Foundations of Biology and Environmental Science: An Open-Access Encyclopedia

Compiled and edited by Nathan Brouwer

2022-01-06

Contents

Chapter 1

Introduction

This website is a compilation of materials open-access and/or free materials on biology, bioinformatics, and environmental science. Most of the pages are entries pages from Wikipedia or entries from the NHGRI's Talking Glossary (https://www.genome.gov/genetics-glossary). Some entries are from other sources, and I've contributed a things I've written myself.

Chapter 2

A

2.1 Alignments

From Sharber, W. Introduction to Sequence Alignments with Biopython. Towards Data Science. Medium. https://towardsdatascience.com/introduction-to-sequence-alignments-with-biopython-f3b6375095db

"When working with biological sequence data, either DNA, RNA, or protein, biologists often want to be able to compare one sequence to another in order to make some inferences about the function or evolution of the sequences. Just like you wouldn't want to use data from data tables where data was in the wrong column for analyses, in order to make robust inferences from sequence data, we need to make sure our sequence data is well organized or "aligned." Unfortunately, sequence data does not come with nice labels, like a date, miles per gallon, or horsepower. Instead, all we have is the position number in the sequence, and that is relative to that sequence only. Luckily, many sequences are highly conserved or similar between related organisms (and all organisms are related to some degree!). If we're fairly certain that we've obtained data from the same sequence from multiple organisms, we can put that data into a matrix that we call an alignment. If you're only comparing two sequences, it's called a pairwise alignment. If you're comparing three or more sequences, it's called a multiple sequence alignment (MSA).

"Using the positions and the identity of each molecule in the sequence, we can infer the relative placement of each molecule in the matrix. Sometimes there will be differences in the sequence, for example, in a position where most sequences are C, we find a sequence with a G. This is referred to as a single nucleotide polymorphism (SNP). In other times, we find that a sequence is missing a molecule that is present in the rest, or a sequence has an extra molecule. The former is a deletion, while the latter is an insertion, together referred to as "indels." When aligning sequences with indels, we must account for these extra or missing

molecules by adding gaps to the remaining sequences. These small differences are usually the interesting parts of sequence data because the variation is how we can make inferences on the function or evolution of the sequence…"

2.2 Talking Glossary: Allele (0.5 min)

Note: This glossary definition focuses on the traditional definition of an allele. More broadly, an allele is *any* genetic variant, whether it is in coding or non-coding DNA, impacts the phenotype of an organism or is neutral, involves a single base or many. This is mentioned in the last line of the abstract.

Abstract: "An allele is one of two or more versions of a gene [or, more generally, a locus]. An individual inherits two alleles for each gene [locus], one from each parent. If the two alleles are the same, the individual is homozygous for that gene [locus]. If the alleles are different, the individual is heterozygous. **Though the term allele was originally used to describe variation among genes, it now also refers to variation among non-coding DNA sequences** [emphasis added]."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/allele.mp3

Note Any time they say "gene", think "locus."

Transcript: ""Allele" is the word that we use to describe the alternative form or versions of a gene. People inherit one allele for each autosomal gene [locus] from each parent, and we tend to lump the alleles into categories. Typically, we call them either normal or wild-type alleles, or abnormal, or mutant alleles."

Leslie G. Biesecker, M.D.

Image (from) https://rarediseases.info.nih.gov/files/glossary/english/allele_sm.jpg

Chapter 3

B

3.1 Talking glossary: Bioinformatics

https://www.genome.gov/genetics-glossary/Bioinformatics

Transcript: "Bioinformatics is a field of computational science that has to do with the analysis of sequences of biological molecules. [It] usually refers to genes, DNA, RNA, or protein, and is particularly useful in comparing genes and other sequences in proteins and other sequences within an organism or between organisms, looking at evolutionary relationships between organisms, and using the patterns that exist across DNA and protein sequences to figure out what their function is. You can think about bioinformatics as essentially the linguistics part of genetics. That is, the linguistics people are looking at patterns in language, and that's what bioinformatics people do–looking for patterns within sequences of DNA or protein."

Christopher P. Austin, M.D.

Chapter 4

C

4.1 Talking Glossary: Candidate gene

National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Candidate-Gene

Introduction: National Human Genome Research InstituteA candidate gene is a gene whose chromosomal location is associated with a particular disease or other phenotype. Because of its location, the gene is suspected of causing the disease or other phenotype.

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/cancer.mp3

Transcript: A candidate gene is a gene whose chromosomal location fits with a particular disease or phenotype that you're looking for. An example of this is when you're doing any type of linkage analysis and you're trying to find the disease gene that's associated with that particular disease. You use what we call genetic markers, and the markers will tell you, okay, based on what you see, the recombination frequency, etc., the gene has to be between marker X and marker Y. And what that means is that this distance between marker X and marker Y constitutes your candidate region, and all the genes in that region will be a candidate gene. And now the next thing for you to do is then to look individually at each of those genes to see if they have the mutation that's associated with the disease that you're looking for.

Milton English, Ph.D.

4.2 CATH database

Modified from Wikipedia https://en.wikipedia.org/wiki/CATH_database

The CATH Protein Structure Classification database is a free, publicly available online resource that provides information on the evolutionary relationships of protein domains. It was created in the mid-1990s by Professor Christine Orengo and colleagues including Janet Thornton and David Jones (2), and continues to be developed by the Orengo group at University College London. CATH shares many broad features with the **SCOP** resource, however there are also many areas in which the detailed classification differs greatly (3, 4, 5, 6).

