamJ, and MamY proteins that could explain these results.

Describe an experiment or series of experiments to test this model. (Fluorescent protein fusions are potentially very useful here, but if you choose a different approach, that's great, too.) State:

- your hypothesis, and which aspect of your model it tests
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

LECTURE 12: CRITICAL READING (BACTERIAL CELL STRUCTURE)

EXPECTATIONS

As a reminder, to prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
    - What is the central **question** of this paper?
    - Is the experimental design clear and appropriate to address that question?
    - Do you understand the methods used?
    - Are the data clearly presented, with appropriate statistics?
    - Do you agree with the conclusions the authors came to based on their data?
    - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Shiratori *et al.* (2019) "Phagocytosis-like cell engulfment by a planctomycete bacterium." Nature Communications 10:5529.

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

LECTURE 13: PROTEIN SECRETION

INTRODUCTION

Proteins are synthesized in the cytoplasm, but many of them need to be targeted to other places in the cell in order to function, including the cytoplasmic membrane, the periplasm, the outer membrane, the extracellular space, and, for many pathogens, into host cells. This is called *protein secretion*, and bacteria have a wide variety of mechanisms to accomplish this goal. In this lecture, I will describe the basic operation of most of the currently known systems for protein secretion in bacteria. Since most protein secretion systems are large multi-protein complexes, I will also discuss genetic methods for detecting and studying protein-protein interactions *in vivo*.

THE SEC PATHWAY

The general secretion (Sec) pathway is conserved among bacteria, archaea, and eukaryotes, and is a very common mechanism by which proteins are transported across the cytoplasmic membrane. In Gram-negative bacteria, this results in substrate proteins being translocated into the periplasm or into the inner membrane itself, while in Gram-positive bacteria substrate proteins are excreted into the extracellular environment, modified to attach them to the cell surface, or inserted into the inner membrane. Sec substrates include many key metabolic proteins, as well as a variety of virulence factors. More than a third of the bacterial proteome is found in the cell envelope, and a large proportion of those proteins utilize the Sec pathway to reach their final destinations.

The Sec system (Figure 13.1) translocates unfolded proteins, which are recognized by the presence of an N-terminal *signal sequence*. For proteins destined to be secreted into the periplasm or further exterior cellular compartments, the newly synthesized proteins are recognized and maintained in an unfolded state by the SecB *chaperone* protein. SecB targets the substrate to the ATPase SecA, which binds to the SecYEG protein channel in the inner membrane. ATP hydrolysis by SecA provides the energy for secretion of the substrate through SecYEG. A *signal peptidase* (also called LepB) proteolytically cleaves off the signal sequence, and the substrate protein folds into its final tertiary structure in the periplasm.

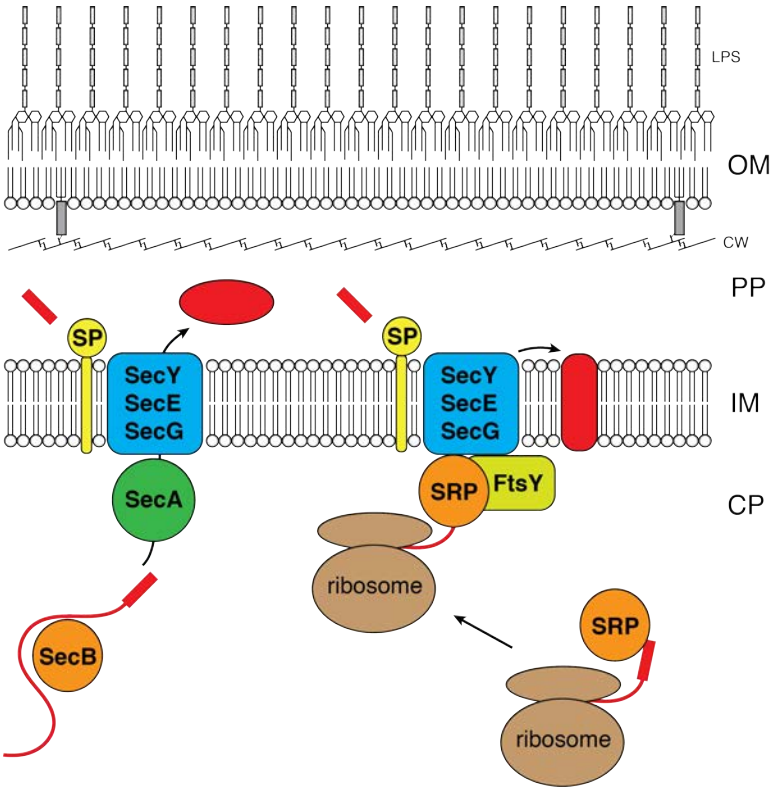


Figure 13.1. The Sec pathway. Proteins targeted to the periplasm (PP) or for secretion out of the cell are expressed with an N-terminal signal sequence (red rectangle) and are maintained in an unfolded state by the chaperone SecB. SecB delivers substrates to SecA, an ATPase which provides the energy to secrete the unfolded substrate through the SecYEG channel in the inner membrane (IM). The signal sequence is removed by signal peptidase (SP; also called LepB), and the secreted protein folds in the periplasm. Proteins targeted to the inner membrane have an N-terminal signal sequence that is recognized as soon as it is translated by the signal recognition particle (SRP; 4.5S RNA and the Ffh protein), which recruits the docking protein FtsY. This complex interacts with SecYEG and the protein is translocated into the IM as it is being translated.

Integral inner membrane proteins exported by the Sec system are recognized by a different mechanism. The signal sequence for these proteins is recognized as soon as it is translated by the ***s**ignal **r**ecognition **p**article* (SRP), a riboprotein complex between 4.5S RNA (encoded by the *ffs* gene) and the Ffh protein. The SRP, along with the docking protein FtsY, assembles a complex between the ribosome and SecYEG, so that these proteins are simultaneously translated and transported, with protein synthesis providing the energy needed to drive secretion. There appears to be a mechanism by which integral inner membrane proteins are secreted out the "side" of the SecYEG channel, so that they end up inserted fully into the membrane.

Although Figure 13.1 illustrates the Sec system in a Gram-negative bacterial envelope, as mentioned above, Gram-positive bacteria also contain this secretion system. In fact, they often have two distinct copies of SecY or sometimes of the entire Sec system, each of which secretes a different set of proteins. Since there is no outer membrane in Gram-positive bacteria, in order to prevent proteins from diffusing away from the cell, Gram-positives use membrane-anchored enzymes called *sortase*s that recognize a conserved LPXTG motif and covalently attach secreted proteins containing that motif to the peptidoglycan cell wall.

Since many inner membrane proteins are required for bacterial growth, the Sec system is essential, although some individual components (e.g. SecB, SecG, signal peptidase) can be knocked out without immediately lethal effects.

THE TAT PATHWAY

Unlike the Sec pathway, which secretes unfolded proteins, the twin arginine translocation (Tat) pathway secretes folded proteins. This is necessary when the proteins to be secreted contain cofactors, metal ions, or post-translational modifications that can only be synthesized in the cytoplasm or for substrates that are secreted as multi-protein complexes. Phospholipase C enzymes, which are virulence factors secreted by *Pseudomonas aeruginosa*, *Mycobacterium tuberculosis*, and other pathogens, are one example of enzymes secreted via the Tat pathway.

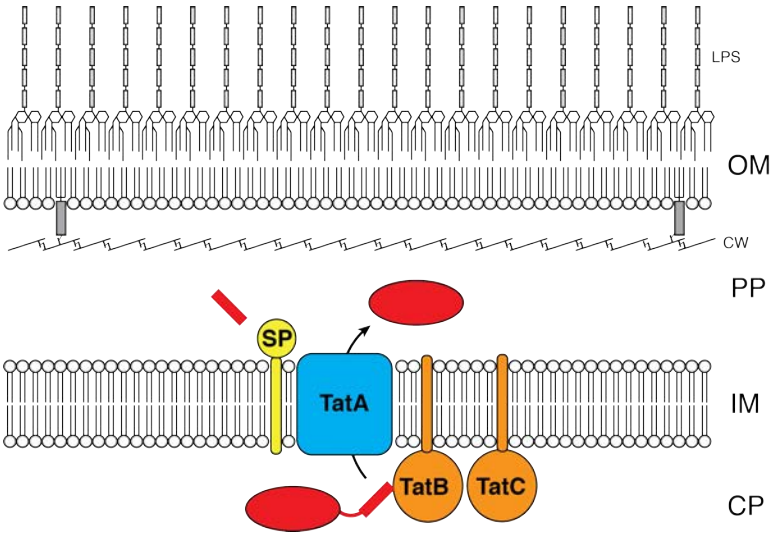


Figure 13.2. The Tat pathway. Folded Tat substrate proteins have an N-terminal signal sequence (containing a twin-arginine motif) that is recognized and bound by the TatB and TatC proteins, then secreted through the TatA channel. The signal sequence is removed by signal peptidase (SP; LepB).

Tat pathway substrates have an N-terminal signal sequence that contains the Ser-Arg-Arg motif that gives the system its name. This signal sequence is recognized by TatB and TatC, which deliver the substrate to the TatA pore, through which it is secreted into the periplasm or (in some cases) the inner membrane. The signal sequence is cleaved by the same LepB signal peptidase used in the Sec system. No ATP is needed for Tat-dependent secretion, and the energy necessary appears to be derived entirely from the *proton motive force* (we will discuss bacterial energetics in **Lecture 16**).

As with the Sec system, both Gram-negative and Gram-positive bacteria contain the Tat secretion system, and homologous systems are found in archaea and some eukaryotes (although not animals). Perhaps surprisingly, the *tatABC* genes are not essential (at least in *E. coli*), although many Tat substrates **are** redox cofactor-containing proteins required for anaerobic respiration (also see **Lecture 16**).

TYPE I SECRETION

Type I secretion systems (T1SS) are found in a wide variety of Gram-negative bacteria and secrete proteins across both membranes in a single step, without involving either the Sec or Tat pathways. The substrates secreted by T1SS include many toxins (*e.g.* HlyA from uropathogenic *E. coli* and MARTX from *V. cholerae*) nutrient acquisition proteins (*e.g.* the iron scavenger HasA from *Serratia marcescens*), and antimicrobial effectors (*e.g.* colicin V from *E. coli*). T1SS are therefore important for virulence in many pathogens.

T1SS consist of three components: an **i**nner **m**embrane **c**omponent (IMC), a periplasmic **m**embrane **f**usion **p**rotein (MFP), and an **o**uter **m**embrane **f**actor (OMF). The IMC is an ATPase of the ABC transporter protein family, the MFP forms a channel that spans the width of the periplasm, and the OMF (usually a protein called TolC) forms a pore in the outer membrane. T1SS substrates are recognized by the IMC and MFP by the presence of a repeated GGXGXD motif in their C-terminus and are transported through the T1SS complex in an unfolded state. Energy for this process is provided by the ATPase activity of the IMC.

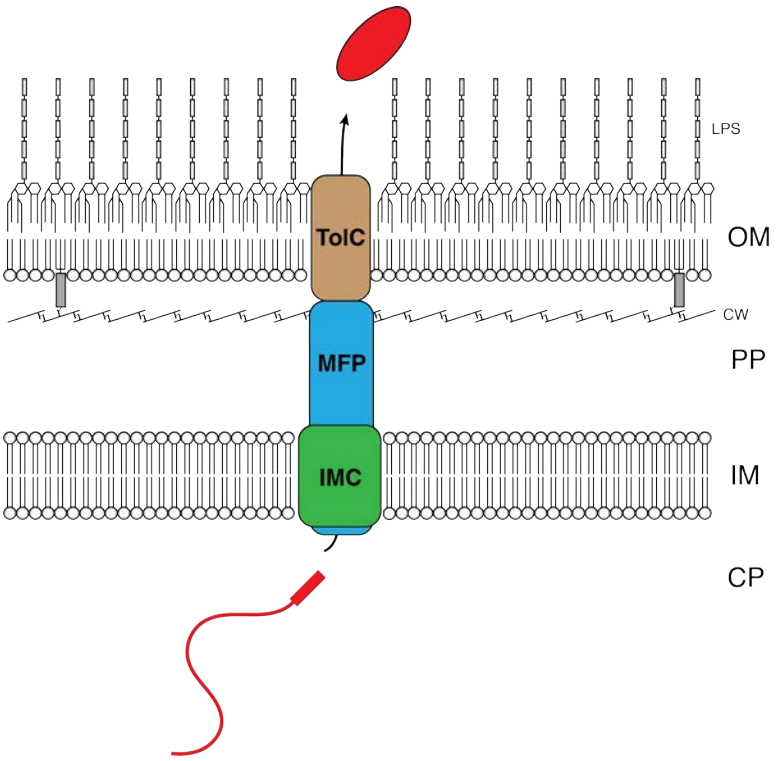


Figure 13.3. Type I secretion. T1SS consist of three components: an inner membrane component (IMC) which is a member of the ABC transporter ATPase family, a membrane fusion protein (MFP) which creates a channel spanning the periplasm, and an outer membrane factor (usually TolC). T1SS substrates are secreted in an unfolded state, and are typically recognized by a C-terminal GGXGXD repeat motif, which is not cleaved off of the secreted protein.

T1SS are homologous to RND efflux pump proteins, which use a related mechanism and the same TolC OMF to secrete small molecules (such as antibiotics) out of the cell.

TYPE II SECRETION

Most Gram-negative bacteria encode homologs of the type II secretion system (T2SS), which exports folded proteins from the periplasm into the extracellular environment. T2SS substrates must be secreted through the inner membrane by either the Sec or Tat pathways, and it is not currently known how the T2SS recognizes its substrates within the periplasm, although presumably there are structural features that act as targeting signals. T2SS substrates include a wide variety of toxins (*e.g.* cholera toxin, *Pseudomonas aeruginosa* exotoxin A) and enzymes (*e.g.* proteases, lipases, cellulases, nucleases, *etc.*), so T2SS are important for both virulence and adaptation to diverse environments.

The T2SS is a large protein complex, consisting of up to 15 different proteins. These are called Gsp (**g**eneral **s**ecretion **p**athway) proteins in *E. coli*, which is the nomenclature I will use here, but in other species they are given other names, including "Xcp" in *Pseudomonas* and "Eps" in *Vibrio*. The outer membrane pore, or *secretin*, is composed of a *dodecamer* of the GspD protein. The lipoprotein *pilotin* GspS recruits GspD to the outer membrane and assists in its proper assembly there. GspF, GspL, and GspM form a complex in the inner membrane which is tethered to the secretin by GspC. A cytoplasmic ATPase, GspE, associates with this complex and provides the energy necessary to drive secretion.

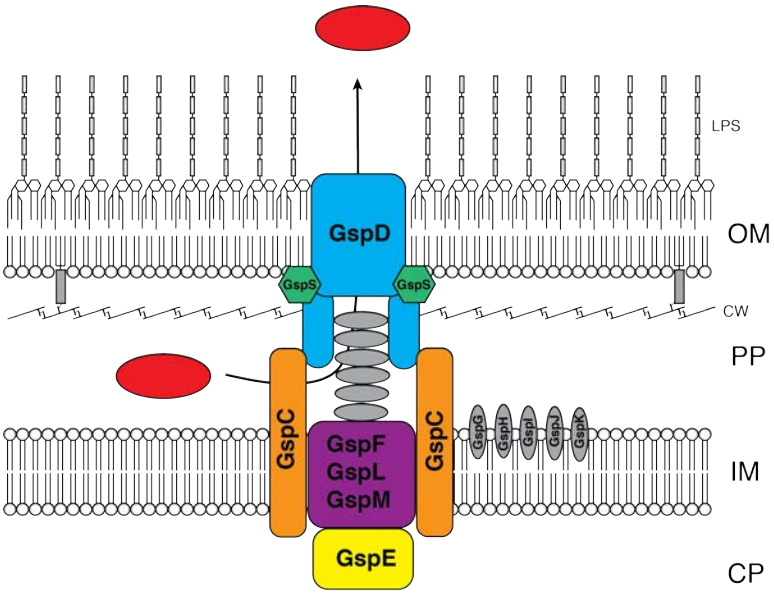


Figure 13.4. Type II secretion. Folded substrates in the periplasm are pushed out of the cell through the GspD secretin channel by polymerization of GspGHIJK monomers into a pseudopilus. The energy for pseudopilus assembly and disassembly is provided by ATP hydrolysis by GspE.

Secretion by the T2SS is driven by the assembly and disassembly of a short polymeric *pseudopilus* within the channel formed by the GspCD proteins. The pseudopilus is composed of the major *pseudopilin* GspG and the minor pseudopilins GspHIJK, and is thought to act as a piston that physically pushes substrate proteins out of the cell. T2SS are very closely related to *type 4 pili* (involved in *adherence* and *twitching motility*; **Lectures 14** and **15**, respectively), but unlike a true *pilus*, the pseudopilus of T2SS never extends past the cell surface.

TYPE III SECRETION

Type III secretion systems (T3SS) have been extensively studied for more than 20 years, and are instrumental in the pathogenesis of many disease-causing Gram-negative bacteria. T3SS deliver effector proteins directly into eukaryotic cells, and these effectors can have dramatic effects on the physiology of host cells. For example, in *Salmonella enterica*, two virulence-associate T3SS deliver at least 6 proteins into epithelial cells that remodel the host cell cytoskeleton and membrane, resulting in the uptake of *Salmonella* into non-phagocytic cells. Additional T3SS effector proteins orchestrate the formation of a "***S****almonella*-**c**ontaining **v**esicle" or SCV, in which *Salmonella* replicates. Other pathogens, including *Shigella, Yersinia,* pathogenic *E. coli*, *Vibrio*, and *Pseudomonas* species also deliver multiple effectors and toxins into host cells with T3SS, and there is extensive evidence for the horizontal transfer of T3SS between species during the evolution of different pathogens.

The T3SS is a very large protein complex (often called a "needle complex", because of its syringe-like structure and function), comprised of as many as 30 proteins, that spans both membranes of the Gram-negative cell envelope and extends well beyond the outer membrane. The proteins indicated in Figure 12.5 use the *Salmonella* nomenclature, but homologous proteins in other species have often been given other names. Effector proteins are exported in a single step as unfolded proteins that pass through the channel in the center of the PrgI needle, driven by the ATPase activity of InvC. It is not clear how the T3SS recognizes substrates for export, but most of them are associated with specific chaperone proteins that keep them unfolded in the cytoplasm and these chaperones associate with the SpaA, OrgA, and OrgB proteins, which are thought to be involved in substrate recognition. It is also not clear how the T3SS is able to detect that it has come in contact with a suitable host cell, although this presumably has something to do with conformational changes in the needle upon interaction with host cell surface receptors.

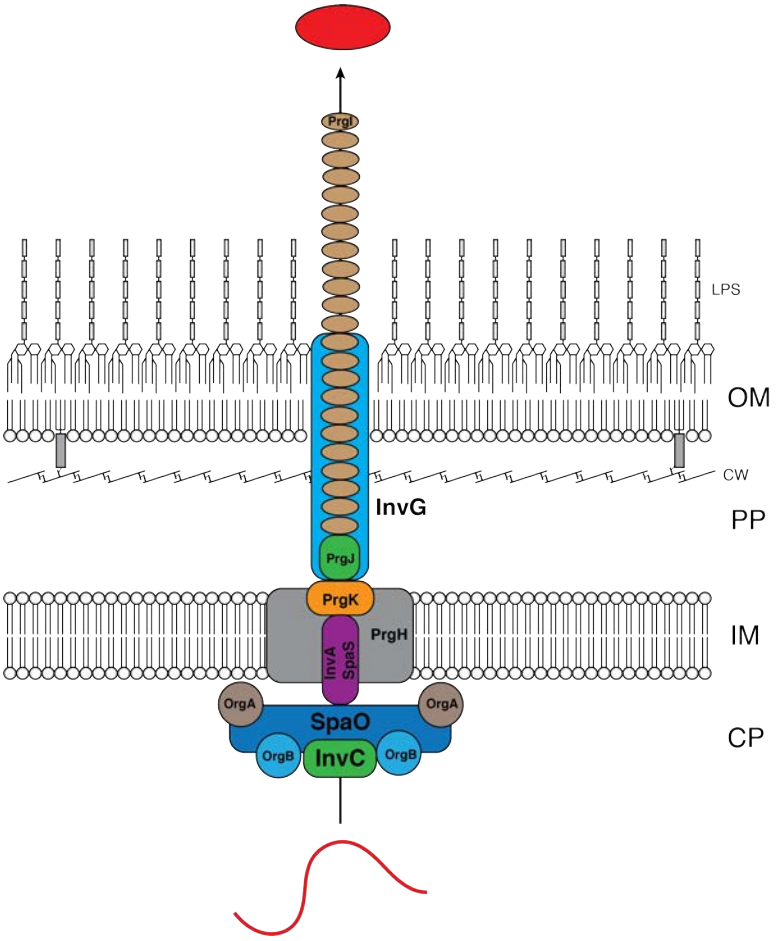


Figure 13.5. Type III secretion. Unfolded substrates are exported in a single step through the needle complex and directly into the cell membrane or cytoplasm of a eukaryotic cell. The energy for export comes from ATP hydrolysis by InvC.

Notably, the T3SS is very closely related to bacterial flagella (see **Lecture 15** for more on flagella and motility). The core proteins and structure of the T3SS share extensive homology with the flagellar basal body and PrgI is closely related to the flagellar hook protein. The T3SS lacks the motor proteins that rotate flagella, but the mechanism of export for PrgI and effector proteins is very similar to that for flagellin subunits.

TYPE IV SECRETION

Type IV secretion systems (T4SS) are widely distributed among Gram-negative and Gram-positive bacteria as well as the archaea. They are involved in toxin and effector protein secretion in a few pathogenic bacteria (e.g. *Bordetella pertussis* and *Legionella pneumophila*). They are unique among the secretion systems discussed in this chapter in that they can mediate the translocation of DNA into adjacent cells (pro- or eukaryotic) as well as proteins. In fact, many naturally-occurring plasmids encode T4SS, and T4SS are the most common mechanism by which bacterial conjugation occurs (recall **Lecture 8**), so they are also referred to as *competence pili*. Because of this, T4SS underlie a lot of antibiotic resistance gene spread.

Remarkable recent work from Ankur Dalia's lab (linked here) has shown how T4SS pili in *Vibrio cholerae* are directly involved in DNA uptake from the media. They used fluorescent microscopy to observe the binding of the tips of pili to extracellular double-stranded DNA, then observed the pili being retracted, pulling the DNA into the cell, where it was expressed.

The nomenclature for T4SS is mostly derived from that of the T4SS encoded by the Ti plasmid of the plant pathogen *Agrobacterium tumefaciens*, which is conjugated into host cells to deliver genes that lead to the formation of crown gall tumors. A large complex of proteins (called VirB1-11 and VirD4) spans both membranes (of Gram-negative cells), with VirB2 and VirB5 forming the long, filamentous pilus and its tip, respectively. VirB4, VirB11, and VirD4 are ATPases, and are presumably involved in the assembly and retraction of the pilus and the secretion of effector proteins and DNA.

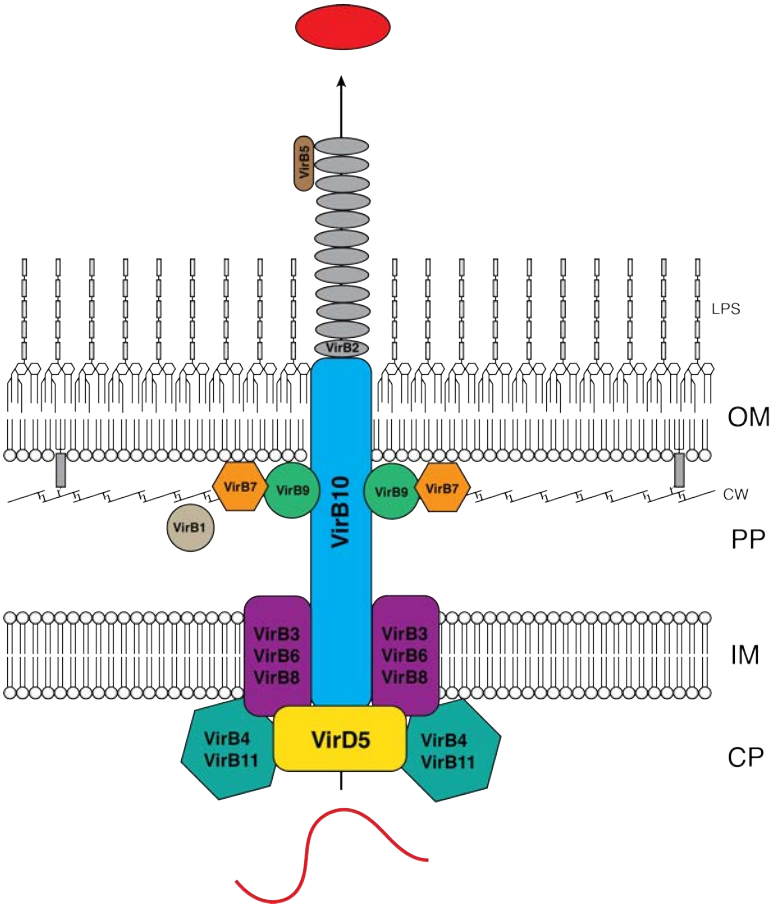


Figure 13.6. Type IV secretion. Protein or DNA substrates are secreted through the pilus, which is a dynamic structure that extends and retracts, driven by the ATPase activities of VirB4, VirB11, and VirD5.

Be very careful not to confuse the T4SS with type 4 **pili**, which are involved in adherence and twitching motility (**Lectures 14** and **15**, respectively). As mentioned above, type IV pili are related to the **T2SS**, not the T4SS. This is particularly tricky since both "type 4" systems involve pili, but they are otherwise quite different. There's really no excuse for this to be so confusing, but the two nomenclatures were established independently of one another and it's too late to disentangle them in the literature at this point.

TYPE V SECRETION

Compared to the complex machinery involved in T2, T3, and T4SS, type V secretion systems (T5SS) are refreshingly simple. Also called *autotransporters*, T5SS proteins secrete themselves across the outer membrane with no other specialized components necessary other than those necessary for all outer membrane protein assembly. They have a β-barrel domain which inserts into the outer membrane, forming a channel through which the remainder of the protein (the *passenger domain*) is extruded into the extracellular space. Many T5SS proteins contain a protease domain which cleaves and releases the passenger domain, but some remain displayed on the cell surface.

Most of the well-studied T5SS substrates are virulence factors, and include a variety of toxins and other proteins. These include the VacA vaculolating cytotoxin from *Helicobacter pylori*, the IcsA actin-assembling protein from *Shigella flexneri*, the *Neisseria gonorrhoeae* immunoglobulin A protease, and the large autoaggregation protein from *E. coli* called antigen 43 (encoded by the *flu* gene, see **Lecture 14**). Since T5SS are dedicated to moving proteins across the outer membrane, they are found only in Gram-negative bacteria.

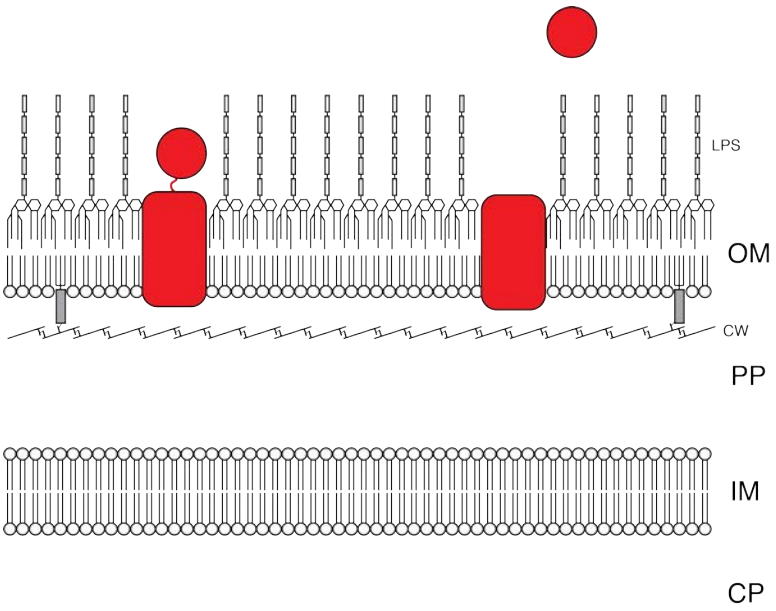


Figure 13.7. Type V secretion. The β-barrel domain of T5SS substrates assembles in the outer membrane (OM), and extrudes the passenger domain out of the cell. In some cases, the passenger domain is released by intrinsic protease activity of the T5SS protein.

T5SS proteins are always secreted across the inner membrane by the Sec pathway, and rely on the periplasmic chaperones and Bam complex which assemble integral outer membrane proteins. There are variants of T5SS which are slightly more complicated and involve two or sometimes three proteins (the *two-partner* and *chaperone-usher secretion systems*), but the core mechanism remains extremely similar.

TYPE VI SECRETION

Type VI secretion systems (T6SS) are another very large multiprotein complex, but they make up for it by being, for my money, the coolest of the protein secretion mechanisms. T6SS are evolutionarily related to the tails of *contractile phage*, and essentially work by launching a protein-tipped harpoon that physically penetrates adjacent bacterial or eukaryotic cells.

Discovered in *Pseudomonas aeruginosa* in 2006 by Joseph Mougous, working in the lab of John Mekalanos, T6SS are widely conserved among Proteobacteria and deliver *effector proteins* that are typically toxins. Some bacteria use T6SS to deliver proteins into eukaryotic cells, including immune-modulating proteins from *Salmonella* and *Legionella*, the VgrG2b microtubule disruptor from *P. aeruginosa*, phagosome escape factors by *Francisella tularensis*, or anti-amoeba effectors in *Xanthomonas citri*, but most T6SS seem to be involved in competition between bacteria and carry anti-bacterial toxins. These kinds of T6SS loci typically encode *immunity proteins* that prevent the T6SS-encoding cells from being killed by their own toxins, and T6SS often mediate a kind of "self-recognition" among very otherwise closely related bacterial strains.

Like contractile phage tails, the T6SS sheath (composed mostly of the TssBC proteins) does not require ATP in order to contract, and appears to be triggered by the physical impact of the surface-exposed portion of the T6SS with another cell. Indeed, another common term for T6SS is "CDI", or ***c****ontact-****d****ependent* ***i****nhibition*. The TssBC tail sheath can be very long, and can extend across the entire width of a bacterial cell (a micron or more). In most species, after the "harpoon" is launched, the baseplate and tail sheath are disassembled by the ClpV ATPase, so that the Tss protein monomers can be reassembled into new T6SS complexes, making T6SS very dynamic structures within the bacterial cell.

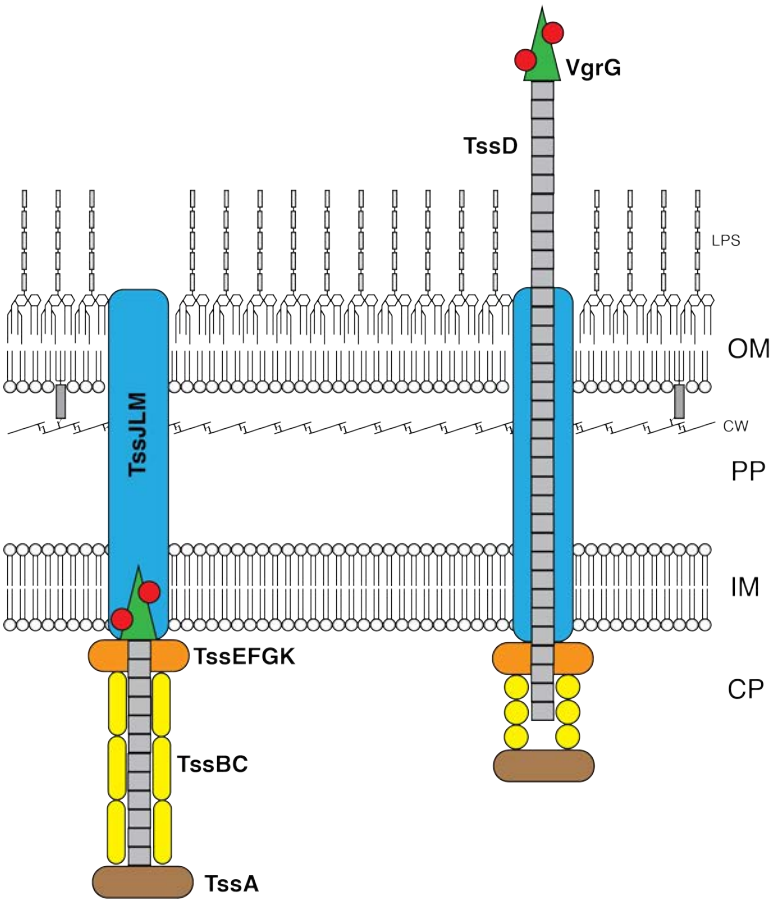


Figure 13.8. Type VI secretion. Effector proteins are bound to VgrG and are physically driven into target cells along with the TssD shaft by the contraction of the TssBC sheath.

TYPE VII SECRETION (ESX)

Unlike the rest of the secretion systems in this chapter, type VII secretion systems (T7SS, also called ESX systems, which purportedly stands for "6-kDa early secretory antigenic target protein family secretion", although don't ask me how **that** acronym works) are only found in monoderm Gram-positive and mycobacteria. First identified in 2003 in *Mycobacterium bovis*, T7SS are best characterized in pathogenic mycobacteria, and are required for virulence of *M. tuberculosis*, for example, where the EsxG and EsxH secreted proteins are involved in damaging host phagosomes. However, it seems clear that T7SS are more versatile and can be involved in nutrient acquisition, conjugation, development (in *Streptomyces* spp.), and other general transport processes. The roles of the T7SS homologs found in more conventional Gram-positive bacteria (*e.g. B. subtilis, S. aureus,* or *Listeria monocytogenes*) have not been well-studied to date.

The core channel of T7SS through the inner membrane is formed from the EccBCDE proteins. EccA is a cytoplasmic ATPase which associates with the pore, and is probably responsible for providing the energy necessary for secretion. There is a cytoplasmic chaperone (EspG) which delivers substrates to the complex, and MycP is a membrane-associated protease involved in substrate processing. T7SS substrates are recognized by a relatively large N-terminal signal sequence consisting of a helix-turn-helix structure followed by a conserved YxxxD/E motif. The EspC or EsxEF proteins forms a filamentous pore that penetrates the thick outer membrane layers of the mycobacterium and provides a channel for secreted substrates to escape the cell, and an additional surface-associated protease called PecA removes the remaining signal sequences.

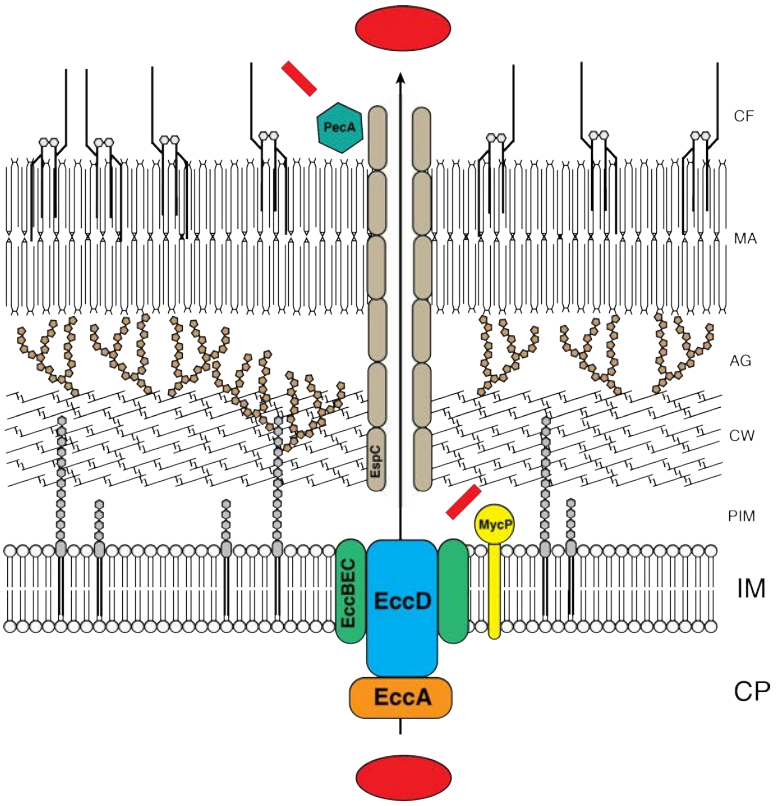


Figure 13.9. Type VII secretion. Substrate proteins are exported through the EccBECD channel in the inner membrane, powered by ATP hydrolysis by EccA. They pass through the outer layers of the cell through a filamentous EspC / EsxEF pore, and signal sequences are removed by proteases MycP and PecA.

T7SS remain relatively poorly understood, and understanding their mechanism, substrates, and function in bacterial physiology and virulence are very active areas of research.

TYPE VIII SECRETION (CURLI)

Type VIII secretion systems (T8SS), best characterized in *Enterobacteriacea* (e.g. *E. coli* and *Salmonella*), are dedicated to the secretion and assembly of *curli*, which are proteinaceous fibers attached to the outer surface of the cell. Curli are unusual in that they are *functional amyloid proteins* that form a characteristic extremely stable cross-β-sheet secondary structure. Like other amyloids (including the amyloid-β and tau proteins involved in human Alzheimer's disease), curli are extremely sticky and prone to aggregation, and are important for biofilm formation and adherence. *Salmonella* use curli to bind to Teflon, stainless steel, and the surfaces of eukaryotic cells, which contributes to the pathogenesis and spread of disease-causing strains. We will discuss biofilms in detail in **Lecture 14**.

The T8SS (Figure 13.10) consists of three proteins (called CsgEFG in *E. coli* and AgfEFG in *Salmonella*) that form a complex in the outer membrane. The curli monomer itself (CsgA or AgfA) is secreted into the periplasm by the Sec pathway, then exported across the outer membrane by the T8SS. Once outside the cell, the curli monomers rapidly aggregate and form curli fibers, which are normally anchored to the outer membrane by the minor curli monomer CsgB/AgfB, which nucleates the aggregation of CsgA.

Curli-producing strains also express a dedicated periplasmic chaperone called CsgC that prevents CsgA aggregation in the periplasm. Interestingly, work by Margery Evans (in Matthew Chapman's lab at the University of Michigan) has shown that *E. coli* CsgC also prevents the amyloid aggregation of variants of human α-synuclein associated with neurodegeneration in Parkinson's disease.

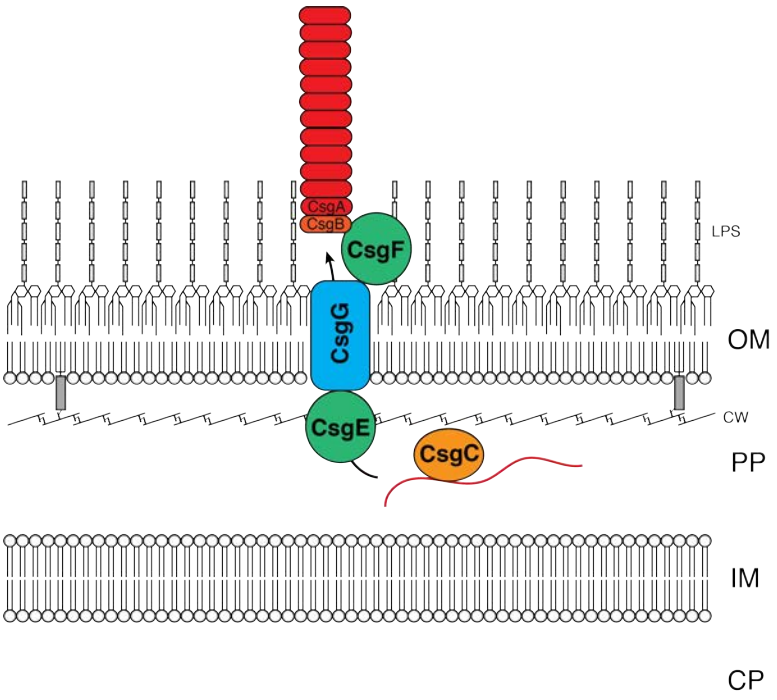


Figure 13.10. Type VIII secretion. Curli monomers (CsgA and CsgB) are secreted into the periplasm by the Sec pathway. While in the periplasm, they are maintained in an unfolded state by the chaperone CsgC. The CsgEG complex is required for secretion of curli, while CsgF is required for attachment of CsgB and CsgA to the outer membrane.

TYPE IX SECRETION

Type IX secretion systems (T9SS) are found only in certain species of the *Bacteroidetes* phylum, and are best characterized in the periodontal pathogen *Porphyromonas gingivalis*, where T9SS is responsible for secretion of more that 30 proteins. Among these are several proteases, collectively called gingipains, which are key virulence factors that degrade host proteins, including cytokines, antibodies, fibrinogen, and collagen, leading to bleeding, tissue destruction, and bone loss in the gums of infected patients. Gingipain accumulation in the brain has been associated with Alzheimer's disease, although the exact mechanism by which they enter the brain is not yet clear. You should probably take this as further incentive to floss regularly, though.

The T9SS is a complex structure, consisting of at least 18 proteins, not all of which have known functions. They include the inner membrane proteins PorL and PorM, the periplasmic proteins PorN, PorK, and PorW, and the outer membrane or surface-associated proteins Sov, PorQ, PorP, PorT, PorV, PorU, and PorZ, in addition to some other proteins that have not yet been given formal names. However, it is clear that the T9SS is a large protein complex that spans the inner and outer membranes, with the outer membrane complex including at least 7 integral membrane β-barrel proteins.

T9SS substrates are secreted to the periplasm by the Sec system, at which point they fold into their final structure and are recognized by the T9SS based on a C-terminal signal sequence. By a currently unknown mechanism, the T9SS translocates its substrates across the outer membrane, at which point the C-terminal signal sequence is removed by the sortase-like protease PorU. Many T9SS substrates are then covalently modified by the attachment of an anionic LPS lipid group, which causes the resulting proteins to be anchored into the outer leaflet of the outer membrane. The current model is that the energy necessary for type IX secretion comes from the proton motive force across the inner membrane, and is transduced across the periplasm by PorLM.

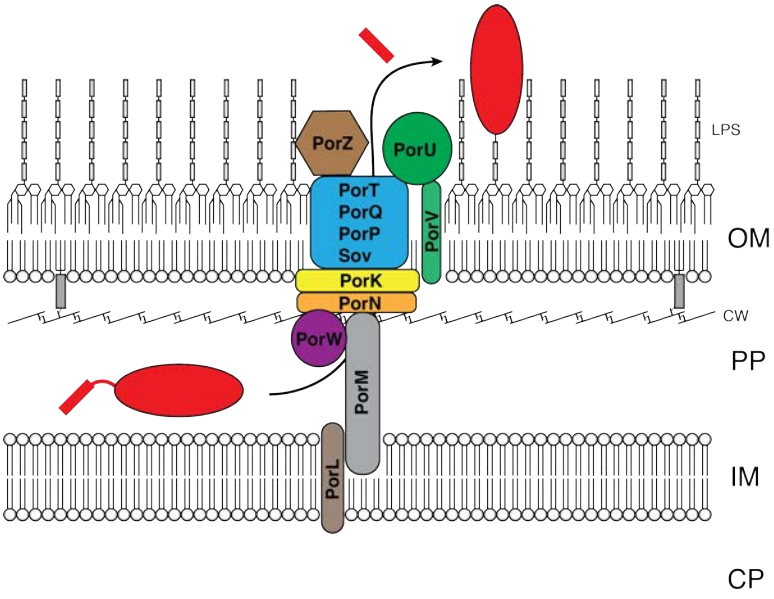


Figure 13.11. Type IX secretion. Substrates are translocated into the periplasm by the Sec system. After removal of the N-terminal Sec signal peptide, a C-terminal T9SS signal peptide directs the folded protein to the T9SS apparatus. After translocation across the outer membrane, the C-terminal signal sequence is cleaved off by the sortase PorU and, for many substrates, the protein is covalently attached to LPS for anchoring to the cell surface.

T9SS are also required for a remarkable mechanism of *gliding motility* in the non-pathogenic environmental species *Flavobacterium johnsoniae*, which we will discuss in more detail in **Lecture 15**.

DISCUSSION PROBLEM SET #25: TOXIN SECRETION BY *PHOTORHABDUS LUMINESCENS*

Photorhabdus luminescens is a Gram-negative insect pathogen which secretes a large (324 kDa) soluble protein toxin that induces apoptosis in cultured insect cells. The gene encoding this protein is not encoded in an operon with any other genes of known function.

The following methods are available for *P. luminescens*:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
plasmids can be introduced by conjugation	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
compatible transposon	√

Describe a series of observations and experiments that will allow you to identify the mechanism by which the *P. luminescens* toxin is secreted. State:

- observations you can make to narrow down the possible mechanisms
- experiments you will use to further narrow the possibilities

- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

GENETIC METHODS TO DETECT PROTEIN-PROTEIN INTERACTIONS *IN VIVO*

As we have seen, protein secretion often involves large protein complexes. It is certainly possible to directly measure the interactions of two or more purified proteins *in vitro*, but this is primarily a bacterial **genetics** class, and in this section I will introduce you to genetic methods that can be used to study the interactions of proteins *in vivo*. These mostly involve the construction of fusion proteins with different kinds of reporters, the most common of which depend on a general principle called *protein fragment complementation*.

Protein fragment complementation depends on the observation that some proteins can be divided into two or more amino acid chains that individually have no activity, but when brought into close proximity to one another are able to interact and form an active protein. The individual reporter fragments are expressed as fusions with the proteins of interest (typically called the "bait" and "prey" proteins), and any resulting activity is interpreted as evidence that the proteins of interest physically interact (or are at least close to one another) in the cell. This kind of assay was first developed in yeast by Stanley Fields and Ok-kyu Song (in this paper), who referred to their system as *two-hybrid screening*, a term which is still commonly used for protein fragment complementation procedures. In fact, in many cases, the interactions between bacterial proteins have been screened by expressing those proteins in the yeast two-hybrid system.

However, here I will discuss three protein fragment complementation systems used in bacteria: the "bacterial two-hybrid screen" (described in this paper), which uses adenylate cyclase as a reporter gene, the split murine dihydrofolate reductase (mDHFR) system (described here), and the use of tripartite split GFP (described here). I will also briefly describe a very recently-developed system with fluorescent reporters that can be used to study the interaction of proteins in the cell envelope, which is difficult with the more conventional systems.