4.2.1 References

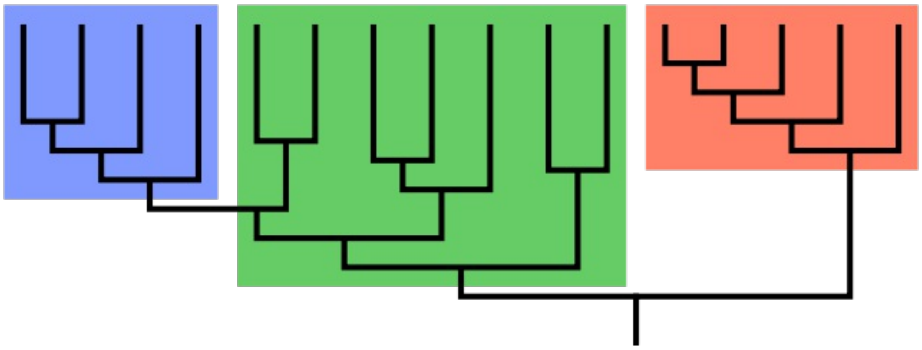
1. Dawson, NL; Lewis, TE; Das, S; Lees, JG; Lee, D; Ashford, P; Orengo, CA; Sillitoe, I (28 November 2016). "CATH: an expanded resource to predict protein function through structure and sequence". Nucleic Acids Research. 45 (D1): D289–D295. doi:10.1093/nar/gkw1098. PMC 5210570. PMID 27899584.
2. Orengo, CA; Michie, AD; Jones, S; Jones, DT; Swindells, MB; Thornton, JM (1997). "CATH – a hierarchic classification of protein domain structures". Structure. 5 (8): 1093–1109. doi:10.1016/S0969-2126(97)00260-8. ISSN 0969-2126. PMID 9309224.
3. "CATH: Protein Structure Classification Database at UCL". Cathdb.info. Retrieved 9 March 2017.
4. "CATH". Cathdb.info. Retrieved 9 March 2017.
5. "CATH Database (@CATHDatabase)". Twitter. Retrieved 9 March 2017.
6. Pearl, F. M. G. (2003). "The CATH database: an extended protein family resource for structural and functional genomics". Nucleic Acids Research. 31 (1): 452–455. doi:10.1093/nar/gkg062. ISSN 1362-4962. PMC 165509. PMID 12520050.
7. "Tools". cathdb.info. Retrieved 18 December 2016.

4.3 Clade

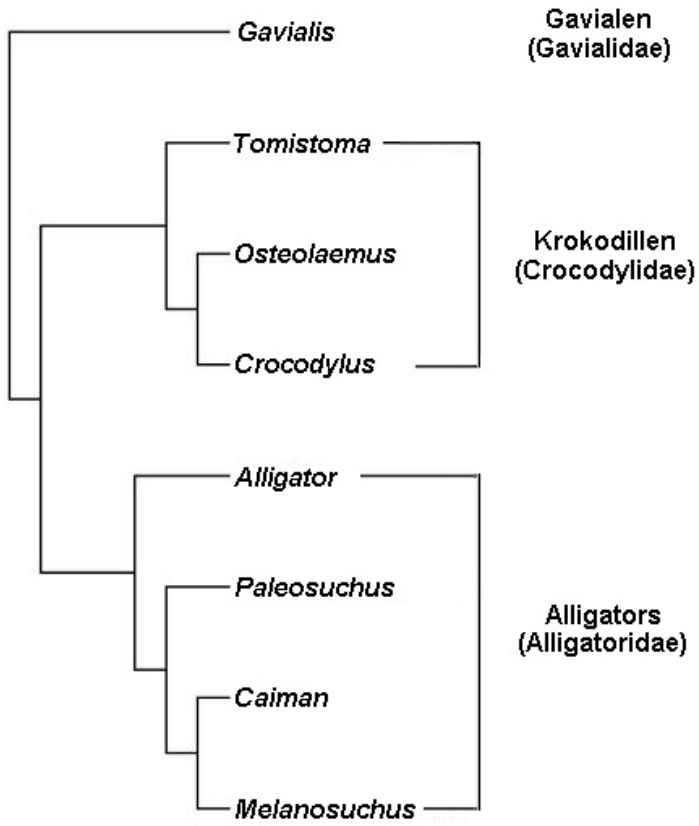
Adapted from Wikipedia https://en.wikipedia.org/wiki/Clade

A clade, also known as a monophyletic group, is a group of organisms that are monophyletic —that is, composed of a common ancestor and all its descendants .[4] . The common ancestor may be an individual, a population , a species (extinct or extant), and so on. Clades are nested, one in another, as each branch in turn splits into smaller branches. These splits reflect evolutionary history as populations diverged and evolved independently. The term "clade" is also used with a similar meaning in other fields besides biology, such as historical linguistics.

A clade is by definition monophyletic , meaning that it contains one ancestor (which can be an organism, a population, or a species) and all its descendants.[note 1] [12] [13] The ancestor can be known or unknown; any and all members of a clade can be extant or extinct.



Over the last few decades, the cladistic approach has revolutionized biological classification and revealed surprising evolutionary relationships among organisms.[5] Taxonomists have worked to make the taxonomic system reflect evolution. Taxonomists therefore try to avoid naming taxa that are not clades; that is, taxa that are not monophyletic . Some of the relationships between organisms that the molecular biology arm of cladistics has revealed are that fungi are closer relatives to animals than they are to plants, archaea are now considered different from bacteria , and multicellular organisms may have evolved from archaea.[6]



Many commonly named groups, rodents and insects for example, are clades because, in each case, the group consists of a common ancestor with all its descendant branches. Rodents, for example, are a branch of mammals that split off after the end of the period when the clade Dinosauria stopped being the dominant terrestrial vertebrates 66 million years ago. The original population and all its descendants are a clade. The rodent clade corresponds to the order Rodentia, and insects to the class Insecta. These clades include smaller clades, such as chipmunk or ant , each of which consists of even smaller clades. The clade "rodent" is in turn included in the mammal, vertebrate and animal clades.

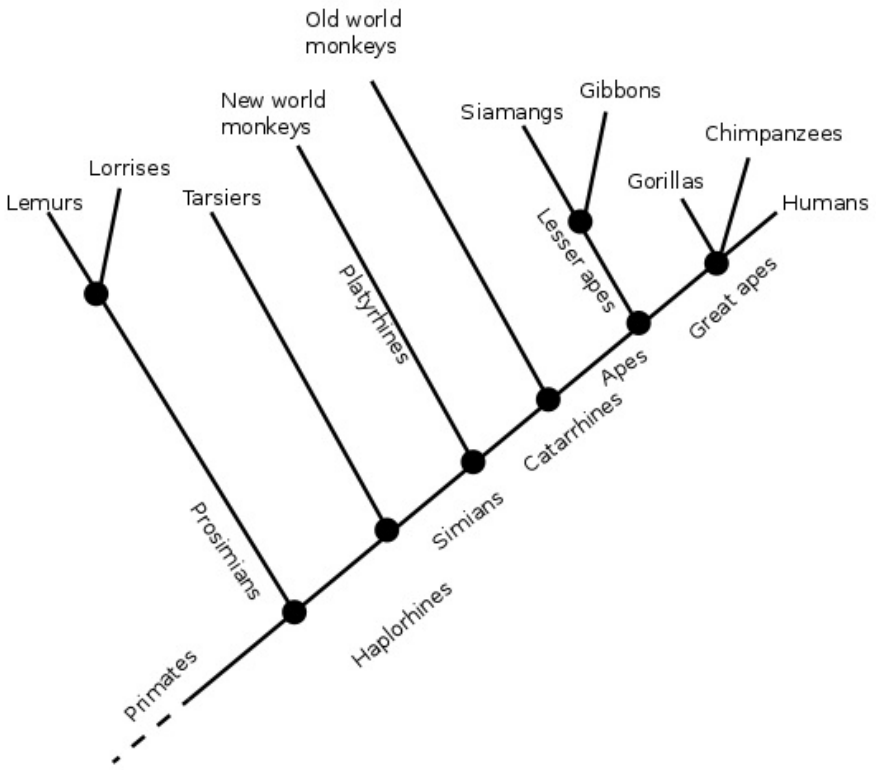
4.3.1 Clades and phylogenetic trees

The science that tries to reconstruct phylogenetic trees and thus discover clades is called phylogenetics . The results of phylogenetic analyses are tree-shaped diagrams called phylogenies; they, and all their branches, are phylogenetic hypotheses.[14]

4.3.2 Terminology

The relationship between clades can be described in several ways:

1. A clade located within a clade is said to be nested within that clade. In the diagram, the hominoid clade, i.e. the apes and humans, is nested within the primate clade.
2. Two clades are sisters (sister groups sister clades) if they have an immediate common ancestor. In the diagram, lemurs and lorises are sister clades, while humans and tarsiers are not.



4.3.3 In popular culture

An episode of Elementary is titled “Dead Clade Walking” and deals with a case involving a rare fossil.

4.3.4 References

1. Wells, John C. (2008). Longman Pronunciation Dictionary (3rd ed.). Longman. ISBN 978-1-4058-8118-0. “clade”. Merriam-Webster Dictionary. Retrieved 19 April 2020. Martin, Elizabeth; Hin, Robert (2008). A

Dictionary of Biology. Oxford University Press.
2. Cracraft, Joel; Donoghue, Michael J., eds. (2004). "Introduction". Assembling the Tree of Life. Oxford University Press. p. 1. ISBN 978-0-19-972960-9.
3. Palmer, Douglas (2009). Evolution: The Story of Life. Berkeley: University of California Press. p. 13.
4. Pace, Norman R. (18 May 2006). "Time for a change". Nature. 441 (7091): 289. Bibcode:2006Natur.441..289P. doi:10.1038/441289a. ISSN 1476-4687. PMID 16710401. S2CID 4431143.
5. Dupuis, Claude (1984). "Willi Hennig's impact on taxonomic thought". Annual Review of Ecology and Systematics. 15: 1–24. doi:10.1146/annurev.es.15.110184.000245.
6. Huxley, J. S. (1957). "The three types of evolutionary process". Nature. 180 (4584): 454–455. Bibcode:1957Natur.180..454H. doi:10.1038/180454a0. S2CID 4174182.
7. Huxley, T.H. (1876): Lectures on Evolution. New York Tribune. Extra. no 36. In Collected Essays IV: pp 46-138 original text w/ figures
8. Brower, Andrew V. Z. (2013). "Willi Hennig at 100". Cladistics. 30 (2): 224–225. doi:10.1111/cla.12057.
9. "Evolution 101". page 10. Understanding Evolution website. University of California, Berkeley. Retrieved 26 February 2016.
10. "International Code of Phylogenetic Nomenclature. Version 4c. Chapter I. Taxa". 2010. Retrieved 22 September 2012.
11. Envall, Mats (2008). "On the difference between mono-, holo-, and paraphyletic groups: a consistent distinction of process and pattern". Biological Journal of the Linnean Society. 94: 217. doi:10.1111/j.1095-8312.2008.00984.x.
12. Nixon, Kevin C.; Carpenter, James M. (1 September 2000). "On the Other"Phylogenetic Systematics"". Cladistics. 16 (3): 298–318. doi:10.1111/j.1096-0031.2000.tb00285.x. S2CID 73530548.

4.4 Talking Glossary: Cloning (0.5 min)

National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Cloning

Introduction: "Cloning is the process of making identical copies of an organism, cell, or DNA sequence. Molecular cloning is a process by which scientists amplify a desired DNA sequence. The target sequence is isolated, inserted into another DNA molecule (known as a vector), and introduced into a suitable host cell. Then, each time the host cell divides, it replicates the foreign DNA sequence along with its own DNA. Cloning also can refer to asexual reproduction."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/cloning.mp3

Transcript: "Cloning is a word we use to describe a molecular process of making millions or billions of copies of a single molecule. It's different from the uses of the terms"cellular cloning" or "organism cloning" that are used in the reproductive genetics universe. We use molecular cloning to amplify, or make many copies of, genes or proteins or other micro molecules that amplifies the signal and allows us to study these molecules in a laboratory".

Leslie G. Biesecker, M.D.

4.5 Talking Glossary: Codominance (0.5 min)

Abstract: "Codominance is a relationship between two versions of a gene. Individuals receive one version of a gene, called an allele, from each parent. If the alleles are different, the dominant allele usually will be expressed, while the effect of the other allele, called recessive, is masked. In codominance, however, neither allele is recessive and the phenotypes of both alleles are expressed."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/codominance.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/codominance.jpg Transcript

Codominance means that neither allele can mask the expression of the other allele. An example in humans would be the ABO blood group, where alleles A and alleles B are both expressed. So if an individual inherits allele A from their mother and allele B from their father, they have blood type AB.

Suzanne Hart, Ph.D.

4.6 Comparative genomics

Adapted from Wikipedia

https://en.wikipedia.org/wiki/Comparative_genomics

Comparative genomics is a field of biological research in which the genomic features of different organisms are compared.[2][3] The genomic features may include the DNA sequence, genes, gene order, regulatory sequences, and other genomic structural landmarks.[3] In this branch of genomics, whole genomes resulting from genome projects are compared to study basic biological similarities and differences as well as evolutionary relationships between organisms.[2][4][5] The major principle of comparative genomics is that common features of two organisms will often be encoded within the DNA that is evolutionarily conserved between them.[6] Therefore, comparative genomic approaches start with making some form of alignment of genome sequences and looking for similar sequences in the aligned genomes and checking to what extent those sequences are conserved.

Based on these, genome and molecular evolution are inferred and this may, in turn, be put in the context of, for example, evolution, ecology, or pathogenicity.

Comparative genomics is now a standard component of the analysis of every new genome sequence.[2][8] With the explosion in the number of genome projects due to the advancements in DNA sequencing technologies[9] Comparative genomics has revealed high levels of similarity between closely related organisms, such as humans and chimpanzees, and, more surprisingly, similarity between seemingly distantly related organisms, such as humans and yeast.