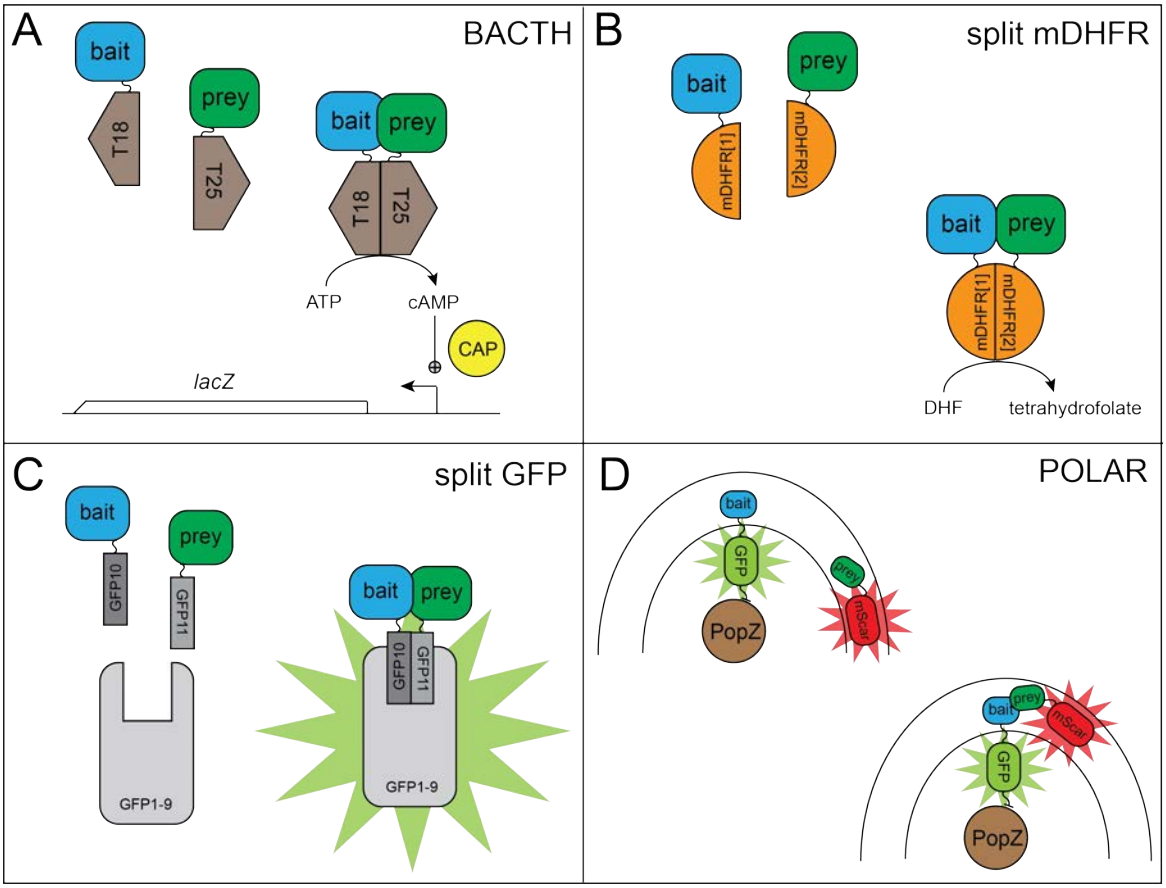


Figure 13.12. Genetic tools for detecting protein-protein interactions in bacteria. A) Bacterial two-hybrid screen, using split adenylate cyclase as a reporter, with (usually) a colorimetric *lacZ* expression readout. B) Split murine dihydrofolate reductase reporter, with trimethoprim-resistant tetrahydrofolate production as an output. C) Split GFP reporter, with fluorescence signal as an output. D) PopZ-linked apical recruitment assay, with colocalization of fluorescent GFP and mScarlet at the cell pole as an output signal.

The original and most commonly-used **bac**terial **t**wo-**h**ybrid screen (abbreviated "BACTH"; Figure 13.12 A) depends on the division of adenylate cyclase into two inactive fragments, called "T18" and "T25". Adenylate cyclase, encoded by the *cya* gene, is an enzyme which generates the second messenger cyclic AMP (cAMP). *E. coli cya* null mutants are unable to ferment lactose or maltose, because expression of the promoters for the *lac* and *mal* operons requires the transcriptional activator CAP (**c**atabolite **a**ctivator **p**rotein, also called "CRP", for **c**yclic AMP **r**eceptor **p**rotein, since it acts as a repressor at some promoters), and CAP is only active when bound to cAMP. The formation of active adenylate cyclase from T18 and T25 can be screened for on indicator plates (*e.g.* LB containing the colorimetric lactose analog X-gal) or selected for on media with maltose as a sole carbon source. BACTH is usually used as a colorimetric screen, since weak interactions may not result in production of enough cAMP to allow growth, but do result in the formation of visibly colored colonies on indicator plates.

The most common application of BACTH is to look at a single pairwise protein-protein interaction, for which the bait and prey proteins are cloned individually as T18 and T25 fusions (recall the methods from **Lecture 7** on cloning). Some very large screens of *interactomes* (the set of protein-protein interactions in a cell) have been performed this way. but this is very labor-intensive, requiring individually cloning thousands of genes. If you are interested in identifying new interaction partners of a particular protein in a more elegant way, you can clone the bait protein as one fusion fragment and clone a library of random fragments of genomic DNA into the plasmid for the other fragment. Most of the resulting clones will, of course, encode protein fragments that are out of frame with the fusion, but if you generate a large enough library you can screen for proteins that interact with the bait protein.

Split mDHFR (Figure 13.12 B) is a system that allows more sensitive protein fragment complementation **selections** than BACTH. DHFR is an enzyme that is required for the synthesis of the essential cofactor folate, without which *E. coli* cannot grow. Bacterial DHFR is inhibited by the antibiotic trimethoprim, but mammalian DHFR (including mDHFR from mice) is not. Therefore, when the inactive fragments of mDHFR (called "mDHFR[1]" and "mDHFR[2]") are brought

together in *E. coli* to form active mDHFR, the resulting protein confers trimethoprim resistance, which is easily selected for. Stronger interactions generally result in resistance to higher concentrations of trimethoprim.

BACTH screening uses a colorimetric readout, which is useful at a population or per-colony basis, but for many screening applications, especially those that use microscopy or flow cytometry, fluorescence is more sensitively or more easily detected. Tripartite split GFP (Figure 13.12 C) can be used in these cases. In this assay, individual strands of the GFP β-barrel (called "GFP10" and "GFP11") are fused to the bait and prey proteins with short, flexible linkers and expressed in the same cell as the remainder of GFP ("GFP1-9"). None of the individual fragments of GFP are fluorescent, but when GFP10 and GFP11 are brought close together, they associate into a structure that can assemble with GFP1-9 into complete, active GFP. This method also has the advantage that the GFP10 and GFP11 fragments are quite small (20 and 21 amino acids, respectively), and therefore less likely to interfere with the structure or interactions of the bait and prey proteins.

One limitation of the above methods, especially using adenylate cyclase or mDHFR, is that they are dependent on the reporter proteins being in the cytoplasm. Many protein complexes, as we have seen in this chapter, assemble in the membranes and periplasm. A recent publication from Thomas Bernhardt's lab (linked here) describes "POLAR" (**Po**pZ-**l**inked **a**pical **r**ecruitment; Figure 13.12 D) a clever approach for studying protein interactions in the cell envelope which I suspect will be extremely useful.

POLAR takes advantage of the *Caulobacter crescentus* PopZ protein, which, when expressed in *E. coli*, localizes to the poles of the bacterial cell. Bait proteins are expressed in PopZ-expressing *E. coli* as fusions with both a PopZ-binding tag and GFP. This results in the recruitment of the bait protein, which can be either a cytoplasmic or an inner membrane protein, to the poles of the cell, where it can be visualized by fluorescence microscopy. The prey protein is then expressed in the same cells fused to the red fluorescent protein mScarlet. The prey can be present in the cytoplasm, the inner membrane, the periplasm (as shown in Figure 13.12 D), or even the outer membrane. If the bait and prey proteins interact, the GFP and mScarlet fluorescence will colocalize at the cell poles. Otherwise, the mScarlet signal will be distributed throughout the cell.

Note that all of the above methods test for protein-protein interactions **in *E. coli***, even when examining proteins from other organisms. This is not typically a problem, and is sometimes a benefit, both because *E. coli* plasmids are straightforward to engineer and because there will be less interference from "native" protein interactions. However, if a particular protein of interest is poorly expressed or toxic in *E. coli*, this may present a challenge.

Finally, in some cases, protein interactions can also be studied more directly using genetics. Protein-protein interactions are determined by the surface properties of proteins, with interacting surfaces having complementary shapes and chemical properties. For example, there may be a positively charged amino acid in one protein that interacts with a negatively charged amino acid on the surface of its interaction partner, or two proteins may have complementary patches of surface-exposed hydrophobic amino acids. Missense mutations of key amino acids forming the interaction surfaces can disrupt these interactions.

If missense mutations in one protein are found that prevent complex formation, it can often be interesting to look for suppressors of those mutations in the **other** protein (recall the discussion of intergenic suppressors in **Lecture 3**). For example, if mutating a positively-charged Arg residue to Glu in one protein prevents interaction, perhaps there is a negatively-charged amino acid in the other protein that can be mutated into a positively-charged residue to restore the interaction. This would be strong evidence that the charge-charge interaction at that site is important for the interaction of the two proteins involved. These kinds of mutants could be made by site-directed mutagenesis, if there was structural or other evidence to implicate specific amino acids, or by random mutagenesis, if there is a screenable or selectable phenotype for the interaction between the proteins in question.

DISCUSSION PROBLEM SET #26: IDENTIFYING CHAPERONE BINDING PARTNERS

The Cag T4SS of *Helicobacter pylori* is responsible for secretion of several different effector proteins that alter host cell gene expression and cytoskeletal structure, and are necessary for pathogenesis. For one of these effector proteins, CagA, successful secretion requires the presence of the cytoplasmic chaperone protein CagF, which binds to CagA and recruits it to the T4SS.

It is currently unknown whether CagF is involved in secretion of any other T4SS substrates in *H. pylori*.

The following methods are available for *H. pylori*:

growth in pure culture	✓

can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposon	✓
oligo-directed recombineering	✓

Describe an experiment to identify additional binding partners of CagF in *H. pylori*. State:

- the independent and dependent variables
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiment, and how you will interpret them

---

LECTURE 14: CAPSULE AND BIOFILMS

INTRODUCTION

Most, if not all, bacteria are able to synthesize and secrete complex polysaccharide carbohydrate polymers. This includes the GlcNAc-MurNAc polymer that forms the backbone of peptidoglycan (**Lecture 10**), but also a very wide range of other ***e****xtracellular* ***p****oly****s****accharide*s (EPS) with diverse functional roles. *Capsule*s are EPS that (usually) remain attached to the surface of the bacterial cell, while *slime*s and *gum*s are EPS that are released into the extracellular environment. EPS are also a key component of bacterial *biofilm*s, which are multicellular communities attached to surfaces.

In this chapter, we will discuss the synthesis of EPS, the importance of capsule in pathogenesis, how bacteria adhere to surfaces, and the structure, formation, and regulation of biofilms. I will note that the topics of this chapter are the subjects of research by many labs here at UAB, including those of Drs. Moon Nahm, Janet Yother, Carlos Orihuela, Jessica Scoffield, Ed Swords, and Megan Kiedrowski, all of whom are, of course, far more expert than I in this area.

Thanks in particular to Drs. Janet Yother and Megan Kiedrowski for their suggestions and comments on this chapter, which helped improve it immensely.

POLYSACCHARIDE BASICS

Strictly speaking, *carbohydrate*s are organic compounds composed only of carbon, hydrogen, and oxygen, but for biological purposes, the term is mostly synonymous with sugars, which are 3- to 7-carbon carbohydrates distinguished by the arrangement and chirality of their various hydroxyl groups, and their polymers. Simple sugars, or *monosaccharide*s, are the monomers that make up *polysaccharide*s, including EPS.

The most common sugars in biological polysaccharides are *hexoses* (6 carbons) and *pentoses* (5 carbons), which are typically found as rings in biological systems. Some examples are shown in Figure 14.1. The *anomeric carbon* of a monosaccharide is numbered "1". The other carbons in the ring are numbered consecutively from the anomeric carbon. In an α sugar, the substituent groups on the anomeric carbon and the carbon on the other side of the oxygen atom in the sugar ring are facing **opposite** directions, while in a β sugar, they are facing the **same** direction. The substituent group of the anomeric carbon is the one whose orientation can change.

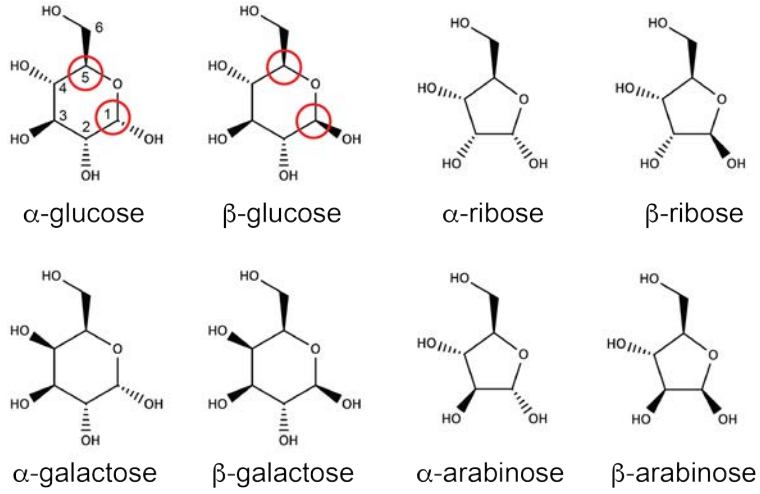


Figure 14.1. A few common hexoses and pentoses found in bacterial polysaccharides, with each shown in its α and β form. Carbon numbering starting at the anomeric carbon is illustrated for α-glucose, and the carbons whose substituents determine the difference between α- and β-glucose are circled in red.

In polymers, monosaccharides are linked by *glycosidic bonds*. When the two linked carbons have the same relative stereochemistry, that is called an α-glycosidic bond. When they have opposite stereochemistries, it is a β-glycosidic bond. The nomenclature for describing specific glycosidic bonds indicates the stereochemistry of the bond as well as the carbons in the two sugars being linked. The disaccharide lactose, for example, contains a β-1,4 linkage between the hexoses galactose and glucose (Figure 14.2).

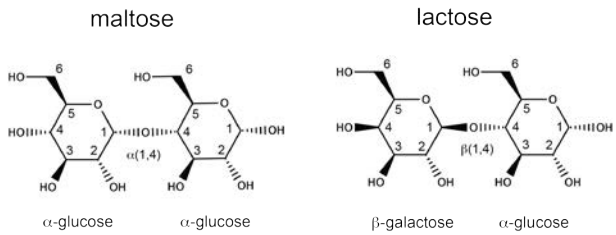


Figure 14.2. Examples of disaccharides containing different kinds of glycosidic bonds.

In many biologically important polysaccharides, the monosaccharide subunits are chemically modified. Common modifications are the replacement of a hydroxyl group with an amino group to form an *amino sugar* and *acetylation*, the addition of a CH_3CO- group to either a hydroxyl group (*O-acetylation*) or an amino group (*N-acetylation*)(Figure 14.3). These and other modifications can have dramatic effects on the chemical properties of a sugar or polysaccharide.

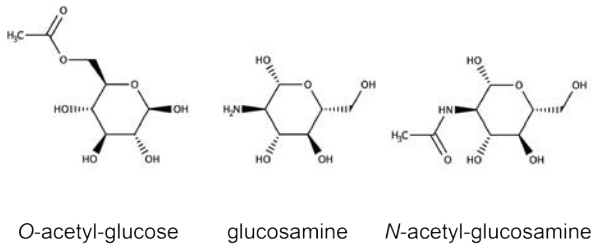


Figure 14.3. Common sugar modifications found in bacterial polysaccharides, illustrated on a glucose monosaccharide.

There is a nearly unlimited amount of diversity possible in polysaccharide biochemistry. Any number of different monosaccharides, with different modifications, can be linked by different kinds of glycosidic bonds, in linear or branched forms, and with widely varying lengths. This imparts tremendous functional and chemical diversity, but also makes the study and categorization of polysaccharides a complicated topic.

Luckily, most polysaccharides are constructed from repeating subunits, so we can categorize them at least partially by reducing them to their minimal repeat unit. The simplest polysaccharides have repeat units of a single monosaccharide: cellulose is β-1,4-linked glucose, and amylose is α-1,4 linked glucose (note that even this seemingly minor difference results in polymers with dramatically different chemical properties). The repeat unit of the polysaccharide component of peptidoglycan is a disaccharide: β-1,4-linked *N*-acetyl-glucosamine and *N*-acetyl-muramic acid. However, many EPS have more complicated, often branching repeat units (Figure 14.4).

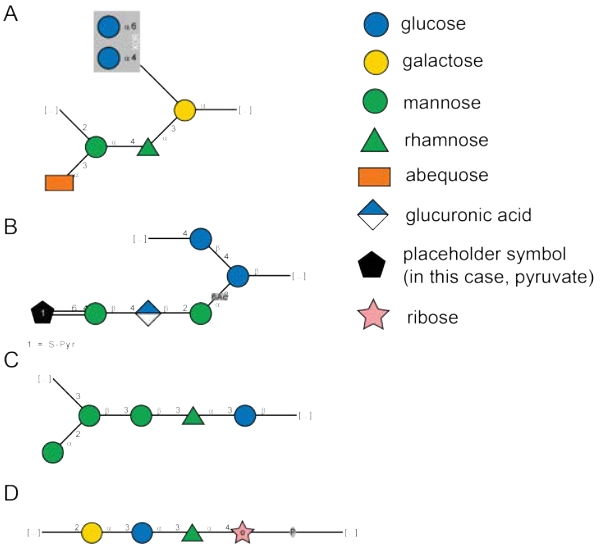


Figure 14.4. Examples of the polymeric repeat units of some bacterial polysaccharides. (A) A *Salmonella enterica* O-antigen, with an example of a sugar that can have either of two possible glycosidic linkages. (B) Xanthan gum from *Xanthomonas campestris*, with an example of an acetylated sugar. (C) The *Pseudomonas aeruginosa* Psl secreted polysaccharide. (D) A *Streptococcus pneumoniae* type 6B capsular antigen. Schematics obtained from the Carbohydrate Structure Database.

The Carbohydrate Structure Database is an invaluable resource that stores all of the known biological carbohydrate structures, although it suffers from a somewhat outdated design and user interface. Nevertheless, it does provide a comprehensive system for visualizing and reporting polysaccharide structures. Some examples drawn from that database are shown in Figure 14.4.

The function of biological polysaccharides is determined not only by their structure, of course, but also by their localization. Obviously, peptidoglycan is an integral component of the cell wall, and LPS and O-antigen are major constituents of the outer leaflet of the Gram-negative outer membrane (**Lecture 10**). EPS are more peripherally associated with the cell. EPS that are physically attached to the surface of the bacteria cell are generally referred to as *capsule*. In Gram-positive bacteria, capsules are often covalently bound to peptidoglycan, similar to wall teichoic acids. Surprisingly, in Gram-negative bacteria, the mechanism(s) by which capsules are attached to the cell surface are not well understood. EPS that are released from the cell into the supernatant are sometimes called either *slime* or *gum*, depending on their physical properties. (The food-safe thickening agent xanthan gum is a secreted EPS from *Xanthomonas campestris*, for example.)

Note that, because EPS form the outermost layers of bacterial cells, the mammalian immune system is very good at recognizing and differentiating between subtly different polysaccharides. *Serotyping* is a method of distinguishing bacteria by what antibodies recognize them, and the antigens involved are very often EPS. As you will see, many bacterial EPS have names that reflect this (*e.g.* the O- and K-antigens of *E. coli*, Vi-antigen of *Salmonella*, *etc.*).

There are four main pathways by which polysaccharides are synthesized in bacteria. We will discuss two of them in some detail, and mention the others in passing before moving on to exploring the functions of EPS in host-microbe interactions and the formation of surface-attached bacterial communities.

WZY-DEPENDENT EPS SYNTHESIS PATHWAY

The most prevalent type of EPS synthesis pathway depends on a class of enzymes called Wzy polymerases. This includes the pathways for the synthesis of most enterobacterial O-antigens, the Psl polysaccharide from *Pseudomonas aeruginosa*, and most *Streptococcus pneumoniae* capsule types, among many, many others. It's also very closely related to the pathway by which peptidoglycan is synthesized, as we will see shortly.

In this pathway, polysaccharide repeat units are assembled in the cytoplasm on a phospholipid carrier molecule called undecaprenyl phosphate (Und-P)(Figure 14.5), which is inserted into the cytoplasmic face of the inner membrane.

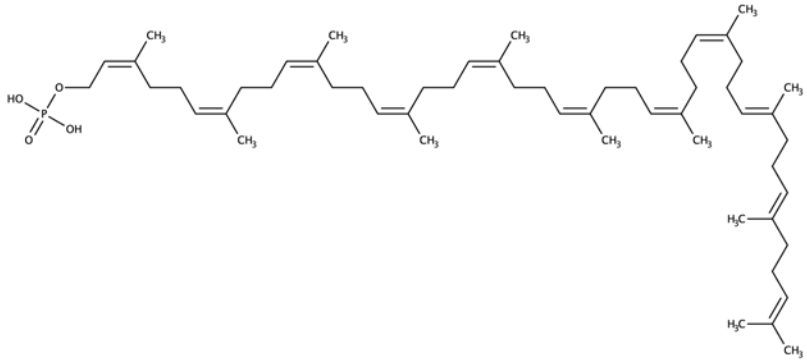


Figure 14.5. The structure of undecaprenyl phosphate (Und-P), the lipid carrier on which Wzy-dependent polysaccharides are assembled.

The Wzy polymerase dependent EPS synthesis pathway is illustrated schematically in Figure 14.6. The "Wz*" nomenclature used here indicates protein homolog families. Each polysaccharide synthesis pathway has its own dedicated equivalents.

The monosaccharide subunits for polysaccharide assembly are in the form of nucleotide diphosphate (NDP) sugars, and assembly of the repeat unit begins with a membrane-bound polyprenol phosphate phosphoglycosyl transferase that takes Und-P and a specific NDP-sugar to generate an Und-PP-monosaccharide molecule, releasing nucleotide monophosphate (NMP). Subsequent sugars are added to the repeat unit by sequentially-acting cytoplasmic glycosyltransferase enzymes, until the complete repeat unit is assembled.

At that point, the completed repeat units, still anchored into the membrane by the lipid carrier, are moved through the membrane to the **outer** leaflet by a Wzx *flippase*. At that point, the Und-PP-repeat unit is the substrate for the Wzy polymerase, which assembles chains of repeat units. The polysaccharide chain is constructed while still anchored to an

Und-PP lipid carrier, and as each new repeat unit is added, the attached Und-PP is released. In many pathways, the length of the polysaccharide product, by which I mean the number of repeat units, is controlled by the Wzz protein.

In Gram-negative bacteria, completed EPS must be exported through the outer membrane, which is accomplished by a protein complex whose central member is called Wza. In Gram-positive bacteria, assembled EPS may be released from the cell completely or may be covalently attached to the peptidoglycan cell wall by the Wzd/Wze proteins.

There are additional proteins that regulate EPS synthesis and export, as well as dedicated enzymes for modification of NDP-sugar precursors, but these differ from pathway to pathway.

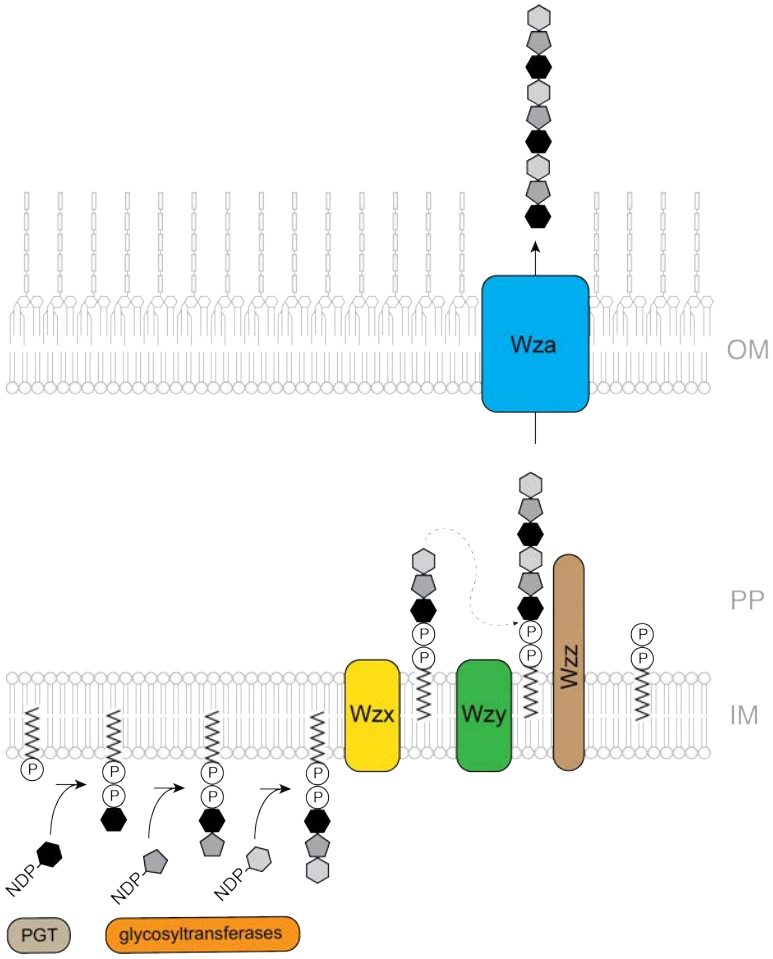


Figure 14.6. Schematic of Wzy-dependent EPS biosynthesis in a Gram-negative bacterium. EPS biosynthesis is initiated on an undecaprenyl phosphate (Und-P) carrier by a PGT (polyprenyl phosphate phosphoglycosyltransferase) and repeat units are synthesized by sequential addition of monosaccharides by glycosyltransferases. Completed Und-PP-linked repeat units are flipped across the membrane by a Wzx flippase and added to growing polymer chains by a Wzy polymerase. Polymer length is often regulated by Wzz. In Gram-negative bacteria, released EPS polymers are exported through the outer membrane via the transporter Wza.

The O-antigen of LPS (**Lecture 10**) is synthesized by a Wzy-dependent mechanism in most Gram-negative bacteria, and is covalently attached to the LPS core in the periplasm by an enzyme called WaaL before the complete LPS is flipped to the outer leaflet of the outer membrane. It is worth noting here that K-12 strains of *E. coli*, including most laboratory strains, do not synthesize O-antigen, due to a mutation (*rfb-50*) in the O-antigen synthetic pathway. Among other things, this makes K-12 considerably more sensitive to hydrophobic antibiotics than strains that **do** synthesize O-antigen.

POLYSACCHARIDE GENE NOMENCLATURE

I want to take a moment here to try to explain the very confusing state of gene nomenclature in the bacterial polysaccharide world. The fundamental problem is that, for each different polysaccharide, there is usually a different, specific set of Wzx, Wzy, and Wzz proteins, each individual bond between two monosaccharides has to be formed by its own specific glycosyltransferase, and individual monosaccharides may need to be modified by specific

acetyltransferases, aminotransferases, or other enzymes. This results in a vast number of genes and proteins, all of which need names.

For example, assembly of the *E. coli* enterobacterial common antigen (ECA), which has a relatively simple linear three-sugar repeat unit, requires three glycosyltransferases (WecA, WecG, and WecF), six sugar-modifying enzymes (WecB, WecC, WecD, WecE, RffH, and RffG), and dedicated WzxE flippase, WzyE polymerase, and WzzE length determination proteins. For ECA, at least the names of the flippase, polymerase, and length determination protein contain "Wzx", "Wzy", and "Wzz", but this is by no means universal. Many polysaccharide synthesis gene names follow the "*w****" format (proposed in this paper in 1996), but others may use *cps, cap, eps, rfb*, or a variety of other gene symbols, and there's very little in the way of standardization from one species to another.

On the other hand, for the most part, all of the genes necessary for synthesis of a particular polysaccharide are **usually** encoded in the same operon or genetic locus in bacterial genomes, and it is possible to examine the sequence of those genes and get a pretty good idea of how many glycosyltransferases, *etc.* are involved in any particular pathway, which is helpful.

PEPTIDOGLYCAN SYNTHESIS PATHWAY

The main difference between the peptidoglycan synthesis pathway and other Wzy-dependent polysaccharide synthesis pathways is that the pentapeptide cross-linking stem (**Lecture 10**) is assembled on UDP-MurNAc (by the MurA-F enzymes) before the phosphoglycosyltransferase MraY attaches that modified sugar to Und-P, forming an intermediate called *Lipid I*.

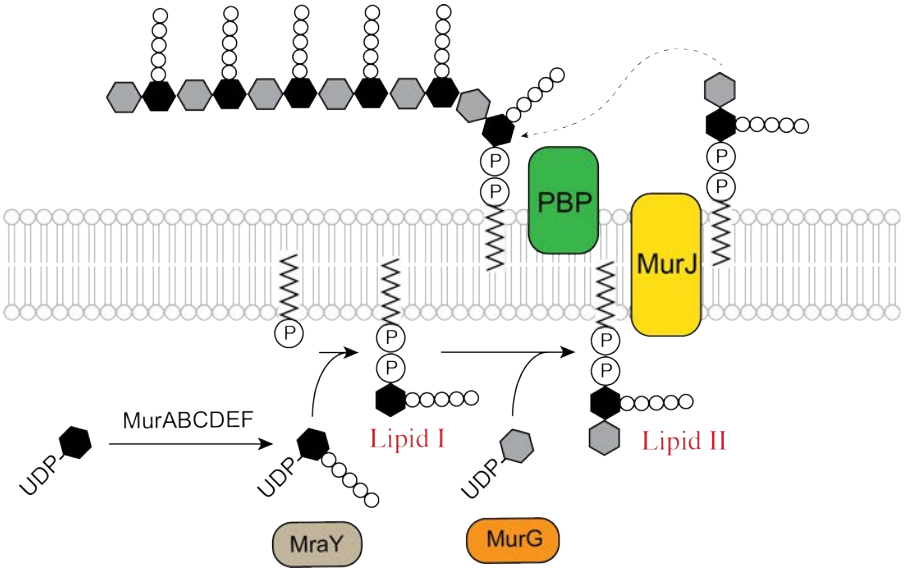


Figure 14.7. Schematic of peptidoglycan polysaccharide backbone synthesis. UDP-MurNAc (black hexagons) is synthesized and the pentapeptide stem (white circles) added by the MurA-F enzymes. The phosphoglycosyltransferase MraY combines this with Und-P to generate Lipid I. The glycosyltransferase MurG adds GlcNAc (grey hexagons), generating Lipid II, which is flipped across the membrane by the flippase MurJ. Lipid II is added to growing peptidoglycan chains by the glycosyltransferase activity of PBPs (penicillin-binding proteins).

To form the repeat unit of peptidoglycan, the glycosyltransferase MurG adds UDP-GlcNAc to Lipid I, forming the intermediate *Lipid II*, which is flipped across the membrane by MurJ. The assembly of peptidoglycan chains is then catalyzed by the glycosyltransferase activity of penicillin-binding proteins (PBPs), releasing Und-PP.

Peptidoglycan synthesis is the major drain on the cell's fairly limited pool of Und-P, and the recycling of Und-PP after the repeat unit has been transferred to the growing polysaccharide chain is essential for continued bacterial growth and survival, as well as to supply Und-P for other polysaccharide synthesis pathways. *E. coli* has three partially redundant Und-PP phosphatases, BacA, YbjG and PgpB, but it is currently unknown how the Und-P generated by these enzymes returns to the inner leaflet of the inner membrane.

SYNTHASE-DEPENDENT EPS SYNTHESIS PATHWAY

In contrast to the Wzy polymerase pathways that assemble complex polysaccharides attached to a lipid carrier molecule, some polysaccharides are directly assembled as linear glycans by synthase-translocase catalytic enzyme complexes. These are called "synthase-dependent" systems, and the basic components of such a system are illustrated in Figure 14.8.

In these pathways of EPS synthesis, a synthase-copolymerase complex in the cell membrane constructs linear polysaccharides from NDP-monosaccharide precursors at the same time that the polysaccharide chain is translocated through the membrane. Once outside of the cell, there may be a variety of accessory proteins that modify some or all of the monomers in the polysaccharide chain, including acetylases, epimerases, methylases. Finally, in Gram-negative bacteria, there are dedicated translocon proteins that allow the completed polysaccharide to cross the outer membrane. There do not seem to be any consistent naming conventions for the component proteins of synthase-dependent polysaccharide pathways.

The synthase is the enzymatically-active component of the inner membrane complex, while the copolymerases seem to play important roles in activation and regulation of synthase activity.

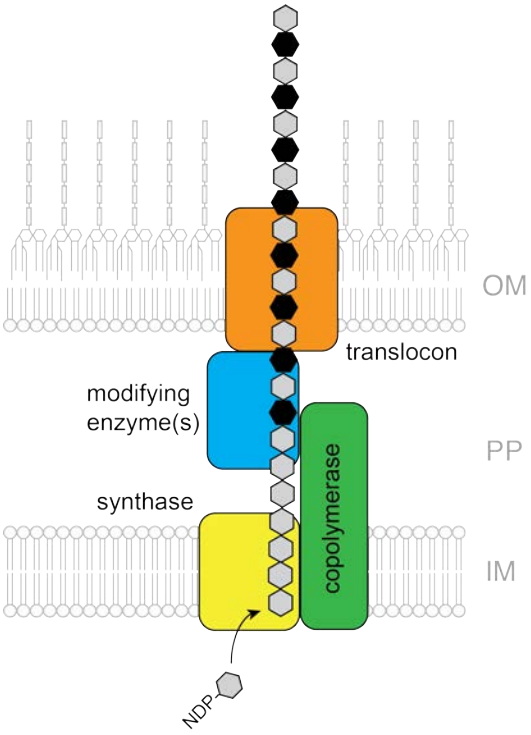


Figure 14.8. Schematic of a generic synthase-dependent EPS biosynthesis complex in a Gram-negative bacterium. The synthase is the enzymatically-active member of the polymerization complex, and the copolymerase is required for activity and regulation of the synthase.

As you might imagine, the polysaccharides assembled by synthase-dependent pathways are, on average, not as structurally complex as those assembled by Wzy-dependent polymerases. In fact, probably the most abundant EPS assembled by a synthase-dependent pathway is cellulose, which is simply a linear chain of β-1,4-linked glucose. This and some other examples of EPS synthesized by this pathway are illustrated in Figure 14.9.

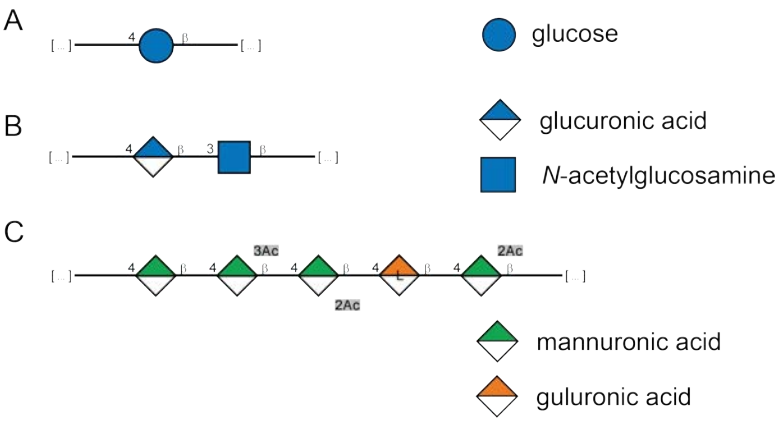


Figure 14.9. Examples of polysaccharides synthesized by synthase-dependent EPS biosynthesis. (A) Cellulose, synthesized by a very wide range of bacteria. (B) Hyaluronan, synthesized by *Streptococcus dysgalactiae* subsp. *equisimilis*. (C) Alginate, synthesized by *P. aeruginosa*. Schematics obtained from the Carbohydrate Structure Database.

Many EPS assembled by synthases are important for biofilm formation (as we will see below), while others play key roles in bacterial pathogenesis. These include hyaluronan synthesized by group A streptococci, which is structurally identical to a major polysaccharide synthesized by their mammalian hosts and allows these bacteria to evade recognition by the immune system, as well as alginate, a thick, sticky polysaccharide produced by *P. aeruginosa* which is a major virulence determinant in patients with cystic fibrosis.

ALTERNATIVE EPS SYNTHESIS PATHWAYS

There are two other major pathways by which EPS are assembled, which we will not discuss in detail here due to space limitations. This should not be taken to mean that the EPS made by these pathways are not important!

Some Gram-negative bacteria synthesize capsular glycolipids by a pathway that is characterized by the cytoplasmic synthesis and polymerization of a polysaccharide on a phosphatidylglycerol lipid carrier and export via a dedicated ATP-dependent ABC transporter (**Lecture 16**). Unlike the previous pathways, the polysaccharides are completely assembled in the cytoplasm before export, and tend to be composed of relatively simple, unbranched repeat units. These include capsular antigens from *E. coli*, *Neisseria meningitis, Haemophilus influenza*e, and *Salmonella typhi*.

Finally, some bacteria simply secrete extracellular transglycosylases so that EPS biosynthesis happens entirely outside of the cell. Glucansucrases polymerize starch or sucrose into a variety of α-linked glucans, and fructansucrases assemble β-linked fructans. The products are more heterogenous than those synthesized by other pathways, but are particularly prominent products made by cultures of lactic acid bacteria. In fermented foods, EPS synthesized this way can have significant effects on food texture, while in biofilms (see below), they can play important roles in adhesion and colonization, for example by the dental pathogen *Streptococcus mutans*.

FUNCTIONS OF CAPSULES

Capsules are EPS attached to the surfaces of bacteria, and are common among many different species. In different species and strains they have different compositions, different thicknesses, and different mechanisms of synthesis, but their physiological roles fall generally into the categories of protection and/or adherence. We will discuss adherence more generally below, and will focus here on their role in protecting the cell against stress.

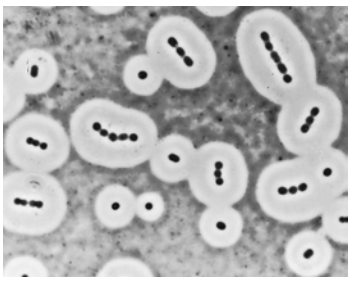


Figure 14.10. An India ink stain of *Acinetobacter calcoaceticus*, allowing visualization of the capsule as areas lacking stain under a light microscope. From Taylor & Juni (1961) J Bacteriol 81(5):688-693.

Capsules protect bacterial cells from abiotic stresses, like desiccation, UV light, and oxidative stress, but they have been more extensively studied for their roles in resistance to biotic stresses. This includes antibiotics and antimicrobial peptides, which may not be able to penetrate through the capsule, bacteriophage, whose receptors may be concealed under the capsule, as well as resistance to consumption by protists or phagocytic cells of the animal immune system.

Immune recognition of capsular polysaccharides, as I mentioned above, is the basis of serotyping, and capsule is one of the main targets of the antibodies the adaptive immune system generates against many bacterial pathogens. Different strains of the same species often produce very different capsular polysaccharides, and capsule biosynthesis operons are often horizontally transferred from one strain to another. Selective pressure by the immune system drives the evolution of new, serologically distinct capsules.

Some bacteria are capable of *antigenic variation*, in which the same strain encodes two or more capsular polysaccharide synthesis pathways and is able to switch among them. This is a mechanism for evading the immune system, and allows infections to persist much longer, since antibodies need to be generated against all of the possible capsule types encoded by the infecting strain.

It should be noted, however, that while capsule's role in pathogenesis has received a lot of attention, many commensal bacteria **also** produce capsule, and those capsules seem to play important roles in establishing productive host-microbe interactions. The mechanism(s) by which they might do so, and how and whether the immune system can distinguish between capsules produced by pathogens and those produced by commensals remains poorly understood.

DISCUSSION PROBLEM SET #27: *STREPTOCOCCUS PNEUMONIAE* CAPSULE SHEDDING

UAB Microbiology has a long-standing tradition of expertise in the biology of the pathogen *Streptococcus pneumoniae* (also called "the pneumococcus"). *S. pneumoniae* is a leading cause of pneumonia, and its capsule is a key determinant of its pathogenesis. There are at least 100 *S. pneumoniae* capsule serotypes, most of which were identified by UAB's own Dr. Moon Nahm, whose lab the NIH declared a "national treasure" in 2016.

Nearly all of the capsule serotypes of *S. pneumoniae* are synthesized by Wzy-dependent pathways. The most important exception is serotype 3, which is able to cause extremely serious invasive disease. The serotype 3 capsule is synthesized by a synthase-dependent pathway. Carlos Orihuela (also of UAB Microbiology) has recently written a good review of the properties of serotype 3 *S. pneumoniae*, if you're interested in more details.

The type 3 capsule is not covalently bound to the cell wall, and is shed copiously into the medium, a feature that contributes to the high virulence of type 3 strains. The Wzy-dependent capsule types, however, are covalently linked to peptidoglycan by a glycosidic bond to GlcNAc. Surprisingly, those capsule polysaccharides are **also** shed into the medium, albeit in lower quantities. This shedding does not appear to be due to membrane or cell wall turnover, cell lysis, or random breaks in the carbohydrate chains, but the mechanism by which they **are** released is not known.

The following methods are available for *S. pneumoniae*:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent*	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
compatible transposon	√
oligo-directed recombineering	√

Describe an experiment to determine the mechanism by which Wzy-dependent *S. pneumoniae* capsule types are shed into the media. State:

- a model to explain this phenomenon
- the hypothesis derived from that model that your experiment will test
- the independent and dependent variables of your experiment
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiment, and how you will interpret them

* "Can be made competent" kind of sells *S. pneumoniae* short. They are highly naturally competent, and can easily be transformed by mixing cells with DNA. This, of course, is why Oswald Avery's experiments to determine the nature of the genetic material worked with *S. pneumoniae*. Natural competence, along with a tendency to autolyse and release their cellular contents in stationary phase, contributes heavily to the horizontal exchange of capsule biosynthesis loci among *S. pneumoniae* strains.

BIOFILMS

In most environments, the vast majority of microbial cells are not free-swimming *planktonic cells*, but are found attached to surfaces, often in multi-cellular aggregates called *biofilms*. The main exception to this rule is the open ocean, although some authors argue that the relatively high concentration ($\sim 10^5 - 10^6$ per milliliter) of bacteria in the top millimeter of the water column (the *bacterioneuston* or *surface micro-layer*), many of which are found in small aggregates, constitutes a kind of loose biofilm. The same kind of debate about whether any aggregate of bacteria constitutes a "biofilm" is actively discussed in medical fields, as well. Of course, nearly all of our knowledge about bacterial physiology comes from studies of planktonic cells, which are more homogenous in their growth rates and far easier to measure and manipulate. This is a problem for those of us trying to figure out what bacteria are doing in the "real world". It is, for example, very clear that biofilm formation is important in the virulence of many pathogens.

In this second half of the chapter, we will discuss the properties and development of biofilms, how bacteria sense and adhere to surfaces, and some of what is known about the regulation of biofilm formation. As you will see, the extracellular polymers produced by bacteria (including EPS) play an important role in biofilm biology.

Before I begin, though, I want to make a couple of important points. One is that most studies of the molecular biology of biofilm formation have been done using pure cultures and mostly with a fairly limited set of model organisms (notably *Pseudomonas aeruginosa*, for which you can largely thank George O'Toole, a pioneer in the field). Nearly all bacteria can form biofilms of some kind, and actual biofilms usually contain multiple species of bacteria and often other microbes. The second point is that, while biofilms have general properties in common, as we will see, the **details** of how bacteria attach to surfaces, what polymers they secrete to bind the biofilm together, and the effects of biofilm growth on their metabolism can vary widely, even between different strains of the same species. There is nothing simple about biofilms, and our understanding of their biology is actually fairly limited.

Biofilms form at interfaces, including those between liquids and solids, liquids and gases, and solids and gases. They can be macroscopically visible, and the names we give to different kinds of biofilms is often due to their visible appearance. Probably the best-studied kind of biofilm is the slimy kind that forms along solid-liquid surfaces, and most of the molecular biology we will discuss below was done on this kind of biofilm. Biofilms at liquid-gas interfaces are called *pellicles*. You could argue that bacterial colonies on an agar plate are a kind of biofilm forming at a solid-gas interface, and certainly many kinds of bacteria form elaborate multicellular structures under these conditions. When bacteria clump together so that the aggregate of bacteria itself is the surface to which additional bacteria adhere, it is called *autoaggregation* or *flocculation*, and individual aggregates are called *flocs*. In a biofilm, bacterial cells are encased in a *matrix* composed of biopolymers, a species- and strain-specific combination of polysaccharides, proteins, and eDNA.

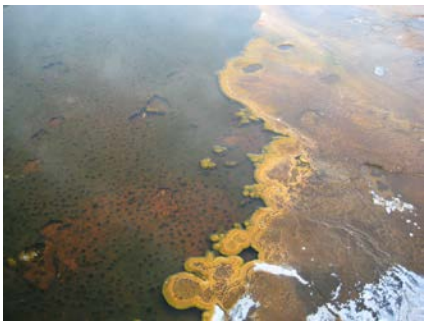


Figure 14.11. A microbial mat, a kind of macroscopic biofilm, on the surface of a hot spring in Yellowstone National Park, Wyoming (Image from Wikimedia Commons, taken by Penny Higgens.)

There is no hard and fast rule about how many adherent bacteria you need to constitute a "biofilm". In reality, there is a continuum between planktonic cells, single adherent bacteria, microcolonies, *etc.* all the way up to very large and complex microbial mats many centimeters thick.

MEASURING BIOFILM FORMATION

There are a variety of ways of visualizing and quantifying biofilm formation in the lab. The challenge is making measurements of biofilm formation quantitative instead of qualitative, but there are a couple of approaches that are commonly used to do that. Some images of biofilms visualized with different techniques are shown in Figure 14.12.

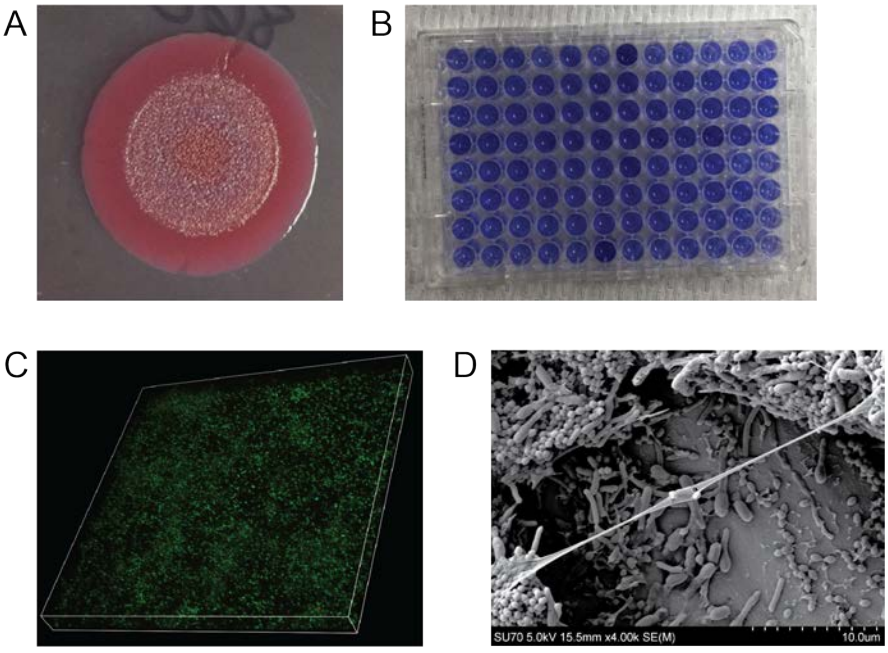


Figure 14.12. (A) "Wrinkled colony" *E. coli* biofilm on an agar plate containing the dye Congo red. (B) 96-well plate with crystal violet dye used to quantify biofilm formation. (C) 3-D confocal micrograph of an *E. coli* biofilm on a glass coverslip, stained with green fluorescent dye. Images in A-C by Rhea Derke and Leanna Crafford (Gray lab). (D) Scanning electron micrograph of a polymicrobial biofilm on a stainless steel surface. Image from Wikimedia Commons, taken by Krzysztof A. Zacharski.