The medical field benefits from the study of comparative genomics. Vaccine development in particular has experienced useful advances in technology due to genomic approaches to problems. In an approach known as reverse vaccinology, researchers can discover candidate antigens (proteins in pathogens that can be targeted by the immune system) for vaccine development by analyzing the genome of a pathogen or a family of pathogens.

1* Darling A.E.; Miklós I.; Ragan M.A. (2008). "Dynamics of Genome Rearrangement in Bacterial Populations". PLOS Genetics. 4 (7): e1000128. doi: 10.1371/journal.pgen.1000128. PMC 2483231. PMID 18650965. open access 1. Touchman, J. (2010). "Comparative Genomics". Nature Education Knowledge. 3 (10): 13. 1. Xia, X. (2013). Comparative Genomics. SpringerBriefs in Genetics. Heidelberg: Springer. doi:10.1007/978-3-642-37146-2. ISBN 978-3-642-37145-5. S2CID 5491782. 1. Russel, P.J.; Hertz, P.E.; McMillan, B. (2011). Biology: The Dynamic Science (2nd ed.). Belmont, CA: Brooks/Cole. pp. 409–410. 1. Primrose, S.B.; Twyman, R.M. (2003). Principles of Genome Analysis and Genomics (3rd ed.). Malden, MA: Blackwell Publishing. 1. Hardison, R.C. (2003). "Comparative genomics". PLOS Biology. 1 (2): e58. doi:10.1371/journal.pbio.0000058. PMC 261895. PMID 14624258. open access 1. Ellegren, H. (2008). "Comparative genomics and the study of evolution by natural selection". Molecular Ecology. 17 (21): 4586–4596. doi:10.1111/j.1365-294X.2008.03954.x. PMID 19140982. S2CID 43171654. 1. Koonin, E.V.; Galperin, M.Y. (2003). Sequence - Evolution - Function: Computational approaches in comparative genomics. Dordrecht: Springer Science+Business Media.

4.7 Talking Glossary: Contig (0.75 min)

https://www.genome.gov/genetics-glossary/Contig

Abstract: "A contig–from the word"contiguous"–is a series of overlapping DNA sequences used to make a physical map that reconstructs the original DNA sequence of a chromosome or a region of a chromosome. A contig can also refer to one of the DNA sequences used in making such a map."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/contig.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/narration/contig.mp3

Transcript: "A chromosome is a very long molecule of DNA. And it is very hard to study it at once, so what researchers do is they break it into smaller pieces and they sequence each one of those individual pieces first, and then they attempt to put it together to reconstruct the original chromosome sequence. A contig is the physical map, which results from putting together several little overlapping bits of DNA into a longer sequence. The contig is the physical map resulting from taking small pieces of DNA that overlap and putting them together into a longer sequence."

Belen Hurle, Ph.D.

4.8 Talking Glossary - Foundations 1 Review: Crossing over

Abstract: "Crossing over is the swapping of genetic material that occurs in the germ line. During the formation of egg and sperm cells, also known as meiosis, paired chromosomes from each parent align so that similar DNA sequences from the paired chromosomes cross over one another. Crossing over results in a shuffling of genetic material and is an important cause of the genetic variation seen among offspring."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/crossing_over.mp3

Transcript: "Crossing over is a biological occurrence that happens during meiosis when the paired homologs, or chromosomes of the same type, are lined up. In meiosis, they're lined up on the meiotic plates, [as they're] sometimes called, and those paired chromosomes then have to have some biological mechanism that sort of keeps them together. And it turns out that there are these things called chiasmata, which are actually where strands of the duplicated homologous chromosomes break and recombine with the same strand of the other homolog. So if you have two Chromosome 1s lined up, one strand of one Chromosome 1 will break and it will reanneal with a similar breakage on the other Chromosome 1. So that then the new chromosome that will happen will have part of, say, the maternal Chromosome 1 and the paternal Chromosome 1, where maternal and paternal means where that person got their Chromosomes 1s from, their one or their two. Therefore, the child that's formed out of one of those Chromosome 1s now has a piece of his or her grandmother's Chromosome 1 and a piece of his or her grandfather's Chromosome 1. And it's this crossing over that lets recombination across generations of genetic material happen, and it also allows us to use that information to find the locations of genes."

Joan E. Bailey-Wilson, Ph.D

Photo: https://i.irp.nih.gov/pi/0010100622.jpg

Chapter 5

D

5.1 Talking Glossary: Deletion mutation (1.5 min)

https://www.genome.gov/genetics-glossary/Deletion

Abstract: "Deletion is a type of mutation involving the loss of genetic material. It can be small, involving a single missing DNA base pair, or large, involving a piece of a chromosome."

Note: Among evolutionary biologists, "deletion" usually means deletion of a single base; larger deletions are specified, e.g. "chromosomal deletion." In order to recognize a deletion, there's needs to be some frame of reference, such as a consensus sequence, a reference genomic sequence, or other individuals in a multiple sequence alignment (MSA). Deletions of all sizes can be useful for constructing phylogenies.

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/deletion.jpg Example of single base and chromosomal deletion.

Transcript: "Deletion really means that something is missing. And as a geneticist talking about deletion it means something is missing of the genetic material. And it can be something small, just a base pair; it can be something larger; it can be part of a gene; it can be even larger; it can be an entire gene; or yet larger again, it can be part of the chromosome. And depending upon what it is, you have to look at it in different ways. You can find a deletion in a chromosome just by doing a cytogenetic or chromosome analysis, or a deletion in a gene you can find out by sequencing the DNA. So when you have a deletion, depending upon the size, it can have different effects. What was the most surprising to me was that just by having a deletion of one base pair, you can have the most severe birth defect, and sometimes by missing an entire chromosome, you don't even

see all that much compared to just having a deletion of a small base pair. Different deletions can lead to different findings, and they can affect just behavior; they can affect how a child, how a person looks; they can affect a very severe problem that the child may die at birth; or they can affect something that just has to do with eye color, hair color, with weight or height of the person."

Maximilian Muenke, M.D.

5.2 Talking Glossary: DNA sequencing (0.75 min)

Introduction: "DNA sequencing is a laboratory technique used to determine the exact sequence of bases (A, C, G, and T) in a DNA molecule. The DNA base sequence carries the information a cell needs to assemble protein and RNA molecules. DNA sequence information is important to scientists investigating the functions of genes. The technology of DNA sequencing was made faster and less expensive as a part of the Human Genome Project."