The simplest way to measure a biofilm is simply to take a picture of it. If one strain makes a lot of biofilm (at the gas-liquid interface in a test-tube, for example) and another does not, then it may be very easy to just show what it looks like. This is qualitative, but if there's a big enough difference, that might be fine.

A common assay to quantify differences in biofilm formation between strains uses the purple dye crystal violet. Cultures are grown, often in 96-well plates, under conditions where biofilms are expected to appear. The liquid media are then removed, and the plates are rinsed thoroughly to remove any cells that are not firmly adhered to the plastic. Next, crystal violet is added, which stains the adherent biofilm. Excess dye is rinsed away (this assay is very messy!), then resolubilized in ethanol. The dye in the resulting extract can be quantified in a spectrophotometer. Here is a link to a video protocol showing this assay. The crystal violet assay has pretty high variability, but does a reasonable job of

quantifying total biofilm mass. It does not give any information about more subtle structural differences between biofilms or distinguish between living cells and matrix materials.

Fluorescent microscopy, and especially *confocal microscopy*, which generates a 3D image by taking images at several different focal planes, is very useful for obtaining more subtle assessments of biofilm structure. Cells in a biofilm can be stained with various fluorescent dyes or engineered to express fluorescent proteins. There is a variety of software available designed to analyze and extract quantitative data from these kinds of images. Some students in my lab have been getting reasonable results lately using BiofilmQ, but there are lots of other options, including Comstat, Imaris, and Volocity (those last two cost money, while BiofilmQ and Comstat are both free).

The highest resolution images of biofilms are obtained with electron microscopy, but it is difficult to extract quantitative information from electron micrographs, and the samples need to be fixed in such a way that the cells will certainly be dead when the images are taken.

Some labs that study biofilms intensively use sophisticated *microfluidics* technology which can maintain carefully-controlled flow rates, media composition, and growth conditions under constant microscopic observation. These kinds of experimental setups allow the best observations of biofilm formation and development over time, but require specialized equipment.

BIOFILM DEVELOPMENT

The growth of a biofilm goes through a number of developmental stages, which are illustrated in Figure 14.13:

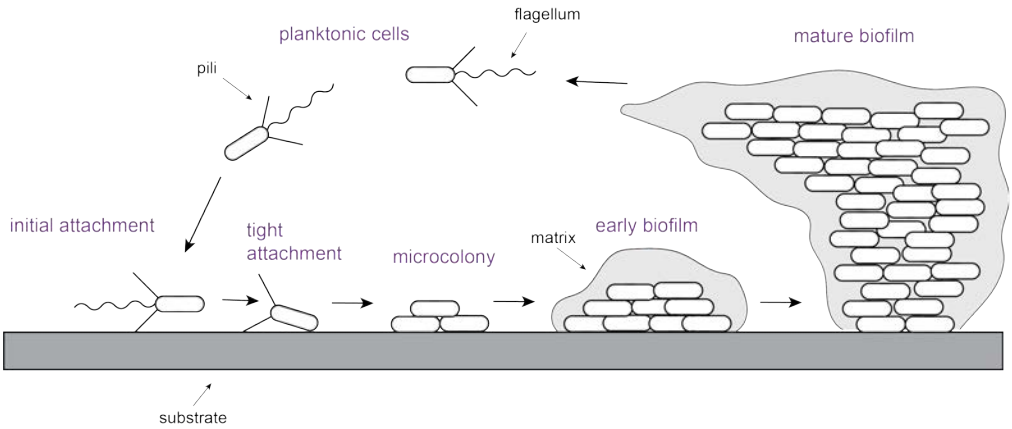


Figure 14.13. The developmental stages of a single-species biofilm. Planktonic cells swim freely until they encounter a surface, which may be sensed by a variety of mechanisms, commonly involving flagella and / or pili. Initial attachment is reversible, but is followed by irreversible tight attachment Attached cells divide to form microcolonies, and produce extra-cellular matrix polymers to become an early, or immature biofilm. Mature biofilms contain many more cells, often have complex 3-dimensional structures, and are able to shed planktonic cells to begin the cycle again.

Planktonic cells, if they are motile, swim through the liquid phase of their environment until they encounter a surface. They may recognize that this has happened by a variety of mechanisms, including changes in the properties of liquid flow, distortions in their cell shape, increased resistance to flagellar rotation (because the surface gets in the way), or by binding of adhesive pili (usually, but not always type 4 pili) to the surface. This initial encounter is reversible, and cells may spend some time close to the surface before swimming away. Non-motile bacteria (*e.g. Staphylococcus* and *Streptococcus* spp.) have a simpler attachment process which mostly involves settling by gravity or fluid flow.

However, on the path to a biofilm, the cell eventually becomes more permanently attached to the surface. It generally loses any motility appendages it might have had, and begins to replicate, forming a *microcolony*. As this continues to grow, matrix compounds are produced and the cells begin to enter into a state that is more biofilm-like. As it matures, the biofilm grows in height and width, supported by the matrix and forming more complex structures.

Finally, *biofilm dispersal* is the release of planktonic cells from the biofilm, or even the complete reversal of biofilm formation, and involves the active modification and breakdown of matrix components. See this review for more information. Dispersal is a regulated process, and since bacteria in biofilms are more resistant to antibiotics and disinfectants than planktonic cells, there has been quite a bit of interest in developing drugs that stimulate dispersal for use in combination with chemicals that kill bacterial cells.

ADHERENCE

The first step in biofilm formation is adherence. Bacteria can adhere to abiotic surfaces, like rocks, metal, and plastic, as well as biotic ones, like plant leaves, chitin, and animal epithelia (both internal and external).

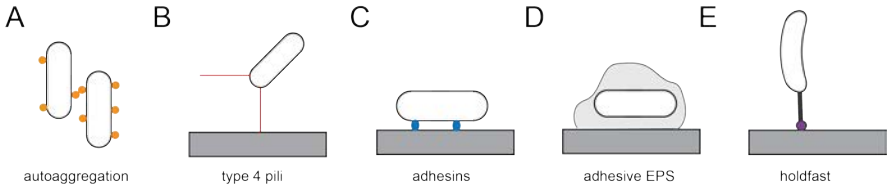


Figure 14.15. Different bacterial structures involved in adhesion. (A) Autoaggregation proteins bind to each other, causing bacteria to clump together. (B) Type 4 pili bind to surfaces and retract, pulling cells closer to the surface. (C) Adhesins are bacterial surface proteins that bind to surfaces directly, often to a specific receptor. (D) Some EPS are sticky, forming a glue that attaches cells to surfaces. (E) The holdfast of *Caulobacter* and similar polar adhesion foci in other α-proteobacteria combine adhesive EPS and proteins into a very strong attachment point.

Bacteria can adhere to surfaces by simple physical interactions, like charge-charge interactions (bacteria are usually negatively charged) or hydrophobic interactions, but also produce a variety of *adhesins*, which are molecules produced by bacteria that mediate the attachment of those bacteria to surfaces. They include proteins and carbohydrates, and can either bind to specific receptors or be just generally sticky.

When the surface that bacteria are adhering to is made up of other bacteria, and especially when it is other bacteria of the same species, this depends on specific *autoaggregation* adhesins. One example is *E. coli* "antigen 43", a 1099-amino acid autotransported (T5SS) protein encoded by the *flu* gene (which stands for "**flu**ffing", a description of the flocculent phenotype of antigen 43-expressing cells). The interaction between antigen 43 molecules on adjacent *E. coli* cells is tight and specific. Many bacteria autoaggregate, and this can be an important step in the early development of biofilms.

We discussed curli in **Lecture 13**, and these secreted amyloid proteins play an important adhesive role in mature biofilms produced by enterobacteria. Curli are very sticky and extremely stable, the exact properties that make amyloid aggregates in the brain so toxic in human neurodegenerative diseases. For a biofilm, however, this reinforces the matrix and makes it extremely resistant to physical stress.

Other adhesin proteins are divided generally into *fimbrial* and *non-fimbrial adhesins*. Fimbrial adhesins are fibrous and extend well away from the cell, and are also usually called pili. There are several types, but many bacteria use retractable type 4 pili (recall that these are homologous to type 2 secretion systems, **Lecture 13**). The initial attachment of *C. crescentus* swimmer cells, for example, depends on the type 4 Tad pili, which bind non-specifically to surfaces. Other kinds of pili can also be involved in attachment. The type 1 pili of uropathogenic strains of *E. coli* (UPEC), which are not retractile, but are exported by a type 5 chaperone-usher secretion mechanism (**Lecture 13**) are tipped with an adhesin called FimH. FimH binds tightly and specifically to mannose sugars on glycoproteins found on mammalian epithelial cell surfaces.

Non-fimbrial adhesins are a very diverse group of proteins, and again, range from non-specific "sticky" proteins to proteins with extremely specific binding partners.

Some Gram-positive and mycobacteria seem to export some proportion of the metabolic enzyme glyceraldehyde-3-phosphate dehydrogenase (GapDH, see **Lecture 17**) to their surfaces, where it is involved in adherence to epithelial and endothelial cells and host proteins like plasminogen and fibrinogen. This kind of dual protein function is called *moonlighting*, and this particular case has required particularly extensive evidence to convince the scientific community that it's real, since GapDH's role as a central metabolic enzyme is so important.

Another, much more specialized example of a non-fimbrial adhesin is intimin, a surface protein produced by **e**ntero**p**athogenic and **e**ntero**h**emorrhagic ***E. c****oli* strains (EPEC and EHEC, respectively). Intimin forms a very specific bound complex with a protein called Tir (**t**ranslocated **i**ntimin **r**eceptor). Tir is a substrate of the type 3 secretion system (T3SS, **Lecture 13**) of EPEC and EHEC, and is injected from the bacteria into host cells, where it is displayed on the host cell surface. These strains therefore deliver their own adhesin receptor into the cells to which they want to bind, allowing them to form extremely tight attachment sites.

Of course, polysaccharide EPS also play roles in attachment, since many EPS can act as, essentially, glues to stick cells to surfaces and each other. We've already discussed the *Caulobacter* holdfast, which contains a polysaccharide that is one of the strongest biological adhesives ever described. Surprisingly, the repeat unit structure of the holdfast EPS is not currently known (polysaccharide biochemistry is difficult!), but recent work from Sean Crosson's lab at the University of Chicago has shown that it contains an 1,4-linked backbone of glucose, mannose, *N*-acetylglucosamine, and xylose monosaccharides that is decorated with branches at the C-6 positions of glucose and mannose.

As another example, *Pseudomonas aeruginosa* produces several EPS, including Psl, Pel, and alginate, all of which play different roles in adhesion and biofilm structure. Different strains produce different amounts of these polysaccharides, with most laboratory strains and wound isolates producing Psl and Pel, but very little alginate, while isolates from the lungs of cystic fibrosis patients produce very large amounts of alginate, but very little Pel. Psl is anchored to the bacterial cell surface, is found throughout *P. aeruginosa* biofilms, and is involved in both autoaggregation and attachment to surfaces. Pel, on the other hand, is required for formation of pellicle biofilms, is localized to the periphery and vertical "mushroom stalk" structures of surface-attached *Pseudomonas* biofilms, and is positively charged, which allows it to interact with extracellular DNA in the biofilm matrix.

If you want to explore this topic in a little more depth and see another example of the variety of polysaccharides bacteria use for adhesion and biofilm structure, here is a good review of EPS and biofilms in staphylococci.

STRUCTURAL AND METABOLIC PROPERTIES OF BIOFILMS

Bacteria in biofilms are much more resistant to physical stresses, disinfectants, antibiotics, predation by protists, and attack by the immune system. Some of this has to do with the physical properties of the biofilm and its matrix, but there are also important metabolic and regulatory differences between planktonic and biofilm-grown bacteria that contribute to stress tolerance.

The matrix components we discussed above as "adhesins" are often also important structural components that give biofilms their physical properties and contribute to the ability of bacteria in biofilms to resist stress. EPS and other adhesins interact with cells, surfaces, and each other, as well as with other compounds. An important component of the matrix of many biofilms is **e**xtracellular DNA (eDNA), first described by Cynthia Whitechurch in 2002 (in this paper). As mentioned above, positively-charged EPS can cross-link with eDNA to form a tight, gel-like matrix, and non-specific DNA-binding proteins can play a similar role. In some cases, eDNA seems to be released from dead cells in the biofilm, but in other cases chromosomal DNA is actively secreted by bacteria. Biofilms can also become *mineralized*, with the deposition of carbonate minerals solidifying the matrix into a rock-like state. This is clinically relevant not only in terms of the calcification of dental plaque but also the formation of kidney stones, which can be the product of mineralized biofilms produced by uropathogenic bacteria. The more cross-linked and impermeable the matrix, the more difficult it is for predators, immune cells, and disinfectants to penetrate or remove the biofilm from a surface.

The exact composition of a particular biofilm matrix depends on the strain(s) and species of the bacteria that live in it, as does the arrangement and interactions between the organism(s) involved. Since most biofilms are *polymicrobial*, there is a lot of complexity to deal with in "real" biofilms. See this paper for a very cool visualization of polymicrobial biofilms formed by the oral microbiota, to get an idea of what we're dealing with here. Some labs (including ones here at UAB) are beginning to address this complexity with laboratory models in which biofilms composed of two or more genetically-manipulatable species are studied, and this is an exciting area of contemporary research.

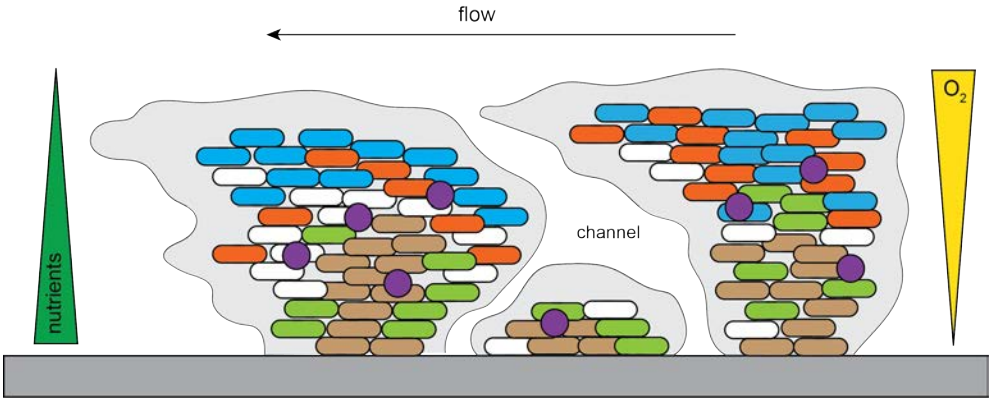


Figure 14.14. The structure of a polymicrobial biofilm, shaped by gradients of nutrients, oxygen, and flow. Different bacteria are found in different parts of the biofilm, with (for example) aerobes (blue and orange) more abundant near the surface and anaerobes (brown and green) near the substrate. Channels through the biofilm allow the exchange of molecules among different parts of the biofilm or the removal of waste products. Biofilms often contain dead or metabolically-inactive cells (white).

Bacteria growing in biofilms experience a number of gradients different from those generally experienced by planktonic cells, and unlike planktonic cells, bacteria in a biofilm are not able to move to respond to chemical gradients (we will address motility in **Lecture 15**). The cells near the top of a biofilm may experience higher oxygen levels, higher flow rates, greater light intensity, or greater exposure to environmental toxins, while cells near the bottom may experience

lower levels of these stimuli, but higher levels of nutrients or available electron acceptors, depending on the species and the nature of the substrate to which they are attached. Biofilms in liquid environments often contain channels through which liquids can circulate, allowing nutrients and waste products to diffuse through the biofilm, but the permeability and density of the matrix can vary dramatically.

The physical properties of the biofilm matrix are protective, but that is not the only reason that bacteria in biofilms are resistant to stress. Unlike planktonic cells, which can be in a rapidly-replicating exponential growth phase, bacteria in biofilms grow and replicate much more slowly, if at all. This is analogous to the stationary phase of planktonic cultures, and slow-growing cells are usually **much** more resistant to killing by antibiotics and other stress conditions.

This is a complex topic, but the general concept is that antibiotics typically kill bacteria by inhibiting cellular processes essential for **growth**, like RNA polymerase, ribosomes, or cell wall synthesis. If the cell isn't growing and dividing, then inhibition of these processes isn't nearly as big a deal, and the cell can degrade, export, or wait out the antibiotic exposure. This is called *tolerance*.

Persister cells are an extreme example of this phenomenon which can be observed even in planktonic cultures. A small proportion of the cells in a culture contain very low levels of ATP (**Lecture 16**), are not metabolizing, and are therefore not susceptible to killing by antibiotics. They can then recover and regrow after the antibiotics are removed. Tolerance is distinct from *antibiotic resistance*, which results from heritable genetic mutations. The cultures regrown from a sample of persister cells will be just as sensitive to antibiotics as the original strain. As many as 1% of the cells in a biofilm or stationary phase culture may be persisters.

DISCUSSION PROBLEM SET #28: BIOFILMS IN THE *VIBRIO CHOLERAE* LIFE CYCLE

The causative agent of cholera, *Vibrio cholerae*, has two distinct phases to its life cycle. In brackish water, *V. cholerae* forms biofilms on the chitinous exoskeletons of copepods, a kind of microscopic marine crustacean, and doesn't cause any particular harm to those organisms. However, when *V. cholerae* is ingested by a human, it grows to extremely high density in the small intestine, produces a potent toxin, and causes devastating "rice water stool" diarrhea, in which a patient loses up to 20 L of water per day. There is good evidence that *V. cholerae* forms biofilm-like aggregates in the human gut.

V. cholerae senses chitin (a polysaccharide not found in mammals, but common in the cell walls of fungi and the exoskeletons of crustaceans and insects) with a two-component regulatory system driven by the histidine kinase ChiS. ChiS activation leads to upregulation of a variety of genes for chitin utilization, including a type 4 pilus that mediates attachment to chitin. The chitin-regulated pilus is required for biofilm formation on chitin, but not for virulence in mouse models of cholera.

The following methods are available for *V. cholerae*:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
generalized transducing phage (not commonly-used, though, as far as I can tell)	√
compatible transposons	√

oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

You wonder if there is any overlap between the pathways involved in biofilm formation in the two very different environments inhabited by *V. cholerae*.

Describe an experiment or series of experiments to address this question and identify genes or proteins involved in biofilm formation by *V. cholerae* A) on crustacean shells, B) in the mammalian gut, and C) in both environments. State:

- your hypothesis and how your experiment(s) will test that hypothesis
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiment(s), and how you will interpret them

---

REGULATION OF BIOFILM FORMATION

There are complex dedicated regulatory pathways that control the formation and dispersal of biofilms, and we will only be able to touch on a couple of these briefly. The first is quorum sensing, which we mentioned in **Lecture 11**, and which plays an important role in biofilm regulation in a wide range of Gram-negative and Gram-positive bacteria. To review, bacteria synthesize and export quorum sensing signal molecules or *autoinducers*, and when those signals reach a threshold concentration in the medium they activate expression of specific genes.

Quorum sensing systems are a mechanism by which bacteria sense cell **density**, and bacteria in a biofilm are in physical contact with each other, which is essentially the maximum possible density. The diffusion of quorum sensing signaling molecules may also be limited by the permeability of the matrix, increasing the local concentration. Quorum sensing is involved in biofilm formation in *Pseudomonas, Vibrio, Staphylococcus*, and a variety of other bacteria. For the sake of space, I'm going to focus on quorum sensing in Gram-negative biofilms, but you should be aware that the role(s) of quorum sensing in Gram-positive biofilms is complicated and quite different.

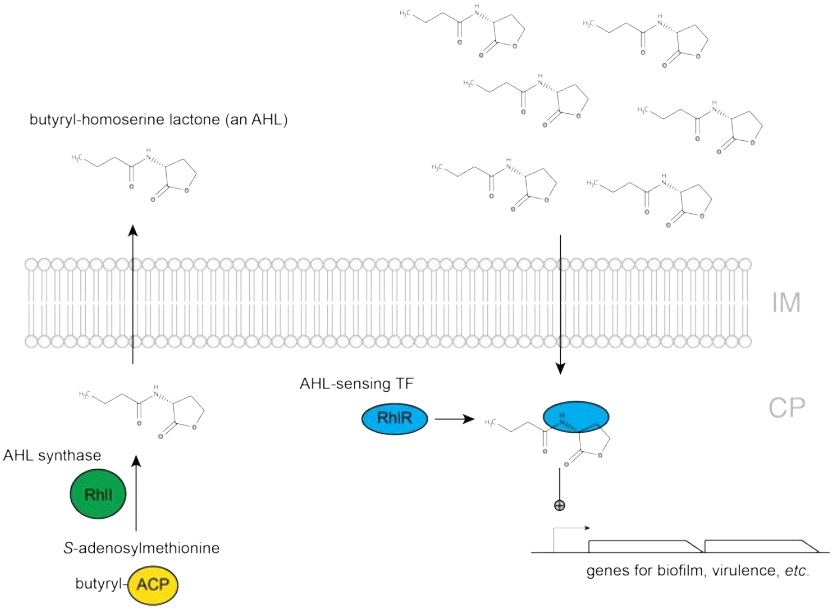


Figure 14.16. An example of an **a**cyl **h**omoserine **l**actone (AHL)-dependent quorum sensing signaling system, in this case the RhlI/RhlR system of *Pseudomonas aeruginosa*. AHL synthases synthesize their cognate AHL from *S*-adenosylmethionine and an acylated **a**cyl **c**arrier **p**rotein (ACP), in this case with a 3-carbon butyryl group. AHLs diffuse freely through cell membranes, and once they reach a threshold concentration they bind to AHL-sensing transcription factors (in this case, RhlR) that regulate expression of genes involved in population behaviors, including biofilm development.

The initial work describing quorum sensing was done in *Vibrio* (now *Aliivibrio*) *fischeri* by Bonnie Bassler's lab, and in that Gram-negative organism, one of the systems under quorum sensing control is the expression of *luciferase*, which produces light. In *A. fischeri*, the signaling molecule *N*-3-oxohexanoyl homoserine lactone is synthesized by the enzyme LuxI and detected by the transcription factor LuxR. Homologous systems using different **a**cyl **h**omoserine **l**actone (AHL) signaling molecules are common in Gram-negative bacteria. *P. aeruginosa*, for example, produces four distinct AHLs (one of which, butyryl-homoserine lactone, is illustrated in Figure 14.16), each of which regulates a different set of genes involved in biofilms, virulence, and other population-level behaviors.

Due to the generally hydrophobic nature of AHLs, they typically pass freely through cell membranes, so the concentration of AHL inside the cell is the same as that outside of the cell. This is not universally true for all combinations of AHLs and cell membranes, though. One of the **other** *P. aeruginosa* AHLs (*N*-3-oxohexanoyl homoserine lactone, generated and sensed by LasI/LasR) requires a dedicated efflux pump (MexAB-OprM) to be excreted from the cell. Note that this is the same molecule produced and sensed by LuxI/LuxR of *A. fischeri*, in which organism it diffuses freely through the membrane.

Many Gram-negative and some Gram-positive bacteria produce a conserved quorum signaling molecule called ***a**uto**i**nducer* **2** (AI-2; a furanosyl borate diester) which appears to function as a signal of overall species-nonspecific bacterial population density, and is also involved in regulating genes required for biofilm production. As mentioned in **Lecture 11**, Gram-positive bacteria usually use peptide signals for quorum sensing.

For an in-depth review of quorum sensing in biofilms, I recommend this paper.

Another important regulator of biofilm production is the second messenger (**Lecture 4**) cyclic-di-GMP (c-di-GMP). This molecule is widely conserved among bacteria, and is involved in regulating developmental processes in diverse species. It was first discovered in 1987 in *Komagataeibacter xylinus*, where it is required for cellulose production and formation of a pellicle biofilm (see Discussion Problem Set 7 in **Lecture 3**).

C-di-GMP is synthesized from GTP by *guanylate cyclases*, which contain a characteristic GGDEF motif, and degraded by specific phosphodiesterases, containing either HD-GYP or EAL domains. It is fairly common for GGDEF and EAL domains to occur in the same protein, indicating that these proteins can both synthesize and degrade c-di-GMP.

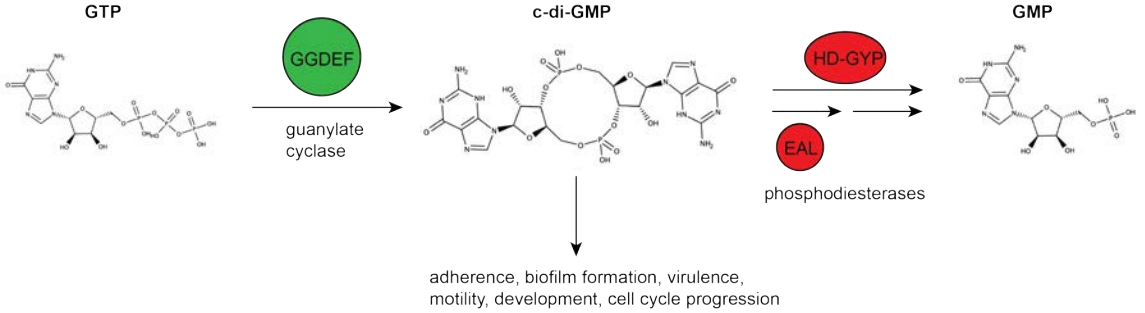


Figure 14.17. Cyclic di-GMP is synthesized from GTP by guanylate cyclases, which contain a characteristic GGDEF motif. C-di-GMP interacts with effector proteins that regulate a variety of cellular functions, as discussed in the text. C-di-GMP is degraded to GMP by specific phosphodiesterases, of which there are two types: those containing an HD-GYP domain, and those with an EAL motif.

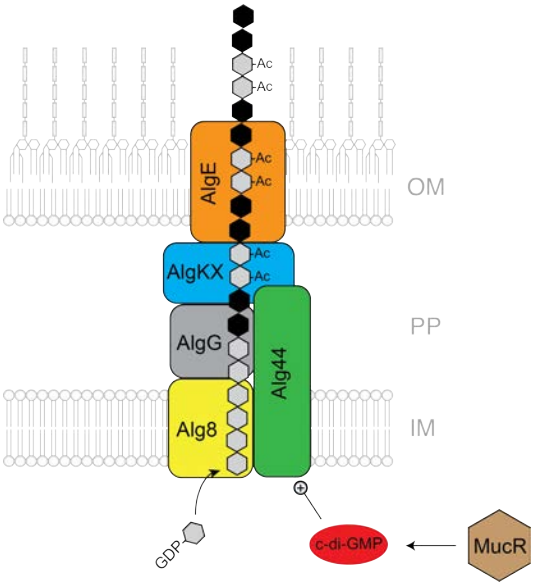
Guanylate cyclases have a variety of sensor domains which respond to environmental stimuli to activate c-di-GMP production. C-di-GMP is then bound by different effector molecules, whose activity is modulated by that binding. This can regulate bacterial processes at transcriptional, post-transcriptional, and post-translational levels (**Lecture 4**).

Many of the effectors regulated by c-di-GMP are relevant to biofilm formation. In *E. coli*, for example, c-di-GMP-bound YcgR inhibits the flagellar motor, inhibiting motility at the same time that c-di-GMP binds to BcsA, a protein involved in cellulose synthesis and to PdeR and DgcM, which ultimately lead to expression of CsgD , the transcription factor that activates the *csg* operon encoding curli. This theme of c-di-GMP repressing motility while inducing attachment / biofilm is conserved in many bacterial systems. C-di-GMP is involved in regulation of type 4 pili in many bacteria, including *Myxococcus*, *Vibrio*, *Pseudomonas, Clostridium*, and *Caulobacter*, and is required for production of specific EPS in many biofilm-forming bacteria.

For a review of the diverse roles of c-di-GMP in bacterial regulatory networks, this paper is a great place to start, with an older, but more exhaustive review here.

DISCUSSION PROBLEM SET #29: CYCLIC-DI-GMP SIGNALING SPECIFICITY

Cystic fibrosis-associated strains of *Pseudomonas aeruginosa* produce large amounts of alginate using a synthase-dependent EPS synthesis pathway encoded by the *alg* locus.



Alg8 and Alg44 polymerize GDP-mannuronate (grey hexagons), some of which are epimerized into gluluronate (black hexagons) by AlgG or acetylated by AlgK or AlgX as it crosses the periplasm to the translocase AlgE. C-di-GMP produced by the diguanylate cyclase MucR is bound by Alg44 and activates polymerization. Note that alginate synthesis is regulated at multiple additional stages, and this figure is a dramatic simplification of the process.

C-di-GMP production by MucR is required for alginate production (*i.e.* a *mucR* null or enzymatically-inactive mutant does not produce alginate), but the genomes of *P. aeruginosa* strains also encode 15 or 16 **other** GGDEF domain-containing guanylate cyclases.

The following methods are available for *P. aeruginosa*:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
generalized transducing phage	√
compatible transposons	√
oligo-directed recombineering	√

CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

Propose a model to explain why c-di-GMP synthesis by MucR **specifically** is required for alginate production in *P. aeruginosa*. Describe an experiment or series of experiments to test your model. State:

- your model
- your hypothesis and how your experiment(s) will test that hypothesis
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiment(s), and how you will interpret them

LECTURE 15: MOTILITY

INTRODUCTION

Not all bacteria are able to move under their own power, but in many environments the ability to do so (*motility*) provides a significant advantage. In this chapter, which is in many ways the opposite of the previous section on adherence and biofilms, I will discuss several different mechanisms by which different bacterial species move. I will also cover the basics of how bacteria direct their movement towards or away from particular stimuli, a phenomenon generally called *chemotaxis*.

OVERVIEW OF BACTERIAL MOTILITY

Motility is essential for the survival of many bacteria both in the environment and for the ability of many pathogens to cause disease. It allows cells to spread into new environments, either as individuals or as populations. The mechanisms of motility are varied, and different mechanisms are necessary in different environments, depending on the viscosity of the medium, whether the cells are moving as individuals or as a group, whether they are moving through a liquid or along a surface, and the nature of that surface. In many organisms,

As we will see, motility depends on complex multi-protein machinery, and these are among the largest and most complicated structures found in bacterial cells. (An individual flagellum consists of 30 or so different proteins, with copy numbers ranging from 1 to 30,000, a total length of as much as 15 μm, and a molecular mass of more than 1000 MDa, and some species can have dozens of flagella per cell.) In the interests of space, we will not go into great detail about the genes and proteins that make up any individual motility complexes here, although we **will** briefly discuss the molecular machinery underlying chemotaxis in *E. coli*.

SWIMMING

In relatively low-viscosity liquid environments, many bacteria are able to actively swim using *flagella*. Flagella are fairly rigid helical filaments (either right- or left-handed, depending on the species) that rotate to act as propellers to push or pull bacterial cells through the medium. This is unlike eukaryotic flagella, which are completely unrelated and move in a whip-like manner.

Different bacterial species have different numbers and arrangements of flagella (Figure 15.1). Some species are *monotrichous*, and have a single *polar flagellum* at one end. Others are *amphitrichous* and have one flagellum at each pole. Species with *lophotrichous* and *amphilophotrichous* flagella have multiple polar flagella at one or both poles, respectively, and those with *peritrichous* flagella have multiple flagella distributed around the cell surface. *E. coli*, the organism in which flagellar motility is best studied, has an average of 4 - 6 peritrichous flagella per cell. In some Gram-negative bacteria, the flagella are *sheathed*, which means they are covered by a lipid bilayer derived from the outer membrane. *Aliivibrio fischeri* (formerly *Vibrio fischeri*), the bioluminescent symbiont of Hawaiian bobtail squid, has recently been shown to coil its single polar flagellum tightly around itself to "burrow" through narrow channels filled with mucus to enter the light organ of its host.

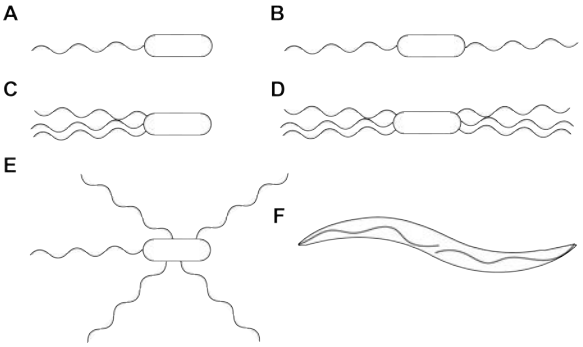


Figure 15.1. Representative arrangements of bacterial flagella, including (A) polar or monotrichous, (B), amphitrichous, (C) lophotrichous, (D) amphilophotrichous, and (E) peritrichous flagella, as well as (F) a very poorly-drawn rendition of the internal periplasmic flagella of spirochetes.

The *flagellar basal body* is a specialized type III secretion system (see **Lecture 13**) that is dedicated to the export of flagellar proteins. The basal body and attached flagellum is spun by *motor proteins* (called MotA and MotB) that are held in place by interactions with the cell wall and use either a proton or sodium gradient to power the rotation (see **Lecture 16**). The flagellum is essentially a very small rotary electric motor. MotAB makes up the *stator*, or stationary part of the motor, and turns the *rotor* made up of the rest of the basal body. The speed of rotation of bacterial flagella is variable, but generally appears to be tuned to maintain constant torque under physiological conditions. The flagella of

archaea are not homologous to bacterial flagella, and are evolutionarily related to type 4 pili, but are also rotating helical filaments, so function similarly in cellular propulsion.

The maximum swimming speed of different bacteria varies a great deal, and is also affected by the viscosity of the medium they are moving in (unsurprisingly, bacteria move slower in more viscous solutions). *E. coli*, whose cells are about 1 µm long, has a maximum speed of about 30 µm per second, while the freshwater predatory bacterium *Bdellovibrio bacteriovorus* moves at about 160 µm per second. Some marine bacteria have been reported to swim at up to 400 µm per second, which may be an adaptation to life in the very dilute, sparsely-populated ocean environment.

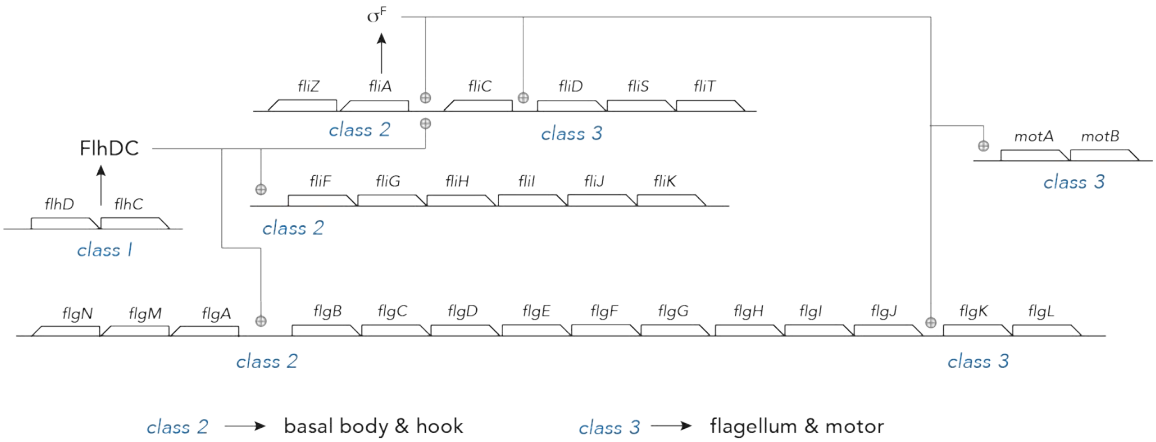
Flagella are too thin to see with normal light microscopy (~ 20 nm in diameter), but can be stained or visualized in a variety of ways. Most of these, unfortunately, involve killing the cell. However, many recent studies have taken advantage of a technique in which the *flagellin* protein (which makes up the body of the flagellar filament) is mutated to contain an extra cysteine residue. A fluorescent dye attached to a cysteine-reactive group (usually a maleimide) can then be added to fluorescently label the flagella of living cells. Analogous methods have been applied to visualize many different proteinaceous cell surface structures.

Swimming motility in the spirochetes is unlike that of any of the bacteria mentioned above, although it does still depend on flagella. Spirochete cells are long, corkscrew-like spirals that can burrow through extremely viscous environments, notably the connective tissue of animal hosts. Spirochetes cause a number of important diseases (*e.g.* leptospirosis, Lyme disease, and syphilis), and their ability to penetrate host tissues is key to their pathogenesis. Spirochetes have amphitrichous flagella anchored near the cell poles (up to hundreds, depending on the species), but the flagella are wrapped close to the cell **inside** the periplasm (sort of shown in Figure 15.1 F, although that is an appallingly bad drawing). As they rotate, they turn the whole cell body, drilling it through the tissue. The flagella of spirochetes are also sheathed in a complex, asymmetric protein layer not found in other bacteria, and which appears to be involved in maintaining a supercoiled structure important for determining overall cell shape and driving motility.

DISCUSSION PROBLEM SET #30: REGULATION OF MOTILITY IN *SALMONELLA TYPHIMURIUM*

Salmonella typhimurium is motile during log phase growth, but non-motile during stationary phase (appropriately enough).

The expression of the many genes encoding components of flagella in *S. typhimurium* is regulated in several stages. First, the expression of the "class 1" flagellar *flhDC* operon, encoding the transcription factor FlhDC, is under both positive and negative control by multiple regulators (including repression by the osmotic stress regulator OmpR and activation by the quorum-sensing regulator QseB and the cAMP-sensing regulator CRP).



When FlhDC is expressed, it activates transcription of the "class 2" operons encoding components of the flagellar basal body and hook, as well as the alternative sigma factor σ^F and its corresponding anti-sigma factor FlgM (see **Lecture 4**). Expression of the *fliA* gene encoding σ^F is also under the control of other transcription factors, including repression by the nitrite-sensitive repressor NsrR and the curli operon regulator CsgD, and activation by σ^F itself.

When the basal body is completed, it becomes a functional T3SS (see **Lecture 13**) and exports FlgM, releasing σ^F and allowing σ^F-dependent expression of the "class 3" flagellar genes, including flagellin (which is exported through the

basal body and assembled into the flagellum itself) and the MotAB motor proteins responsible for rotating the flagellum.

S. typhimurium is pretty closely related to *E. coli*, and you have access to all of the tools of molecular genetics:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
generalized transducing phage (P22)	√
compatible transposons	√
oligo-directed recombineering	√
CRISPR and related technologies (e.g. CRISPRi)	√
a genome-wide knockout collection (link)	√

Describe a series of experiments to determine at **which point** in the flagellar regulation cascade the inhibition of motility in stationary phase occurs (e.g. expression of the class 1, class 2, or class 3 genes). State:

- a model to explain this phenomenon
- a hypothesis to test that model
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

SWARMING

The other form of bacterial motility that depends on flagella is *swarming*, which is one way in which bacteria move along solid surfaces. Swarming is a group activity that requires many cells working together, and only occurs on soft, moist surfaces (in the lab, freshly-made rich media plates containing 0.4 – 0.6 % agar). It is often, but not always, associated with the secretion of *surfactants*, which are amphipathic lubricants that reduce the surface tension, and therefore the drag, between bacteria and a solid surface. Many surfactants (like surfactin from *Bacillus subtilis* and serrawettin from *Serratia liquefaciens*) are small, cyclic, non-ribosomally assembled peptides (see **Lecture 18**).

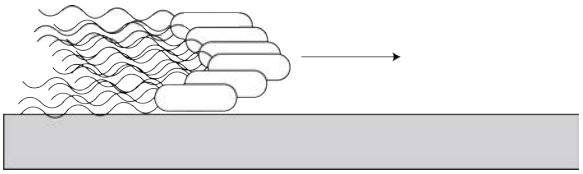


Figure 15.2. Swarming motility is a group behavior in which populations of bacteria move along moist surfaces. It is dependent on the concerted action of many flagella per cell.

One notable feature of swarming is that it requires the interaction of many flagella, so that bacteria often increase the expression and number of flagella they produce under swarming conditions. In some species, including several *Vibrio* spp., that normally produce only a single polar flagellum, an entire second flagellar operon system is expressed for swarming, producing large numbers of peritrichous flagella in addition to the normal polar one.

Swarming bacteria can move very rapidly, at close to the same speed as swimming (*i.e.* about 30 μm per sec for *E. coli*), and swarming cells can spread across the surface of a soft agar plate very quickly. For this reason, many "domesticated" lab strains of bacteria have been selected for the loss of swarming, since fast spreading colonies are inconvenient for the researcher. This is especially noticeable in *Bacillus subtilis* and some *Pseudomonas* strains, where "wild" isolates often form elaborate and beautiful swarms on the surface of growth media.

TWITCHING

Twitching is another type of motility along moist surfaces, but in this case it is not dependent on the presence of flagella. Instead, it depends on the presence of type 4 pili, which, if you recall from **Lectures 13** and **14**, are homologous to the **type II** protein secretion system and are important adhesins cells use to attach to surfaces. Cells extend pili that adhere to surfaces via proteins at their tips, then retract those pili, pulling the cell along like a grappling hook (Figure 15.3). The jerky movement of cells using this system led to the name.

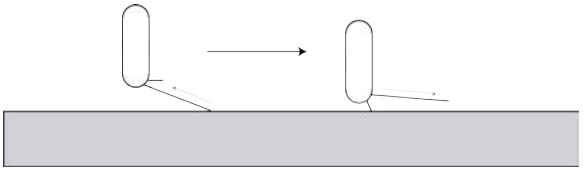


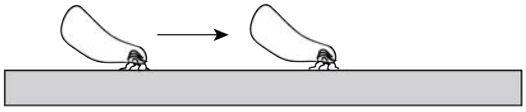
Figure 15.3. Twitching motility in bacteria by extension and retraction of surface-binding type IV pili.

Twitching motility can be performed by individual cells (for example, those of *Pseudomonas aeruginosa*) or by populations (e.g. *Myxococcus xanthus* S- [social] motility, **Lecture 11**). *Pseudomonas aeruginosa* type IV pili are localized to the poles of the cells, so they tend to stand up "on end" when moving via twitching, and move relatively slowly (~0.2 μm per sec).

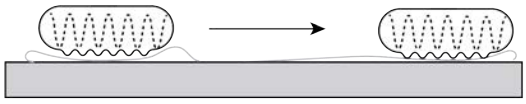
GLIDING

Gliding motility is a catch-all term for bacteria moving smoothly along a surface with no obvious external appendages.

A



B



C

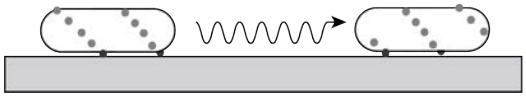


Figure 15.4. Gliding motility in different bacteria. (A) Gliding motility in *Mycoplasma mobile* depends on surface binding by many large proteins that change conformation to pull the cells along a sialic acid-coated surface. (B) Gliding (A- or adventurous-) motility in *Myxococcus xanthus* depends on the production of extracellular polysaccharide slime and a "standing wave" distortion of the cell shape, the physical interactions between which pull the cells forward along the surface. (C) Gliding motility in *Flavobacterium johnsoniae* is driven by T9SS-dependent movement of the adhesin SprB (grey circles) along a helical track on the exterior of the bacterial cell.

Swarming and twitching both require fairly soft, moist surfaces, while gliding bacteria can often move across firmer, drier substrates. Many different kinds of bacteria can do this, and there are at least three radically different mechanisms by which they do so. These are exemplified by gliding motility in the mycoplasmas, in the myxobacteria, and in the *Bacteroidetes* (Figure 15.4).

The gliding motility of mycoplasmas has been best studied in the fish pathogen *Mycoplasma mobile*. *M. mobile* glides at 2 to 4 µm per sec along the surfaces of host cells by repeated binding to, pulling, and release of host sialic acid residues by up to 450 50-nm long "legs" composed of the very large proteins Gli123, Gli349, and Gli521 (the numbers in their names are each protein's mass in kDa). These legs are attached to a cytoskeletal structure called the "jellyfish" (composed of many copies of each of 10 JSP jellyfish-structure proteins) which defines the "front" or "head" of the mycoplasma cell, which as you may remember from **Lecture 10**, have no cell wall and are small, flexible, and blob-like. Gliding proceeds in the direction of the head, as the gliding machinery pulls the cells along the surface. Mycoplasma gliding is powered by ATP hydrolysis catalyzed by the JSP MMOB1670. See this paper for more details and electron microscopy of the "jellyfish" structure.

Myxobacteria, as we discussed in **Lecture 11**, are predatory bacteria with a complex social lifestyle exemplified by the model organism *Myxococcus xanthus*. When moving as a swarm, their S- or social-motility is (as mentioned above) type 4 pili-dependent twitching. However, myxobacteria are also capable of exploring surfaces by gliding as single cells, and this is called A-motility (for **a**dventurous). Both S- and A-motility are required for both efficient predation and myxospore development.

The mechanism of A-motility has been mysterious for some time, but a recent model (described in this paper) has emerged to explain the phenomenon. Genes required for gliding motility were first identified in the late 1970's, but little real progress was made in understanding how A-motility is powered until 2011, when the AglRQS protein complex required for gliding was identified as a proton channel composed of a MotA homolog (AglR) and two MotB homologs (AglQS). This suggested that the energy driving gliding was the proton motive force, and that it was, in some way, descended from or related to the flagellar motor system. Unlike MotA and MotB, however, AglRQS are **not** attached to the cell wall, and in fact, they are able to move freely within the inner membrane (*M. xanthus* is a Gram-negative species).

High-resolution microscopy eventually revealed that the hundreds of individual AglRQS motor complexes in an *M. xanthus* cell are in continuous motion in rotating helices around the length of the cell, with "traffic jam"-like accumulations along the cell's ventral surface. Exactly how this led to gliding motility was obscure, however, until the very recent work cited above showed that, in combination with the production of a thin slime layer underneath the cells, the deformations of cell shape driven by AglRQS create a net forward capillary force or pressure gradient. The exact properties of this force depend on the nature of the surface the cells are on, and it will be fascinating to see if other groups of bacteria take advantage of the same kinds of physical forces for movement along surfaces.

As mentioned briefly in **Lecture 13**, some members of the *Bacteroidetes* phylum are capable of gliding motility via a type IX secretion system-dependent mechanism. This is best studied in *Flavobacterium johnsoniae*, where the T9SS secretes many proteins, including the adhesin SprB. SprB is distributed along a fixed helical track around the outside surface of the cell, and movement of the cell is driven by the movement of SprB along this track. The body of the cell rotates along its long axis as it moves. *F. johnsoniae* glides relatively quickly, at up to 4 µm per sec. The T9SS is also responsible for powering the movement of SprB along the cell surface, and does so by rotating at a constant speed of about 1 Hz, apparently acting analogously to the pinion of a rack and pinion motor (see this paper for more details). The rotation of the T9SS is driven by the proton motive force. Mutants in GldJ, a component of the T9SS, have been isolated that secrete and localize SprB to the cell surface but are not motile, and so may be defective in either rotation or interactions with the "rack" or track component, but more work needs to be done to figure out the detailed mechanism involved. The proteins that make up the track have not yet been identified.