Transcript: "DNA consists of a linear string of nucleotides, or bases, for simplicity, referred to by the first letters of their chemical names–A, T, C and G. The process of deducing the order of nucleotides in DNA is called DNA sequencing. Since the DNA sequence confers information that the cell uses to make RNA molecules and proteins, establishing the sequence of DNA is key for understanding how genomes work. The technology for DNA sequencing was made faster and less expensive as a part of the Human Genome Project. And recent developments have profoundly increased the efficiency of DNA sequencing even further."

Eric D. Green, M.D., Ph.D.

5.3 Sex-Chromosome Dosage compensation

Adapted from Wikipedia

Dosage compensation is the process by which organisms equalize the expression of genes between individuals with different sex chromosome karyotypes (e.g. XX versus XY). Across species, different "sexes" are often characterized by different types and numbers of sex chromosomes . In order to neutralize the large difference in gene dosage produced by differing numbers of sex chromosomes, various evolutionary branches have acquired various methods to equalize gene expression . Because sex chromosomes contain different numbers of genes , different species of organisms have developed different mechanisms to cope with this inequality. Replicating the actual gene is impossible; thus organisms instead equalize the expression from each gene. For example, in humans , XX individuals silence the transcription of one X chromosome of each pair, and

transcribe all information from the other, expressed X chromosome. Thus, human XX individuals have the same number of expressed X-linked genes as do human XY individuals, with both having essentially one X chromosome per cell, from which to transcribe and express genes. This is called X-inactivation. In each XX cell, one of the two X chromosomes is randomly inactivated.

Other lineages have evolved different mechanisms to cope with the differences in gene copy numbers between the sexes that are observed on sex chromosomes. Some lineages have evolved upregulation mechanisms which restores expression of X-specific genes in the heterogametic sex (e.g. XY) to the same levels observed in the ancestor prior to the evolution of the sex chromosome.

5.3.1 Random inactivation of one X in XX individuals

One logical way to equalize gene expression amongst XX and XY that follow a XX/XY sex differentiation scheme would be to decrease or altogether eliminate the expression of one of the X chromosomes in an XX, (homogametic) individual, such that both XX and XY individuals express only one X chromosome. This is the case in many mammalian organisms, including humans and mice.[1 This process involves histone tail modifications, DNA methylation patterns, and reorganization of large-scale chromatin structure. In spite of these extensive modifications, not all genes along the X chromosome are subject to X-inactivation.] Because so many variants are tested, it is standard practice to require the p-value to be lower than 5×10−8 to consider a variant significant.

There are several variations to this case-control approach. A common alternative to case-control GWAS is the analysis of quantitative phenotypic data, e.g. height or biomarker concentrations or even gene expression. Likewise, alternative statistics designed for dominance or recessive penetrance patterns can be used (16] Calculations are typically done using bioinformatics software such as SNPTEST and PLINK, which also include support for many of these alternative statistics (15, 17). GWAS focuses on the effect of individual SNPs. However, it is also possible that complex interactions among two or more SNPs (epistasis) might contribute to complex diseases. Due to the potentially exponential number of interactions, detecting statistically significant interactions in GWAS data is both computationally and statistically challenging. This task has been tackled in existing publications that use algorithms inspired from data mining (18) Moreover, the researchers try to integrate GWA data with other biological data such as protein-protein interaction network to extract more informative results (19, 20).

In addition to the calculation of association, it is common to take into account any variables that could potentially confound the results. Sex and age are common examples of confounding variables. Moreover, it is also known that many genetic variations are associated with the geographical and historical populations in which the mutations first arose (25). Because of this association, studies must take account of the geographic and ethnic background of partici-

pants by controlling for what is called population stratification. If they fail to do so, these studies can produce false positive results (26)

After odds ratios and P-values have been calculated for all SNPs, a common approach is to create a Manhattan plot. In the context of GWAS, this plot shows the negative logarithm of the P-value as a function of genomic location. Thus the SNPs with the most significant association stand out on the plot, usually as stacks of points because of haploblock structure. Importantly, the P-value threshold for significance is corrected for multiple testing issues. The exact threshold varies by study,[27] but the conventional threshold is $5\times10-8$ to be significant in the face of hundreds of thousands to millions of tested SNPs (7, 16, 28] GWAS typically perform the first analysis in a discovery cohort, followed by validation of the most significant SNPs in an independent validation cohort (29]

5.3.2 Results

An early GWAS, conducted in 2005, compared 96 patients with the eye disorder age-related macular degeneration (ARMD) with 50 healthy controls (33). It identified two SNPs with significantly altered allele frequencies between the two groups. These SNPs were located in the gene encoding complement factor H, which was an unexpected finding in the research of ARMD. These findings prompted further functional research towards therapeutical manipulation of the complement system in ARMD (34). Another landmark publication in the history of GWAS was the Wellcome Trust Case Control Consortium (WTCCC) study, the largest GWAS ever conducted at the time of its publication in 2007. The WTCCC included 14,000 cases of seven common diseases (~2,000 individuals for each of coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder, and hypertension) and 3,000 shared controls (15). This study was successful in uncovering many new disease genes underlying these diseases (15, 35).

Since these first landmark GWAS, there have been two general trends (36). One has been towards larger and larger sample sizes. In 2018, several genome-wide association studies are reaching a total sample size of over 1 million participants, including a study of insomnia containing 1.3 million individuals (38). The reason is the drive towards reliably detecting risk-SNPs that have ever smaller effects and lower allele frequencies. Another trend has been towards the use of more narrowly defined continous phenotypes, such as concentrations of blood lipids, proinsulin or similar biomarkers (39, 40). These are called intermediate phenotypes, and their analyses may be of value to functional research into biomarkers (41).

A central point of debate on GWAS has been that most of the SNP variations found by GWAS are associated with only a small increased risk of the disease, and have only a small predictive value.

The effects sizes are considered small because they do not explain much of

the heritable variation. This heritable variation is estimated from heritability studies based on monozygotic twins (44). For example, it is known that 80-90% of variance in height can be explained by hereditary differences, but GWAS only account for a minority of this variance (44). This is sometimes referred to as the problem of the **missing heritability** of GWAS.

5.3.3 Clinical applications

THIS SECTION IS OPTIONAL

A challenge for future successful GWAS is to apply the findings in a way that accelerates drug and diagnostics development, including better integration of genetic studies into the drug-development process and a focus on the role of genetic variation in maintaining health as a blueprint for designing new drugs and diagnostics (45). Several studies have looked into the use of risk-SNP markers as a means of directly improving the accuracy of prognosis. Some have found that the accuracy of prognosis improves (46), while others report only minor benefits from this use (47). Generally, a problem with this direct approach is the small magnitudes of the effects observed. A small effect ultimately translates into a poor separation of cases and controls and thus only a small improvement of prognosis accuracy. An alternative application is therefore the potential for GWAS to elucidate pathophysiology (48).