Many filamentous cyanobacteria are able to move along surfaces, and this has been referred to as "gliding motility" in the older literature, but more recent results show that cyanobacterial "gliding" is powered by type 4 pili and is essentially a form of twitching motility, although they do also have to produce a polysaccharide slime to lubricate the surface they move along. This slime-dependence caused some early models to suggest that the force of slime extrusion was pushing the cells along a surface. This hypothesis has been discarded now that new evidence shows that slime extrusion is not sufficient for motility.

KINKING

Spiroplasma species are cell wall-less bacteria closely related to the mycoplasmas (**Lecture 10**). Many *Spiroplasma* species are symbionts of insects, some are pathogens of crustaceans, and others are plant pathogens, causing diseases like corn stunt disease and citrus stubborn disease. Unlike *Mycoplasma* spp., *Spiroplasma* cells are helical, and they are able to swim in liquid media without flagella or other appendages.

The helical shape of *Spiroplasma* is maintained by a cytoskeletal protein called Fib that polymerizes into a helical bundle called the fibril. It associates with the cell membrane in an interaction that requires an MreB homolog (**Lecture 11**) called MreB5. Dynamic changes in the helicity of the fibril create "kinks" or bends in the helical cells which propagate processively along the length of the cell, creating wave-like motions that propel the cells forward. The exact mechanism by which this is accomplished is unknown, although *Spiroplasma* motility is inhibited by chemicals that disrupt the proton motive force (**Lecture 16**).

CHEMOTAXIS: INTRODUCTION

Chemotaxis, strictly speaking, is the ability of an organism to move along a chemical gradient, either towards chemical *attractants* or away from chemical *repellants*. It is often used as a catch-all term for any movement towards or away from stimuli, though, although this isn't quite technically correct. Many bacteria, especially photosynthetic ones, exhibit *phototaxis*, or movement towards light, motile aerobic bacteria are almost universally capable of *aerotaxis*, movement towards oxygen, and a subset of bacteria exhibit *magnetotaxis*, or movement along magnetic field lines (mentioned in **Lecture 11**). All of these kinds of motility regulation have some features in common, especially with regards to how cells regulate their movements, sense changes in their environments, and adapt to new conditions.

Chemotaxis is important for motile bacteria to spread efficiently, acquire nutrients, and avoid predation, both in environmental bacteria and in bacteria that live in animal hosts. For example, in the human gut, enterohemorrhagic *E. coli* are chemotactically attracted to the human hormone norepinephrine, which is thought to allow the bacteria to move rapidly towards host epithelial cells. Many plant pathogens and symbionts are attracted by chemical compounds secreted by the roots of their plant hosts. There are a vast array of chemicals that either attract or repel different bacteria, and each species has its own repertoire of chemical responses.

In this section, we will focus mostly on the molecular mechanism of chemotaxis in *E. coli*, since that is a well-understood and relatively simple model system. Be aware, as always, that other species have different, and in many cases more complex, pathways to accomplish chemotaxis, but the basic principles laid out here apply broadly.

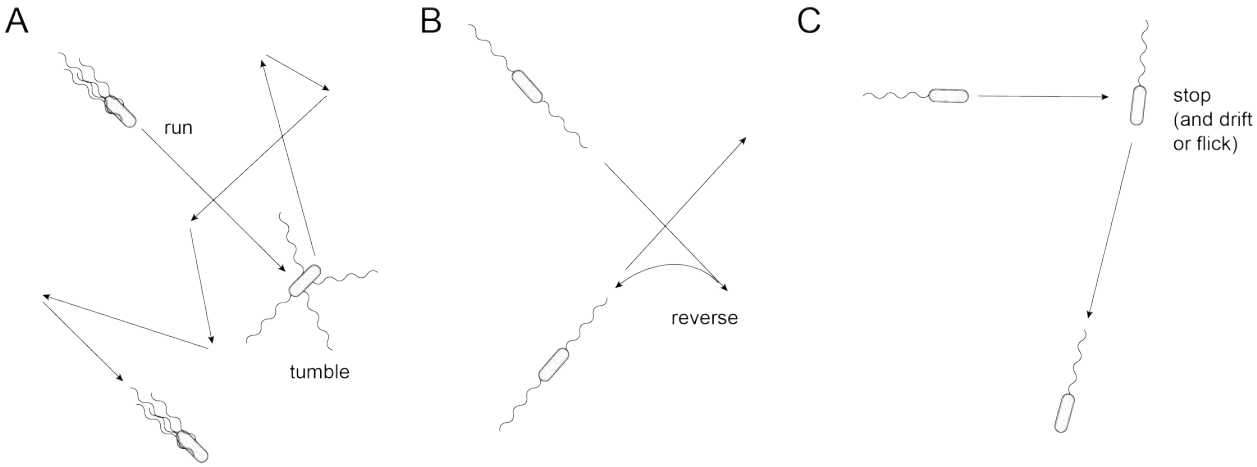


Figure 15.5. Patterns of movement by swimming bacteria. (A) Many bacteria with peritrichous flagella (including *E. coli*) move in more or less linear "runs", interspersed with random "tumbles", distinguished by the direction of rotation of the flagellar motors. (B) Amphitrichous and some monotrichous bacteria move in runs, interspersed with reverses in direction which reorient the bacteria in a random plane. (C) Many monotrichous bacteria intersperse runs with stops, during which they reorient randomly due to Brownian motion. Some also are able to perform a flagellar "flick" to actively switch their direction of travel.

Bacterial movement is characterized by *random walk* patterns, in which the cells alternate periods of linear movement with random reorientations and changes in direction. As shown in Figure 15.5 A for bacteria like *E. coli* with peritrichous flagella, the linear movements are called "runs" and the reorientations are called "tumbles". In the absence of chemotaxis, this results in entirely random movement through three-dimensional space. However, when attractants or repellants are present, bacteria change the proportion of runs and tumbles such that their **average** movement is in a favorable direction. Some bacteria with different arrangements of flagella are not able to actively tumble, and reorient

themselves by reversing direction (Figure 15.5 B) or by simply stopping and allowing Brownian motion to reorient the cells (Figure 15.5 C). Some monotrichous bacteria also perform a "flick" with their flagellum to reorient themselves. On two-dimensional surfaces, the same general idea of interspersing periods of linear movement with random reorientations applies, although constrained, of course, by the surface itself.

HOW CHEMOTACTIC SIGNALS ARE SENSED

But how are these runs and tumbles regulated in response to changes in the environment? Bacteria are, in general, too small to sense chemical gradients over **space** (the length of their "bodies"), so they typically sense changes in their environment over **time**. For *E. coli*, a general rule of thumb is that environmental conditions are constantly being compared to conditions from about 2 seconds ago. If conditions are **more** favorable than they were 2 seconds ago, the length of runs between tumbles increases, biasing the random walk in a favorable direction. If conditions are **less** favorable, the frequency of tumbling increases, to hopefully reorient the cell into moving in a more favorable direction.

In *E. coli*, the default direction of rotation for the flagella is clockwise (CW), and when the flagella are rotating CW, the cells make linear runs. However, the flagellar basal body contains a *motor switch complex*, which includes the FliM protein, that can reverse the direction of rotation to counterclockwise (CCW), resulting in tumbling. The frequency of CCW rotation of the flagella is directly controlled by the phosphorylation state of a response regulator (recall the discussion of two-component regulatory systems in **Lecture 4**) called CheY (Figure 15.6 A). Phosphorylated CheY (CheY-P) interacts directly with FliM, and **the more CheY-P is present in the cell, the more often it will tumble**. The phosphatase CheZ dephosphorylates CheY-P to inactive CheY more or less constitutively, although some data suggests that its activity may also be regulated under some conditions.

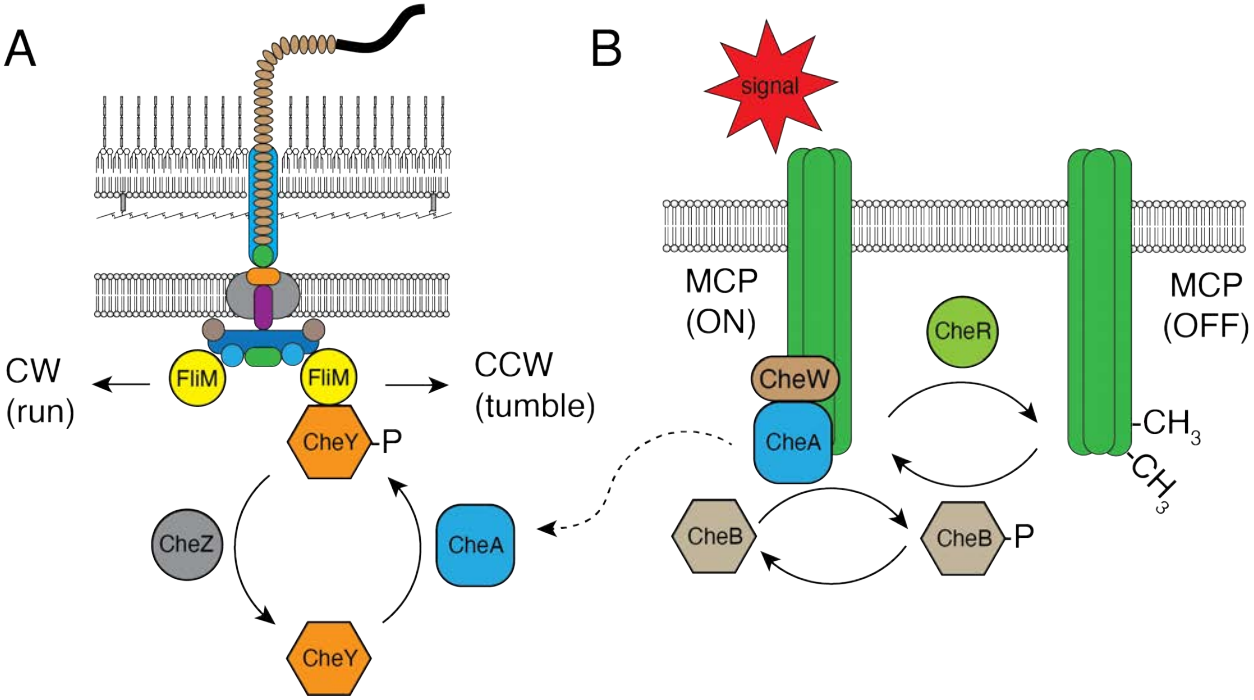


Figure 15.6. Chemotaxis in *E. coli*. (A) Flagella can rotate either clockwise (CW) or counterclockwise (CCW), which bias the cell towards either runs or tumbles, respectively. The switch in direction is regulated by an interaction between the motor switch protein FliM and the response regulator CheY. The histidine kinase CheA phosphorylates CheY and the phosphatase CheZ dephosphorylates CheY-P. Phosphorylated CheY increases the rate of switching to CCW rotation. (B) Methyl-accepting chemotaxis proteins (MCPs) interact with CheA (via the adaptor protein CheW) to drive the phosphorylation of both CheY and the response regulator CheB. Phosphorylated CheB is a demethylase that removes inhibitory methyl groups from MCPs. CheR is a constitutive methyltransferase that continuously methylates MCPs, inactivating them.

The first key question, then, is: How do environmental conditions affect the phosphorylation state of CheY?

The genome of *E. coli* encodes five sensor proteins that indirectly control CheY phosphorylation, and therefore chemotaxis: the **m**ethyl-**a**ccepting **c**hemotaxis (MCP) proteins Tsr, Tar, Trg, and Tap, and the non-methylated MCP homolog Aer. (See the section below for a discussion of the role of MCP methylation in chemotaxis.) Tsr is responsible for sensing the attractants serine, cysteine, alanine, glycine, and warm temperatures (37°C) and the repellants acetate, benzoate, indole, and leucine. Tar is responsible for sensing the attractants aspartate, asparagine, glucose, maltose, and phenol and the repellants nickel and cobalt. Trg senses the attractants ribose and galactose, and Tap is responsible for

attraction towards dipeptides. Aer senses the energy state of the cell, and is responsible for attraction towards terminal electron acceptors (**Lecture 16**), especially oxygen.

The MCPs are integral membrane proteins, and form large clusters at the poles of the bacterial cell. Their interactions with each other are thought to allow amplification of weak signals. Note that different species of bacteria have different repertoires of MCP proteins and therefore are able to sense and respond to different sets of environmental signals. *Azospirillum* sp. B510, a nitrogen-fixing bacterium that is associated with the roots of rice, has 89 different MCPs, for example, and we have very little idea what each of them responds to.

The histidine kinase CheA is responsible for phosphorylating CheY. The activity of CheA is controlled by the MCPs, through an adaptor protein called CheW (Figure 15.6 B). Some signals increase CheA activity, and others decrease it, modulating the total amount of CheY-P present in the cell, and therefore integrating all of the individual chemotactic signals into a single output that controls how frequently the flagella switch to rotating CCW, and therefore the length of runs between tumbles.

ADAPTATION AND MEMORY IN CHEMOTAXIS

The second key question of chemotaxis is this: how do the cells remember what their previous environment was like?

This "memory" or adaptation feature of chemotaxis is important to allow cells to continue moving along gradients across a very wide range of concentrations. The chemotaxis system continuously adjusts back to a baseline state, so that it is sensing **changes** in attractants or repellants, and not just their presence, absence, or a threshold concentration. This is where the methylation of MCPs comes into play.

The constitutively-active methyltransferase CheR adds methyl groups to MCPs, progressively reducing their ability to affect CheA activity (the "OFF" state) (Figure 15.6 B). This is counteracted by the activity of CheB-P, the phosphorylated form of a **second** response regulator that is phosphorylated by CheA. This means that activation of CheA in response to signals has **two** simultaneous effects: increasing tumbling (via production of CheY-P) **and** increasing the sensitivity of the MCP sensors (the "ON" state). CheB-P is rapidly autodephosphorylated, which means that, due to the constitutive activity of CheR, in the absence of activated CheA, the MCPs are constantly adjusted back towards the less-sensitive baseline OFF condition. The rate at which this occurs is responsible for the length of *E. coli*'s "memory", which, as I mentioned above, is about 2 seconds. The net result is that *E. coli* cells can adapt to their current conditions and sense when those conditions change, no matter what the actual concentrations of attractants or repellants might be.

Question for in-class discussion: The O_2 sensor Aer is not methylated. What affect do you expect this to have, and why might this have been selected for?

ACTIN POLYMERIZATION

None of the mechanisms of motility described above function inside of host cells. Several different intracellular pathogens have independently evolved the ability to hijack the eukaryotic cytoskeleton to drive motility within the host cytoplasm. They do this by expressing surface proteins that mimic host actin nucleation factors, causing actin to polymerize around the bacterial cell and pushing themselves rapidly through the host cytoplasm. This is thought to facilitate the spread of pathogens from one host cell to another without exposing themselves to the host immune system. Bacteria that move by actin polymerization move quickly within host cells, but apparently at random, with no known mechanisms of "steering" or chemotaxis.

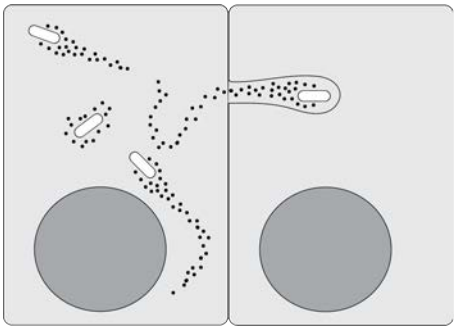


Figure 15.7 Intracellular bacteria moving within and between eukaryotic host cells by actin polymerization. Actin, indicated by black dots, forms comet-like clouds and tails as it polymerizes around and behind bacteria. Bacterial cells that "push" into neighboring host cells are generally able to lyse the layers of host cell membrane with lipases to release themselves into the cytoplasm.

The classic examples of pathogens that move by actin polymerization are *Listeria monocytogenes* and *L. ivanovii*, which cause listeriosis in humans and ruminants, respectively, and *Shigella flexneri*, one of the causative agents of dysentery. However, actin-based motility is also found in *Burkholderia pseudomallei* and *B. mallei,* which cause melioidosis in humans and glanders in horses, respectively, as well as in *B. thailandensis*, a pathogen of fruit flies, in the fish pathogen *Mycobacterium marinum*, and in many species of *Rickettsia*, including those that cause spotted fevers and typhus. Each of these groups expresses different, unrelated proteins that lead to actin polymerization (*e.g.* ActA in *Listeria,* BimA in *Burkholderia*, IcsA in *Shigella*), and indeed, the mechanism by which *M. marinum* stimulates actin polymerization remains unknown.

DISCUSSION PROBLEM SET #31: INTRACELLULAR MOTILITY IN *MYCOBACTERIUM MARINUM*

M. marinum causes skin ulcers and is an economically important pathogen of fish, but unlike other pathogenic mycobacteria, is highly motile inside the host cell cytoplasm. This motility is actin-dependent, and actin tails can be seen forming around motile *M. marinum* cells in appropriately stained host cells, but the genome of *M. marinum* does not encode homologs of any known bacterial or mammalian proteins that stimulate actin polymerization.

The following methods are available for *M. marinum*, which has gained some popularity as a model mycobacterium that grows considerably faster than *M. tuberculosis* and is able to infect zebrafish, amoebas, and cultured human macrophages:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
inducible promoter known	✓
compatible transposon	✓
CRISPR-assisted recombineering	✓

Describe an experiment or series of experiments to identify factors specifically required for driving or regulating actin-dependent motility in *M. marinum*. State:

- a model to explain this phenomenon
- a hypothesis to test that model
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

FLOATING

Several different types of aquatic bacteria are able to adjust their buoyancy by producing *gas vesicles*, allowing them to float upward towards oxygen (for aerobic heterotrophs) or to a particular depth where the light intensity and

wavelength is optimal (for phototrophs). This depth varies among different organisms, and therefore photosynthetic bacteria stratify into layers in the water column.

Gas vesicles are hollow gas-filled shells (or organelles) made up of a few protein components; largely the major protein GvpA (for **g**as **v**esicle **p**rotein) and the structural strengthening protein GvpC. They are typically small, biconical structures that assemble into large, hexagonally-packed "gas vacuoles", which can take up very large proportions of the cytoplasmic volume. Gases simply diffuse into the vesicles from the cytoplasm, so the gas present in a vesicle will depend on the gases present in a cell's environment or produced by its metabolism. The genes for gas vesicle production in cyanobacteria, halophilic archaea, and other types of bacteria are homologous, indicating that they have been spread by extensive horizontal gene transfer during the course of evolution.

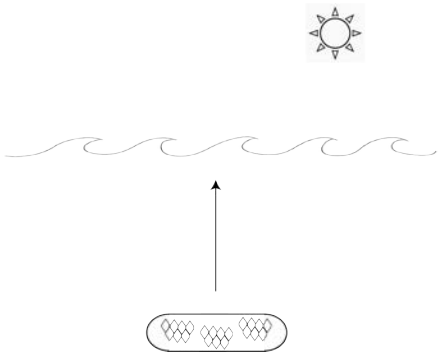


Figure 15.8. Gas vesicles providing buoyancy to an aquatic bacterium, allowing it to move upwards towards oxygen and/or light.

Mikhail Shapiro's lab at Cal Tech has recently developed a system to express gas vesicles in *E. coli* and other species where they are not normally present as an "acoustic" reporter gene to track bacteria inside living human or animal hosts. If enough bacteria are present, gas vesicles can be detected by non-invasive ultrasound, and they've shown that visualizing bacteria this way can give 3-dimensional images with resolution down to 100 µm in mice. Read more here.

SYMBIOTIC MOTILITY INTERACTIONS

Even bacteria that do not produce their own motility machinery can sometimes take advantage of other motile bacteria to move in their environments. There are recent data from a couple of different research groups showing that non-motile *S. aureus* can adhere to and "hitchhike" on motile *E. coli* or *P. aeruginosa* cells (Figure 15.9 A), and that this can contribute to the spread of *S. aureus* during polymicrobial infections.

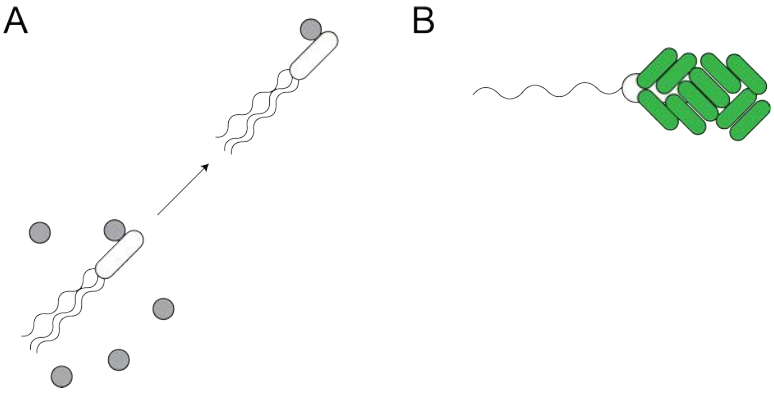


Figure 15.9. (A) Non-motile *S. aureus* (grey circles) can "hitchhike" on motile *E. coli* or *P. aeruginosa* cells. (B) Photosynthetic *C. chlorochromatii* (green) in close symbiotic association with a motile *Candidatus S. mobilis* cell.

In a more complex example, the photosynthetic green sulfur bacterium *Chlorobium chlorochromatii* forms a close symbiotic relationship with the motile heterotroph *Candidatus Symbiobacter mobilis* (a consortium originally called "*Chlorochromatium aggregatum*")(Figure 15.9 B). Between 20 and 70 cells of *C. chlorochromatii* attach to a central *Candidatus S. mobilis* cell, which uses its flagellum to propel the consortium.

DISCUSSION PROBLEM SET #32: PHOTOTAXIS BY A BACTERIAL CONSORTIUM

The "*Chlorobium aggregatum*" consortium as a whole is able to swim towards light, and if it enters a region of darkness, is able to reverse direction back into the light, but, as described above, only *Candidatus Symbiobacter mobilis* has flagella. As you can tell from the prefix "*Candidatus*", the motile partner cannot be cultured in the absence of the photosynthetic partner.

The following methods are available for each species in this consortium (or at least for closely related species):

	C. chlorochromatii	***Candidatus S. mobilis***
growth in pure culture	yes	**no**
can extract DNA/RNA/ protein	yes	yes
complete genome sequence	yes	yes
susceptible to mutagens	yes	yes
can be made competent	yes	**no**
shuttle vectors available	yes	**no**
inducible promoter known	yes (light-activated)	**no**
compatible transposon	yes	**no**

Design an experiment or series of experiments to determine how the two partners in this consortium coordinate phototaxis towards light conditions optimal for *C. chlorochromatii* photosynthesis. State:

- a model to explain this phenomenon
- a hypothesis to test that model
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

LECTURE 16: BACTERIAL ENERGETICS

INTRODUCTION

Living cells must do a substantial amount of work, both physical (*e.g.* flagellar rotation or pilus retraction) and chemical (catalyzing endergonic reactions), and both of these require the input of energy. This chapter is about the forms that energy takes and, to some extent, where that energy comes from. We will discuss the *chemiosmotic theory* and how the *proton motive force* is generated and provides energy for many cellular processes. We will also discuss "*high-energy phosphate bonds*" and how nucleotide phosphates provide the other main source of energy to drive chemical reactions in the cell. Finally, we will cover respiration, fermentation, and photosynthesis, three mechanisms by which bacteria generate and conserve energy.

I suspect much of the core content of this chapter and the next chapter on central metabolism will be review of material you've covered in your undergraduate biochemistry classes. I'll cover the basics, but will try to focus on illustrating aspects of these topics that are specific to prokaryotic systems, which have considerable diversity.

PROTON MOTIVE FORCE

The *chemiosmotic theory* describes how potential energy is stored in biological systems as an electrochemical gradient of ions across a lipid bilayer membrane. Since the ions involved are usually protons (H^+), the energy is also called the ***p****roton* ***m****otive* ***f****orce* (PMF), and is the sum of the force derived from the difference in **concentration** in ions on either side of the membrane and the force derived from the difference in **charge**. Any molecule will tend to diffuse towards a region of lower concentration, and charged molecules will tend to move towards an area of opposite charge.

Lipid bilayer membranes are **not** permeable to charged ions, and the PMF in bacteria is due to cells actively maintaining a higher proton concentration and positive charge **outside** of the cell than **inside**:

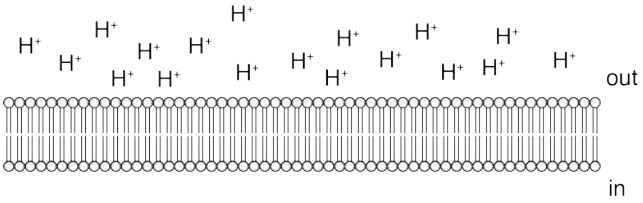


Figure 16.1. The proton motive force is a combination of the force derived from the difference in proton concentration on either side of a membrane and the force from the difference in charge (positive outside, negative inside).

The forces involved can be expressed mathematically, although we will not be worrying about calculating the exact amounts of energy involved in any given system:

$$PMF = \Delta p = \Delta\Psi + \Delta pH$$

The PMF (or Δp) is equal to the difference in charge ($\Delta\Psi$) and the difference in proton concentration (ΔpH). In bacteria living at neutral pH, the ΔpH across the membrane contributes about 70-80% of the PMF, but in bacteria that live in very high or very low pH environments the relative contributions change. In *alkaliphilic bacteria* living at pH 10 or 11, for example, membranes must maintain a very high $\Delta\Psi$ to overcome the negative contribution of ΔpH to the PMF.

Another thing that is good to know is that many bacteria also maintain a **sodium** ion gradient in addition to the PMF. Those Na^+ ions can do all of the same kinds of work that protons can, but require specialized protein components to do so (*i.e.* a transporter that is driven by the PMF cannot derive energy from the Na^+ gradient and *vice versa*).

The potential energy of the PMF is used by bacteria to drive a variety of biological processes, as we will see below, but the first issue to address is how the PMF is generated.

HOW THE PMF IS GENERATED

Bacteria use a variety of mechanisms to generate a PMF, with the relative importance of different mechanisms varying among species. We will discuss some examples here (as illustrated in Figure 16.2). This is by no means an exhaustive list, and we will, in fact, cover some additional pathways that generate PMF later in this chapter. What they all have in common is that protons are removed from the cytoplasm and transported, by one means or another, across the cell

membrane. They use different sources of energy to accomplish this goal. Remember that the PMF is a way of **storing** potential energy, and therefore creating it requires an energy input.

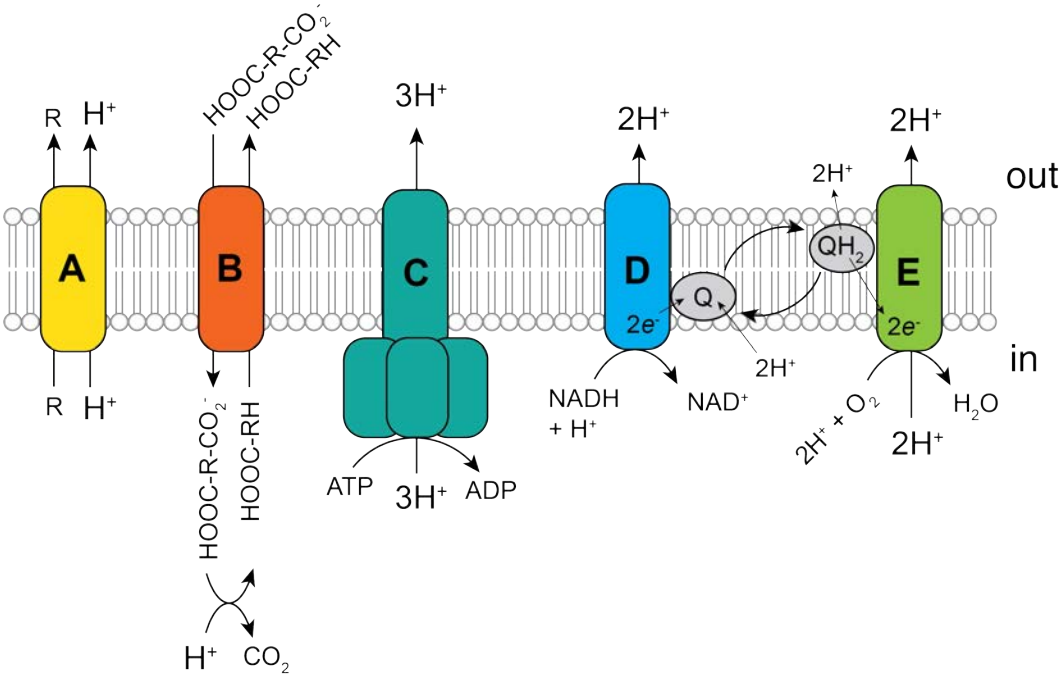


Figure 16.2. Representative mechanisms by which bacteria generate a PMF. (A) Symport of protons with fermentation end products. (B) Coupling decarboxylation of a dicarboxylic acid to antiport. (C) Proton pumping by the F_1F_0 ATPase. (D) Proton pumping and quinone reduction by NADH:quinone oxidoreductase. (E) Proton pumping and reduction of oxygen by cytochrome quinol oxidase.

Figure 16.2A illustrates a very simple mechanism by which some bacteria generate a PMF. If there is a concentration gradient of some other compound (typically a fermentation end product like lactate; see below) with that compound much more concentrated inside the cell than outside, the potential energy of **that** gradient can drive the export of protons through a *symporter* (a transporter that transports two molecules across a membrane in the same direction). This depends, of course, on maintaining the gradient of the other compound. Organisms that use this mechanism to generate a PMF rely on other species in their environment to rapidly degrade that compound. This is a kind of *syntrophy* or metabolic symbiosis.

Figure 16.2B illustrates a less common mechanism for generating a PMF, but it's one that demonstrates an important point. The decarboxylation of a dicarboxylic acid consumes a proton and releases CO_2:

$$HOOC\text{-}R\text{-}CO_2^- + \mathbf{H}^+ \rightarrow HOOC\text{-}R\mathbf{H} + CO_2$$

In combination with an *antiporter* (a transporter that transports two molecules across a membrane in opposite directions) that links import of the dicarboxylic acid to export of the decarboxylated product, the net result is the loss of a proton from the cytoplasmic side of the membrane. Removing protons from the cytoplasm generates PMF even without proton pumping *per se*, since the PMF results from the **difference** in proton concentration on either side of the membrane (and, as mentioned above, the difference in charge).

Figure 16.2C shows a very common mechanism by which bacteria generate a PMF by using the reversible proton-pumping F_1F_0 ATPase (a large protein complex encoded by the *atpCDEFG* operon). This complex can use binding and hydrolysis of ATP (see below) to power the export of 3 protons across the membrane. As we will see shortly, this is a reversible reaction and the same protein complex can use the import of 3 protons to power the synthesis of a molecule of ATP. Which direction the F_1F_0 ATPase runs in a particular organism depends on its metabolic needs at any given time. If it has sufficient ATP but insufficient PMF, then it will run in the proton-pumping direction.

Finally, Figures 16.2D and E illustrate *respiration*, a mechanism by which bacteria link the generation of a PMF to oxidation-reduction reactions by way of an *electron transport chain*. The particular form of respiration illustrated in Figure 16.2 is aerobic respiration, powered by the reduction of oxygen to water, which yields a very large amount of energy. The electron transport chain depends partly on membrane-soluble *electron shuttles* called *quinones* to move electrons and protons between respiratory enzymes. We will discuss respiration in more detail below, but note that

oxidation of one NADH and reduction of one O_2 during aerobic respiration results in direct transport of 6 protons across the membrane and consumption of 2 more in the cytoplasm. Respiration is an extremely potent mechanism for generating PMF and is used by most fast-growing organisms.

HOW THE PMF IS USED

The PMF is used to power a variety of important processes in bacteria. Some examples are shown in Figure 16.3:

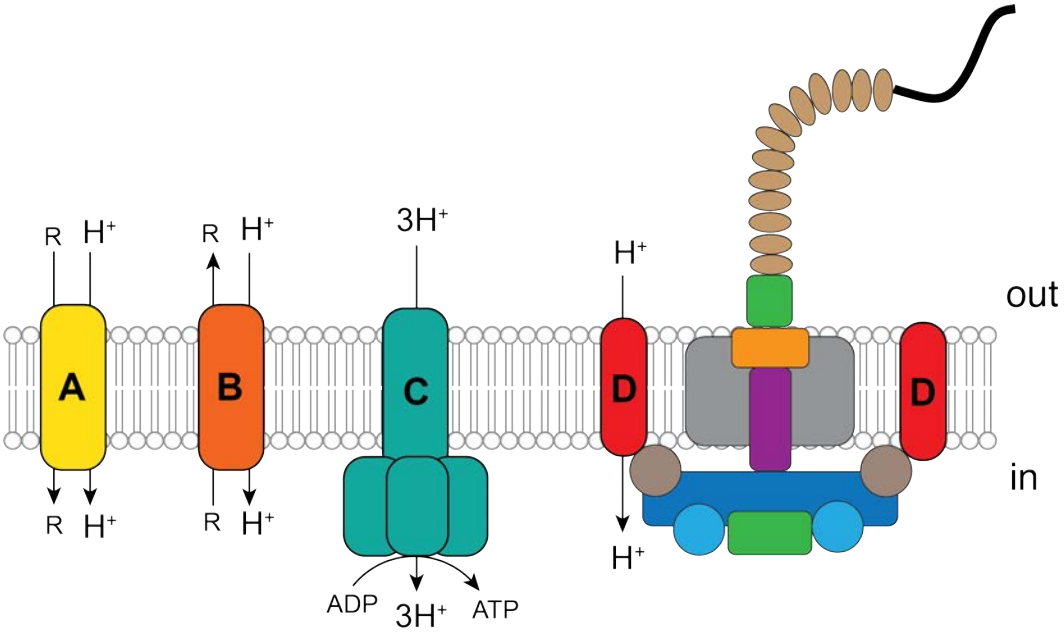


Figure 16.3. Examples of how the PMF is used to power bacterial metabolism. (A) Symporters couple the import of a proton to the simultaneous import of another molecule. (B) Antiporters couple the import of a proton to the simultaneous export of another molecule. (C) F_1F_0 ATP synthase couples the import of 3 protons to the synthesis of ATP. (D) The MotAB motor proteins couple the import of protons to the rotation of flagella.

The transport of small molecules across the cell membrane is often powered by the PMF. As shown in Figure 16.3A and B, H^+ symporters couple the import of a proton to the **import** of another molecule, while H^+ antiporters couple the import of a proton to the **export** of another molecule. In *E. coli*, for example, H^+ symporters are responsible for the import of amino acids, sugars, nucleotides, metals, and a variety of carbon and nitrogen sources (link), while PMF-linked antiporters catalyze the export of various ions, toxins, antibiotics, metabolic byproducts, and metals (link).

As I mentioned above, the F_1F_0 ATPase is reversible, and in fact, its primary function in respiring bacteria is to use the PMF to generate ATP (Figure 16.3C). When operating in this direction, it is often called "ATP synthase". We will discuss how ATP and other nucleotide phosphates are used to power bacterial processes in the next section.

Finally, the MotAB motor proteins use the PMF to power rotation of flagella (**Lecture 15**). This is a major drain on the PMF in motile bacteria, with more than 500 protons needed to power one rotation of a single flagellum in *E. coli*.

See **Lectures 13** and **15** for additional examples of PMF-powered systems in bacteria, and remember that any of these kinds of PMF-powered protein complexes can, in some species, be driven by a Na^+ gradient. Na^+ gradients are generated by different mechanisms in different species, but Na^+ / H^+ antiporters are common, making the generation of the Na^+ gradient dependent on the PMF. There are also examples known of flagellar motors powered by K^+, Ca^{2+}, and Mg^{2+} gradients (in some *Bacillus* and *Paenibacillus* spp.).

Experimentally, the addition of *ionophores* or *uncouplers*, chemicals that bind to cations and allow them to diffuse through membranes, can be used to collapse ion gradients. This can be useful to determine whether the PMF is the driving energy source for a particular process, and since there are ionophores specific for particular cations, they can also be used to determine if H^+, Na^+, or some other cation gradient is involved. **C**arbonyl **c**yanide *m*-**c**hloro**p**henyl hydrazone (CCCP) is a commonly used proton-specific ionophore, and the antibiotic monensin (produced by *Streptomyces cinnamonensis*) is a sodium ionophore. Gramicidin, the first commercially manufactured antibiotic, is an antimicrobial peptide produced by *Brevibacillus brevis* that forms channels in bacterial cell membranes permeable to

both H^+ and Na^+, thereby acting as a general-purpose uncoupler. All of these compounds can be extremely toxic to human cells, so care needs to be taken when using them in the lab.

HIGH-ENERGY PHOSPHATE BONDS

The PMF is a useful source of energy for processes that take place at the cell membrane. The main source of energy in the **cytoplasm** of cells is the hydrolysis of **n**ucleotide **t**ri**p**hosphates (NTPs), usually **a**denosine **t**ri**p**hosphate (or ATP). The same kinds of chemical energy can be derived from other NTPs, and it is fairly common for certain specific enzymes to use GTP instead, but ATP is the most commonly used, and the one we will focus on here.

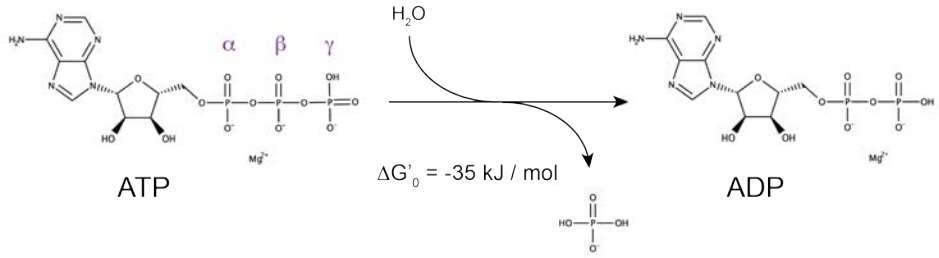


Figure 16.4. The structure of ATP, and its hydrolysis to ADP and P_i. This reaction is exergonic, and release of the γ phosphate results in the release of 35 kJ / mol of free energy.

As shown in Figure 16.4, ATP has three phosphate groups (the α, β, and γ phosphates, in order outward from the ribose sugar of the adenosine nucleoside). The bonds between these phosphate groups are often referred to as "high-energy phosphate bonds", which is a misnomer. The actual chemical bond energy is the same as for any other phosphate bond. What is notable about the phosphate bonds in ATP and other NTPs is their high phosphoryl **group transfer potential**. We will not get into the detailed chemistry here, but essentially this means that hydrolysis of these bonds releases an unusually large amount of free energy (35 kJ / mol). This is largely due to the fact that the negative charges of the sequential phosphate groups repel each other, so that hydrolysis reduces the electrostatic strain within the molecule.

Note also in Figure 16.4 the presence of a Mg^{2+} ion chelated by the phosphates of ATP and ADP. *In vivo*, essentially all NTPs are chelated to magnesium, and *in vitro* all ATP-dependent enzymes require both ATP and Mg^{2+}.

How do enzymes access the free energy of ATP hydrolysis to drive chemical reactions? Some examples of ATP-driven group transfer reactions are shown in Figure 16.5:

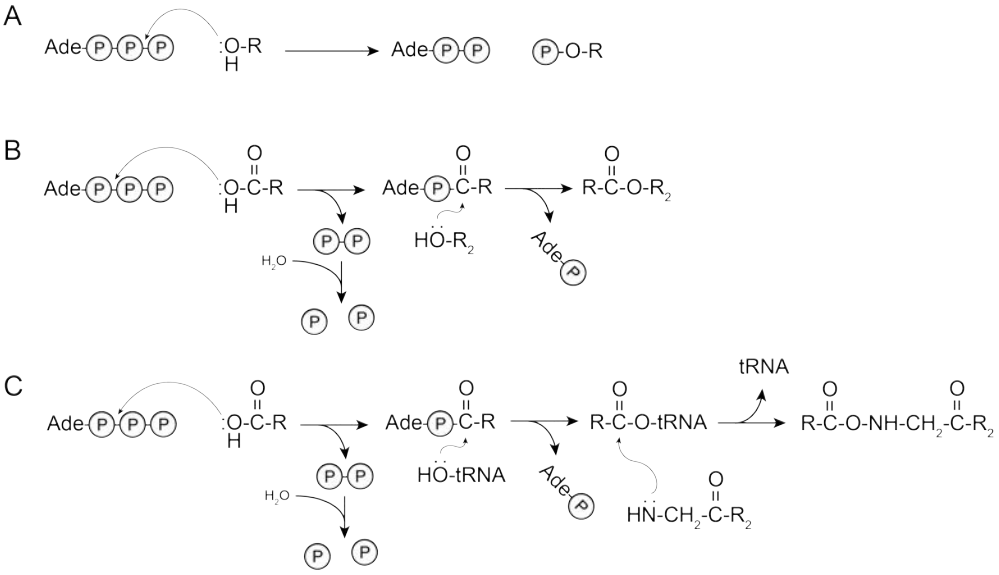


Figure 16.5. Chemical reactions using ATP. (A) Kinases catalyze the group transfer of the γ phosphate of ATP, resulting in ADP and a phosphorylated product. (B) Ester bonds are formed by enzymes that catalyze the attack of a carboxyl group on the α phosphate of ATP, releasing pyrophosphate (PP_i) and forming an AMP derivative. Hydrolysis of PP_i by a pyrophosphatase provides additional free energy to drive the reaction to completion. The AMP is displaced by a hydroxyl group, releasing AMP and the ester product. (C) Formation of a peptide bond, a specific and physiologically important example of the kind of group transfer reaction shown in (B).

Conceptually, the simplest group transfer from ATP is phosphorylation (Figure 16.5A). Enzymes that catalyze phosphorylation reactions are called kinases, and simply direct the attack of a hydroxyl group in their substrate to the γ phosphate of ATP, generating a phosphorylated product and releasing ADP. This is exactly what ATP hydrolysis is, when the R-OH substrate molecule is H_2O (Figure 16.4).

Group transfer reactions can also use ATP to generate other kinds of chemical bonds (Figure 16.5B). In these cases, enzymes catalyze the attack of a carboxyl group on the α phosphate of ATP, releasing pyrophosphate (PP_i) and generating an AMP derivative of the substrate. Hydrolysis of PP_i by a *pyrophosphatase* contributes additional free energy to drive the reaction to completion and make it irreversible. Attack of a hydroxyl, amino, or sulfhydryl group on the AMP-derived product releases AMP and results in an ester, amide, or thioester bond, respectively.

Figure 16.5C shows a very important example of chemical bond formation energized by ATP. The formation of a peptide bond in the active site of the ribosome depends on a series of group transfer reactions, first between ATP and an amino acid, which is transferred to a tRNA, then between the tRNA-derived amino acid and the growing protein chain. Protein synthesis is one of the main consumers of ATP in a growing bacterial cell.

ATP is also important for controlling the activity of many proteins in the cell, even ones that are not enzymes catalyzing biosynthetic reactions. To illustrate, one important example is the ABC (**A**TP-**b**inding **c**assette) transporters, a very widely conserved family of proteins that are responsible for import and export of diverse substrates, including both small molecules and proteins. (Type I protein secretion systems are ABC transporters; **Lecture 13**.)

Figure 16.6 illustrates a generic ABC importer:

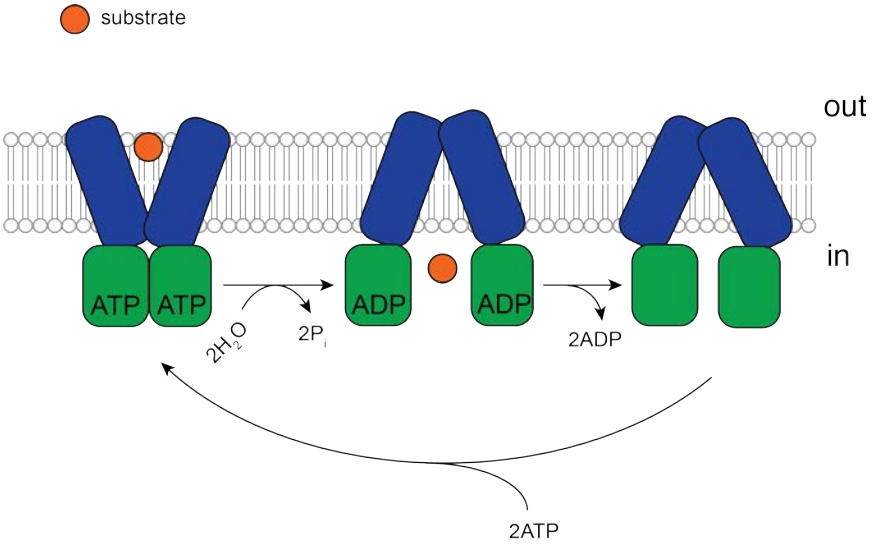


Figure 16.6. The mechanism of an ABC transporter, illustrating how ATP hydrolysis can be linked to protein conformational change.

ABC transport complexes consist minimally of dimers of two proteins: a transmembrane protein and a cytoplasmic nucleotide-binding protein. These form a complex in the membrane which can take two conformations. In one, a cavity in the complex faces **outward** and is open to the periplasm, while in the other, the cavity faces **inward** and is open to the cytoplasm. ATP binding to the nucleotide-binding protein changes the conformation of the transporter complex to face outward, allowing the substrate to enter the cavity. The nucleotide-binding protein then catalyzes hydrolysis of ATP to ADP, and the ADP-bound form of the complex changes conformation to face inward, releasing the substrate into the cytoplasm. The ADP is released and new ATP are bound, allowing the cycle to continue.

E. coli strains have about 70 different ABC transport complexes, responsible for the import of various amino acids, monosaccharides, iron, autoinducer-2, vitamin B_{12}, and other nutrients and for the export of antibiotics and toxic metabolites like putrescine in addition to T1SS-dependent protein secretion. See this review for more information. The repertoire of ABC transporters is obviously species-specific and varies widely between strains, but *E. coli*'s number is not unusual among bacteria. ABC transporters are found in all three domains of life.

Somewhat surprisingly, it's **not** the free energy of ATP hydrolysis that provides the energy for the conformational change in ABC transporters, but rather the energy of ATP **binding**. Hydrolysis simply converts ATP, which binds tightly to the nucleotide-binding protein, to ADP, which binds very weakly. This mechanism is thought to be widely conserved

among proteins that use ATP hydrolysis to control conformational changes, at least among the many that contain homologs of the "Walker A" and "Walker B" ATP-binding motifs found in ABC transporters.

DISCUSSION PROBLEM SET #33: CITRATE TRANSPORT IN *KLEBSIELLA PNEUMONIAE*

Klebsiella pneumoniae is a Gram-negative enterobacterium that is found as a commensal in the human intestine, but which can cause serious disease when it spreads to other parts of the body, which typically occurs in immunocompromised patients. This includes pneumonia (in adults with diabetes or chronic obstructive pulmonary disease, for example) and sepsis (in premature infants).

Unlike its fellow enterobacterium *E. coli*, *K. pneumoniae* is able to take up citrate from its environment for use as a carbon source. This depends on an inner membrane protein called CitS. Small molecule transporters can, as we've discussed above, use a variety of energy sources to drive uptake, including ATP hydrolysis, PMF, or sodium gradients.