One such success is related to identifying the genetic variant associated with response to anti-hepatitis C virus treatment. For genotype 1 hepatitis C treated with Pegylated interferon-alpha-2a or Pegylated interferon-alpha-2b combined with ribavirin, a GWAS (49) has shown that SNPs near the human IL28B gene, encoding interferon lambda 3, are associated with significant differences in response to the treatment. A later report demonstrated that the same genetic variants are also associated with the natural clearance of the genotype 1 hepatitis C virus (50). These major findings facilitated the development of personalized medicine and allowed physicians to customize medical decisions based on the patient's genotype (51).

The goal of elucidating pathophysiology has also led to increased interest in the association between risk-SNPs and the gene *expression* of nearby genes, the so-called **expression quantitative trait loci (eQTL)** studies (52). The reason is that GWAS studies identify risk-SNPs, but not risk-genes, and specification of genes is one step closer towards actionable drug targets. As a result, major GWAS by 2011 typically included extensive eQTL analysis (53, 54, 55). One of the strongest eQTL effects observed for a GWA-identified risk SNP is the SORT1 locus (39). Functional follow up studies of this locus using small interfering RNA and gene knock-out mice have shed light on the metabolism of low-density lipoproteins, which have important clinical implications for cardiovascular disease (39, 56, 57).

5.3.3.1 Atrial fibrillation

This section is optional

For example, a meta-analysis accomplished in 2018 revealed the discovery of 70 new loci associated with atrial fibrillation. It has been identified different variants associated with transcription factor coding-genes, such as TBX3 and TBX5, NKX2-5 o PITX2, which are involved in cardiac conduction regulation, in ionic channel modulation and cardiac development. It was also identified new genes involved in tachycardia (CASQ2) or associated with alteration of cardiac muscle cell communication (PKP2) (58)

5.3.3.2 Schizophrenia

This section is optional

While there is some research using a High-Precision Protein Interaction Prediction (HiPPIP) computational model that discovered 504 new protein-protein interactions (PPIs) associated with genes linked to schizophrenia (59,60) the evidence supporting the genetic basis of schizophrenia is actually controversial and may suffer from some of the limitation of this method of study (61]

5.3.4 Agricultural applications

5.3.4.1 Plant growth stages and yield components

This section is optional

GWAS act as an important tool in plant breeding. With large genotyping and phenotyping data, GWAS are powerful in analyzing complex inheritance modes of traits that are important yield components such as number of grains per spike, weight of each grain and plant structure. In a study on GWAS in spring wheat, GWAS have revealed a strong correlation of grain production with booting data, biomass and number of grains per spike (62).

5.3.4.2 Plant pathogens

This section is optional

The emergences of plant pathogens have posed serious threats to plant health and biodiversity. Under this consideration, identification of wild types that have the natural resistance to certain pathogens could be of vital importance. Furthermore, we need to predict which alleles are associated with the resistance. GWAS is a powerful tool to detect the relationships of certain variants and the resistance to the plant pathogen, which is beneficial for developing new pathogen-resisted cultivars (63).

5.3.5 Limitations

5.3.5.1 Limitations - common errors

This section is optional

GWAS have several issues and limitations that can be taken care of through proper quality control and study setup. Lack of well defined case and control groups, insufficient sample size, control for **multiple testing** and control for **population stratification** are common problems (2). Particularly the statistical issue of multiple testing wherein it has been noted that "the GWA approach can be problematic because the massive number of statistical tests performed presents an unprecedented potential for false-positive results" (2). Ignoring these correctible issues has been cited as contributing to a general sense of problems with the GWAS methodology (64). In addition to easily correctible problems such as these, some more subtle but important issues have surfaced. A high-profile GWA study that investigated individuals with very long life spans to identify SNPs associated with longevity is an example of this (65). The publication came under scrutiny because of a discrepancy between the type of genotyping array in the case and control group, which caused several SNPs to be falsely highlighted as associated with longevity (66). The study was subsequently retracted (67), but a modified manuscript was later published (68).

5.3.5.2 Limitations - fundamental problems (READ THIS ECTION)

READ THIS SECTION

The fundamental assumptions of GWAS and have attracted fundamental criticism, mainly because of their assumption that common genetic variation plays a large role in explaining the heritable variation of common disease (69). Indeed, it has been estimated that for most conditions the SNP heritability attributable to common SNPs is <0.05 (70). This aspect of GWAS has attracted the criticism that, although it could not have been known prospectively, GWAS were ultimately not worth the expenditure (48). GWAS also face criticism that the broad variation of individual responses or compensatory mechanisms to a disease state cancel out and mask potential genes or causal variants associated with the disease (71). Additionally, GWAS identify candidate risk variants for the population from which their analysis is performed, and with most GWAS stemming from European databases, there is a lack of translation of the identified risk variants to other non-European populations (72] Alternative strategies suggested involve linkage analysis (73, 74] More recently, the rapidly decreasing price of complete genome sequencing have also provided a realistic alternative to genotyping array-based GWAS. It can be discussed if the use of this new technique is still referred to as a GWA study, but high-throughput sequencing does have potential to side-step some of the shortcomings of non-sequencing GWA (75]

5.4 Fine-mapping

THIS SECTION IS OPTIONAL

Genotyping arrays designed for GWAS rely on linkage disequilibrium to provide coverage of the entire genome by genotyping a subset of variants. Because of this, the reported associated variants are unlikely to be the actual causal variants. Associated regions can contain hundreds of variants spanning large regions and encompassing many different genes, making the biological interpretation of GWAS loci more difficult. Fine-mapping is a process to refine these lists of associated variants to a credible set most likely to include the causal variant.

Fine-mapping requires all variants in the associated region to have been genotyped or imputed (dense coverage), very stringent quality control resulting in high-quality genotypes, and large sample sizes sufficient in separating out highly correlated signals. There are several different methods to perform fine-mapping, and all methods produce a posterior probability that a variant in that locus is causal. Because the requirements are often difficult to satisfy, there are still limited examples of these methods being more generally applied.