The following genetic tools are available for *K. pneumoniae* (surprisingly, the same P1 transducing phage used for *E. coli* also works in *K. pneumoniae*):

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
generalized transducing phage (P1)	√
compatible transposons	√
oligo-directed recombineering	√
CRISPR and related technologies (e.g. CRISPRi)	√
a genome-wide knockout collection (link)	√

Describe an experiment or experiments to determine the source of energy used to drive citrate import in *K. pneumoniae*. State:

- the hypothesis each experiment is testing
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

AEROBIC RESPIRATION

As we have seen, the PMF is a very important energy source for bacteria, and, when working in concert with the F_1F_0 ATP synthase, is a major driver of ATP synthesis, making it a potent energy source for cellular processes. The most

efficient mechanism for generating PMF is to link proton pumping to oxidation-reduction reactions and electron transport chains, by the process of respiration. We introduced this process in Figure 16.2 above, and in this and the following section we will explore bacterial respiration in somewhat more detail.

In electron transport, electrons flow through a series of *electron carriers* down an energy gradient, so that they move from molecules with low electrode potentials to those with relatively higher electrode potentials. This means that electrons move from *electron donors* (which become oxidized) to *electron acceptors* (which become reduced). The electrical current generated can then be linked to generation of a PMF at specific points in the electron transport chain called *coupling sites*.

A very common electron donor for respiration is the reduced nicotinamide cofactor NADH. We will discuss how cells synthesize NADH in **Lecture 17**. The NAD^+ / NADH redox pair has a standard electrode potential of -320 mV. During aerobic respiration, the *terminal electron acceptor* of the electron transport chain is O_2, which is reduced to H_2O, a redox pair with a standard electrode potential of +815 mV. This electron transport chain therefore has a redox potential of (ΔE_h) of 1.135 V, which is energy that can be used to extrude protons from the cytoplasm at coupling sites. The O_2 / H_2O redox pair has a very high electrode potential, and aerobic respiration is therefore an extremely efficient source of energy. We will discuss other terminal electron acceptors in the next section.

Eukaryotic mitochondria (which are descended from bacteria, of course) have a single pathway for aerobic respiration which generates a constitutively very high PMF for ATP synthesis. You are probably familiar with this pathway from your previous classes. Bacteria have additional pathways, and in fact, many of them have **branched** electron transport chains that allow them to adjust the flow of electrons and pump different numbers of protons depending on which coupling sites they choose. Figure 16.7 shows the options possessed by *E. coli*:

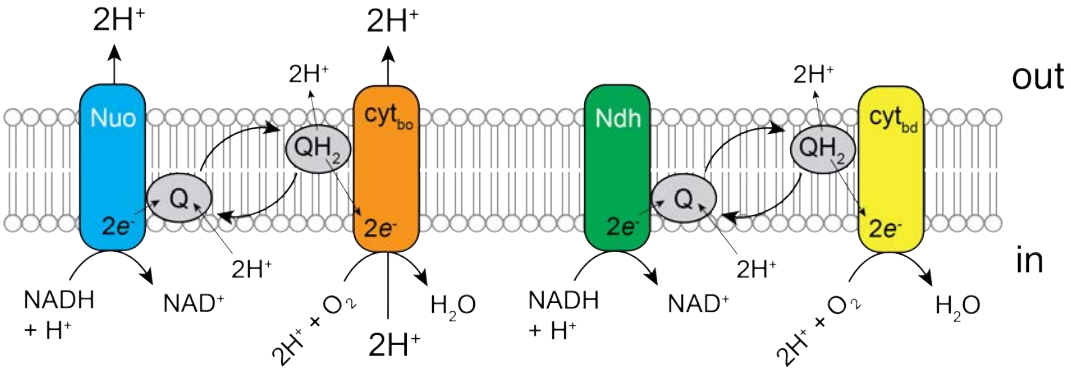


Figure 16.7. Complexes involved in aerobic respiration in *E. coli,* including the NADH:quinone oxidoreductases Nuo (NDH-1) and Ndh (NDH-2) and the cytochrome bo and bd quinol oxidase complexes (encoded by the *cyo* and *cyd* operons, respectively). Ndh and cytochrome bd do not pump protons. Cytochrome bo has a low affinity for O_2, which cytochrome bd has a very high affinity for O_2. Quinones reduced by Nuo or Ndh can be oxidized by either cytochrome bo or bd.

E. coli has NADH:quinone oxidoreductase and cytochrome bo quinol oxidase complexes, encoded by the products of the *nuo* and *cyo* operons, respectively, that are coupling sites and pump protons across the membrane, as discussed above. The "Q loop" of reduced and oxidized quinones is also a coupling site. However, *E. coli* has an additional NADH:quinone oxidoreductase complex, encoded by the *ndh* operon, that is **not** a coupling site and an alternative cytochrome bd quinol oxidase, the product of the *cyd* operon, that is **also** not a coupling site. *E. coli* can mix and match among these complexes depending on its growth conditions, thereby tuning the amount of PMF generated by oxidizing an NADH and reducing an O_2. Cytochrome bo (Cyo) has a low affinity for O_2, and is therefore active at very high O_2 tensions, while the higher-affinity cytochrome bd (Cyd) is active under microaerobic conditions. In general, *E. coli* uses Ndh during aerobic and nitrate respiration, and Nuo during fumarate respiration (see below).

Question for discussion in class: Why would a bacterium ever want to generate less than the maximum possible amount of PMF?

As usual, the pathways present in *E. coli* are not the only pathways found among bacteria more broadly, but the general pattern of oxidation of a reduced cofactor (NADH or $FADH_2$), the passage of electrons through a series of electron carriers (both membrane-soluble quinones and protein-bound flavin, iron-sulfur, and cytochrome cofactors), some of which act as coupling sites to extrude protons to generate PMF, and the final reduction of O_2 to H_2O are conserved among all organisms capable of aerobic respiration.

O_2 and the various ***r**eactive **o**xygen **s**pecies* (ROS) derived from it are potent oxidizing agents that can be quite toxic, but the very high energy yield of aerobic respiration means that it is the energy-generation pathway of choice for many fast-growing bacteria, some of which are *obligate aerobes* that cannot grow without O_2. Aerobic organisms have a variety of mechanisms for resisting and detoxifying ROS, regulated by a broad range of O_2- and ROS-sensing stress response proteins.

ANAEROBIC RESPIRATION

Anaerobic respiration refers to any electron transport pathway in which the terminal electron acceptor is not O_2. There are a vast number of these used by different microbes in different environments, and we will not be able to do more than scratch the surface of the topic here. One key fact, though, is that the standard electrode potentials of all of the *alternative electron acceptors* are lower than that of O_2 / H_2O, so less energy is derived from reduction of those compounds. A very small set of examples of terminal electron acceptors used by some prokaryotes are listed below, with their respective standard electrode potentials:

Fe^{3+} / Fe^{2+}	+771 mV	
NO_3^- (nitrate) / NO_2^- (nitrite)	+421 mV	
fumarate / succinate	+33 mV	
CO_2 / CH_4	-250 mV	(*methanogenesis*, specific to certain archaea)
S^0 / H_2S	-270 mV	

Remember that the energy drawn from respiration is the ΔE_h between the terminal electron acceptor and the NAD^+ / NADH redox pair (standard electrode potential of -320 mV), so you can see from this list that the possible energy yield of anaerobic respiration can vary widely.

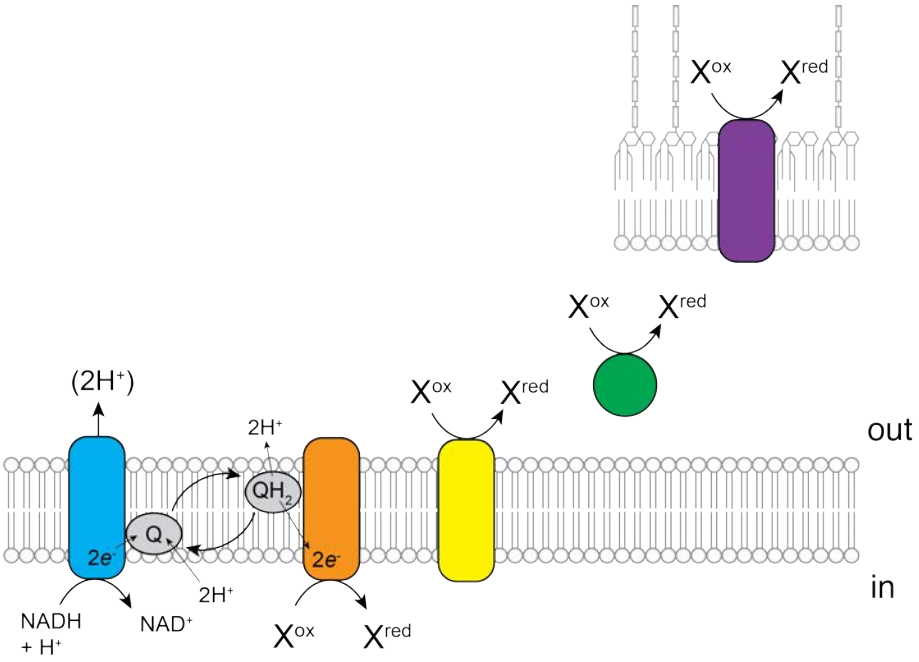


Figure 16.8. Anaerobic respiration, in which oxidation of NADH by a NADH:quinone oxidoreductases (which may or may not be a coupling site) is linked to reduction of any molecule that is not O_2. The respiratory reductases that reduce various terminal electron acceptors used by different bacteria may be in the inner membrane, the periplasm, the outer membrane, or on the surface of the cell.

As shown in Figure 16.8, anaerobic respiration begins with oxidation of NADH by an NADH:quinone oxidoreductase, which may or may not be a proton-pumping coupling site, passes through a quinone intermediate, and ultimately ends up at a reductase that uses the electrons from NADH to reduce the terminal electron acceptor. Depending on the specific terminal electron acceptor, the reductase may be in the inner membrane, the periplasm, the outer membrane, or on the surface of the cell, with additional electron shuttling proteins existing to carry the electrons to their final destination.

Returning once again to *E. coli*, an organism with a conveniently large respiratory repertoire for illustrative purposes, the respiratory nitrate and fumarate reductases are in the inner membrane and carry out their reactions in the cytoplasm. The **t**ri**m**ethyl**a**mine **o**xide (TMAO) and **d**i**m**ethyl**s**ulf**o**xide (DMSO) reductases are also in the inner

membrane, but carry out their reduction reactions in the periplasmic space. Nitrite reductase is a soluble periplasmic protein. None of these proteins is a coupling site, and they do not contribute directly to generating PMF except via the Q loop. Expression of respiratory reductases is typically repressed by the presence of oxygen and stimulated by the presence of their cognate electron acceptors. Nitrate, nitrite, TMAO, and DMSO are all compounds abundant in the mammalian gut, especially during inflammation. Respiration of these compounds gives *E. coli* and related enterobacteria a significant advantage over other bacteria in the intestine during inflammation.

DISCUSSION PROBLEM SET #34: METAL REDUCTION BY *SHEWANELLA ONEIDENSIS*

Shewanella oneidensis is an environmental γ-proteobacterium that is notable for being able to respire using metal oxides as terminal electron acceptors (they can also respire more "normal" compounds like oxygen or sulfate). They are essentially breathing rocks, deriving energy from the reduction of ions like Fe^{3+}, Mn^{4+}, U^{6+}, and Tc^{7+}. The metal oxides containing these ions and/or their reduced products are extremely insoluble solids, which means that *Shewanella* must carry out their terminal electron transfer reactions outside of the cell, using reductases on their cell surfaces, with additional protein complexes needed to move electrons from cytoplasmic NADH across the cell envelope.

When grown anaerobically in the presence of $HAuCl_4$ and $AgNO_3$, *S. oneidensis* uses the gold and silver ions as terminal electron acceptors, producing particles a couple of hundred nanometers in diameter composed of pure Au^0 and Ag^0. These nanoparticles, when purified, are potent antimicrobial and anti-biofilm agents which can be used to kill drug-resistant pathogens.

The following genetic tools are available for *S. oneidensis*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓
oligo-directed recombineering	✓
CRISPR and related technologies (e.g. CRISPRi)	✓
a genome-wide knockout collection (link)	✓

Describe an experiment to identify proteins specifically required for respiration of Au and Ag in *S. oneidensis*. State:

- the hypothesis your experiment is testing
- the independent and dependent variables of that experiment
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

Question for discussion in class: Metal oxide nanoparticles are thought to kill bacteria by a variety of mechanisms, but a primary one is disrupting cell membranes by electrostatic interactions. This is a much more difficult

thing for bacteria to evolve resistance to than antibiotics, which is good for us, but less good for *Shewanella*. What strategies might you hypothesize that *Shewanella* might use to cope with the toxic byproducts of their respiration?

FERMENTATION

Not all bacteria are able to respire, and even many of those that are can often grow in the absence of available terminal electron acceptors. They do so by the process of *fermentation*.

The fundamental problem faced by fermenting organisms is recycling the cell's limited pool of nicotinamide, since, as we will see in **Lecture 17**, central metabolism generates a large amount of reduced NADH. Respiring organisms put that NADH to work generating PMF, but without respiration, NADH and other *reducing equivalents* (like phosphorylated NADPH) must be oxidized by alternative pathways. It is a little tricky to separate the discussion of fermentation from central metabolism, but I will do my best, and some aspects of this will hopefully become more clear after tomorrow's chapter. There are a very wide range of fermentation strategies used by different bacteria, and I will only describe a few of the most common here, but they all are ways to achieve the same goal of having some place to put the electrons that would be fed into an electron transport chain in a respiring organism.

The lactic acid bacteria (*Lactobacillus, Streptococcus, Lactococcus, Leuconostoc, Enterococcus, etc.*) are very abundant and common Gram-positive bacteria that survive almost entirely by fermentation. As their name suggests, they produce abundant lactic acid (lactate) as a fermentation byproduct, by one of two pathways: *homolactic fermentation*, which produces only lactate, and *heterolactic fermentation*, which produces lactate, ethanol, and CO_2. Homolactic fermentation is very simple, and serves to illustrate the principle function of fermentation pretty clearly:

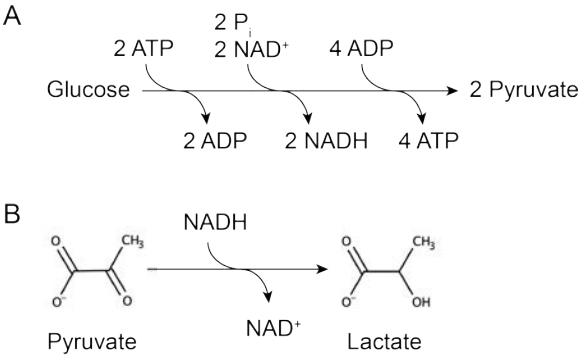


Figure 16.9. Homolactic fermentation. (A) Glycolysis consumes a glucose molecule, generating net products of 2 ATP, 2 NADH, and 2 pyruvate. (B) Lactate dehydrogenase catalyzes the reduction of pyruvate to lactate, with the concomitant oxidation of NADH.

The *substrate-level phosphorylation* steps in glycolysis are the primary source of ATP in homofermentative bacteria, and we will return to them in the next chapter. As shown in Figure 16.9A, glycolysis converts glucose to pyruvate, generating 2 ATP and 2 NADH in the process. In order to regenerate the NAD$^+$ needed for more cycles of glycolysis, NADH is used to reduce pyruvate to lactate (Figure 16.9B), which is secreted from the cell, rapidly acidifying the growth media of bacteria using this pathway.

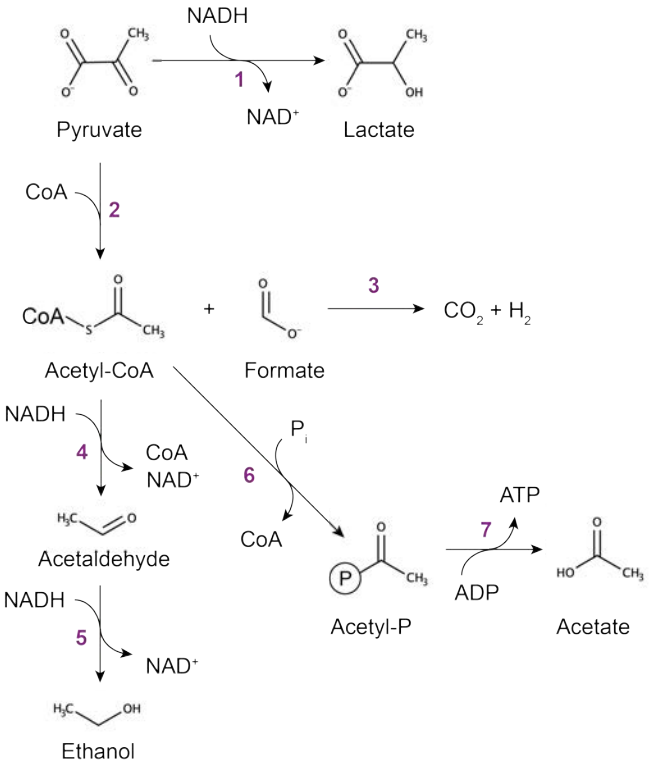


Figure 16.10. Mixed acid fermentation. Pyruvate is reduced to lactate by lactate dehydrogenase (1) or converted to acetyl-CoA and formate by pyruvate-formate lyase (2). Formate is degraded to CO_2 and H_2 gases by formate-hydrogen lyase (3). Acetyl-CoA can either be reduced to acetaldehyde by acetaldehyde dehydrogenase (4) and then to ethanol by alcohol dehydrogenase (5) or converted to acetyl-P by phosphotransacetylase (6). Acetyl-P is used as a phosphate donor for ATP synthesis by acetate kinase (7).

Bacteria that are not able to tolerate the high acidity generated by homolactic fermentation may be able to use a different, more flexible fermentative pathway called *mixed acid fermentation* that reduces pyruvate to a mixture of lactate, formate, acetate, ethanol, CO_2, and H_2 (Figure 16.10). This pathway is found in enterobacteria, including *E. coli*, *Salmonella*, and *Shigella*.

Mixed acid fermenters have lactate dehydrogenase, like homolactic fermenters, but can also use an enzyme complex called pyruvate-formate lyase and a sulfur-containing cofactor called *cofactor A* (just called "CoA", usually) to divide pyruvate into acetyl-CoA and formate. Some mixed acid fermenters use formate-hydrogen lyase to convert formate to CO_2 and H_2 gases, but others just excrete formate. Acetyl-CoA can then follow one of two pathways. It can be reduced in two steps to acetaldehyde and ethanol, consuming an NADH at each stage. It can also be converted into acetyl-phosphate (acetyl-P) by phosphotransacetylase. Acetate kinase can use acetyl-P to generate an ATP and acetate, but does not consume any NADH in the process. Acetyl-CoA is a key intermediate in metabolism, and we will be seeing it again in **Lecture 17**.

For each glucose produced by glycolysis, 2 NADH are produced and must be oxidized. Mixed acid fermenters can balance the different branches of this pathway to produce different ratios of products, including the possibility of some additional ATP that homolactic fermenters cannot produce and a useful acetyl-CoA intermediate. Exactly what products are produced depends on the growth conditions of the organism. For example, secreted acetate and lactate acidify the media, while ethanol does not. However, all of the products of mixed acid fermentation except the gases are toxic to cells at high concentrations.

Butyrate and other ***s**hort **c**hain **f**atty **a**cids* (SCFA) produced by bacterial fermentation in the large intestine are a major carbon source for intestinal epithelial cells and are an important factor in the health of the gut. SCFA have systemic health effects on the host, influencing the gut-brain axis, sleep cycles, epithelial barrier functions, and lowering inflammation. Generally speaking, if your microbiome is fermenting a lot of complex carbohydrates to butyrate in your large intestine, you will be healthier. The most abundant butyrate producing bacteria in the human gut are extremely oxygen-sensitive Gram-positive obligate anaerobes related to the *Clostridia*, including members of the genera *Faecalibacterium, Eubacterium*, and *Roseburia*.

Butyrate fermentation is illustrated in Figure 16.11. Two pyruvates, 2 ATP, and 2 NADH are produced from a glucose by glycolysis in these organisms, as in the fermentations we discussed above.

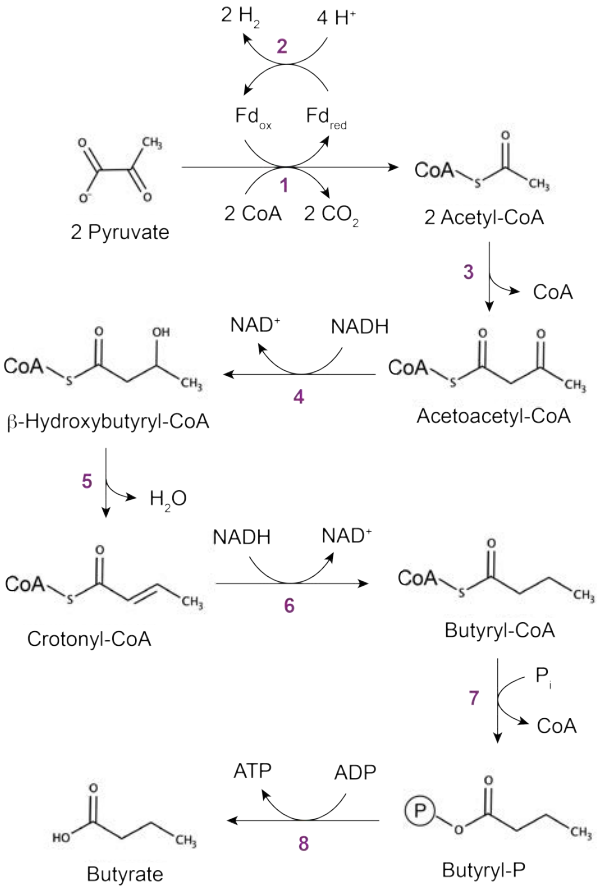


Figure 16.11. Butyrate fermentation. Pyruvate is oxidatively decarboxylated by pyruvate-ferredoxin oxidoreductase (1) to acetyl-CoA and CO_2. Oxidized ferredoxin is reduced by a hydrogenase (2), which produces H_2 gas. Two molecules of acetyl-CoA are condensed into acetoacetyl-CoA by acetyl-CoA acetyltransferase (3), which is then reduced by β-hydroxybutyryl-CoA dehydrogenase (4) to crotonyl-CoA, consuming an NADH. Crotonyl-CoA dehydrogenase (6) consumes another NADH and produces butyryl-CoA. Phosphotransbutyrylase (7) displaces CoA with phosphate, forming butyryl-phosphate, which butyrate kinase (8) uses as a phosphate donor to produce ATP and butyrate.

The two pyruvates are decarboxylated to acetyl-CoA, then condensed to a 4-carbon acetoacetyl-CoA product. Acetoacetyl-CoA is reduced twice, consuming an NADH each time, to yield butyryl-CoA, which, in a direct parallel to the acetyl-CoA to acetyl-phosphate to ATP pathway we saw in mixed acid fermentation, is used to generate an ATP. The resulting butyrate is excreted from the cell, along with 2 CO_2 and 2 H_2 produced in the initial decarboxylation step, which uses a protein called ferredoxin as an electron and hydrogen carrier molecule. Note how the fermentation is balanced, with no NADH left over, and yielding a total of 3 ATP from each molecule of glucose.

I do need to point out that not all fermentation pathways start with glycolysis and its end product pyruvate. For example, heterolactic fermentation begins with the conversion of glucose to xylulose-5-phosphate via the pentose phosphate pathway (**Lecture 17**), and many *Clostridium* species ferment lactate to acetate or propionate.

Since fermentation does not result in proton pumping, fermenting bacteria need to use alternative mechanisms to generate a PMF. The most common is to simply run the F_1F_0 ATP synthase in reverse (Figure 16.2C), although if the organism is living in an environment with another species that very rapidly consumes a secreted fermentation end product, a proton symporter can be used to generate PMF (Figure 16.2A). Obviously, organisms growing by fermentation are much more energy-limited than respiring strains, both in terms of PMF and ATP, and therefore typically grow more slowly and to lower final cell density on a given nutrient source.

PHOTOSYNTHESIS

Most of you are probably most interested in disease-causing bacteria (given the expertise of the UAB Microbiology Department), but I do want to devote some space here to *photosynthesis*, the process by which **light** is used as an

energy source for generating a PMF or reducing equivalents (in this case, usually NADPH rather than NADH). No pathogenic bacteria I'm aware of are photosynthetic, but many environmental bacteria are, and sunlight is, of course, the major source of energy input into the biosphere of Earth. Roughly 50% of the oxygen we breathe is produced by marine photosynthetic cyanobacteria, the same general kind of bacteria which are the ancestors of plant chloroplasts (source of the other 50%). The biochemistry of photosynthesis is very complicated, and I will only touch on the essentials from a few major groups of prokaryotes here.

The simplest photosynthetic pathway depends on a single transmembrane protein called *bacteriorhodopsin* that absorbs light and uses the energy of that light to pump protons across the membrane (Figure 16.12).

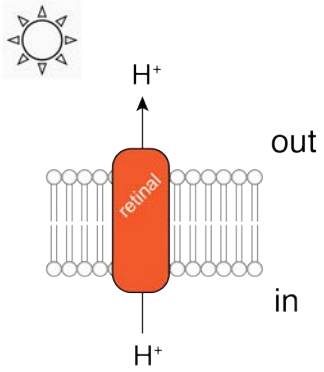


Figure 16.12. Light-driven proton pumping by bacteriorhodopsin, powered by light absorption by the covalently-bound retinal cofactor.

Bacteriorhodopsin was first identified in extremely *halophilic* (salt-loving) archaea belonging to the confusingly-named genus *Halobacterium*, but we now know that homologous systems are found in some bacteria, including the halophilic Bacteroides clade member *Salinibacter ruber* and the cyanobacterium *Gloeobacter violaceus*. Each bacteriorhodopsin protein contains a covalently bound retinal (vitamin A) cofactor. When retinal absorbs a photon, it undergoes a conformational change which drives a change in the structure of the protein, resulting in proton pumping. This is a very direct method of linking light energy to PMF generation, but most photosynthetic organisms use considerably more complex pathways that draw more energy out of each captured photon.

Diverse types of bacteria are able to carry out various forms of *anoxygenic photosynthesis*, which is a general term for photosynthetic pathways that use molecules other than H_2O as an electron donor. These include the well-studied purple non-sulfur bacteria, which are α- and β-proteobacteria belonging to genera like *Rhodospirillum, Rhodobacter, Rhodopseudomonas,* or *Rhodocyclus* ("rhodon" is Greek for "rose", and these bacteria are generally red to purple in color), that use H_2 or organic molecules as electron donors. The purple sulfur bacteria are γ-proteobacteria of the order Chromatiales that use sulfur-containing molecules (largely H_2S and S^0, but also thiosulfate [$S_2O_3^{2-}$]) as electron donors. The green sulfur bacteria are distantly related to the Bacteroidetes and mainly use H_2S as an electron donor. The heliobacteria, including the genera *Heliobacterium, Heliophilum,* and *Heliobacillus*, are spore-forming anaerobic Firmicutes with a photosynthetic pathway similar to, but somewhat simpler than that found in green sulfur bacteria.

To illustrate some of the properties of anoxygenic photosynthesis, we will look at a simplified version of how it works in purple non-sulfur bacteria (Figure 16.13A, B). In these bacteria, the photosynthetic *reaction center* (Figure 16.13A) is a membrane protein complex that contains a bacteriochlorophyll cofactor, which is able to absorb light. When this happens, it powers electron transfer from an organic donor molecule (such as succinate or malate) to a quinone. Very similarly to what we saw earlier when discussing respiration (Figure 16.7), the resulting reduced quinone is oxidized by a cytochrome-containing, proton-pumping membrane protein (Figure 16.13B), resulting in the generation of a PMF. Unlike during respiration, there is cyclic electron flow in anoxygenic photosynthesis, with a soluble cytochrome c protein serving as an electron carrier to return the electrons to the reaction center. The photosynthetic reaction center of purple non-sulfur bacteria is homologous to photosystem II (PSII) of plants and cyanobacteria.

Purple non-sulfur bacteria are typically facultative phototrophs, and can grow well in the dark by respiration, but when they are growing under photosynthetic conditions (anaerobically with light), they produce folded internal membrane structures derived from the inner cell membrane to increase the amount of surface area they have available for photosynthetic reaction centers. This allows them to capture as much light as possible.

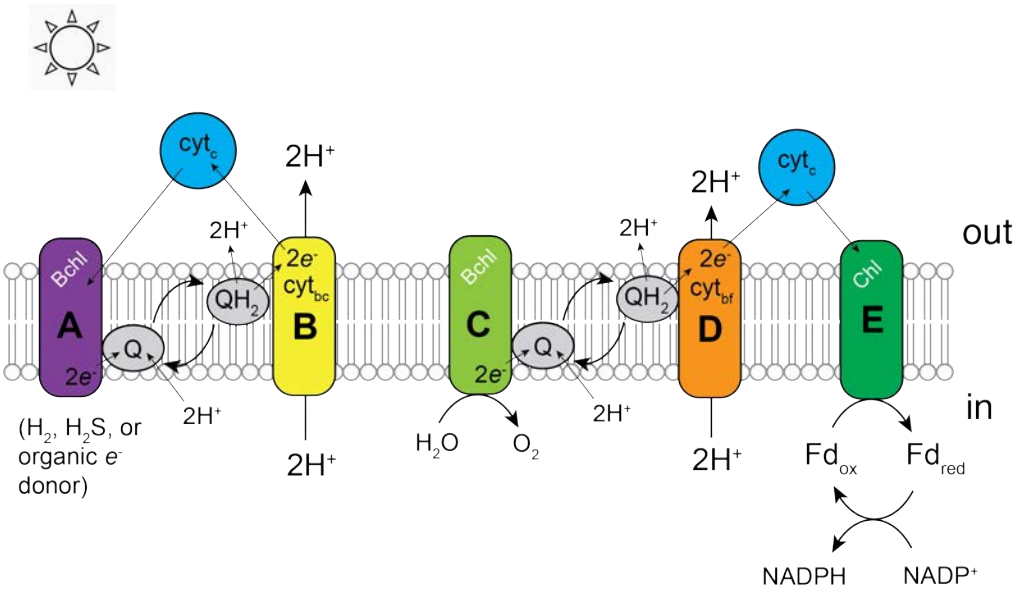


Figure 16.13. Bacterial photosynthesis. In anoxygenic photosynthetic bacteria (A, B), electrons donated from molecules other than water are used to set up a cyclic electron flow, powered by light absorption by the bacteriochlorophyll (Bchl) cofactor of photosystem II (A). This is linked via the quinone pool to a proton-pumping cytochrome bc protein (B). Electrons are returned to (A) by a soluble cytochrome c protein. In oxygenic photosynthesis (C-D), electrons are donated from water, producing O_2, powered by light absorption by the bacteriochlorophyll (Bchl) cofactor of photosystem II (C). This is linked via the quinone pool to a membrane-bound proton-pumping cytochrome bf protein (D) and a soluble cytochrome c protein, with the electrons ultimately being used to reduce the protein ferredoxin in a reaction powered by light absorption by the chlorophyll (Chl) cofactor of photosystem I (E). Reduced ferredoxin is oxidized to produce NADPH.

Oxygenic photosynthesis, on the other hand, is carried out by cyanobacteria and plants, and involves the use of H_2O as an electron donor, producing O_2. This produces lots of energy, but requires a way of dealing with oxygen toxicity, which they usually do at least partly by being capable of aerobic respiration.

The first step is absorption of light by a bacteriochlorophyll-containing photosystem II complex (Figure 16.13C), which draws electrons from water and passes them, via a quinone intermediate, to a cytochrome bf proton pump (Figure 16.13D), generating PMF. So far, this is very similar to what we've seen before. However, from there, the electrons are transferred via a cytochrome c protein to a chlorophyll-containing photosystem I (PSI) complex which, with the absorption of another photon, uses them to reduce ferredoxin. The resulting reduced ferredoxin is used to reduce $NADP^+$ to NADPH. Oxygenic photosynthesis thereby directly generates both PMF and additional reducing equivalents for biosynthetic or other pathways (**Lecture 17**). This is the same mechanism by which plants carry out photosynthesis.

Like chloroplasts and, indeed, most phototrophs, cyanobacteria contain complex internal membrane structures to increase the surface area for photosynthetic reaction centers. These are called *thylakoid membrane*s in cyanobacteria and chloroplasts.

As I said earlier, this is a **very** brief introduction to the world of photosynthesis. There are many more cofactors, light-harvesting molecules, and proteins involved, and many variations on the pathways. Green sulfur bacteria basically have only photosystem I, for example. Hopefully what I've presented here will, at least, give you a place to start if you ever find yourself studying a photosynthetic organism.

Note that, while in many photosynthetic organisms, photosynthesis is linked to *autotrophy* (the ability to incorporate or *fix* CO_2 into organic molecules), this is by no means universally true. Most plants and many cyanobacteria are obligate *photoautotroph*s, but a lot of photosynthetic bacteria are *photoheterotrophs* that obtain carbon from the breakdown of organic molecules, and there are lots of autotrophs that use non-photosynthetic pathways and energy sources to fix CO_2. As fascinating as these pathways are, due to space and time limitations, we will **not** be discussing any of them during this class.

LECTURE 17: CENTRAL METABOLISM

INTRODUCTION

Central metabolism refers to the *catabolic* biochemical pathways by which organisms break down molecules into their component parts and the *anabolic* pathways by which the components of new and growing cells are synthesized. We will explore some of the pathways by which organic molecules are degraded and link these to the energy-generating mechanisms we discussed in the previous chapter. We will also discuss representative pathways by which biochemical building blocks like amino acids and nucleotides are synthesized from their constituent parts, with an attempt to illustrate some of the general principles common to such pathways. The goal is **not** to memorize pathways, but rather to give you a sense of the **patterns** that unify metabolic processes.

Many definitions of "central metabolism" focus entirely on carbon (C) metabolism, but for the sake of branching out a bit and getting away from material you've probably covered several times in prior classes, I'm going to include discussions of some of the pathways of nitrogen (N) and sulfur (S) metabolism as well.

CATABOLISM

When discussing metabolism, there are a few key pathways that we really can't avoid, including *glycolysis*, the *pentose phosphate pathway*, and the ***t**ri**c**arboxylic **a**cid* (TCA) *cycle*. You have almost certainly encountered these before, since they are conserved in eukaryotes and *E. coli*, but let's take a closer look, starting with glycolysis (Figure 17.1).

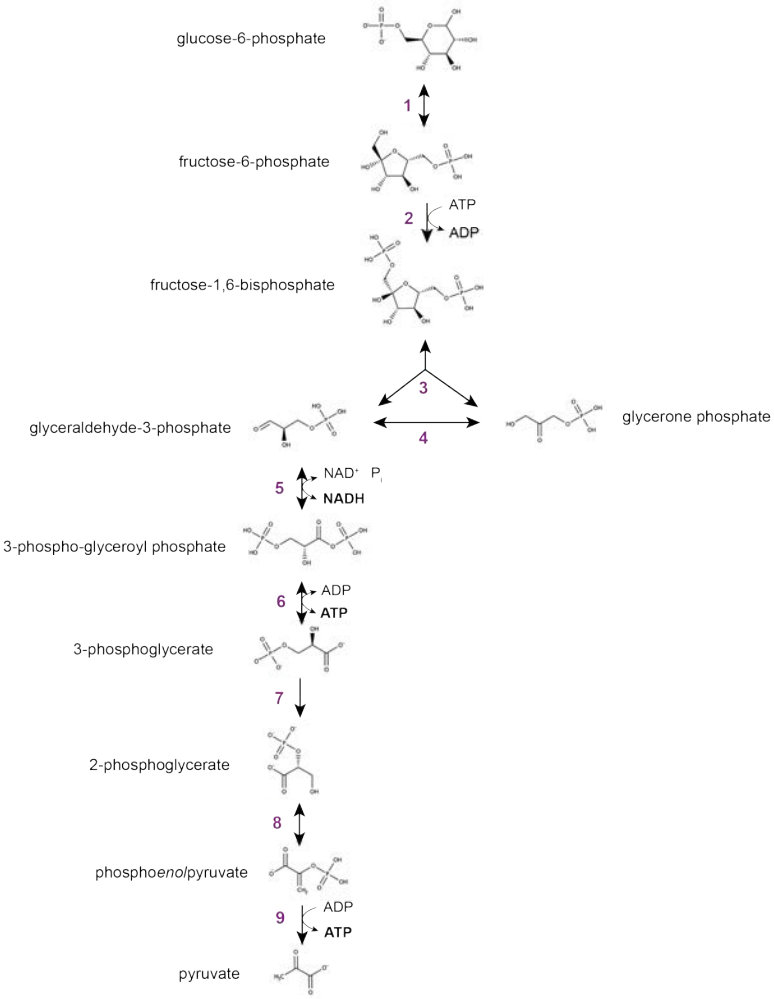


Figure 17.1. Glycolysis. Enzymes: 1, glucose-6-phosphate isomerase (Pgi); 2, 6-phosphofructokinase (Pfk); 3, fructose-bisphosphate aldolase (Fba); 4, triose-phosphate isomerase (Tpi; 5, glyceraldehyde-3-phosphate dehydrogenase (GapDH); 6, phosphoglycerate kinase (Pgk); 7, phosphoglycerate mutase (Gpm); 8, enolase (Eno); 9, pyruvate kinase (Pyk).

Starting with **g**lucose-**6**-**p**hosphate (G6P), glycolysis is catalyzed by a series of enzymes that result in the production of 2 molecules of pyruvate, along with a net 3 ATP and 2 NADH. As we will see below, one of those ATP equivalents is needed for the phosphorylation of glucose to G6P, so the actual energy yield per glucose is 2 ATP. We have already discussed how fermenting organisms recycle the NADH and dispose of the pyruvate generated by glycolysis (**Lecture 16**).

A key enzyme in glycolysis that deserves specific attention is **g**lycer**a**ldehyde-3-**p**hosphate **d**e**h**ydrogenase (GapDH), step 5 in Figure 17.1. Note that this enzyme catalyzes the phosphorylation of **g**lyceraldehyde-**3**-**p**hosphate (G3P) **without** the involvement of an ATP, which is what allows glycolysis to be a net energy-generating process. It is also important to notice that most of the enzymes catalyzing steps in glycolysis are reversible and operate close to equilibrium (**Lecture 4**), so that flux through the pathway is controlled by the activity of just a few enzymes (2, 7, and 9 in Figure 17.1). The active site of GapDH contains an exceptionally oxidation-sensitive cysteine residue, so this enzyme is prone to inactivation during oxidative stress (such as exposure to hydrogen peroxide).

So far, nothing terribly new and exotic, but bacteria have much more metabolic diversity than eukaryotes and glycolysis is not the **only** way to break down glucose to pyruvate. *P. aeruginosa* and many other bacteria use a somewhat different pathway, called the *Entner-Douderoff pathway* (Figure 17.2.).

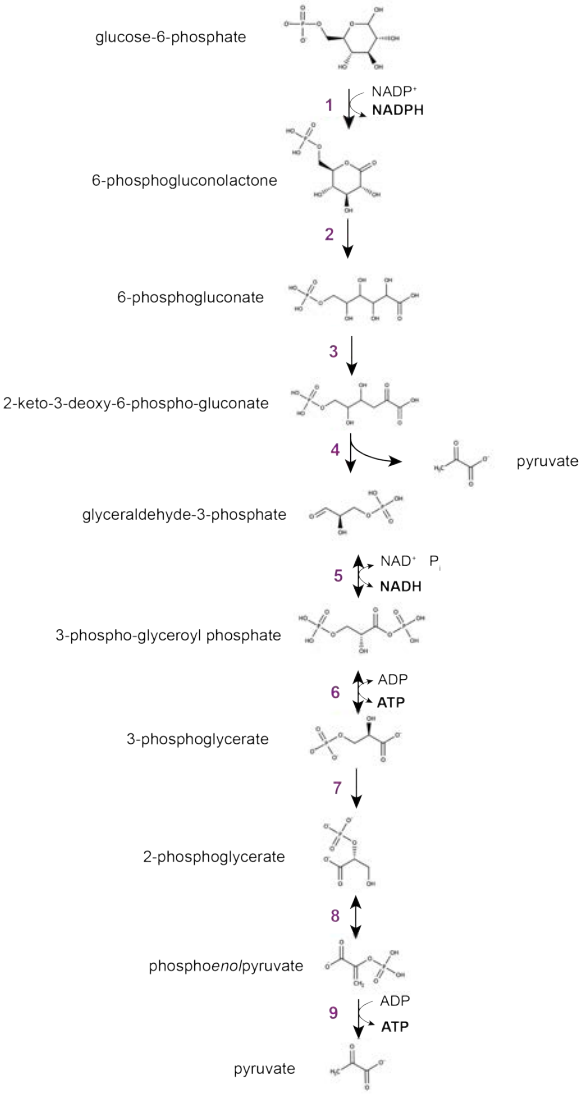


Figure 17.2. Entner-Douderoff pathway. Enzymes: 1, glucose-6-phosphate dehydrogenase (Zwf); 2, 6-phosphogluconolactonase (Pgl), 3, phosphogluconate dehydratase (Edd); 4, 2-keto-3-deoxygluconate-6-phosphate aldolase (Eda); 5, glyceraldehyde-3-phosphate dehydrogenase (GapDH); 6, phosphoglycerate kinase (Pgk); 7, phosphoglycerate mutase (Gpm); 8, enolase (Eno); 9, pyruvate kinase (Pyk).

The Entner-Douderoff pathway converts G6P via four irreversible enzyme steps to one molecule of G3P and one of pyruvate, generating a reduced NADPH in the process. The G3P is oxidized to pyruvate by the same enzymes as in glycolysis, generating 2 ATP and an NADH in the process. The Entner-Douderoff pathway therefore generates **less** ATP than glycolysis, but **more** reducing equivalents for the same amount of pyruvate produced. For a fast-growing, obligately respiring organism like *P. aeruginosa*, this is preferable, since they can generate more ATP using the respiratory electron transport chain than they could by substrate-level phosphorylation (**Lecture 16**).

The Entner-Douderoff pathway combines enzymes from glycolysis and another extremely well-conserved pathway you have almost certainly encountered before: the **p**entose **p**hosphate **p**athway (PPP, Figure 17.3). The PPP is another pathway by which G6P is oxidized to G3P, but in a more complicated way that generates a large amount of NADPH and passes through several intermediates which are precursors of essential metabolites (see below).

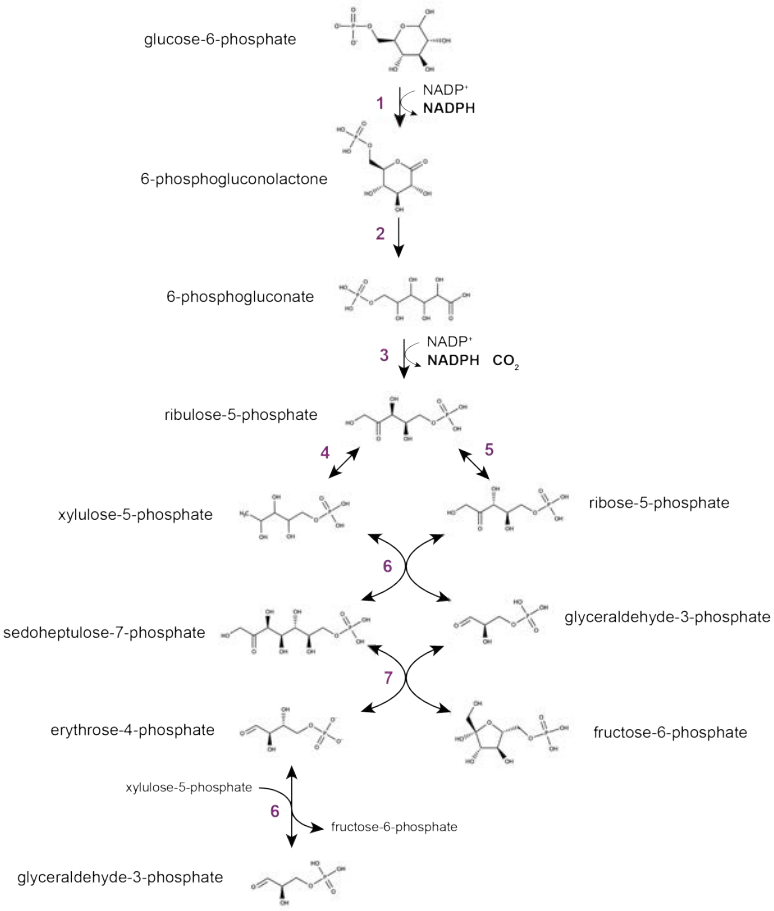


Figure 17.3. The pentose phosphate pathway. Enzymes: 1, glucose-6-phosphate dehydrogenase (Zwf); 2, 6-phosphogluconolactonase (Pgl), 3, phosphogluconate dehydratase (Edd); 4, ribulose-phosphate 3-epimerase (Rpe); 5, ribose-5-phosphate isomerase (Rpi); 6, transketolase (Tkt); 7, transaldolase (Tal).

Enzymes 1, 2, and 3 in Figure 17.3 are the same as the first three enzymes of the Entner-Douderoff pathway (Figure 17.2), with the fate of 6-phosphogluconate as the branch point that distinguishes one pathway from the other.

The net reaction balance of the PPP is:

3 glucose-6-phosphate + 6 $NADP^+$ → 2 fructose-6-phosphate + glyceraldehyde-3-phosphate + 6 NADPH

The **f**ructose-**6**-**p**hosphate (F6P) and G3P are utilized by glycolysis as we've seen above, at least in *E. coli*, so that ultimately this will also yield 5 NADH, 8 ATP, and 5 pyruvate from 3 G6P, which is slightly less of each than the yield of glycolysis alone, with one C lost as CO_2. Note once again how most of the enzymes of this pathway are reversible.

This is all very well for bacteria growing on glucose, of course, but bacteria can use lots of different sugars as carbon sources. Indeed, some bacteria do not have transporters capable of taking up glucose at all. How are these other sugars catabolized? The simple answer is that they are usually broken down into intermediates in one of the above

pathways. We will look at a couple of examples to illustrate, remembering of course that we can't even begin to scratch the surface of the number of pathways that exist (Figure 17.4).

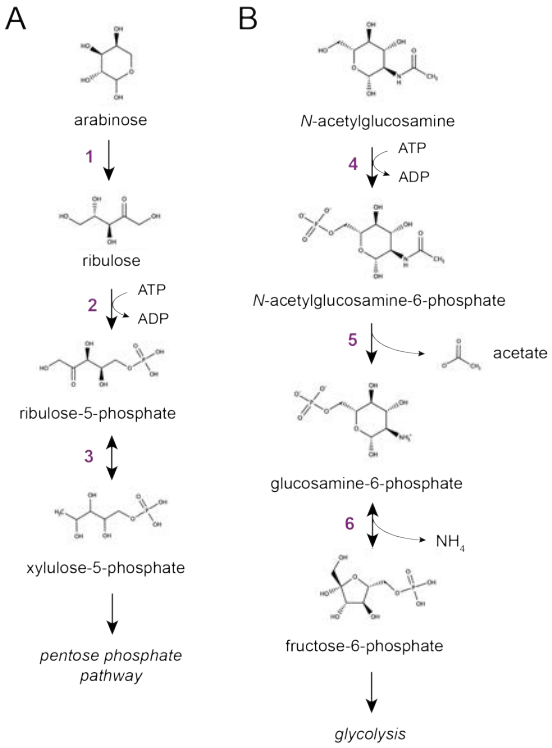


Figure 17.4. Sample monosaccharide utilization pathways. (A) Arabinose utilization. Enzymes: 1, arabinose isomerase; 2, ribulokinase; 3, ribulose-5-phosphate epimerase. (B) *N*-acetylglucosamine (GlcNAc) utilization. Enzymes: 4, *N*-acetylglucosamine kinase; 5, *N*-acetylglucosamine-6-phosphate deacetylase; 6, glucosamine-6-phosphate deaminase.

The 5C sugar arabinose is converted to ribulose, phosphorylated, and fed into the PPP as xylulose-5-phosphate (Figure 17.4A). The peptidoglycan component *N*-acetylglucosamine (GlcNAc, **Lectures 10** and **14**) is phosphorylated, then deacetylated and deaminated, yielding acetate, ammonia, and F6P, which can be degraded by glycolysis (Figure 17.4B).

Polysaccharides are degraded into mono- or disaccharides (by secreted enzymes, as a rule, **Lecture 13**) before uptake into bacteria, and disaccharides are cleaved into their constituent monosaccharides as the first step in their catabolism. LacZ (β-galactosidase), for example, cleaves the disaccharide lactose into glucose and galactose monosaccharides, which are then degraded individually.

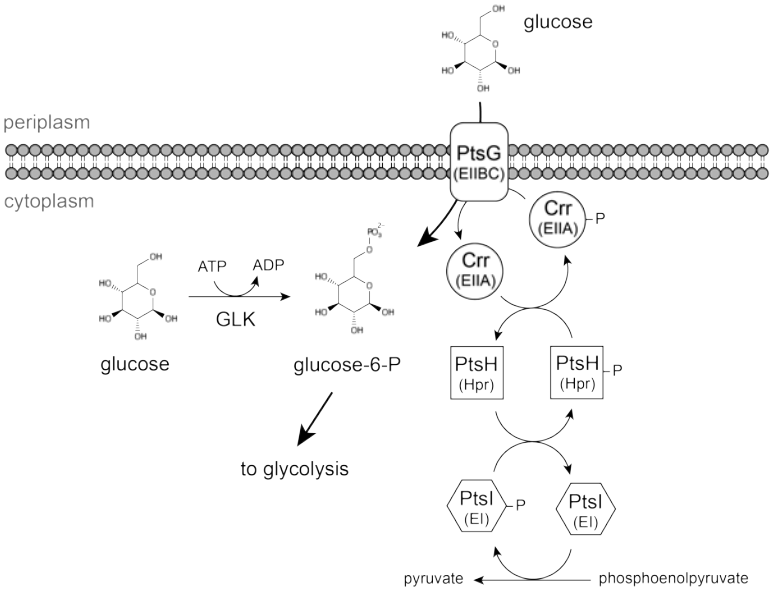
DISCUSSION PROBLEM SET #35: PTS AND NON-PTS SUGAR PHOSPHORYLATION

As we've seen in Figure 17.1 – 17.4, the intermediates of the central metabolic pathways are nearly all phosphorylated. There are a number of reasons for this, including not only the need to derive ATP from substrate-level phosphorylation, but also to make sure that the intermediates stay **inside** the cell. Charged molecules cannot easily cross lipid bilayers, so phosphorylated intermediates are retained in the cytoplasm.

Phosphorylation of sugars after uptake into the cytoplasm serves a second function in maintaining a concentration gradient of the un-phosphorylated sugar with higher concentrations outside the cell than inside, which provides an energetic boost for the transport of those sugars into the cell (**Lecture 16**).

Let's take, as an example, the incorporation and phosphorylation of glucose. Many bacteria contain an enzyme called *glucokinase* (GLK), which uses an ATP to phosphorylate glucose to glucose-6-phosphate, the first intermediate in glycolysis, Entner-Douderoff, and the PPP. This is a very straightforward way to solve this problem, and is typical of the catabolism of many sugars in many species (see Figure 17.4 for a couple of other sugar kinases with analogous roles). However, it is not the **only** way that sugars can be phosphorylated, and indeed, in many bacteria some sugars are phosphorylated **during** transport, by systems called ***p****hospho****t****ransferase* ***s****ystems* (PTS).

The glucose PTS of *E. coli* is illustrated below:



For PTS sugars, the source of the phosphate group is PEP rather than ATP. This phosphate is transferred to the substrate by a series of carrier proteins. The E1 (PtsI) and Hpr (PtsH) proteins are shared among all PTS systems, while the EIIA, EIIB, and EIIC proteins are specific to particular sugars (mono- or disaccharides). The glucose-specific transporters in *E. coli* are, as indicated, called Crr (EIIA) and PtsG (a fusion of EIIB and EIIC), respectively. The EIIC component is an integral membrane protein, and the substrate is phosphorylated as it passes through into the cytoplasm. The EIIA, EIIB, and EIIC components can be separate or can be fused into 1 or 2 combined polypeptides.

PTS systems are common in bacteria, but different species have different repertoires of PTS and non-PTS sugars. *E. coli* has 21 EIIC homologs and *B. subtilis* has 16, although some of these are cryptic, and we don't know what their substrates are.

Xylose is a common 5-carbon sugar, ubiquitous in plants, that serves as a good carbon source for many bacteria. While studying the fermentation of traditional Greek olives, you find that the potential probiotic *Lactobacillus pentosus* depends on xylose for growth in that environment. In the lab, it can also grow on arabinose, which is likely to be fermented by a similar pathway (**Lecture 16**), although you expect there to be separate transporters for xylose and arabinose.

The following methods are available for *L. pentosus* (or at least for closely related species):

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
compatible transposons	√

Describe a set of experiments to determine whether xylose is phosphorylated by a PTS or a non-PTS system in *L. pentosus*, and to identify the gene(s) encoding the xylose transporter in this species. State:

- the hypothesis tested by each experiment

- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

Non-sugar carbon sources with more than 2 carbon atoms per molecule are generally catabolized using a similar strategy, by converting them into intermediates of the glycolytic pathways. Two pathways for glycerol (3C) assimilation, using different reductants and phosphodonors but resulting in the same product (glycerone phosphate), are shown in Figure 17.5 as an example.

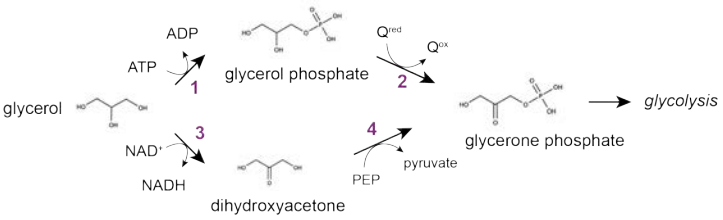


Figure 17.5. Glycerol utilization pathways found in bacteria. Enzymes: 1, glycerol kinase; 2, glycerol-3-phosphate dehydrogenase; 3, glycerol dehydrogenase; 4, dihydroxyacetone kinase.

As we will see shortly, assimilating molecules with fewer than 3 carbon atoms as carbon sources requires somewhat different strategies.

The last highly-conserved central pathway that we need to address is another one found in eukaryotes and *E. coli*, but which is by no means universal among bacteria: the **t**ri**c**arboxylic **a**cid (TCA) or Krebs cycle (Figure 17.6):

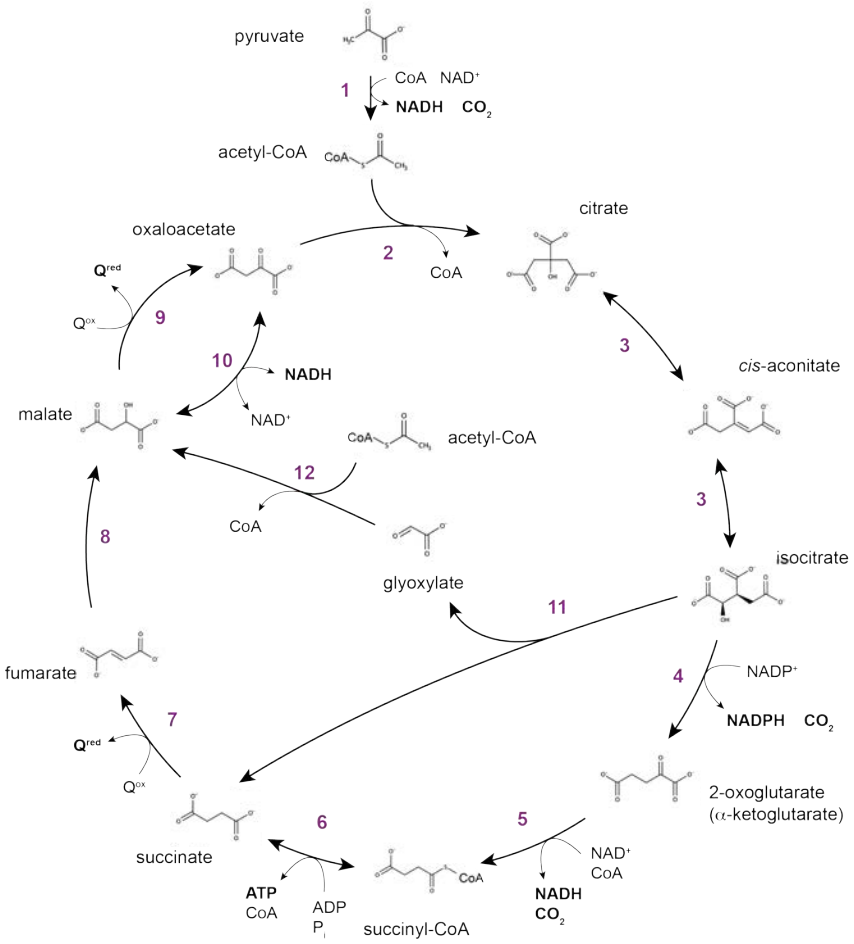


Figure 17.6. The TCA cycle and glyoxylate bypass. Enzymes: 1, pyruvate dehydrogenase (Pdh); 2, citrate synthase (GltA); 3, aconitase (Acn); 4, isocitrate dehydrogenase (Icd); 5, 2-oxoglutarate dehydrogenase; 6, succinyl-CoA synthetase (SucCD); 7, succinate:quinone oxidoreductase (Sdh); 8, fumarase (Fum); 9, malate:quinone oxidoreductase (Mqo); 10, malate dehydrogenase (Mdh); 11, isocitrate lyase (AceA); 12, malate synthase (AceB).

The TCA cycle oxidizes pyruvate to CO_2, generating a little bit of ATP and a large number of *reducing equivalents* (reduced NADH, NADPH, and quinones). In respiring bacteria, those reducing equivalents are converted into PMF using the electron transport chain (**Lecture 16**). Bacteria growing by fermentation may carry out some TCA cycle reactions to generate biosynthetic intermediates (see below), but typically have little flux through the cycle as a whole, and many obligate fermenters lack the TCA cycle entirely.

The first reaction in Figure 17.6, which deserves some additional attention, is decarboxylation of pyruvate, the end product of glycolysis and the Entner-Douderoff pathways, to form acetyl-CoA. In *E. coli,* this is done by the pyruvate dehydrogenase protein complex, which is essential under aerobic conditions. Anaerobically and in many fermentative organisms, pyruvate-formate lyase (**Lecture 16**, Figure 16.10) converts pyruvate to acetyl-CoA, releasing formate instead of CO_2.

A close look at the TCA cycle reveals that there is no net C assimilation from this pathway. A 2C compound (acetyl-CoA) enters the cycle by being condensed with a 4C compound (oxaloacetate) to form a 6C compound (citrate). In the course of the cycle, 2C are lost as CO_2. This is problematic for organisms trying to use the TCA cycle to grow on carbon sources containing only 2C, such as acetate, ethanol, or ethanolamine.

We will look at growth on acetate to illustrate how bacteria can resolve this problem.

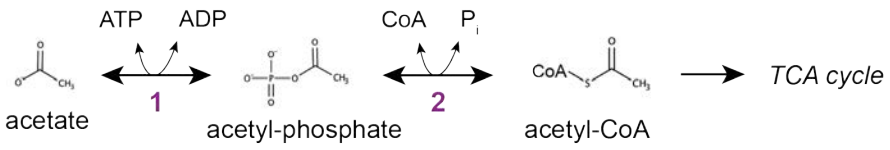


Figure 17.7. Acetate assimilation. Enzymes: 1, acetate kinase; 2, phosphotransacetylase.

The first steps of acetate assimilation, in which acetate is converted to acetyl-CoA, are shown in Figure 17.7.

Longer-chain fatty acids are **also** broken down into acetyl-CoA for catabolism (by β-oxidation, a pathway we do not have time to discuss), so the pathways necessary for growth on acetate are also necessary for growth on lipid carbon sources. (Note that the reactions of this pathway, like so many we've examined in this chapter, are reversible, meaning that acetyl-CoA synthesized by other pathways can, in fact, be used to generate ATP. This happens during mixed acid fermentation, for example; Figure 16.10.)

When growing on carbon sources that yield acetyl-CoA as a breakdown product, *E. coli* and many other bacteria **bypass** the decarboxylation steps of the TCA cycle (reactions 5 and 6 in Figure 17.6), instead using isocitrate lyase (reaction 11 in Figure 17.6) to split isocitrate into succinate (4C) and glyoxylate (2C). Glyoxylate is condensed with another acetyl-CoA to produce a 4C TCA cycle intermediate (malate). This pathway is called the *glyoxylate bypass* or *glyoxylate shunt*, and, as you can see from Figure 17.6, reduces the energy yield of the cycle considerably (by 2 reducing equivalents and 1 ATP).

Not every bacterium that can grow on acetate has the enzymes of the glyoxylate bypass. So, how do **they** solve the problem of carbon loss?

One mechanism is the ethylmalonyl-CoA mutase pathway, found in the purple nonsulfur photosynthetic α-proteobacterium *Rhodobacter sphaeroides*, a species which lacks isocitrate lyase (Figure 17.8). The ethylmalonyl-CoA mutase pathway consumes 3 acetyl-CoA and 2 CO_2 to generate the TCA cycle intermediates malate and succinate, using 11 enzymes to achieve roughly the same outcome as the two enzymes of the glyoxylate bypass, but with the added advantages of generating an ATP and actually **fixing** CO_2 into bioavailable carbon. The ethylmalonyl-CoA pathway consumes an NADPH to do so, and requires the energy-intensive synthesis of the complex cobalt-containing tetrapyrrole cofactor B_{12} (required by enzymes 4 and 9) but as photosynthetic and respiring bacteria, *R. sphaeroides* are not generally limited for energy or reducing power (**Lecture 16**).

The contrast between acetate assimilation in enterobacteria and *R. sphaeroides* is a good example of the diversity of bacterial metabolism. There is no one way to accomplish any particular metabolic goal, and a pathway being conserved in both eukaryotes and *E. coli* is no indication that that pathway is universally conserved among bacteria as a whole.

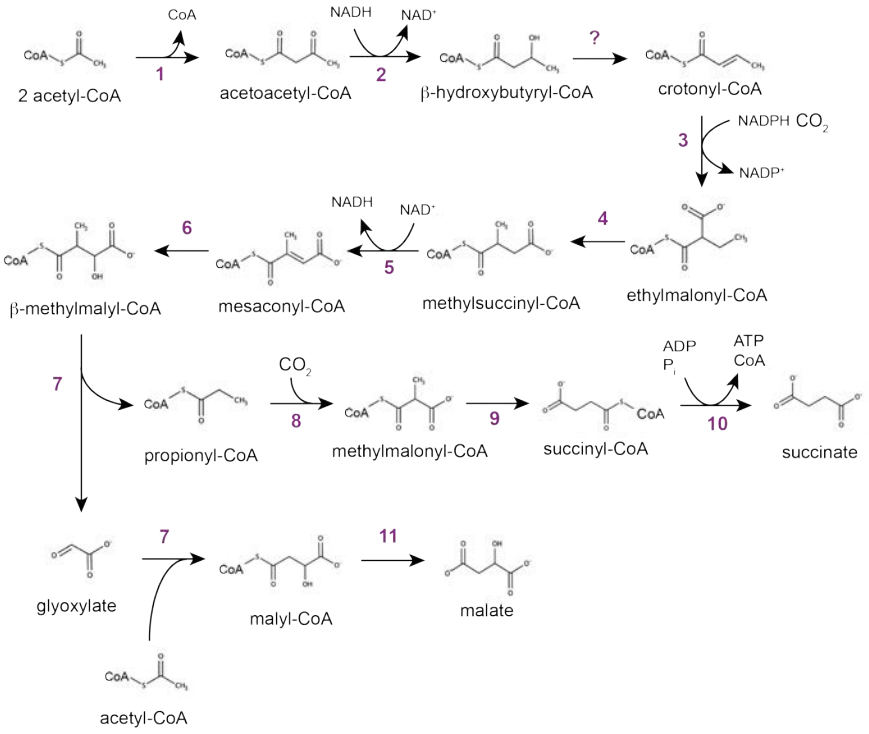


Figure 17.8. Acetate assimilation in *R. sphaeroides via* the ethylmalonyl-CoA mutase pathway. Enzymes: 1, β-ketothiolase; 2, acetoacetyl-CoA reductase; 3, crotonyl-CoA carboxylase/reductase; 4, ethylmalonyl-CoA mutase; 5, methylsuccinyl-CoA dehydrogenase; 6, mesaconyl-CoA hydratase; 7,

β-methylmalyl-CoA / malyl-CoA lyase; 8, propionyl-CoA carboxylase; 9, methylmalonyl-CoA mutase; 10, succinyl-CoA synthetase; 11, malyl-CoA thioesterase. The enzyme that catalyzes conversion of β-hydroxylbutyryl-CoA into crotonyl-CoA in *R. sphaeroides* has not been identified.

Another slightly problematic category of carbon sources are 4-carbon TCA cycle intermediates (succinate, malate, and fumarate). These are preferred carbon sources for many organisms, but since they enter directly into the TCA cycle, bacteria growing on 4C TCA cycle intermediates (or compounds that are broken down **into** 4C TCA cycle intermediates) require a mechanism to counteract the buildup of oxaloacetate (Figure 17.6). Most commonly, this is accomplished by the activity of phospho*enol*pyruvate carboxykinase, which catalyzes the following reaction:

oxaloacetate + ATP → phospho*enol*pyruvate + CO_2 + ADP

PEP can be converted into acetyl-CoA as in glycolysis (Figures 17.1, 17.6), allowing the TCA cycle to continue, or used to generate biosynthetic intermediates, which we will discuss in the second part of this chapter on anabolism. It is also, as noted above, an important phosphodonor in its own right for certain kinase reactions.

Not all carbon sources are sugars or intermediates in central metabolic pathways. Many bacteria can use amino acids as carbon, nitrogen, and sulfur sources, and in fact, in most rich media, these are the primary carbon sources available. **L**ysogeny **B**roth (LB, sometimes incorrectly called "Luria-Bertani medium"), a common rich medium for *E. coli* and other fast-growing bacteria, contains essentially no carbohydrates, for example. There's little point in plowing through all of the possible amino acid catabolic pathways, but I'll present a couple of representatives to illustrate the basic principles, which should not, at this point, be too surprising.

The pattern of carbon sources being broken down into intermediates of one of the central metabolic pathways continues: glutamate, for example (Figure 17.9A), is first converted to aspartate, which is then deaminated to release an NH_4 (ammonia) and generate fumarate, which is catabolized *via* the TCA cycle. Cysteine (Figure 17.9B) is converted into an unstable 2-aminoacrylate intermediate, releasing H_2S (hydrogen sulfide). 2-aminoacrylate spontaneously deaminates, releasing NH_4 and pyruvate, which can then be catabolized by any of the pyruvate-consuming pathways we've discussed, including the TCA cycle.

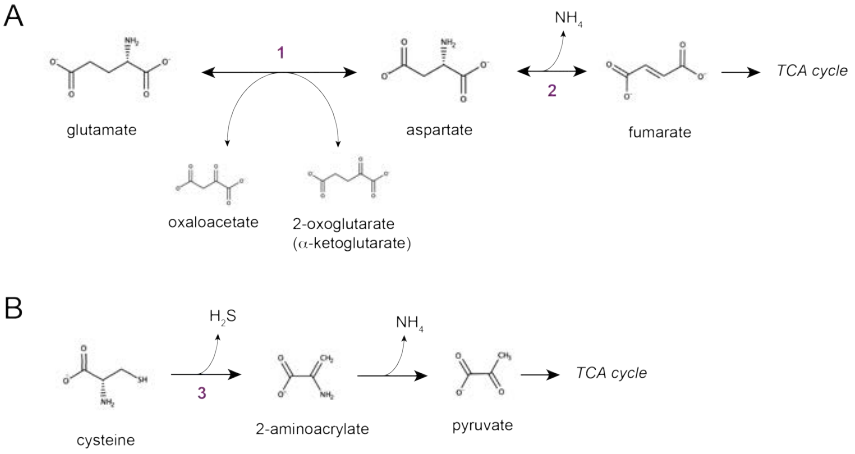


Figure 17.9. Representative amino acid degradation pathways. (A) Glutamate and aspartate degradation. Enzymes: 1, aspartate transaminase; 2, aspartate-ammonia lyase. (B) Cysteine degradation. Enzymes: 3, cysteine desulfhydrase. Deamination of 2-aminoacrylate to pyruvate is spontaneous.

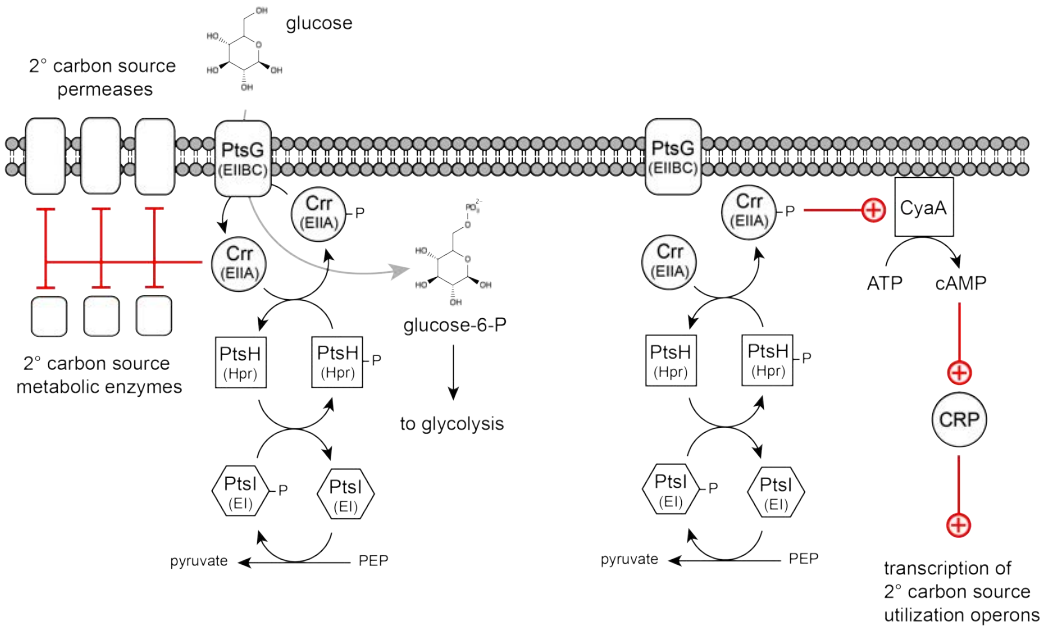
As we will see shortly, NH_4 and H_2S can be used as nitrogen or sulfur sources for biosynthesis or, if they are not needed, secreted out of the cell, where they will make the medium more basic or smellier, respectively.

DISCUSSION PROBLEM SET #36: CARBON CATABOLITE REPRESSION

Bacteria typically prefer to catabolize particular carbon sources, and repress the expression of proteins needed for the catabolism of other, secondary carbon sources when their preferred carbon source (often glucose) is present. This is called ***c**arbon **c**atabolite **r**epression* (CCR) or just "catabolite repression", and is a major global regulatory system in most bacteria.

The mechanisms of CCR have been well-studied in *E. coli* and *B. subtilis*, and to some extent in other bacteria. In both *E. coli* and *B. subtilis*, CCR depends on the glucose-specific PTS (recall Discussion Problem Set #35 above), but in distinctly different ways. The descriptions here are necessarily simplified somewhat, since CCR is a complex process and

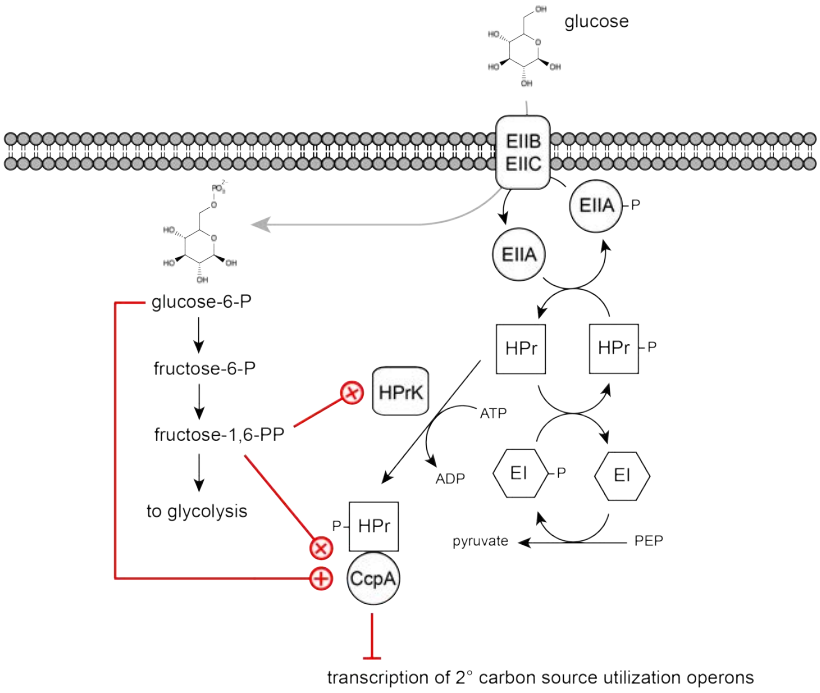
individual secondary carbon sources have their own different variations on the details, but I will try to explain the general principles involved.



In *E. coli*, the presence of glucose results in increased flux through the glucose PTS, with the phosphate group on $EIIA^{Glc}$ (Crr) being transferred to a glucose molecule as soon as it is available. This means that, during growth on glucose, most of the Crr is **dephosphorylated**. Dephosphorylated Crr interacts with and inhibits the activity of a wide range of metabolic enzymes and permeases required for catabolism of secondary carbon sources. This effect, and in particular the prevention of secondary carbon source import during growth on glucose, is called *inducer exclusion*.

In the **absence** of glucose, on the other hand, phosphorylated Crr (Crr-P) accumulates. Crr-P interacts with and **activates** adenylate cyclase, producing the second messenger cyclic AMP (cAMP). The global transcriptional activator CRP, when bound to cAMP, is necessary for transcription of the genes for catabolism of secondary carbon source. Finally, each operon for utilization of a secondary carbon source typically has its own local transcriptional regulation, so that the genes for catabolism of a particular carbon source (*e.g.* lactose or arabinose) are only actually expressed when that particular carbon source is present.

CCR in *B. subtilis* (and most other Firmicutes) also depends on the glucose PTS, but in this case it is the phosphorylation state of the HPr protein that exerts a regulatory effect. When levels of fructose-1,6-bisphosphate are high (indicating high glycolytic activity; see Figure 17.1), the regulatory kinase HPrK phosphorylates HPr on a site **different** from the one used for phosphotransfer to EII. This phosphorylated HPr forms a complex with the transcription factor CcpA, activating it for DNA binding. The CcpA-HPr-P complex is also stabilized directly by the glycolytic intermediates fructose-1,6-bisphosphate and glucose-6-phosphate, meaning that there are at least 3 inputs from glycolysis into CcpA activity.

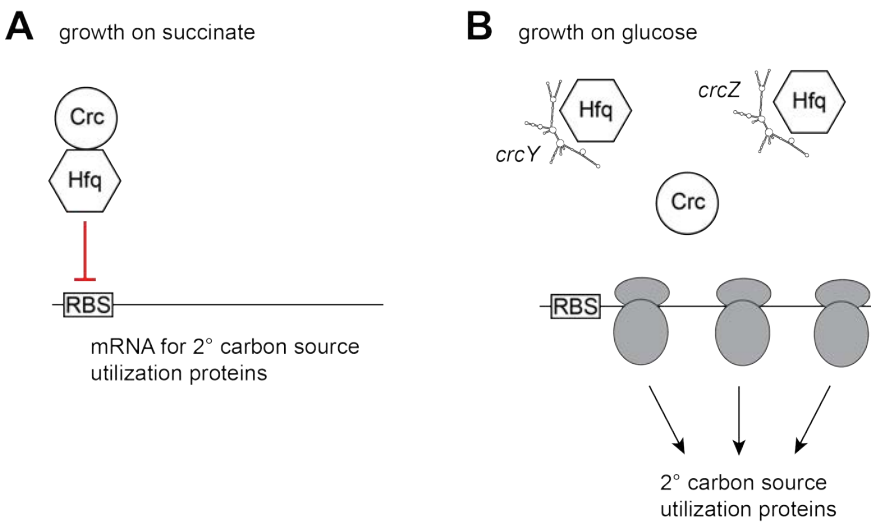


CcpA, unlike CRP, is a **repressor** of gene expression, and, in complex with HPr-P, inhibits the transcription of genes for utilization of secondary carbon sources. In the absence of glucose, the levels of glycolytic intermediates drop, HPrK becomes a phosphatase that dephosphorylates the regulatory site of HPr, the HPr-CcpA complex dissociates, and free CcpA is no longer able to bind to DNA. This frees up the genes for catabolism of secondary carbon sources to be expressed, dependent again on the activity of local regulators specific to each individual carbon source.

CCR does not depend on the PTS in all bacteria. For example, glucose is the preferred carbon source for *Streptomyces*, but *Streptomyces* do not use a PTS to transport glucose, instead depending on a glucose permease and glucokinase. In these bacteria, CCR depends (in a not very well-understood way) on glucokinase instead of on the PTS.

Pseudomonas species, unlike *E. coli*, *B. subtilis*, or *Streptomyces*, prefer to grow on succinate, and have a CCR system that represses utilization of other carbon sources (including glucose) when succinate is present (sometimes called *reverse catabolite repression*). The *Pseudomonas* CCR system is completely unlike that of *E. coli* and *B. subtilis*.

In the presence of succinate, the RNA-binding protein Crc, in complex with Hfq (which we mentioned way back in **Lecture 4**), binds to the ribosome binding sites of mRNAs encoding proteins necessary for utilization of secondary carbon sources, blocking their **translation**. Expression of both Crc and Hfq is constitutive, but in the absence of succinate *Pseudomonas* expresses two sRNAs (*crcY* and *crcZ*) that disrupt the Crc-Hfq-mRNA complex, allowing translation of those proteins, including for example, the enzymes necessary for growth on glucose.



What **isn't** clear in *Pseudomonas* is how the expression of *crcY* and *crcZ* is controlled. How does growth on succinate prevent their expression? Or does the lack of succinate somehow activate their expression?

The following methods are available for *Pseudomonas aeruginosa*:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
generalized transducing phage	√
compatible transposons	√
oligo-directed recombineering	√
CRISPR and related technologies (e.g. CRISPRi)	√
a genome-wide knockout collection (link)	√

Propose a model to explain how growth on succinate regulates *crcY* and *crcZ* expression in *P. aeruginosa*. Describe an experiment or series of experiments to test your model. State:

- your model
- your hypothesis and how your experiment(s) will test that hypothesis
- the independent and dependent variables of each experiment
- both positive and negative controls for each experiment

- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiment(s), and how you will interpret them

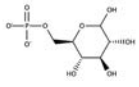
Question for discussion in class: *Mycoplasma* species do not seem to have catabolite repression, and just catabolize everything they're capable of catabolizing simultaneously. Why do you think these particular bacteria might have lost the capacity for catabolite repression?

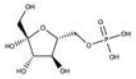
ANABOLISM

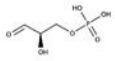
Where catabolism is the process of breaking molecules down into their component parts, anabolism is the process of synthesizing those molecules. *Prototrophic* bacteria are able to synthesize all of the amino acids, nucleotides, lipids, carbohydrates, and other cellular components they need from simple precursors. *Auxotrophic* bacteria must rely on other organisms in their environment to synthesize one or more of the basic building blocks of their cells. However, even prototrophic bacteria will, given the opportunity, often salvage components from their environment rather than expend the energy to synthesize them from scratch.

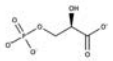
Fortunately for human comprehension, the carbon backbones of **all** biomolecules are derived from just 13 precursor intermediates, which are shown in Figure 17.10. Six of these are intermediates in glycolysis, three from the pentose phosphate pathway, and four from the TCA cycle.

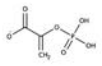
Glycolysis / Entner-Douderoff


glucose-6-phosphate


fructose-6-phosphate

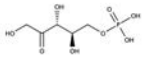

glyceraldehyde-3-phosphate

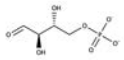

3-phosphoglycerate

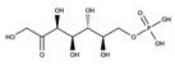

phospho*enol*pyruvate


pyruvate

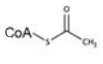
Pentose Phosphate Pathway


ribose-5-phosphate

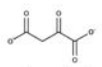

erythrose-4-phosphate

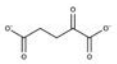

sedoheptulose-7-phosphate

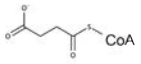
TCA Cycle



acetyl-CoA


oxaloacetate


2-oxoglutarate
(α-ketoglutarate)



succinyl-CoA

Figure 17.10. The 13 precursor intermediates needed to synthesize all biomolecules, and the central pathways from which they are derived.

Any prototrophic organism must, therefore, be able to use a simple carbon source to synthesize all 13 precursor intermediates, as well as having the additional anabolic pathways to construct other biomolecules from those precursors (see below). This is why glycolysis, the pentose phosphate pathway, and the TCA cycle are so central to metabolism, and why so many organisms have all three pathways. Fermentative bacteria (**Lecture 16**), which often lack the TCA cycle, are also notoriously auxotrophic and typically must salvage multiple amino acids, nucleotides, and cofactors from their environments.

While Figure 17.10 indicates that the precursor intermediates derived from glycolysis can also come from the Entner-Douderoff pathway, this is not **quite** true. As you can see in Figure 17.2, fructose-6-phosphate (F6P) is **not** an intermediate in that pathway, so bacteria that employ Entner-Douderoff instead of glycolysis either cannot synthesize the biomolecules derived from F6P (including many polysaccharides; **Lecture 14**) or must employ a separate pathway to generate F6P for biosynthetic purposes (the pentose phosphate pathway will do, since it also produces F6P as a product; Figure 17.3).

I made a point of mentioning above that the catabolic pathways of central metabolism are mostly reversible or cyclic, which means that they mostly allow cells to interconvert the precursor intermediates. That "mostly" requires a little bit of unpacking, though, and reversing central metabolism does sometimes require additional enzymes to convert catabolic pathways into anabolic ones.

In Figure 17.11, for example, you can see that anabolic *gluconeogenesis*, the pathway that synthesizes G6P from pyruvate, uses almost entirely the same set of enzymes as catabolic glycolysis, with two exceptions. In glycolysis, F6P is phosphorylated by phosphofructokinase to form fructose-1,6-bisphosphate, while in gluconeogenesis the opposite reaction is catalyzed by fructose-1,6-bisphosphatase (Figure 17.11 reaction 2). Similarly, in glycolysis, PEP is converted to pyruvate by pyruvate kinase, while in gluconeogenesis, pyruvate is converted to PEP by PEP synthetase (Figure 17.11 reaction 9). When growing on any carbon source that feeds into central metabolism "below" G6P (e.g acetate, glycerol, succinate, glutamate, arabinose, *etc., etc.*), gluconeogenesis is required for synthesis of the upstream precursor intermediates.

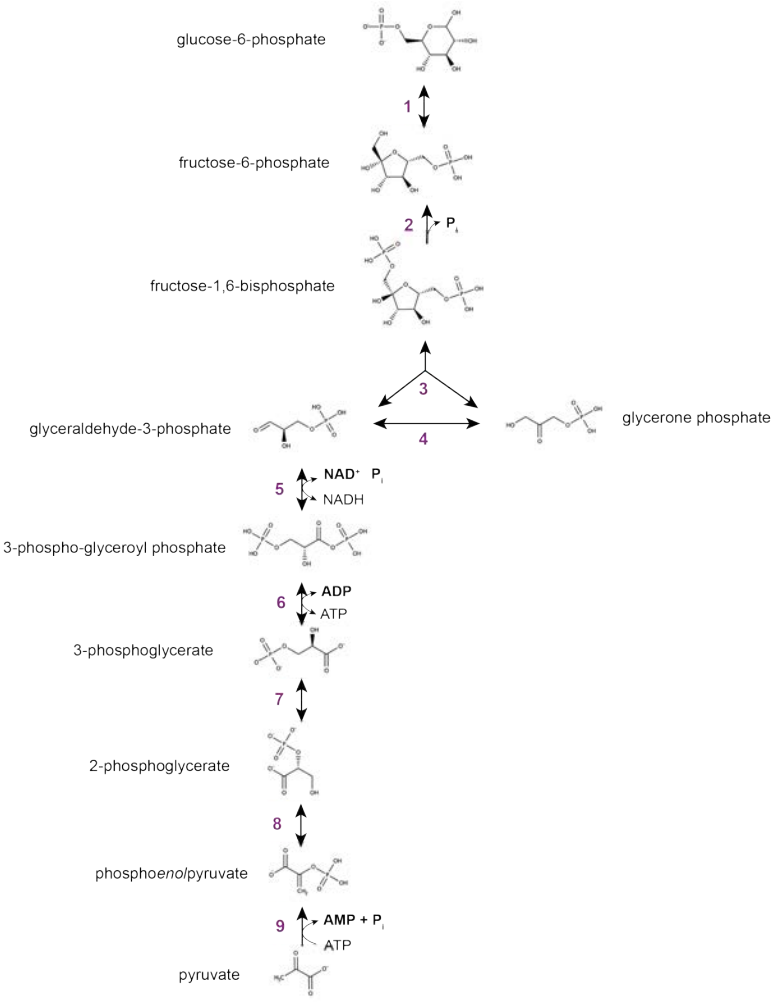


Figure 17.11. Gluconeogenesis. Enzymes: 1, glucose-6-phosphate isomerase (Pgi); 2, fructose-1,6-bisphosphatase (Fbp); 3, fructose-bisphosphate aldolase (Fba); 4, triose-phosphate isomerase (Tpi; 5, glyceraldehyde-3-phosphate dehydrogenase (GapDH); 6, phosphoglycerate kinase (Pgk); 7, phosphoglycerate mutase (Pgm); 8, enolase (Eno); 9, phospho*enol*pyruvate synthetase (PpsA).

All of the precursor intermediates derived from the pentose phosphate pathway are made by reversible enzymes using G3P and F6P as inputs (Figure 17.3), so no additional enzymes are required to run that pathway in reverse. The TCA cycle (Figure 17.6) is cyclic, so also does not need to run in reverse to generate precursor intermediates, although many of the enzymes involved **are** reversible, so in some cases, parts of the cycle can flow in the opposite of the usual direction. (For example, in organisms that do not have the complete set of TCA cycle enzymes.)

In the last portion of this chapter, I will explore a few anabolic pathways in detail, to show how the precursor intermediates are used as the basis of essential cellular components like sugars, lipids, amino acids, and nucleotides. As

with catabolism, I will only be able to touch on a few examples, and the underlying principles are more important that the details of each pathway.

G6P and F6P are the precursors of many other sugars. For biosynthetic purposes, G6P is usually first activated by conversion into a nucleotide-bound form (usually **u**ridine **di**phosphate, forming UDP-glucose; Figure 17.12), then modified by other enzymes. As an example, the synthesis of UDP-galactose is illustrated in Figure 17.12. Notably, UDP-glucose is also a precursor of the LPS core (**Lecture 10**), and various nucleotide-activated sugars are the substrates for many pathways of polysaccharide biosynthesis (**Lecture 14**).

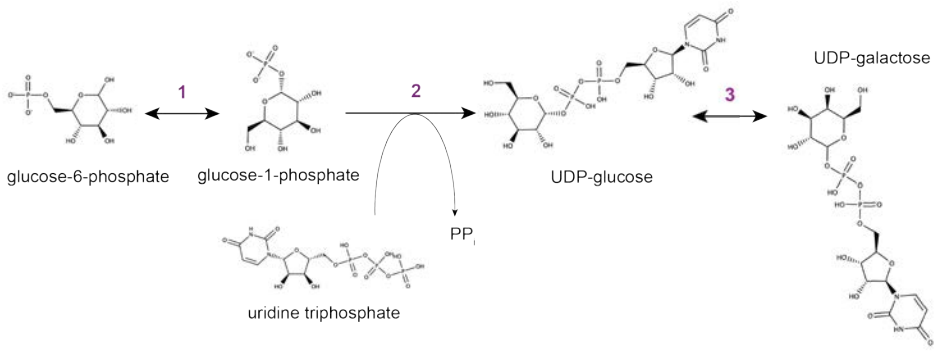


Figure 17.12. Galactose biosynthesis from G6P. Enzymes: 1, phosphoglucomutase; 2, UTP:glucose-1-phosphate uridylyltransferase; 3, UDP-glucose 4-epimerase.

Fatty acids and other lipids are synthesized from acetyl-CoA, as shown in Figure 17.13. There is first an initiation pathway (Figure 17.13A), in which two acetyl-CoA are combined to generate an acetoacetyl chain, and an elongation cycle (Figure 17.13B) in which the products of the initiation pathway are concatenated into longer fatty acid chains.

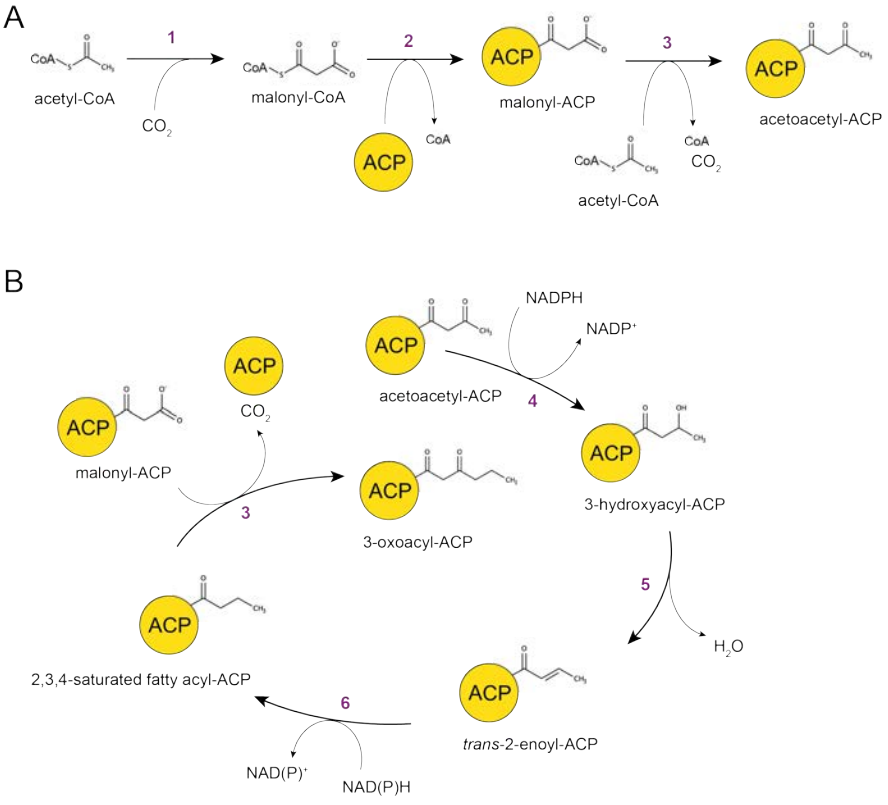


Figure 17.13. Fatty acid biosynthesis from acetyl-CoA. (A) Initiation and (B) elongation cycle. ACP = acyl-carrier protein. Enzymes: 1, acetyl-CoA carboxyltransferase; 2, ACP *S*-malonyltransferase; 3, β-ketoacyl-ACP synthetase; 4, 3-oxoacyl-ACP reductase; 5, 3-hydroxyl-ACP dehydratase; 6, enoyl-ACP reductase.

Note that the growing fatty acid chain is bound to an ***a****cyl-****c****arrier* ***p****rotein* (ACP) while it is being synthesized, and that each turn of the elongation cycle adds 2 carbons to the fatty acid. ACP is needed to keep the hydrophobic lipid soluble in the cytoplasm and accessible to the synthesis enzymes. Once the fatty acid is complete, it is cleaved off of the ACP. Fatty acids of different lengths are generated with different numbers of elongation cycles, and additional enzymes are involved in inserting double bonds and other modifications to generate different fatty acid types (**Lecture 10**). Fatty acid biosynthesis consumes a large number of reducing equivalents (mostly NADPH).

Synthesizing amino acids requires incorporation of N and S atoms in addition to the C backbones derived from the precursor intermediates. This requires reduced N and S donors, typically NH_4 and H_2S, which are initially used to synthesize glutamate and cysteine, respectively (Figure 17.14). Many organisms can reduce other inorganic S sources in their environments (sulfate, sulfite, thiosulfate, *etc.*) into H_2S for incorporation into cysteine. Glutamate is derived from the precursor intermediate 2-oxoglutarate, and cysteine from 3-phosphoglycerate and acetyl-CoA, by way of serine.

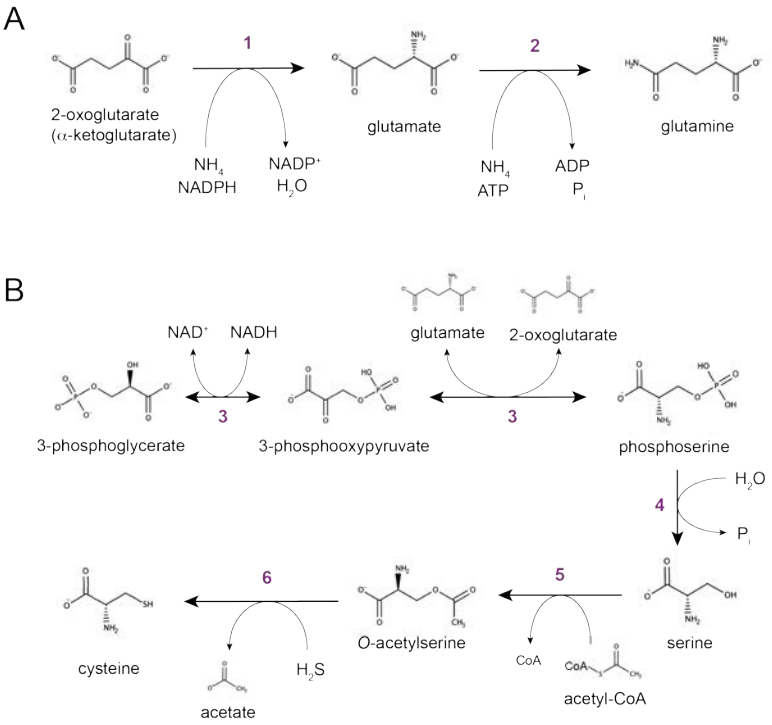


Figure 17.14. Examples of important amino acid biosynthetic pathways. (A) Glutamate and glutamine biosynthesis. Enzymes: 1, glutamate synthase; 2, glutamine synthetase. (B) Serine and cysteine biosynthesis. Enzymes: 3, 3-phosphoglycerate dehydrogenase (yes, it really does both of these reactions); 4, 3-phosphoserine aminotransferase; 5, phosphoserine phosphatase; 6, serine acetyltransferase; 7, *O*-acetylserine sulfhydrylase.