Chapter 6

H

6.1 Talking Glossary: Haploid (1.25 min)

https://www.genome.gov/genetics-glossary/haploid

Abstract: "Haploid is the quality of a cell or organism having a single set of chromosomes. Organisms that reproduce asexually are haploid. Sexually reproducing organisms are diploid (having two sets of chromosomes, one from each parent). In humans, only their egg and sperm cells are haploid."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/haploid.mp3

Transcript: "Haploid refers to a cell or an organism that has only a single set of chromosomes. This is to be contrasted with diploid."Di" means two, of course. So most animal cells and plant cells are diploid. Then they're diploid in part because they got one chromosome from their mother and one chromosome from their father, therefore making them diploid. A haploid cell only has one set of chromosomes, and most of the time that refers to the so-called sex cells, either eggs or sperm. And these are a critical transition from a diploid cell to a haploid cell to allow normal reproduction to occur, so that when these two haploid cells come together with a single set of genetic information–single chromosomes–they can come together into a so-called zygote made of when the egg cell and the sperm cell come together that then reconstitutes a diploid cell, which can then become a new individual."

Christopher P. Austin, M.D.

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/haploid.jpg

6.2 Talking Glossary: Halplotype

A haplotype is a set of DNA variations, or polymorphisms, that tend to be inherited together. A haplotype can refer to a combination of alleles or to a set of single nucleotide polymorphisms (SNPs) found on the same chromosome. Information about haplotypes is being collected by the International HapMap Project and is used to investigate the influence of genes on disease.

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/haplotype.mp3

A haplotype is in its most general sense referring to a set of DNA variations along a chromosome that tend to be inherited together because they're very close together. They get inherited together because they're not generally crossovers or recombinations between these markers or between these different polymorphisms because they are very, very close. So a haplotype can refer to a combination of alleles in a single gene, or it could be alleles across multiple genes. It could be single nucleotide polymorphisms that are not in a gene but are in-between genes. Basically, it just means that these are variations in the DNA that are so close together that they tend not to recombine, and therefore tend to be passed down through the generations together. And the International HapMap Project has given us a very excellent tool to detect these regions of haplotypes that are passed together and to use those in genetic studies.

Joan E. Bailey-Wilson, Ph.D.

6.3 Talking Glossary: Heterozygous

https://www.genome.gov/genetics-glossary/heterozygous (Links to an external site.)

Abstract: "Heterozygous refers to having inherited different forms of a particular gene from each parent. A heterozygous genotype stands in contrast to a homozygous genotype, where an individual inherits identical forms of a particular gene from each parent."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/heterozygous.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/heterozygous.jpg

Transcript: "Heterozygous is a state of having inherited different forms of a particular gene from each one of your biological parents. Now, by different forms we generally mean that there are different portions of the gene where the sequence is different. They may be inconsequential portions of the gene, or they may in fact be pretty important portions of the gene. That doesn't really matter for our discussion today. The word"heterozygous" simply means that your biological mother and your biological father, when they contributed

their copies of a particular gene to you, they did so in a way so that the DNA sequence is slightly different. It can be different at one point in the gene, or it can be different at dozens and dozens of different points in the gene. Now, a heterozygous genotype stands in contrast to a homozygous genotype. And in the case of a homozygous genotype, we're talking about a case where we've gotten identical forms of a particular gene from each biological parent. That is, if we were to read along the DNA sequence that mom gave you and the DNA sequence that dad gave you, we would find absolutely, positively no differences in that gene or in the region of the gene that we're concerned about. "Heterozygous" meaning different, "homozygous" meaning the same."

Amalia S. Dutra, Ph.D.

Amalia Dutra, Ph.D, is a Uruguayan genetic biologist known for being part of the team that mapped the human genome

6.4 Homology:L When things are homolgous AND analogous!

By: Dr. Brouwer

The topics of homologous and analogous structures are often discussed in relation to convergent evolution and homoplasy. It can be difficult to keep track of the exact definitions and when each term applies; in this short reading, I'll reiterate the difference between homologous and analogous features and discuss an important facet that is often not brought up when these things are discussed.

In this Mometrix video – like many videos on the topic – they discuss homologous structures by comparing the arm of a human, the wing of a bat, the flipper of a whale, and the front leg of a cat (1:31). All of these anatomical structures contain the exact same bones, though they are used for different forms of movement.

To be more precise and clearer, instead of calling these arms, fins etc, they should call them forelimbs. At the level of bone structure, these forelimbs are homologous. Their homology is defined based on similar development and morphology and has nothing to do with function. To me, to refer to these different organisms' forelimbs as arms, flippers etc. brings in aspects of function which aren't relevant to consideration of homology and can cause confusion when you start thinking about analogous structures and convergent evolution.

When they discuss analogous structures, they show the wings of bats, birds, and butterflies. Bat wings and bird wings are analogous structures. (Comparing vertebrate and insect wings is a trivial example but a useful starting point). Vetebrate wings have the same function - flight - but have key differences. For example, bats create the wing using membranes derived from skin tissue (epidermis), while birds use feathers (which are derived from scales). Moreover, bats

spread out the membranes of their wings with their fingers, while birds spread out their feathers by making the feathers themselves stiff. These structures are therefore analogous when we consider them from the perspective of anatomical function. Therefore, while discussion of homology doesn't require considering function, discussion of analogy does. In contrast, the flippers of whales and dolphins are not analogous because they share a common ancestor with flippers, and they function the same.

From the perspective of natural selection, bat wings and bird wings are also an example of convergent evolution. Starting from different initial types of organisms - terrestrial rodents and terrestrial dinosaurs, respectively - both bats and birds converged on the ecological strategy of flight. That is, they converged on the strategy of using their forelimbs as wings. (You could say birds and butterflies converged on the strategy of flight but this is trivial since they do very different things while flying; bats and birds, however, often compete for the same resources, are eaten by similar predators, etc).

But wait - in our above example of forelimbs we had humans, whales, cats and birds, and we said these were homologous. These are all tetrapods, and bats are tetrapods too, so shouldn't bat forelimbs also be homologous to these others? Yes, bat forelimbs are homologous to all these others. But we just said that bat wings and bird wings are analogous! Yes, that's true too. At the level of the forelimb bone structure, bird and bat forelimbs are homologous. But at the level of the functional wings, they are analogous because they achieve their function - key to consideration of analogy and convergent evolution - very differently.

Sadava et al (11th edition) puts it this way

> "Any features shared by two or more species that have been inherited from a common ancestor are said to be homologous. Homologous features may be any heritable trait, including DNA sequences, protein structures, anatomical structures, and even some behavior patterns. For example, all living vertebrates have a vertebral column, as did the ancestral vertebrate. Therefore the vertebral column is judged to be homologous in all vertebrates."…similar traits may evolve independently in different lineages, a phenomenon called convergent evolution. For example, although the wing bones [I say forelimbs] of bats and birds are homologous, having been inherited from a common tetrapod ancestor, the wings of bats and birds are not homologous because they evolved independently from forelimbs of different nonflying ancestors. Functionally similar structures that have independent evolutionary origins are called analogous characters." Sadava et al (11th edition, pg 451).