Glutamate is by far the most abundant metabolite in bacterial cells (about 100 mM in the cytoplasm of *E. coli*), and either glutamate or glutamine are the N donors for most biosynthetic pathways (note glutamate donating an amino group in reaction 3 in Figure 17.14). Cysteine is the S donor for most biosynthetic pathways that require S, and while cysteine itself is not especially abundant in most cells, cysteine-derived low molecular weight thiols, such as the tripeptide glutathione (glutamine-cysteine-glycine), are present at millimolar concentrations as reducing agents that help maintain the redox potential of the cytoplasm.

As a last example of a biosynthetic pathway, we'll look at the somewhat more complicated mechanism by which nucleotides are assembled, and specifically at the pathway that leads to the pyrimidine nucleotide cytidine-5'-triphosphate (CTP) (Figure 17.15). This is mostly to illustrate that anabolic pathways can become quite involved, with large numbers of enzymes needed to convert precursor intermediates into their final products.

The pyrimidine base is assembled from bicarbonate, glutamine, and aspartate (itself derived from an oxaloacetate and a glutamate; not shown), and then combined with the ribose-5-phosphate derivative 5-phospho-ribose-1-diphosphate (PRPP) to form the nucleotide precursor orotidine-5'-phosphate. Orotidine-5'-phosphate, in turn, is decarboxylated to form uridine-5'-phosphate (UMP), the precursor of all the pyrimidines in DNA and RNA (CTP, dCTP, TTP, dTTP) and an important player in carbohydrate biosynthesis in its own right (see Figure 17.12 and **Lecture 14**).

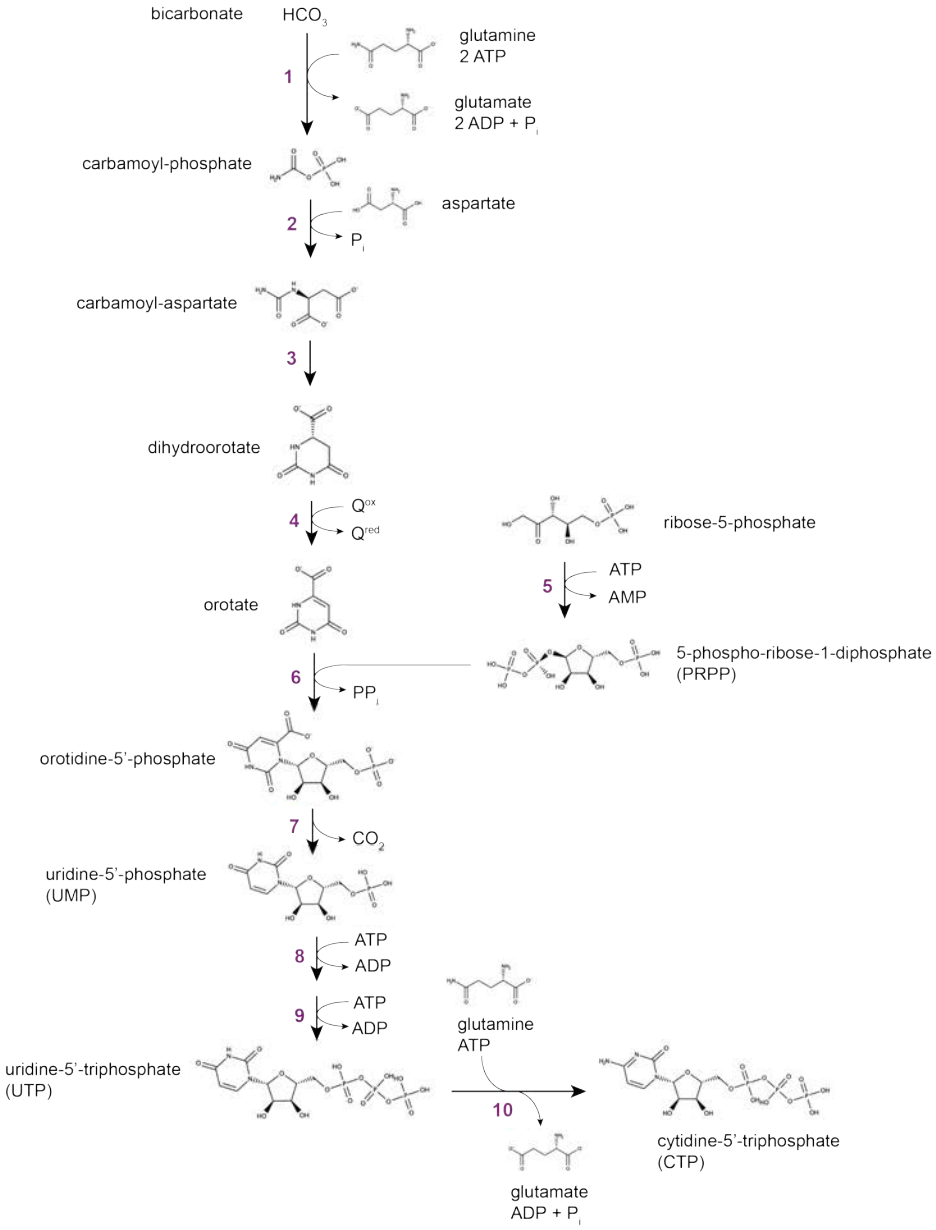


Figure 17.15. Pyrimidine nucleotide biosynthesis. Enzymes: 1, carbamoyl-phosphate synthetase; 2, aspartate carbamoyl transferase; 3, dihydroorotase; 4, dihydroorotate dehydrogenase; 5, ribose-phosphate diphosphokinase; 6, orotate phosphoribosyltransferase; 7, orotidine-5'-phosphate decarboxylase; 8, UMP kinase; 9, UDP kinase; 10, CTP synthetase.

PRPP is an important biosynthetic intermediate, and is a precursor of not only pyrimidines, but purines, the amino acids histidine and tryptophan, and the nicotinamide cofactors (NAD and NADP), among other compounds.

LECTURE 18: SECONDARY METABOLISM

INTRODUCTION

Secondary metabolism (a subset of anabolism) encompasses the synthesis of an enormous range of biological molecules that do not make up the bulk constituents of cells (*i.e.* are not the precursors of proteins, RNA, DNA, cell walls, capsules, or membrane lipids). These include a wide range of biologically interesting but non-essential compounds like antibiotics, quorum sensing autoinducers, electron shuttles, siderophores, and many more.

In this chapter, we will examine the pathways by which a representative set of these *secondary metabolites* are synthesized, although the nature of the topic means that this will necessarily be a **very** incomplete survey.

WHAT IS SECONDARY METABOLISM?

The original terminology for secondary metabolism was drawn from the fact that these molecules are species- or strain-specific, as opposed to the "primary metabolites" that are found in all cells (**Lecture 17**). Some people have argued that "secondary metabolites" are important to the organisms that produce them, and they should be called "specialized metabolites" instead, but I will use the more commonly-accepted term here. Another term you will encounter in the field is "*natural products*", an extremely vague category which includes essentially any molecule produced by a living organism. It is useful in comparison to the (usually) simpler molecules generated by organic chemists or to human-modified variants of naturally occurring compounds.

Essentially all bacteria produce at least a few secondary metabolites, and there is great diversity in the metabolite repertoire of strains within a single species. Different strains of *B. subtilis* produce widely differing sets of surfactants and anti-fungal compounds, for example, and different strains of *E. coli* produce different metal-chelating siderophores, which can contribute to their ability to cause infections.

Secondary metabolites have a vast range of biological functions, and we actually have very little idea of what the physiological function of many bacterial metabolites might be, even those which have turned out to be extremely useful for human purposes, say as anti-cancer or anti-inflammatory drugs. (More than 30% of all FDA-approved drugs are natural product derivatives.) *Cryptic metabolites* are secondary metabolites with no known function. You will often see secondary metabolites with known effects on cells referred to as *bioactive compounds*, but of course presumably all metabolites are "bioactive" in some context for the organisms that produce them.

Most known secondary metabolites are synthesized by *biosynthetic gene clusters* (BGCs), which are genetic loci dedicated to the production of a specific secondary metabolite. These can be very large (100 kb or more), and contain multiple operons with complex regulation. *Silent BGCs* are BGCs that are not expressed under laboratory conditions, a problem we will return to later in the chapter. The grouping of secondary metabolite synthesis genes into BGCs means that production of secondary metabolites can fairly easily be horizontally transferred between strains and species.

Of course, as microbiologists, the secondary metabolites that probably have the most direct relevance to our work are antibiotics. Most of the families of antibiotics in current use are derivatives of natural products, mostly from different species of actinomycetes, especially members of the genus *Streptomyces*. The evolution of pathogenic bacteria resistant to clinical antibiotics is a problem which has led to a great deal of discussion and concern, with wide interest in discovering novel secondary metabolites that may be useful as new antimicrobial treatments. Kim Lewis, director of the Antimicrobial Discovery Center at Northeastern University, has recently written a thorough, thoughtful perspective article in Cell on the history and current state of antibiotic discovery science, and explains several reasons to be optimistic about our ability to overcome the looming antibiotic resistance crisis.

There are many resources available for the study of secondary metabolism. The website secondarymetabolites.org is a convenient clearinghouse for many useful bioinformatic tools and databases.

PHENAZINE SYNTHESIS

Some secondary metabolites are synthesized by pathways very much like the anabolic central metabolic pathways we discussed in **Lecture 17**. As our representative example, we will look at the synthesis of pyocyanin, the "blue phenazine" that lends a characteristic greenish tinge to stationary phase cultures of *Pseudomonas aeruginosa*.

The initial steps of pyocyanin synthesis are the same as those for the aromatic amino acids phenylalanine, tryptophan, and tyrosine, and begin with the precursor intermediates phospho*enol*pyruvate and erythrose-4-phosphate (Figure 17.10). Phenazine synthesis branches off from those amino acid synthesis pathways in the fate of chorismate (Figure 18.1, which is somewhat simplified to make everything fit on the page). The basic conjugated three-ring phenazine scaffold (in the form of phenazine-1-caboxylic acid) is formed by combination of two chorismate molecules, and subsequent enzymes modify this scaffold to form specific phenazines, in this case, of course, pyocyanin.

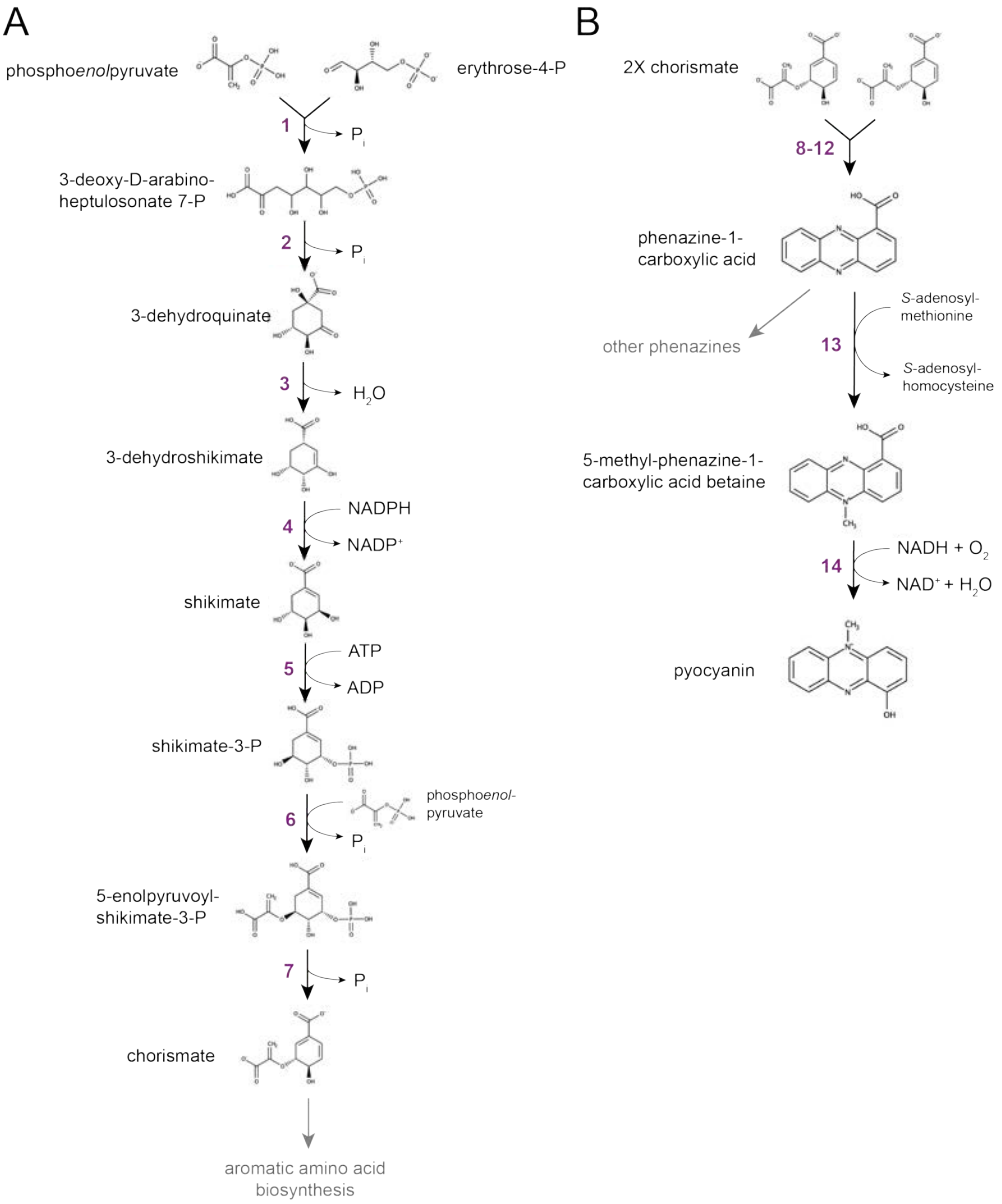


Figure 18.1. Synthesis of the phenazine pyocyanin in *Pseudomonas aeruginosa*. A = shared steps with aromatic amino acid biosynthesis; B = phenazine synthesis pathway. Enzymes: 1, 3-deoxy-7-phosphoheptulonate synthase (AroGFH); 2, 3-dehydroquinate synthase (AroB); 3, 3-dehydroquinate dehydratase (AroD); 4, shikimate dehydrogenase (AroE); 5, shikimate kinase (AroKL); 6, 3-phosphoshikimate-1-carboxyvinyltransferase (AroA); 7, chorismate synthase (AroC); 8 – 12, five enzymatic steps (PhzEDFBG); 13, phenazine-1-carboxylate N-methyltransferase (PhzM); 14, 5-methylphenazine-1-carboxylate 1-monooxygenase (PhzS).

Pyocyanin is a particularly interesting metabolite because it illustrates the difficulty of assigning "a" physiological function to secondary metabolites. Pyocyanin is often described as a virulence factor, and indeed it **is** highly toxic to human cells and makes a major contribution to tissue damage during *P. aeruginosa* infections, but this is only one of its functions, and (from the point of view of *P. aeruginosa*), probably not the most important.

In work pioneered by the lab of Dianne Newman at CalTech, the roles of pyocyanin and other phenazines in *P. aeruginosa* biofilms have recently become more clear. *P. aeruginosa* is an obligate respiring organism (**Lecture 16**), but characteristically forms thick biofilms during infections (**Lecture 14**). Oxygen is not able to penetrate to the inner layers of the biofilms, but phenazines can diffuse freely through the matrix and cell membranes. Phenazines are redox-active electron shuttles, and *P. aeruginosa* can use oxidized pyocyanin as a terminal electron acceptor for respiration (**Lecture 16**). The resulting reduced pyocyanin diffuses towards the outer surface of the biofilm, where it is oxidized by O_2. Phenazines therefore allow deeply-buried cells in biofilms to respire O_2 "at a distance".

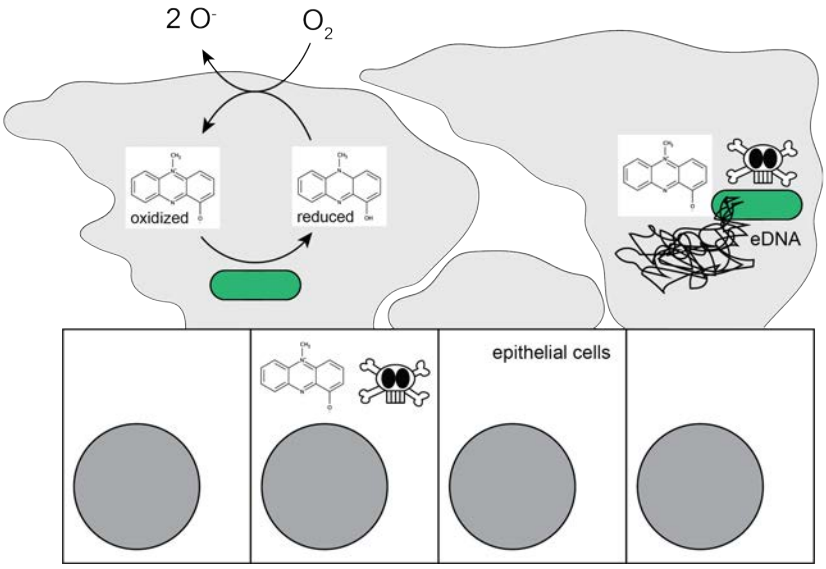


Figure 18.2. Distinct functions of pyocyanin in *P. aeruginosa* biofilms and infections. Redox cycling of pyocyanin allows deeply-buried *P. aeruginosa* to respire, but starved bacteria may be killed by pyocyanin toxicity, releasing eDNA to reinforce the biofilm matrix. Pyocyanin is also a potent toxin that kills host cells, contributing directly to pathogenesis.

That's not all that pyocyanin does, though. The most abundant polymer in the matrix of a *P. aeruginosa* biofilm is eDNA, which is released from bacteria by lysis of part of the population. Pyocyanin also plays a key role in this process. When *P. aeruginosa* is starved for carbon (and therefore for ATP; **Lecture 16**), *P. aeruginosa* can no longer cope with the toxic effects of pyocyanin, resulting in the lysis and death of about 90% of the starved cells, release of eDNA, and stabilization of the biofilm as a whole.

Pyocyanin clearly plays a dynamic physiological role in *P. aeruginosa* biofilms and infections, with different functions under different growth conditions and stages. There is no particular reason to think that a given metabolite, secondary or otherwise, ever has just one physiological function.

ACYL-HOMOSERINE LACTONE SYNTHESIS

The AHL quorum sensing signal molecules (**Lecture 14**) are another large class of secondary metabolites whose structures vary widely among bacteria. However, their synthesis proceeds by a fairly simple pathway.

The hydrophobic acyl portion of the AHL is synthesized on ACP by the same kind of pathway used to synthesize other lipids, which we have already discussed (Figure 17.13). An AHL synthase enzyme then catalyzes the reaction of that acyl-ACP with *S*-adenosylmethionine, yielding ACP, 5'-methyl-thioadenosine, and the desired AHL (Figure 18.3).

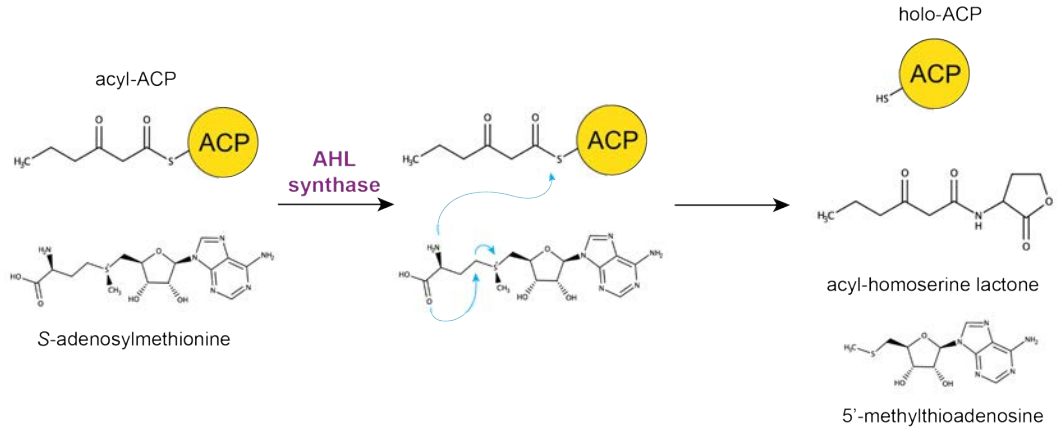


Figure 18.3. Acyl-homoserine lactone synthesis by AHL synthase, using an acyl-ACP and *S*-adenosylmethionine as substrates.

***S*-adenosyl*m*ethionine** (often abbreviated SAM or AdoMet) is a common cosubstrate involved many kinds of anabolic reactions, most notably in methyl group transfer reactions. (EcoCyc lists 88 known reactions that consume SAM in *E.*

coli K-12, a strain that lacks AHL-based quorum sensing pathways.) All cells contain SAM, which is synthesized from a methionine and an ATP. The reaction catalyzed by AHL synthases is an unconventional SAM-dependent reaction, and not at all typical of the kinds of reactions SAM is normally involved in.

POLYKETIDE SYNTHASES

The really complicated secondary metabolites, though, are synthesized by two types of modular biosynthetic machines: the ***p****oly****k****etide* ***s****ynthases* (PKS) and the ***n****on-****r****ibosomal* ***p****eptide* ***s****ynthetases* (NRPS), which construct complex molecules out of simple carboxylic acid and amino acid components, respectively. The *megasynthase* enzymes involved in some PKS and NRPS pathways are among the largest proteins synthesized by bacteria, with the largest known being proteins over 2 MDa in molecular weight, more than 17,000 amino acids long, encoded by genes spanning more than 50 kb (> 1% of an entire bacterial genome). Both PKS and NRPS pathways are broadly conserved across the tree of life, and are capable of synthesizing a vast array of compounds. Here, we will touch briefly on how each of these pathways typically works in bacteria, with the understanding that, once again, I am **vastly** oversimplifying the true complexity and diversity of mechanisms that exist.

Like fatty acid synthases (**Lecture 17**), PKS construct their products from acyl-carrier protein (ACP)- or CoA-bound carboxylic acid subunits. These can include malonyl-ACP (as for fatty acid synthesis), but also a variety of other acyl-ACPs, like acetyl-ACP, benzoyl-ACP, and ethylmalonyl-ACP.

PKS are made up of large, highly modular enzymes. They contain several ACP-like domains, on which the products are assembled, as well as additional enzymatic domains that carry out key steps in synthesis. These include:

acyltransferases (abbreviated AT)
β-ketoacylsynthetases (KS)
ketoreductases (KR)
dehydratases (DH)
enoylreductases (ER)
methyltransferases (MT)
cyclases (CYC)
aromatases (ARO)
thioesterases (TE)

In any PKS, the first step is transfer of an initiating acyl group from acyl-ACP to the first ACP domain of the PKS by an AT domain. Each subsequent acyl-CoA building block is added by the next module's KS and AT domains in an elongation stage, transferring the growing polyketide chain to the next ACP domain. Optionally, there may be any combination of DH, ER, MT, CYC, and ARO domains which chemically modify the resulting product at each step. Once the polyketide reaches the final ACP domain, the TE domain cleaves it off of the PKS, often including a cyclization step.

The minimal PKS therefore consists of the following domain structure, with a minimal "module" enclosed in brackets:

AT-ACP-[KS-AT-ACP]-TE

The initiating AT domain loads the first acyl group onto the first ACP domain, each subsequent module adds another acyl group, and the final TE domain releases the product. The resulting product will then almost always be modified further to generate the final metabolite, which can involve many additional enzymatic steps.

There are many types and variations of PKS, of which we only have space to cover a couple. In bacterial type I PKS, like the very well-studied one responsible for synthesis of the antibiotic erythromycin by *Saccharopolyspora erythraea* (Figure 18.4A), the PKS modules are arranged in one or a few enormous megasynthases and synthesis proceeds linearly from N-terminal to C-terminal ACPs. While erythromycin happens to be constructed by addition of 6 sequential methylmalonyl-CoA groups to a malonyl-ACP starter, the acyl groups added to growing type I PKS substrates can vary widely.

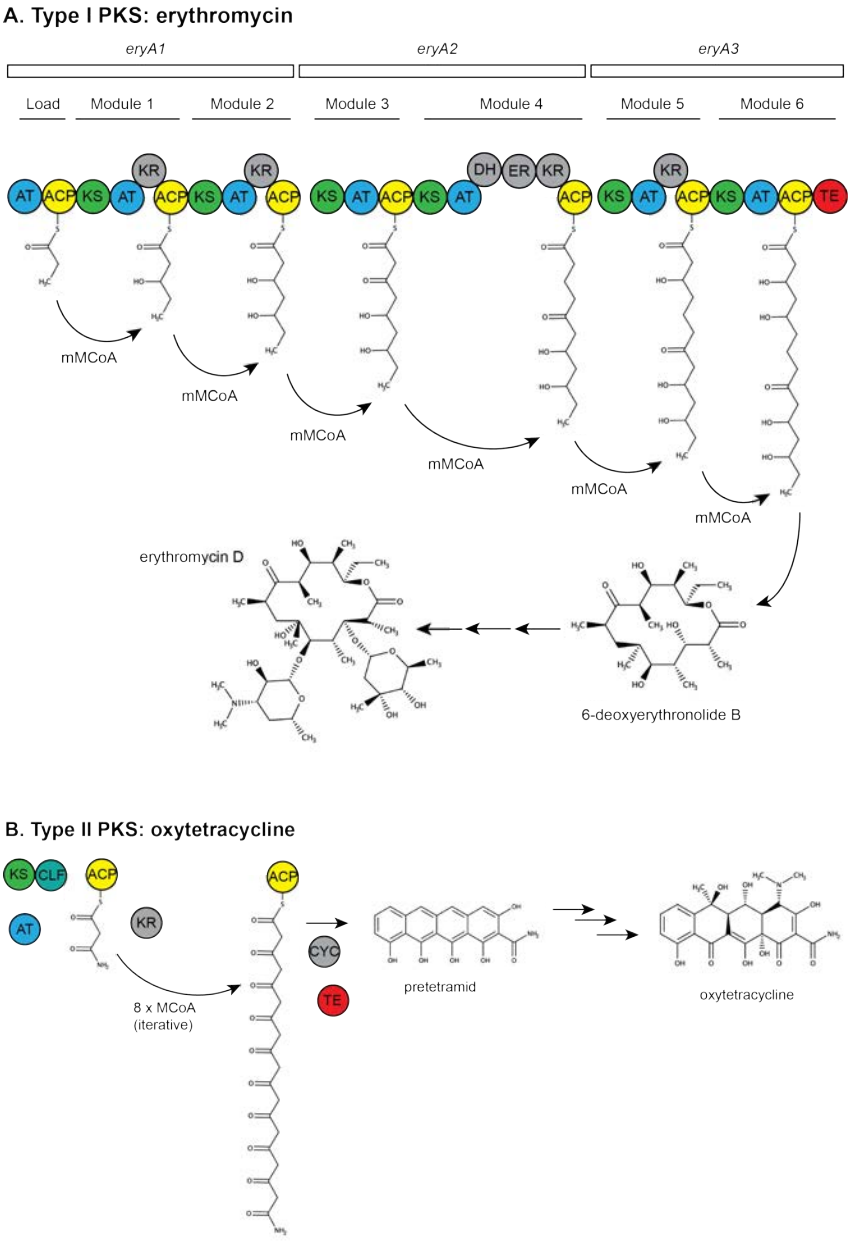


Figure 18.4. Polyketide synthase-dependent secondary metabolism. (A) Erythromycin biosynthesis by linear type I PKS, encoded by the *eryA1, eryA2,* and *eryA3* genes. (B) Tetracycline biosynthesis by iterative type II PKS. Abbreviations: AT, acyltransferase; ACP, acyl carrier protein; KS, β-ketoacylsynthetase, KR, ketoreductase; DH, dehydratase; ER, enoylreductase; TE, thioesterase; CLF, chain length factor; CYC, cyclase; mMCoA, methylmalonyl-CoA; MCoA, malonyl-CoA.

Bacterial type II PKS, on the other hand, like fatty acid synthesis (Figure 17.13), are **iterative**, in that they re-use the same enzyme modules repeatedly to progressively add the **same** acyl-CoA unit to a growing polyketide chain. In bacterial type II PKS, the individual domains are **not** typically part of the same polypeptide, although they do form multi-enzyme complexes, and there is an additional protein, related to KS domains, called ***c****hain* ***l****ength* ***f****actor* (CLF) that determines how many cycles will proceed before the TE protein releases the product.

Iterative type II PKS is illustrated in Figure 18.4B with the pathway for synthesis of oxytetracycline by *Streptomyces rimosus*, which begins with a malonymate group and proceeds with successive addition of 8 malonyl-CoA groups.

NON-RIBOSOMAL PEPTIDE SYNTHETASES

PKS use acyl-CoA subunits as the building blocks of secondary metabolites. NRPS use amino acids instead, generating peptides between 3 and 15 amino acids long. However, the modular logic of the two synthetic pathways is very similar.

There are both linear (type A) NRPS megasynthases, analogous to type I PKS, and iterative (type B) NRPS pathways, analogous to type II PKS, and NRPS pathways are also constructed from a modular set of enzymatic domains, in this case:

peptidyl carrier protein domains (abbreviated PCP)
adenylation domains (A); specific for particular amino acids
condensation domains (C)
thioesterases (TE)

There are also, as for PKS, optional modification domains, like cyclases (CY) or epimerases (E), that occur in some pathways.

In a minimal NRPS, an A domain transfers an activated amino acid to the first PCP domain, in much the same way that tRNA synthetases attach amino acids to tRNA. Note that NRPS are not limited to the 20-ish proteinaceous amino acids, and there are A domains specific for a very wide range of both D- and L-amino acids. The C domain catalyzes the formation of the peptide bond between each consecutive amino acid.

A minimal NRPS therefore has the following domain architecture:

A-PCP-[C-A-PCP]-TE

The synthesis of the *B. subtilis* surfactant surfactin is catalyzed by a linear type A NRPS (Figure 18.5), which also incorporates a fatty acid chain in the first module.

Type A NRPS: surfactin

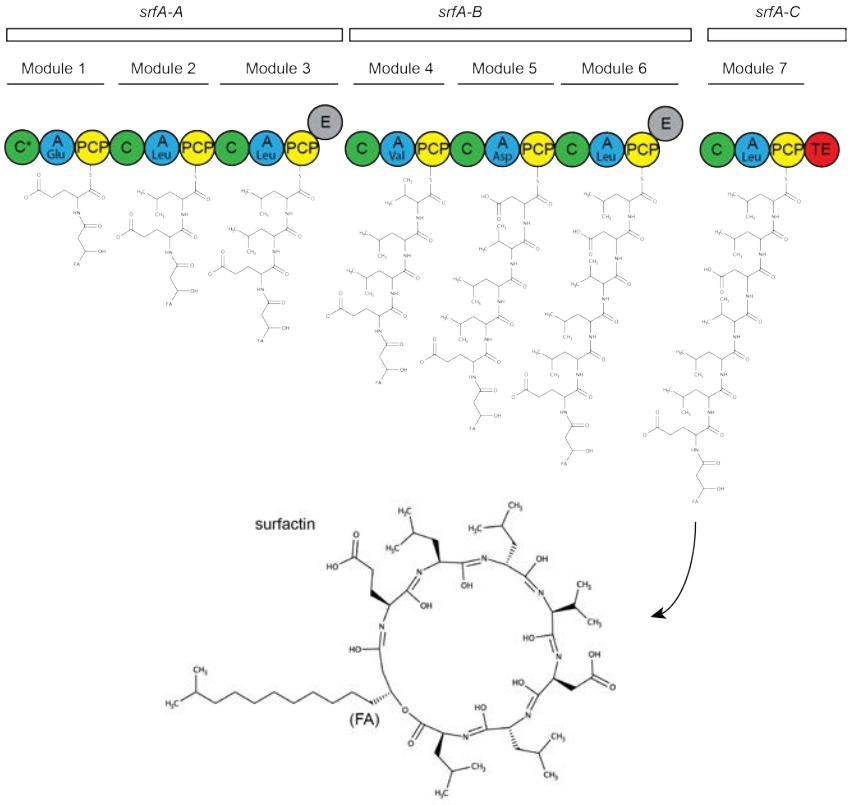


Figure 18.5. Surfactin biosynthesis by linear type A non-ribosomal peptide synthesis, encoded by the *srfA-A, srfA-B,* and *srfA-C* genes. The first C domain, indicated with an asterisk (C*) transfers a fatty acid to the glutamate added by the first adenylation domain. (B) Enterobactin biosynthesis by iterative type B NRPS. Abbreviations: C, condensation domain; PCP, peptidyl carrier protein; A, adenylation domain (amino acid-specific); E, epimerase domain; TE, thioesterase; FA, fatty acid; Glu, glutamate; Leu, leucine; Val, valine; Asp, aspartate.

The synthesis of the *E. coli* siderophore enterobactin is catalyzed by an iterative type B NRPS (Figure 18.6), combining three dihydroxylbenzoyl-seryl amino acid dimers into a cyclic hexameric peptide.

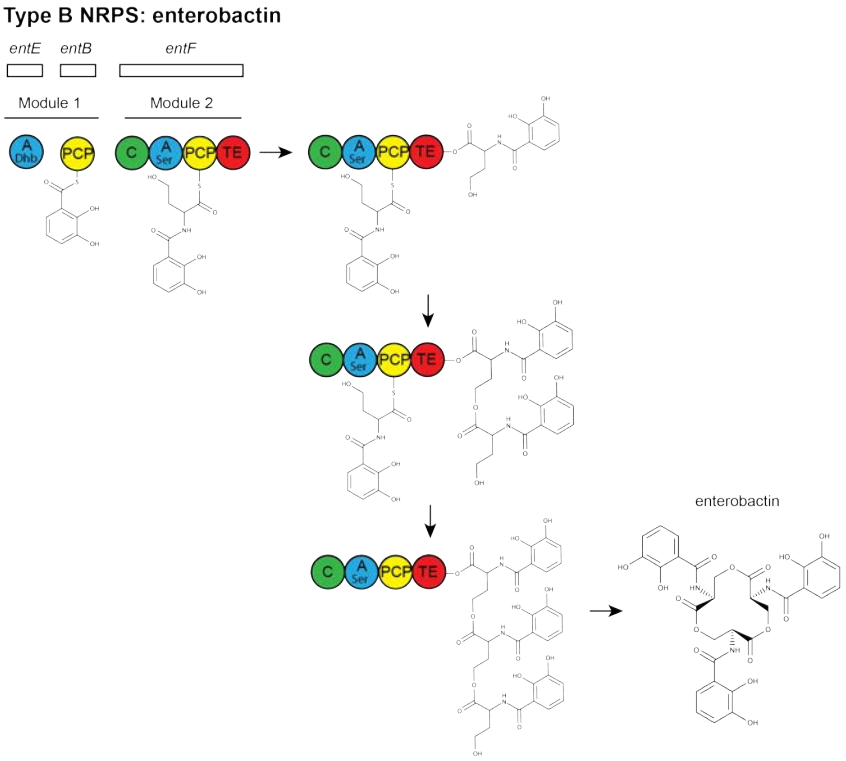


Figure 18.6. Enterobactin biosynthesis by iterative type B non-ribosomal peptide synthesis, encoded by the *entE, entB,* and *entF* genes. Abbreviations: C, condensation domain; PCP, peptidyl carrier protein; A, adenylation domain (amino acid-specific); TE, thioesterase; Dhb, 2,3-dihydrobenzoic acid; Ser, serine.

Not **all** peptide-derived secondary metabolites are synthesized by NRPS. Some are made by processing fragments of proteins transcribed and translated in the normal way. These secondary metabolites (including nisin, which we looked at in Discussion Problem Set 8) and many quorum sensing signal peptides are called *RiPPs* (**ri**bosomally synthesized and **p**ost-translationally modified **p**eptides).

I will note again that not all PKS and NRPS systems are as simple as the ones illustrated above. There are type III PKS, which are iterative megasynthase-type enzymes and non-linear type C NRPS, which skip or repeat individual modules, as well as variations and combinations of all of the different possibilities.

Indeed, because of the similarity between the PKS and NRPS biosynthetic strategies, it turns out to also be possible to **combine** PKS and NRPS modules in a single hybrid PKS/NRPS biosynthetic pathway. This is fairly common, and is how compounds like bleomycin, bacillaene, and myxothiazol are assembled. The PKS and NRPS modules may be on separate polypeptides or combined in hybrid megasynthases that have both PKS and NRPS modules.

The highly modular nature of PKS and NRPS synthesis pathways is central to their ability to generate diverse natural products, and presumably encourages the rapid evolution of new secondary metabolites. Certainly, human bioengineers have put a lot of effort into rearranging and swapping out modules to engineer pathways that synthesize novel compounds, some of which have been extremely valuable.

DISCUSSION PROBLEM SET #37: ACTIVATING SILENT BIOSYNTHETIC GENE CLUSTERS

One of the most widely-discussed problems in natural product discovery is that of cryptic and silent BGCs. The genomes of most actinomycetes and of many other bacteria encode BGCs whose products are unknown and which are not expressed under laboratory growth conditions. Many researchers consider these cryptic BGCs an untapped source of natural products that may have useful therapeutic properties.

Photorhabdus luminescens is a Gram-negative bacterium that, in symbiosis with a nematode host, is a pathogen of insects. Its genome encodes at least 23 large BGCs, mostly of the NRPS type, but the products of only 5 of these have been identified (two siderophores, two antibacterial compounds, and one compound which repels ants and birds that might eat an infected insect).

The following genetic tools are available for *P. luminescens*:

growth in pure culture	✓
can extract DNA/RNA/protein	✓
complete genome sequence	✓
susceptible to mutagens	✓
can be made competent	✓
shuttle & suicide vectors available	✓
inducible promoters known (arabinose-inducible)	✓
selectable & counter-selectable markers available	✓
compatible transposons	✓
oligo-directed recombineering	✓

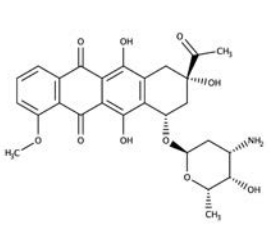
Describe an experiment to determine whether the secondary metabolites produced by any of the silent BGCs of *P. luminescens* might be useful antibiotics. State:

- the hypothesis your experiment is testing
- the independent and dependent variables of that experiment
- both positive and negative controls
- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

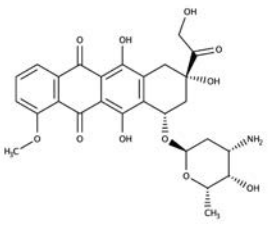
P. luminescens is the source of the most commonly used bacterial luciferase reporter (the *luxCDABEG* operon). It is not especially obvious why this insect pathogen is so strongly bioluminescent, but occasionally human wounds can become infected with *Photorhabdus*, especially during hypothermia (*Photorhabdus* does not survive well at normal human body temperature). Legend has it that after the Battle of Shiloh during the American Civil War, some soldiers' wounds lit up at night with a faint blue "angel's glow", that scientists now hypothesize might have been due to infestation with *Photorhabdus*-harboring nematodes. The glowing wounds were reported to heal faster, which may have been due to the release of anti-microbial compounds from *Photorhabdus*. It's a nice story, but I don't know that I'd count on it as a treatment plan!

DISCUSSION PROBLEM SET #38: PHAGE-INHIBITING SECONDARY METABOLITES

Karen Maxwell's lab at the University of Toronto has recently reported that some secondary metabolites produced by *Streptomyces* species function to protect bacteria against infection by bacteriophage. Most of the metabolites they identified in their initial screen for this activity were members of a chemical family called the anthracyclines, including daunorubicin and doxorubicin, which are currently in clinical use as anti-cancer drugs.



daunorubicin



doxorubicin

Anthracylines are DNA intercalating agents whose anti-cancer effects are thought to be due to inhibition of topoisomerase in rapidly-replicating cancer cells. The doses needed to prevent phage infection of bacteria are much lower than the doses of these drugs that are toxic to the bacteria themselves. Other *Streptomyces* metabolites that affect DNA (like the DNA cleavage stimulator bleomycin) have no effect on phage replication.

Strikingly, daunorubicin and doxorubicin not only prevent *Streptomyces*-specific phages from killing *Streptomyces* strains, but also prevent the unrelated λ, T5, T6, and T7 phages from propagating in *E. coli*, suggesting that they function by interfering with a mechanism common to all phage.

The following genetic tools are available for *S. peucetius*, the species from which both daunorubicin and doxorubicin were originally isolated:

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√

Many more tools are available for the model organism *S. coelicolor*, which does not produce either daunorubicin or doxorubicin (it produces a variety of other secondary metabolites, including the antibiotics actinorhodin and methylenomycin, the anti-malarial undecylprodigiosin, and the anti-fungal perimycin):

growth in pure culture	√
can extract DNA/RNA/protein	√
complete genome sequence	√
susceptible to mutagens	√
can be made competent	√
shuttle & suicide vectors available	√
selectable & counter-selectable markers available	√
compatible transposons	√
oligo-directed recombineering	√
CRISPR and related technologies (e.g. CRISPRi)	√
a genome-wide knockout collection (link)	√

And, of course, you can do anything you want, genetically speaking, in *E. coli* and its phages. Both daunorubicin and doxorubicin are commercially available in pure form.

Describe an experiment or experiments to determine the mechanism by which anthracyclines protect bacteria against phage infections. State:

- a model to explain how daunorubicin and doxorubicin might inhibit phage propagation
- the hypothesis your experiment is testing
- the independent and dependent variables of that experiment
- both positive and negative controls

- a description of how you will construct any necessary strains and plasmids
- potential outcomes of your experiments, and how you will interpret them

---

LECTURE 19: CRITICAL READING (BACTERIAL METABOLISM)

EXPECTATIONS

As a reminder, to prepare for any journal club discussion of a paper, you should do the following:

1. Read the whole paper, including all the figures and supplemental data.
2. Make notes of:
    - What is the central **question** of this paper?
    - Is the experimental design clear and appropriate to address that question?
    - Do you understand the methods used?
    - Are the data clearly presented, with appropriate statistics?
    - Do you agree with the conclusions the authors came to based on their data?
    - What additional experiments (if any) do you think would be helpful?

Remember that your grade in this class depends on your preparation for and participation in class discussion, so be sure that you have read the paper and understand the figures. If you have questions, you are free to ask me or talk among yourselves before class.

CRITICAL READING PAPER

Price-Carter *et al.* (2001) "The alternative electron acceptor tetrathionate supports B_{12}-dependent anaerobic growth of *Salmonella enterica* serovar Typhimurium on ethanolamine or 1,2-propanediol." J Bacteriol 183(8):2463-75

In class, I will start by making a short presentation of background information to help put this paper in context. Then I will have slides prepared for each figure in the paper (including Supplemental Figures), and each of you will take turns presenting individual figures to the rest of the class and leading discussion of that figure. We will finish with a discussion of the paper as a whole.

You may also find the following publication interesting to help explain the importance of the paper above, although we will not be discussing it in detail in class:

Thiennimitr *et al.* (2011) "Intestinal inflammation allows *Salmonella* to use ethanolamine to compete with the microbiota." Proc Natl Acad Sci USA 108(42):17480-5.

SUMMARY OF EXPERIMENTAL DESIGN PRINCIPLES

This page is simply a compilation of the rules for experimental design discussed in the Scientific Process sections of the previous chapters. **This is the single most important thing I want you to take away from this class.** Proposing good experiments to test valid hypotheses is the key to good science, the central element of grant writing, and is something you will have to do for your qualifying exam (and the rest of your career), so it's an important skill to cultivate.

OBSERVATIONS

When describing a set of observations that you plan to make, you should explain:

- What will you be measuring, and how will you measure it?
- Is it a qualitative or a quantitative measurement?
- When and how often will you measure it?

MODELS

When proposing a model, it should:

- incorporate all of the available data
- propose a mechanism that explains the behavior of the system
- make testable predictions about the system being studied

HYPOTHESES

When proposing a hypothesis:

- it should test a specific aspect of a model
- it should be falsifiable
- you should be able to propose a set of observations that can be used to test that hypothesis

EXPERIMENTS

When designing an experiment, you should:

- define the dependent and independent variables
- explain what you will measure and how (*i.e.*, what observations will you make?)
- describe both positive and negative controls
- describe the possible outcomes of the experiment and what they would mean for your hypothesis
- state whether the experiment will determine correlation or causation

ALTERNATIVE APPROACHES

There is more than one way to answer any scientific question. You should be able to:

- design more than one distinct experiment to test a given hypothesis
- be able to explain the strengths and weaknesses of each approach

GRADING RUBRIC FOR WEEKLY QUIZ PROBLEMS

Each week you will be responsible for solving an experimental design problem on your own, and this page describes how those quizzes will be graded.

Each week's problem is worth **20 points** in total, and will involve examining a data set, proposing a model to explain those data, and then designing an experiment to test a hypothesis based on that model. Each week you also have the possibility of earning **1 bonus point** for a particularly clever, creative, or elegant solution to the experimental problem.

	4 points	**3 points**	**2 points**	**1 point**	**0 points**
Model	Incorporates all of the data available and proposes a mechanism.	Lose 1 point each for: • does not incorporate all available data • does not propose a mechanistic explanation of the data • is biologically implausible			No model proposed.
Hypothesis	Falsifiable hypothesis that tests a specific and important aspect of the model.	Lose 1 point each for: • hypothesis is not falsifiable • hypothesis does not test a specific aspect of the model • hypothesis does not test an important part of the model			No hypothesis proposed.
Experimental Design	Will effectively test the hypothesis, is practical, and has clearly-described observations.	Lose 1 point each for: • will not effectively test the stated hypothesis • unclear description of the observations or measurements necessary • is not practical with standard laboratory techniques			No experiment described.
Controls	All necessary positive and negative controls clearly described.	Lose 1 point each for: • no positive controls • no negative controls • missing controls essential to the interpretation of the proposed experiment			No controls described.
Interpretation of Results	All possible results described, with explanations of the meaning of those results for the model. State clearly whether expected results establish correlation or causation.	Lose 1 point each for: • only some possible results described • unclear description of interpretation • not stating clearly or correctly whether results establish correlation or causation			No possible results described.

Note: Week 2's quiz will be graded slightly differently, since it will be a genetic engineering problem rather than an experimental design problem. The rubric for that quiz will be included in the quiz itself, and will be somewhat simpler.

GLOSSARY

2-dimensional gel electrophoresis	a largely obsolete method for direct quantification of proteins that works by separating proteins by both size and isoelectric point
abstract	a short summary of a paper
acetylation	modification of a molecule by addition of an acetyl group
acidophilic bacteria	bacteria that live at very low pH (< 3)
acyl carrier protein	a protein with a CoA cofactor which is used as the platform for synthesis of fatty acids or polyketides
adherence	the process by which bacteria stick to surfaces
adhesin	a protein or polysaccharide produced by a bacteria specifically for the purpose of attaching to a surface
aerobic respiration	a respiratory pathway in which the terminal electron acceptor is oxygen
aerotaxis	movement towards air, or oxygen
alignment	a visual representation of homology between DNA, RNA, or protein sequences
alkaliphilic bacteria	bacteria that live at very high pH (> 10)
allele	a version of a gene, typically differing from other alleles by only a small number of mutations
allele number	a notation used to distinguish between different mutations of the same gene
allosteric effector	a molecule that controls the activity of a protein by interacting with it at a site distant from its active site
allostery	a regulatory mechanism by which a molecule controls protein activity by non-covalently binding to a site that is not the active site of that protein
alternative electron acceptors	catch-all term for terminal electron acceptors in anaerobic respiratory pathways
alternative sigma factors	sigma factors responsible for recognizing promoters other than those recognized by the housekeeping sigma factor; often involved in stress response or development
amidase	an enzyme that breaks an amide bond
amino sugar	a monosaccharide modified by attachment of an amino group
amphilophotrichous	multiple flagella at both poles
amphipathic	a molecule that has both hydrophobic and hydrophilic functional groups
amphitrichous	a single flagellum at each pole
anabolism	the biosynthesis of molecules from precursor intermediates
anaerobic respiration	any respiratory pathway in which the terminal electron acceptor is not oxygen
anomeric carbon	the carbon of a monosaccharide adjacent to the oxygen atom and whose orientation varies in α and β versions of a sugar
anoxygenic photosynthesis	photosynthetic pathways which use a molecule other than H_2O as an electron donor

anti-Shine-Dalgarno sequence	the sequence of the 3' end of the 16S ribosomal RNA, which binds to the ribosome binding site in mRNA
anti-terminator	a regulator that prevents transcription termination
antibiotic resistance	acquisition of mutations that allow bacteria to detoxify or export antibiotics
antibiotic resistance cassette	a gene encoding a product that confers antibiotic resistance, along with all of the additional sequences needed to ensure its expression
antigenic variation	the ability of some bacteria to alternate between expression of different surface components, as a method of avoiding the immune system
antiporter	a transporter that transports two molecules across a membrane in opposite directions
asymmetric cell division	a developmental process in which the two daughter cells are genetically identical but have different phenotypes
attachment site	(also "*att* site") the specific DNA sequence at which lysogenic bacteriophage (and some transposons) insert themselves into their host chromosomes
attractant	a chemical or other stimulus that causes bacteria to move towards higher concentrations of that stimulus
autoaggregation	(also "flocculation") the ability of bacteria to stick to other cells of the same species and strain
autoinducer	the secreted signal molecule secreted by cells in quorum sensing regulatory systems
autoinducer 2	(also "AI-2") a specific boron-containing quorum sensing molecule produced by many Gram-negative and some Gram-positive bacteria
autolysin	an enzyme that disrupts the cell envelope of the organism whose genome encodes it
autotransporters	another name for type V secretion systems; proteins that catalyze their own transport across the outer membrane of Gram-negative bacteria
autotroph	an organism that can fix CO_2 into organic carbon
auxotroph	a mutant that requires a particular nutrient
bacterial artificial chromosome	(also "BAC") a plasmid based on the F factor that can be used to clone very large inserts
bacterioneuston	(also "surface micro-layer") the top millimeter of the ocean, inhabited by a relatively dense population of bacteria
bacteriophage	a virus that infects bacteria
bacteriorhodopsin	a transmembrane protein that absorbs light and uses the energy of that light to pump protons across the membrane
bactofilin	a family of filamentous cytoskeletal proteins found only in bacteria
binning (of data)	grouping quantitative data into categories for ease of analysis
bioactive compounds	secondary metabolites that have a known effect of some kind on cells
biochemistry	the study of the physical properties of biological molecules
biofilm	a multi-cellular bacterial community attached to a surface
biofilm dispersal	the regulated process by which planktonic cells are released from biofilms

biosynthetic gene cluster	(also "BGC") genetic loci dedicated to the production of a specific secondary metabolite
blunt end	a double strand break with no sticky ends, produced by some restriction enzymes
bradytroph	a mutant that grows slowly without a particular nutrient
brute-force approach	an inelegant, labor-intensive experimental design
capsule	the polysaccharide coating of many bacteria, generally attached to the cell surface and often important for recognition by the adaptive immune system
carbohydrate	organic compound composed only of carbon, hydrogen, and oxygen
carbon catabolite repression	(also "CCR" or "catabolite repression") a regulatory system that represses expression of genes for catabolism of non-preferred carbon sources in the presence of preferred carbon sources
catabolism	the breakdown of molecules, yielding their component parts and energy
causation	proof that one phenomenon directly leads to another
cDNA library	a pool of plasmids containing many different cloned DNA inserts derived by reverse transcription from an organism's mRNA
cell biology	the study of the molecular mechanisms that determine the shape, size, division, differentiation, and development of (bacterial) cells
cell envelope	the boundary separating the inside of a cell from the outside of a cell, consisting of membrane(s), cell wall structures, S-layers, *etc.*
cell wall	a rigid polymeric shell outside of a cell's cytoplasmic membrane that maintains the cell's shape and integrity
chain length factor	a component of iterative PKS pathways that determines how many long the final product will be
chaperone	a protein that binds to other proteins and modulates their folding or solubility
chaperone-usher secretion system	a variation on type V protein secretion systems that involves three proteins
chemical mutagen	a chemical that damages DNA, resulting in mutations
chemically competent cells	bacteria treated (often by rinsing in cold $CaCl_2$ followed by brief heat shock) to make them capable of taking up DNA directly from their environment (transformation)
chemiosmotic theory	a biophysical theory that describes how potential energy is stored in biological systems as an electrochemical gradient of ions across a lipid bilayer membrane
chemoautotroph	a non-photosynthetic organism that can fix CO_2 into organic carbon
chemotaxis	the ability of an organism to move along a chemical gradient
chimeric protein	see "protein fusion"
ChIP-seq	(also "chromatin immunoprecipitation sequencing") an *in vivo* technique to identify all of the genomic binding sites of a DNA binding protein using next generation sequencing
chromosome	a large DNA molecule containing essential gene(s) and usually present in single copy
cistron	an obsolete synonym for gene

cloning	incorporating a gene into a plasmid for expression
cloning strain	a strain of *E. coli* (usually) that contains mutations to improve competence and plasmid stability, making molecular cloning easier
cloning vectors	plasmids used to express genes in bacteria
codon optimization	changing the sequence of a gene so that it uses only the most abundant codon for each amino acid; species specific
codon usage	a measure of how well particular codons are translated in a given organism or how frequently they occur in a given genome
cofactor A	(also "CoA") a thiol-containing cofactor involved in many metabolic reactions
compatibility group	see "origin of replication"
competence pili	a subset of type IV secretion system pili that are involved in DNA uptake
competent cells	bacteria capable of taking up DNA directly from their environment (transformation)
complementation analysis	an experimental design that establishes genetic causation by removing and replacing individual genes
conditional phenotype	a phenotype that is only observed under specific growth conditions
confocal microscopy	microscopy technique that constructs a 3D image from sequential images taken in different focal planes
conjugation	DNA transfer between cells *via* pili; requires *tra* factors, an origin of transfer, and physical contact between cells
consensus sequence	the most common or average sequence for a particular gene or locus
conserved residues or nucleotides	(also "conservation") protein, RNA, or DNA sequence features that do not change (or change slowly) over evolutionary time
constitutive promoter	a promoter that is always active and expresses genes under its control at a constant level
constitutively active	always expressed or functioning at a constant level
constitutively inactive	never expressed or never functional
contact-dependent inhibition	another term for type VI protein secretion
contractile phage	bacterial viruses whose tails contract upon contact with a host cell, driving DNA injection
control	a treatment included in an experiment to make sure that the experiment is working as intended
copy number	how many of a DNA molecule (typically a plasmid) are present per cell
correlation	the observation that two or more phenomena appear or change together
corresponding author	the person who gets contacted about a paper if there are any questions, typically the head of the lab where the work was done
cortex	the thick peptidoglycan cell wall layer of an endospore
cos site	a site that allows a plasmid to be packaged in λ phage particles

cosmid	a plasmid with a *cos* site
cotranscribed	genes adjacent to each other on the chromosome, and transcribed in the same direction
cotransduction frequency	how often two genes or mutations will be transferred simultaneously by transduction, a function of transducing phage packaging size and the distance between the genes or mutations on the chromosome
counter-selectable marker	a gene encoding a product which allows you to select for cells that don't contain that gene; a conditionally lethal gene
coupling sites	points in a respiratory electron transport chain at which a PMF is generated
crescentin	(or "CreS") a cytoskeletal protein responsible for the curved shape of *Caulobacter crescentus* cells, homologous to eukaryotic intermediate filaments
CRISPR	(also "clustered regularly interspaced short palindromic repeats") a system that uses short guide RNAs to direct the activity of a nuclease (usually Cas9) to specific sites in a DNA (or RNA) molecule
CRISPR array	the series of repetitive DNA sequences that incorporate guide RNAs in natural CRISPR systems
CRISPR-associated proteins	the various proteins that are part of natural CRISPR systems; Cas9 nuclease is the most important for biotechnological purposes
cryptic metabolites	secondary metabolites with no known function
curli	proteinaceous fibers attached to the outer surface of Gram-negative bacteria, a kind of functional amyloid
cytoplasm	the aqueous interior of a living cell
cytoskeleton	filamentous proteins involved in cell shape, chromosome segregation, *etc.*
data	high-quality, carefully recorded observations
daughter cell	the cells derived from a bacterial cell division event
defective prophage	see "stable lysogen"
degeneracy	the fact that multiple codons can encode the same amino acid
degron	protein sequences recognized by proteases as signals for protein degradation
deletion	the removal of DNA sequence from a gene
dependent variable	the variable(s) measured by the experimenter during an experiment
derepression	the effect of inactivating a negative regulator
development	a process by which genetically-identical cells express different phenotypes
diderm	a bacterial cell with two lipid membranes (an outer and an inner), typically Gram-negative
differentiation	a developmental process in which a proportion of cells in a population change their phenotypes
divergently transcribed	genes adjacent to each other on the chromosome, but transcribed in opposite directions

divisome	the protein complex responsible for the process of cell division
DNA ligase	an enzyme that joins two DNA molecules together
DNA methylase	(see "restriction methylase") an enzyme that methylates specific sequences in DNA
DNA microarray	a direct method to detect RNA by hybridizing it with an array of oligo probes of known sequence; largely obsolete
DNA recombination	see "homologous recombination"
domain	a structurally-conserved element of a protein, usually with a specific function
double-crossover recombination	a recombination event that requires two independent homologous recombinations, such as integrating a linear DNA fragment into a circular chromosome
downstream gene	a gene encoded 3' of the gene being discussed on an mRNA
duplication	a mutation that results in multiple copies of a DNA sequence
effector proteins	secreted proteins that have effects on other cells, often host cells or competing bacteria
electron acceptors	molecules that become reduced during respiration
electron donors	molecules that become oxidized during respiration
electron shuttles	(also "electron carriers") small molecules that can be oxidized and reduced to carry electrons between proteins
electron transport chain	a pathway that links oxidation reduction reactions in a bilayer membrane to generation of a PMF
electronic table of contents	a service that emails you the list of papers published in a journal when each issue becomes available
electrophoretic mobility shift assay	(also "EMSA" or "gel shift assay") a direct measurement of the binding affinity of a protein for a nucleic acid molecule, using gel electrophoresis to separate bound and unbound nucleic acids by size
electroporation	a method for transformation in which cells are mixed with DNA and subjected to an electric shock
ELISA	(also "enzyme-linked immunosorbent assay") an assay that uses immobilized antibodies to detect and quantify antigenic substrates
elongasome	the protein complex responsible for organizing synthesis of new cell wall material as the cell grows, specifically in rod-shaped bacteria
endonuclease	a nuclease that cleaves within a DNA or RNA molecule
endonuclease cleavage site	a DNA or RNA sequence that is recognized by an endonuclease
endopeptidases	proteases that break peptide bonds within proteins
endospore	a spore formed inside of the mother cell, typical of Gram positive *Bacillus* and *Clostridium* species
endotoxin	see "lipopolysaccharide"
enrichment	a procedure that increases the proportion of mutants of interest in a population

Entner-Douderoff pathway	catabolic pathway that breaks down glucose into pyruvate
enzyme activity assay	a direct biochemical measurement of protein activity, specifically for proteins that catalyze chemical reactions
epigenetic	modifications of DNA or other cellular components that result in a (usually heritable) change in phenotype without a change in the DNA sequence
episome	see "plasmid"; obsolete
epitope tag	a short peptide sequence that can be fused with proteins of interest to allow their detection or purification with commercially available antibodies
essential gene	a gene that cannot be knocked out; encodes a function the cell depends on
exonuclease	a nuclease that degrades a DNA or RNA molecule from one end
exotoxin	a toxic protein or compound secreted by a bacterial cell
experiment	a test of the effects of a specific manipulation on a system
extracellular DNA	(also "eDNA") DNA released from bacteria for functional purposes, for example as a component of a biofilm matrix
extracellular polysaccharides	(also "EPS") polymers of sugars produced by bacteria and secreted out of the cell
f1 origin	a site that allows a plasmid to be packaged as concatenated single-stranded DNA when the host bacterium is infected with bacteriophage f1
false negative result	an erroneous result that looks like nothing happened when something did
false positive result	an erroneous result that looks like something happened when it did not
falsifiable	a property of a useful hypothesis – can it be proved wrong?
fatty acid	a carboxylic acid with a long, hydrophobic hydrocarbon chain; a linear lipid
fermentation	a metabolic pathway in which the byproducts of initial catabolic pathways are reduced using the reducing equivalents generated during those pathways
fimbrial adhesin	fibrous adhesins that extend well away from the cell, usually called pili
first author	typically the person who did most of the experiments on a paper; may have multiple "first authors" who contributed equally to the work
flagella	helical rotary filaments used by many bacteria for motility
flagellar basal body	a specialized type III secretion system dedicated to the export and assembly of flagella
flagellin	the protein monomer that makes up the bulk of the flagellar filament
flippase	an enzyme that moves a lipid from one side of a bilayer membrane to the other side
floc	a clump of bacteria stuck to each other in suspension
fluidity	a property of lipids that depends on their melting temperature
fluorophore	a fluorescent molecule (either a small molecule or a protein)
forespore	the specialized cell that will become an endospore during sporulation
frameshift mutation	insertion or deletion of 1 or 2 nucleotides (or any number not divisible by 3)

fruiting body	a multicellular structure formed during development (*e.g.* in myxobacteria) which contains spores or other resting cells
FtsZ	bacterial tubulin homolog essential for cell division in most bacteria
functional amyloid proteins	proteins that aggregate into characteristically extremely stable, sticky β-sheet rich structures, often involved in adherence and biofilm formation
functional redundancy	two genes products that carry out the same or overlapping functions
functional RNA	RNA that is not mRNA; includes ribosomal RNA, transfer RNA, small regulatory RNAs, and ribozymes
fusion protein / tag	see "protein fusion"
gain-of-function mutation	a mutation that gives a gene product new or enhanced abilities
gas vesicles	bacterial organelles that contain gas, involved in floating
gene	a DNA sequence encoding a functional product
gene knockdown	artificially reducing the expression of a gene without constructing a null mutation; useful for studying essential genes, for example
gene knockout	see "null mutation"
gene product	an RNA or protein encoded by a gene
generalized transducing phage	phage which are able to package random fragments of DNA from the chromosome of their host cell into virus particles
genes of unknown function	genes with no currently known role in the cell
genetic toolkit	ways to put new DNA into an organism or to change the DNA that it already has
genetics	the science of how heritable characteristics are passed from one organism to another
genome	the complete DNA sequence of a cell
genomic library	a pool of plasmids containing many different cloned inserts derived from an organism's genomic DNA
genotype	the sequence of the genome of an organism
germinate	the process by which a spore turns into a growing vegetative cell
gliding motility	any of several mechanisms by which bacteria move along surfaces without obvious external appendages
global regulator	a regulator that controls many genes or gene products from around the genome
glucokinase	an enzyme that phosphorylates glucose to glucose-6-phosphate
gluconeogenesis	anabolic pathway for the synthesis of glucose; mostly the reverse of glycolysis
glycolysis	catabolic pathway that breaks down glucose into pyruvate
glycosidic bonds	covalent bonds between monosaccharides in a polysaccharide
glycosyltransferase	an enzyme that adds a monosaccharide to one end of a polysaccharide chain
glyoxylate bypass	(also "glyoxylate shunt") catabolic variation of the TCA cycle that bypasses several steps to allow catabolism of 2-carbon compounds

Gram stain	a method for differential staining of bacteria
Gram-negative	a cell that stains pink in the Gram stain, often a diderm bacterium
Gram-positive	a cell that stains purple in the Gram stain, often a monoderm bacterium
guide RNA	a short sequence that serves to direct Cas9 nuclease to a specific target site
gum	a secreted polysaccharide that is sticky
hairpin	a DNA or RNA structure that is folded into a small, stable loop
heterocyst	a terminally-differentiated cell specialized in nitrogen fixation formed by filamentous cyanobacteria
heterolactic fermentation	a fermentative pathway in which pyruvate is reduced to lactate, ethanol, and CO_2
hexose	a monosaccharide containing 6 carbons
high-energy phosphate bonds	the phosphate bonds in ATP and other NTPs, notable for their high phosphoryl group transfer potential
histidine kinase	an enzyme that phosphorylates a histidine, but more specifically usually refers to the sensor component of a two-component regulatory system
holdfast	the sticky anchor at the end of the stalk in *Caulobacter crescentus*
Holliday junction	the crossover point between two homologous DNA sequences that is the essential intermediate in homologous recombination
homolactic fermentation	a fermentative pathway in which pyruvate is reduced to lactate
homologous recombination	a DNA repair mechanism that allows the exchange of sequences from one DNA molecule to another; requires sequence homology
homologs	(also "homologous genes" or "homologous proteins") genes with a common evolutionary ancestor, inferred from sequence homology
homology	a measure of how similar two DNA, RNA, or protein sequences are
horizontal gene transfer	the acquisition of genetic material from a phylogenetically distant organism
host range	the list of different species a particular plasmid can replicate in
housekeeping sigma factor	the most abundant sigma factor in the cell, and the one responsible for recognizing most promoters
hypothesis	a prediction made by a scientific model, a possible answer to a scientific question
immunity proteins	proteins encoded by bacteria which prevent those strains from killing themselves, for example with type VI secretion systems or bacteriocins
immunoblot	see "western blot"
impact factor	the number of citations of papers in a journal over the previous 2 years, divided by the number of papers published in that journal in that time
in-frame	denotes DNA sequences whose codons are lined up with each other so that a continuous protein is produced from them during translation
incompatible plasmids	plasmids with the same origin of replication and / or the same selectable marker
independent variable	the variable(s) changed by the experimenter during an experiment

inducer	a compound that can be added to cells to control the activity of an inducible promoter
inducer exclusion	a mechanism of catabolite repression in which the import of alternative carbon sources is inhibited in the presence of preferred carbon sources
inducible promoter	a promoter that can be turned on or off by the addition of inducers; this term is usually used in reference to promoters in plasmids
initiating nucleotide	the first nucleotide of a transcribed RNA
inner leaflet	of a lipid bilayer; the phospholipids in the half of the membrane facing the cytoplasm (or periplasm of an outer membrane)
insertion	the addition of extra DNA sequence into the chromosome
integral membrane protein	a protein that is embedded in a lipid bilayer and crosses it at least once
interactome	the complete set of protein-protein interactions in an organism
intragenic suppressor	a second mutation in a mutated gene that reverses the phenotype of the mutant
intrinsic terminator	a stable, GC-rich stem-loop RNA structure, followed by several uracil residues, that leads to transcription termination
ionophore	chemical that bind to cations and allows them to diffuse through membranes, can be used to collapse ion gradients
isogenic strains	strains that are identical except for the specified mutations
isozymes	non-homologous enzymes in the same organism that catalyze the same reaction
journal club	a group meeting in which papers from the (usually) recent scientific literature are discussed in detail
kilobase pair	1,000 base pairs
kinase	an enzyme that adds phosphate groups to a substrate
knockout collection	a complete set of null mutants in a particular strain, each lacking one non-essential gene
leader peptide	a short protein encoded at the beginning of an operon, often as part of a transcriptional attenuation regulatory mechanism
lethal mutation	a mutation that kills the cell
linkage mapping	an obsolete method of determining the location of mutations by how often different genes are cotransduced by generalized transducing phage
linked marker	a selectable marker located in the genome close to a mutation of interest
Lipid A	the innermost lipid-disaccharide component of lipopolysaccharide
lipid bilayer membrane	a biological membrane made up of amphipathic lipids that assemble into sheets with their hydrophobic regions inside
lipopolysaccharide	(or "LPS") the complex sugar-lipid molecule that makes up most of the outer leaflet of the outer membrane in diderm bacteria
lipoprotein	a protein with a covalently-attached lipid group (a post-translational modification)
lipoteichoic acid	(or "LTA") polymer of repeating sugars and phosphate groups anchored in the cell membrane of monoderm bacteria by lipid groups

local regulator	a regulator that controls only a small number of genes or loci, often including the regulator itself
localized mutagenesis	random mutagenesis of a single gene or locus, as opposed to the entire genome
locus	a location on a chromosome; could be a gene, an operon, a regulatory site, *etc.*
locus tag	a unique identifier for a gene, used in genome sequencing projects
lophotrichous	multiple flagella at one pole
lysogeny	a bacterial cell containing a prophage
lysogenic phage	a bacteriophage able to integrate itself into the chromosome of a host cell
lytic transglycosylase	an enzyme that breaks the bonds between MurNAc and GlcNAc in peptidoglycan
magnetotactic bacteria	bacteria that are able to orient themselves along magnetic field lines
magnetotaxis	movement along magnetic field lines
mass spectrometry	a powerful technique for determining the molecular weight of molecules
material transfer agreement	(also "MTA") paperwork necessary to transfer research materials from one university to another
matrix	the extracellular components (polysaccharides, eDNA, and proteins) of a biofilm
megabase pair	1,000,000 base pairs
megasynthase	a very large modular enzyme for synthesis of secondary metabolites by PKS, NRPS, or hybrid PKS/NRPS pathways
merodiploid	a strain that contains two copies of a gene (often one on the chromosome and one on a plasmid, but potentially both in the chromosome), usually two different alleles
metabolic flux	a measurement of how active a particular enzyme or pathway is within a cell
metabolite	a small molecule produced by a cell or used as an intermediate in a cellular pathway
metabolome	the set of all small molecules (metabolites) in a cell
metabolomics	methods for measuring large numbers of metabolites in a cell simultaneously
metagenome	the DNA sequences of a community of organisms
metatranscriptome	mRNA sequences derived from a community of organisms
methylation	covalent addition of a methyl group to a protein or DNA molecule
Michaelis constant	(also "K_m") the concentration of substrate at which an enzyme's reaction rate V is half of V_{max}
microcolony	a small group of cells clumped together on a surface, can be a precursor of biofilm formation
microfluidics	a technology that examines bacteria in very small volumes and with precisely-controlled flow conditions
mineralization	the the deposition of carbonate minerals in a biofilm, solidifying the matrix into a rock-like state
minimal media	growth media that contains only the compounds a particular species needs to grow

minireview	a short review, either giving a brief introduction or reporting recent progress in a field
missense mutation	a mutation of an amino acid encoding codon to a different amino acid encoding codon
mixed acid fermentation	a fermentative pathway in which pyruvate is reduced to lactate, formate, acetate, ethanol, CO_2, and H_2
model	a mechanistic explanation of a system, based on data from observations and experiments
model organism	an easily-studied species, the properties of which are used to infer the properties of less easily-studied (or just less studied) organisms
molecular biology	(see "molecular genetics")
molecular genetics	genetics with an understanding of the biochemical nature of genes
monocistronic	an mRNA encoding one gene
monoderm	a bacterial cell with one lipid membrane, typically Gram-positive
monophyletic	a group of organisms descended from a single common ancestor
monosaccharide	(also "sugar") 3- to 7-carbon carbohydrates distinguished by the arrangement and chirality of their various hydroxyl groups
monotrichous	a single flagellum at one pole
mother cell	the cell that nurtures the formation of an endospore during sporulation
motor proteins	proteins that convert proton motive force into rotary motion of the flagellar basal body
motor switch complex	the components of the flagellar basal body which are involved in changing the direction of rotation of the motor
MreB	bacterial actin homolog
mRNA stability	how long a particular mRNA remains in the cell before being degraded
multicopy suppressor	a gene that reverses the phenotype of a mutation in a different gene when overexpressed
multiple alignment	an alignment of more than two sequences
multiple cloning site	(also "MCS") a small region of a plasmid with several closely spaced restriction sites
murein	see "peptidoglycan", somewhat outdated term
mutagen	a treatment that damages DNA, resulting in mutations
mutagenesis	the act of making mutations in an organism
mutant	an organism containing a mutation
mutant hunt	an experiment intended to identify mutations that affect a particular phenotype
mutation	a change in the DNA sequence of an organism
mutation rate	how quickly mutations accumulate in a population
mutator strain	a bacterial strain defective in DNA repair; useful for random mutagenesis of plasmids
mycolic acid	long-chain, extremely hydrophobic lipid unique to the mycobacteria

myxospores	stress-tolerant resting cells produced during development of myxobacteria
N-acetylation	modification of a molecule by addition of an acetyl group to a nitrogen atom
N-acylation	covalent addition of acyl groups to lysine residues in proteins
natural products	any molecule produced by a living organism, usually means secondary metabolites
naturally competent cells	bacteria capable of taking up DNA directly from their environment (transformation) without special treatment
negative control	a control that tests for the possibility of false positive results in an experiment
negative regulator	a regulator that represses the system being studied
next generation sequencing	(also "NGS") any of a variety of methods of DNA sequencing that read the sequence very large numbers of (typically) very short DNA fragments
non-fimbrial adhesin	adhesins that are not pili
non-ribosomal peptide synthetase	(also "NRPS") a biosynthetic pathway that construct complex molecules out of simple amino acid components
nonsense mutation	a mutation of an amino acid encoding codon to a stop codon
northern blot	a direct method to detect RNA by probing with radioactively labeled oligos; obsolete
nuclease	an enzyme that degrades DNA or RNA by breaking the bonds between nucleotides
nucleoid occlusion	the prevention of Z-ring formation in some bacteria in regions of the cell containing large amounts of DNA
null mutation	a mutation that inactivates a gene product
O-acetylation	modification of a molecule by addition of an acetyl group to an oxygen atom
O-antigen	the outermost polysaccharide chain component of lipopolysaccharide
observation	a measurement of some feature of the objective universe
oligonucleotide	(also "oligo") a short, artificially synthesized DNA molecule
open reading frame	the protein-coding sequence of a gene
operator sequence	the DNA sequence to which a regulator binds
operon	several genes encoded on the same mRNA
origin of replication	(also "*ori*" or "*oriC*") the site which determines the ability of a plasmid to replicate within a cell, it's copy number, and host range
origin of transfer	(also "*oriT*") a DNA sequence allowing a plasmid to be mobilized by conjugation
orthologs	(also "orthologous genes" or "orthologous proteins") homologs in different genomes
outer leaflet	of a lipid bilayer; the phospholipids in the half of the membrane facing the outside of th cell
overexpression strain	a strain for use with overexpression vectors, optimized for very high level production of cloned gene products
overexpression vector	a plasmid specifically designed to allow very high level production of a cloned gene product

oxygenic photosynthesis	photosynthetic pathways which use a H_2O as an electron donor
pairwise alignment	an alignment between two sequences
paralogs	(also "paralogous genes" or "paralogous proteins") homologs in the same genome
parent strain	see "wild-type"; could also denote a strain from which a particular mutant strain was constructed
passenger domain	the domain of a T5SS autotransporter which is passed through the outer membrane
pellicle	a biofilm that forms at a liquid-gas interface
penicillin-binding protein	(also "PBP") see "transpeptidase"
pentose	a monosaccharide containing 5 carbons
pentose phosphate pathway	(also "PPP") catabolic pathway that breaks down glucose into fructose-6-phosphate and glyceraldehyde-3-phosphate
peptide	a short protein
peptidoglycan	polymer of MurNAc and GlcNAc sugars, cross-linked by peptides; the main component of the bacteria cell wall
percent identity	what percentage of positions in an alignment of two homologous protein or nucleic acid sequences contain the same amino acid or nucleotide in both sequences
percent similarity	what percentage of positions in an alignment of two homologous proteins contain amino acids with similar chemical properties in both sequences
peripheral membrane protein	a protein that is associated with a lipid membrane, but not embedded in it
periplasm	(also "periplasmic space") the space between the inner and outer membranes of a diderm bacterium
peritrichous	flagella distributed across the cell surface
permissive temperature	for temperature sensitive mutants, the temperature at which the gene functions
persister cells	a proportion of a bacterial population, that due to their low levels of ATP, are able to survive and eventually recover from antibiotic treatment
phage recombinase	a highly efficient recombinase derived from a lysogenic bacteriophage
phagemid	a plasmid with an f1 origin
phenomenon	a measurable event in objective reality
phenotype	the measurable physical properties of an organism
phenotypic heterogeneity	a property of many clonal bacterial populations, where proportions of a genetically identical community will express different phenotypes
phosphatase	an enzyme that removes phosphate groups from a substrate
phospholipid	a hydrophobic lipid that contains one or more hydrophilic phosphate groups at one end
phosphorelay	a signaling network related to two-component regulatory systems, but containing more than two components
phosphorylation	covalent addition of a phosphate group to a molecule

phosphotransferase	an enzyme that transfers a phosphate group, some are involved in phosphorelay-type signaling pathways
phosphotransferase system	(also "PTS") sugar transport system that phosphorylates the substrate as it is imported
photoautotroph	a photosynthetic organism that can fix CO_2 into organic carbon
photoheterotroph	a photosynthetic organism that does not fix CO_2 into organic carbon (they use organic carbon sources instead)
photosynthesis	the process by which light is used as an energy source for generating a PMF or reducing equivalents
phototaxis	movement towards light
phylogenetic tree	a visualization of the evolution of organisms from common ancestors, typically based on the sequence homology of highly conserved genes
phylogeny	the evolutionary relationship between organisms or genes, inferred from homology
physiology	for the purposes of this class, the structure, metabolism, energetics, and development of bacteria as living organisms
pilot experiment	a quick experiment, meant to test the practicality of a more complex experiment
pilus	(plural "pili") a fiber or tube-like structure in which DNA is transferred from one bacterial cell to another (conjugation), or is involved in attachment to surfaces
planktonic cells	bacteria not currently attached to a surface
plasmid	a small DNA molecule capable of replicating in a bacterial cell
plasmid library	a pool of plasmids containing many different cloned inserts
plasmid map	a visual representation of a plasmid, with indications of important features and sites
pleiotropic phenotype	multiple, apparently unrelated phenotypes resulting from a single mutation
point mutation	a change in a single nucleotide in a genome
polar flagellum	a flagellum located at the pole of a rod-shaped bacterial cell
polarity	the fact that mutations of one gene in an operon can have effects on the expression of downstream genes in that operon
polycistronic	an mRNA encoding several genes
polyketide synthase	(also "PKS") a biosynthetic pathway that constructs complex molecules out of simple carboxylic acid components
polymerase chain reaction	(also "PCR") a very common method that uses DNA polymerase to amplify large amounts of a specific DNA molecule *in vitro*
polymicrobial	a biofilm or infection containing bacteria of more than one species
polyphyletic	a group of organisms descended from different ancestors
polysaccharide	a polymer of monosaccharides
porin	an integral membrane protein, usually in the outer membrane of diderms, that forms a channel through the membrane
positive control	a control that tests for the possibility of false negative results in an experiment

positive regulator	a regulator that activates the system being studied in response to a signal
post-translational modification	(also "PTM") a covalent modification of a protein that affects its activity
predatory bacteria	bacteria that obtain nutrients by attacking and killing other bacteria
predatory journal	a journal with no scientific standards that exists solely to make money
predictive power	the ability of a model to predict the behavior of reality
prestige journal	a journal which only publishes "high-impact" science; *Nature, Science, Cell*, *etc.*
primary literature	published papers directly reporting the results of scientific research
primer	see "oligonucleotide"
product inhibition	a property of some enzymes, whose reactions are slowed by high concentrations of product
programmed cell death	a developmental pathway which ends in the death of a cell
promoter	a DNA sequence that binds RNA polymerase and, potentially, regulators to control transcription of a gene
prophage	a bacteriophage that is integrated into a bacterial chromosome
prostheca	a membrane-enclosed, cytoplasm-containing bacterial appendage
protease	an enzyme that breaks peptide bonds in proteins
protein	a linear chain of amino acids, encoded by an mRNA and produced by a ribosome
protein fragment complementation	a technique for studying protein-protein interactions *in vivo* by dividing a reporter into two or more segments that only regain activity when brought into physical proximity
protein fusion	a single polypeptide encoded by sequence derived from more than one gene, or a protein artificially modified to add a small peptide sequence to its C- or N-terminal end
protein secretion	transport of proteins to cellular compartments other than the cytoplasm, including outside of the cell entirely
protein stability	how long a particular protein remains in the cell before being degraded
proteome	the complete set of proteins in a cell
proteomics	methods to quantify the entire set of proteins in a cell
proton motive force	(also "PMF") the energy source derived from the difference in H^+ concentration on either side of the cell membrane
prototroph	a strain that does not require a particular nutrient (compare to auxotroph and bradytroph)
pseudomurein	a peptidoglycan-like cell wall polymer found in some archaea
pseudopilus	a short pilus homolog involved in type II protein secretion systems
pulse-chase experiment	an experiment that briefly labels proteins and then follows their stability over time
pupylation	a posttranslational modification added to proteins in actinobacteria to direct their degradation by the bacterial proteasome
pyrophosphatase	an enzyme that hydrolyzes pyrophosphate to two orthophosphates

qRT-PCR	(also "quantitative reverse transcriptase PCR") a direct method to detect RNA by reverse transcribing it to DNA and amplifying it by PCR
qualitative measurement	a measurement that results in a categorical (non-numerical) value
quantitative measurement	a measurement that results in a numerical value
quinones	membrane-soluble electron shuttles essential for respiration (among other reactions)
quorum sensing	a mechanism depending on secretion of soluble signal molecules (autoinducers) that bacteria use to control gene expression in response to culture density
radiation	electromagnetic energy or energetic particles that damage DNA, resulting in mutations
random mutagenesis	any of a variety of methods of making mutations throughout a DNA molecule with no (or little) predetermined targeting
random walk	a pattern of movement that depends on random changes in direction
rare codons	codons that are not translated efficiently in an organism due to low numbers of tRNAs for that codon
reaction center	a chlorophyll-containing light-absorbing protein complex central to photosynthesis
reactive oxygen species	(also "ROS") H_2O_2, superoxide, singlet oxygen, and oxygen radicals; potent and toxic oxidants derived from oxygen
read-through transcription	transcription from the promoter of one gene that drives (often unwanted) expression of a downstream gene
recombinant DNA	a DNA molecule constructed with sequences from two or more different organisms
recombinase	an enzyme or enzyme system that catalyzes homologous recombination
recombination	see "homologous recombination"
recombineering	a method of constructing chromosomal mutations using phage recombinases and PCR products or oligos as templates
reducing equivalents	NADH, NADPH, reduced quinones, and other sources of electrons in a cell
repellant	a chemical or other stimulus that causes bacteria to move towards lower concentrations of that stimulus
replica printing	using a sterile piece of velvet as a printing block to transfer colonies to several different plates; useful for screens
replisome	the protein complex that catalyzes replication of the chromosome
reproducibility	a desirable property of experiments: they give the same result each time
respiration	a mechanism by which bacteria link the generation of a PMF to oxidation-reduction reactions by way of an electron transport chain
response regulator	the effector protein of a two-component regulatory system or phosphorelay
restriction enzyme	a nuclease that makes double strand breaks in or near a specific sequence in a DNA molecule
restriction methylase	a DNA methylase that blocks the activity of a particular restriction enzyme
restriction site	the DNA sequence recognized by a restriction enzyme
restrictive temperature	for temperature sensitive mutants, the temperature at which the gene does not function

reverse catabolite repression	catabolite repression in which the preferred carbon source is not glucose
reverse transcription	the production of DNA from an RNA template by reverse transcriptase
revertant	a mutation that reverses the phenotype of a different mutation
review	a paper summarizing previous research on a particular topic
Rho-dependent transcription termination	transcription termination driven by the Rho protein, which recognizes single-stranded RNA with no ribosomes attached
Rho-independent transcription termination	transcription termination at intrinsic terminators
ribonuclease	a nuclease that specifically degrades RNA
ribosomally synthesized and post-translationally modified peptides	(also "RiPPs") peptide-based secondary metabolites synthesized by processing of a ribosomally-translated protein
ribosome binding site	(also "RBS") a short AG-rich sequence required for ribosomes to interact with mRNA and start translation
ribosome profiling	an indirect method to measure protein abundance using next-generation sequencing to quantify the proportion of each mRNA in a cell which is bound by ribosomes
riboswitch	a regulator formed entirely from RNA structures in an mRNA
RNA sequencing	(also "RNA-seq") a direct method to detect RNA by next-generation sequencing
rotor	what is turned by a stator, in the case of flagella, the flagellar basal body
S-layer	a protein or glycoprotein layer making up the outer surface of a cell envelope
sacculus	the purified cell wall of a bacterium, with no cell inside it
Sanger sequencing	a common and inexpensive way of sequencing several hundred to 1000 bp of DNA
saturation	for lipids, the number of double bonds present; a fatty acid with no double bonds is "saturated", while a fatty acid with several is "polyunsaturated"
scientific literature	the whole body of published scientific work
scientific method	a systematic approach to uncover truths about objective reality
screen	a mutant hunt in which each cell or colony must be individually analyzed to determine whether it contains a mutation of interest
second messenger	a small molecule that allosterically regulates multiple proteins, often produced in response to stressful changes in the cell's environment
secondary metabolism	anabolic pathways for the biosynthesis of molecules that are not the precursors of proteins, RNA, DNA, cell walls, capsules, or membrane lipids
secondary metabolites	molecules that are not the precursors of proteins, RNA, DNA, cell walls, capsules, or membrane lipids
secondary mutation	(see "revertant" and" multicopy", "intra-", and "intergenic suppressor") a mutation that is selected for by the presence of a primary mutation
selectable marker	a gene encoding a product which allows you to select for cells containing that gene; most often a gene for antibiotic resistance

selection	a mutant hunt in which the wild-type dies and only mutants of interest survive
septum	the barrier that forms between two dividing cells
sequence logo	a visual representation of an alignment in which the relative frequency of particular nucleotides or amino acids is represented by letter size
serotype	classification system for bacteria based on reactivity to specific antibodies
serotyping	a method of distinguishing bacterial strains by the antibodies they react with
Shine-Dalgarno sequence	see "ribosome binding site"
short chain fatty acids	(also "SCFA") acetate, propionate, butyrate, and other small carboxylic acid lipids
shuttle vector	a plasmid used to move genes from one species to another, may have separate origins of replication for each species
sigma factor	(also "sigma subunit") a small protein component of RNA polymerase that determines the promoter sequence that will be bound
signal peptidase	an enzyme that cleaves signal peptides off proteins
signal recognition particle	(also "SRP") riboprotein complex that, with the Sec secretion system, is required for assembly of integral membrane proteins in the inner membrane
signal sequence	an N-terminal protein sequence that is recognized by cellular export machinery and directs the cell to secrete the protein
silent biosynthetic gene cluster	a BGC that isn't expressed under laboratory growth conditions
silent mutation	a mutation of an amino acid encoding codon to a different codon encoding the same amino acid
single nucleotide polymorphism	see "point mutation"
single-crossover recombination	a recombination event that requires only one homologous recombination event, such as integrating a circular plasmid into a circular chromosome
site-directed mutagenesis	(also "targeted mutagenesis") constructing a specific mutation at a specific site in a DNA molecule
slime	a secreted polysaccharide that is slippery
society journal	a journal published by a scientific professional society
sortase	an enzyme that covalently attaches secreted proteins to peptidoglycan in the cell wall
specialized transducing phage	lysogenic phage which are able to package some DNA from near their site of insertion into the chromosome of their host cell into virus particles
sphaeroplast	a cell from which the cell wall (and outer membrane, if it has one) has been removed
spontaneous mutagenesis	random mutations resulting from natural mistakes made by DNA polymerase during replication
spore	a metabolically-inactive, highly stress-tolerant resting state cell formed by some bacteria
spore coat	the outermost layers of an endospore, consisting of modified peptidoglycan and protein layers

sporulation	a process by which bacteria form metabolically inactive, stress-tolerant spores
sRNA	(also "small non-coding RNA") a regulatory RNA that interacts with mRNA to change its expression, often by targeting it for degradation
stable lysogen	a DNA element incorporated into the bacterial chromosome that is derived from a lysogenic bacteriophage, but lacks the ability to re-enter the lytic lifecycle
stator	the MotAB motor proteins that turn the flagellar basal body
sticky end	a staggered double strand break produced by some restriction enzymes
stringent response	a widely conserved starvation stress response pathway depending on the second messenger (p)ppGpp
subcloning	a protocol in which a DNA fragment from one plasmid is moved into another plasmid by restriction digestion and ligation
substrate analog	a non-natural molecule that can be acted on by an enzyme, often resulting in products that are easier to measure than the natural products
substrate inhibition	a property of some enzymes, whose reactions are slowed by high concentrations of substrate
substrate-level phosphorylation	reactions that produce ATP without being linked to electron transport (for example, during glycolysis)
suicide vector	a plasmid which can be introduced into a species, but does not replicate there, or one whose replication can be blocked under certain conditions (see temperature-sensitive origin of replication)
super-resolution microscopy	any of a variety of microscope technologies that allow visualization of individual fluorescent molecules in bacterial cells
swarm	a motile group of bacteria, moving across a surface
swarming	flagellum-dependent group motility along a surface
symporter	a transporter that transports two molecules across a membrane in the same direction
synthetic lethality	two genes which can be knocked out individually, but not simultaneously
syntrophy	metabolic symbiosis in which one organism consumes the metabolic endproduct of another organism
teichoic acid	polymer of repeating sugars and phosphate groups that makes up a substantial portion of the cell envelope of monoderm bacteria
temperature-sensitive mutant	a mutant that grows at low temperature, but not at high temperature; typically due to mutations that destabilize essential proteins
temperature-sensitive origin of replication	an origin of replication that only functions at low temperature, typical of some suicide vectors
Ter macrodomain	the region of the bacterial chromosome where DNA replication terminates
terminal differentiation	a developmental process in which a proportion of cells in a population change their phenotypes irreversibly
terminal electron acceptor	the last molecule to be reduced in a respiratory electron transport chain
terminator	a sequence which stops transcription

testable	see "falsifiable"
thylakoid membrane	an internal membrane structure, rich in photosynthetic reaction centers, in cyanobacteria
tolerance	the ability of non- or slowly-growing bacteria to survive treatment with antibiotics
tra functions	genes encoding the machinery that allows transfer of plasmids with an appropriate *oriT* by conjugation
transconjugant	a cell that has incorporated DNA delivered by conjugation
transcription	the production of mRNA from a DNA template by RNA polymerase
transcription elongation	the activity of RNA polymerase actively producing mRNA
transcription factor	a protein that binds to the promoter of a gene to control its transcription
transcription initiation	the process by which RNA polymerase begins transcribing a gene into mRNA
transcription termination	the process by which RNA polymerase releases DNA and stops transcribing
transcriptional activator	a transcription factor that increases transcription of a gene
transcriptional attenuation	a regulatory mechanism in which an mRNA can take on more than one structural conformation, one of which is an intrinsic terminator
transcriptional pause site	a DNA or RNA sequence where RNA polymerase briefly stops producing mRNA
transcriptional reporter fusion	an indirect method to measure transcription by placing an easily-measured gene product under control of a promoter of interest
transcriptional repressor	a transcription factor that reduces transcription of a gene
transcriptional start site	the point in a promoter sequence where RNA polymerase begins producing mRNA
transcriptome	the entire set of mRNAs in a cell
transcriptomics	methods to quantify the entire set of mRNAs in a cell
transductant	a cell that has incorporated DNA derived from a transducing phage
transduction	DNA transfer between cells mediated by bacteriophage
transformant	a cell that has incorporated DNA delivered by transformation
transformation	bacterial cells taking up DNA directly from their environment
transition	a mutation of a purine (A or G) to a purine or of a pyrimidine (T or C) to a pyrimidine
translatability	a measure of how easily an mRNA is translated into protein in a particular organism
translation	the production of protein from an mRNA template by ribosomes
translation elongation	the activity of ribosomes actively producing protein
translation initiation	the process by which ribosomes bind to mRNA and begin producing protein
translational reporter fusion	an indirect method to measure translation by placing an easily-measured gene product under control of the promoter and translation initiation signals of a gene of interest
transpeptidase	enzyme that catalyzes the formation of peptide crosslinks in peptidoglycan

transposon	(also "insertion element") a DNA sequence capable of inserting itself into another DNA sequence, often at random
transposon library	a pool of transposon mutants, each cell containing only one transposon, but with a total of tens or hundreds of thousands of different insertion sites
transposon sequencing	(also "Tn-seq", "INSeq", "TraDIS", or "HITS") a technique that uses next-generation sequencing technology to identify all of the insertion sites in a transposon library
transversion	a mutation of a purine (A or G) to a pyrimidine (T or C) or vice versa
treadmilling	by cytoskeletal proteins; rapidly polymerizing at one end while simultaneously depolymerizing at the other end
treatment	see "independent variable"
tri-carboxylic acid cycle	(also "TCA cycle" or "Krebs cycle") catabolic cycle that breaks down pyruvate to CO_2
twitching motility	a form of motility along surfaces dependent on the sequential attachment and retraction of type 4 pili
two-hybrid screening	see "protein fragment complementation"
two-partner secretion system	a variation on type V protein secretion systems that involves two proteins
Tyndallization	a method of sterilizing liquids by briefly boiling them on several consecutive days
uncoupler	chemical that bind to cations and allows them to diffuse through membranes, can be used to collapse ion gradients
untranslated region	the parts of an mRNA which do not encode protein; often include regulatory elements
UP element	AT-rich sequence upstream of the -35 site of a promoter that increases transcription 30 to 70-fold
upstream gene	a gene encoded 5' of the gene being discussed on a polycistronic mRNA
vector	see "plasmid"
vector-only control	a type of negative control in which a strains containing an empty plasmid is compared to the same plasmid containing a gene of interest
vegetative cell	a growing bacterial cell, as opposed to a spore
wall teichoic acid	(or "WTA") polymer of repeating sugars and phosphate groups covalently attached to the cell wall of monoderm bacteria
western blot	a direct method of detecting proteins using antibodies specific to those proteins
wild-type	a strain that does not contain a particular mutation of interest
Z-ring	the ring of proteins, organized by FtsZ, that is the necessary precursor to formation of the divisome and initiation of cell division