6.5 Talking Glossary: Homozygous

https://www.genome.gov/genetics-glossary/homozygous (Links to an external site.)

Abstract: "Homozygous is a genetic condition where an individual inherits the same alleles for a particular gene from both parents."

Audio: https://www.genome.gov/sites/default/files/media/audio/2019-04/homozygous_narration.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/homozygous.jpg

Transcript: "Homozygous describes the genetic condition or the genetic state where an individual has inherited the same DNA sequence for a particular gene from both their biological mother and their biological father. It's often used in the context of disease. We talk about a situation where an individual has inherited a mutant allele or an error in DNA sequence from their mother and they have inherited the identical mutant allele from their father. We would then say that individual is homozygous for that mutation. They have two identical copies of the deleterious version of that gene and, as a result, they are then going to be predisposed to the genetic condition the gene codes for."

Amalia S. Dutra, Ph.D.

Chapter 7

I

7.1 Talking Glossary: Insertion mutation (1.25 min)

genome.gov/genetics-glossary/Insertion

Abstract: "Insertion is a type of mutation involving the addition of genetic material. An insertion mutation can be small, involving a single extra DNA base pair, or large, involving a piece of a chromosome."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/insertion.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/insertion.mp3

Transcript "Insertion really means that something has been stuck in there. And again, as a geneticist, when we think of an insertion, we think of a piece of DNA, and that can be small or large, being stuck in at a place where it really doesn't belong. So an insertion of just one base pair could lead to something that we call a frameshift. It shifts the reading of the three-base pair code and by that can throw off the entire protein, and by that can lead, for example, to a birth defect. Insertion can be larger, that, for example, there is an insertion of three base pairs, and then it will not throw off the frame, or it will not lead to a frameshift, and potentially is less harmful than having the insertion of just one base pair. And of course you can have an insertion of huge pieces of DNA. They can be so large that you could actually see it on the chromosome analysis, where all of the smaller insertions you would see only by sequencing the stretch of DNA."

Maximilian Muenke, M.D.

For an interview with Dr. Muenke, see: https://www.genome.gov/player/wyo8AF_3nz8/PL1ay9ko4A8sk0o9O-YhseFHzbU2I2HQQp

7.2 Intrinscially disordered protein

Adapted from Wikipedia https://en.wikipedia.org/wiki/Intrinsically_disordered_proteins

An **intrinsically disordered protein** (IDP) is a protein that lacks a fixed or ordered three-dimensional structure (2, 3, 4) typically in the absence of its macromolecular interaction partners, such as other proteins or RNA. IDPs range from fully unstructured to partially structured and include **random coil**, **molten globule**-like aggregates, or flexible linkers in large **multi-domain** proteins. They are sometimes considered as a separate class of proteins along with **globular**, **fibrous** and **membrane** proteins (5).

The discovery of IDPs offers support against the idea that three-dimensional structures of proteins must be fixed to accomplish their biological functions. The dogma of rigid protein structure has been questioned due to the increasing evidence of dynamics being necessary for the protein machines. Despite their lack of stable structure, IDPs are a very large and functionally important class of proteins. Many IDPs can adopt a fixed three-dimensional structure after binding to other macromolecules. Overall, IDPs are different from structured proteins in many ways and tend to have distinctive function, structure, sequence, interactions, evolution and regulation (6).

7.2.1 Abundance

It is now generally accepted that proteins exist as an ensemble of similar structures with some regions more constrained than others. IDPs occupy the extreme end of this spectrum of flexibility.

Bioinformatic predictions indicated that intrinsic disorder is more common in genomes and proteomes than in known structures in the protein database. Based on DISOPRED2 prediction, long (>30 residue) disordered segments occur in 2.0% of archaean, 4.2% of eubacterial and 33.0% of eukaryotic proteins (10) including certain disease-related proteins (11).

7.2.2 Disorder annotation

Intrinsic disorder can be either annotated from experimental information or predicted with specialized software. Disorder prediction algorithms can predict Intrinsic Disorder (ID) propensity with high accuracy (approaching around 80%) based on primary sequence composition, similarity to unassigned segments in protein x-ray datasets, flexible regions in NMR studies and physico-chemical properties of amino acids.

7.2.3 Disorder databases

Databases have been established to annotate protein sequences with intrinsic disorder information. The **DisProt database** contains a collection of **manually curated** protein segments which have been experimentally determined to be disordered. MobiDB is a database combining experimentally curated disorder annotations (e.g. from DisProt) with data derived from missing residues in X-ray crystallographic structures and flexible regions in NMR structures.

7.2.4 Predicting IDPs by sequence

Separating disordered from ordered proteins is essential for disorder prediction. One of the first steps to find a factor that distinguishes IDPs from non-IDPs is to specify biases within the amino acid composition. The hydrophilic, charged amino acids (A, R, G, Q, S, P, E and K) have been characterized as disorder-promoting amino acids, while order-promoting amino acids (W, C, F, I, Y, V, L, and N) are hydrophobic and uncharged. The remaining amino acids (H, M, T and D) are ambiguous, found in both ordered and unstructured regions (2). A more recent analysis ranked amino acids by their propensity to form disordered regions as follows (order promoting to disorder promoting): W, F, Y, I, M, L, V, N, C, T, A, G, R, D, H, Q, K, S, E, P (43).

This information is the basis of most sequence-based predictors. Regions with little to no secondary structure, also known as NORS (NO Regular Secondary structure) regions (44) and low-complexity regions can easily be detected. However, not all disordered proteins contain such low complexity sequences.

7.2.5 Prediction methods

Determining disordered regions from lab methods is very costly and time-consuming. Due to the variable nature of IDPs, only certain aspects of their structure can be detected, so that a full characterization requires a large number of different methods and experiments. This further increases the expense of IDP determination. In order to overcome this obstacle, computer-based methods are created for predicting protein structure and function. It is one of the main goals of bioinformatics to derive knowledge by prediction. Predictors for IDP function are also being developed, but mainly use structural information such as linear motif sites (4, 45). There are different approaches for predicting IDP structure, such as neural networks or matrix calculations, based on different structural and/or biophysical properties.

Many computational methods exploit sequence information to predict whether a protein is disordered (46). Notable examples of such software include IUPRED and Disopred. Different methods may use different definitions of disorder. Meta-predictors show a new concept, combining different primary predictors to create a more competent and exact predictor.

Due to the different approaches of predicting disordered proteins, estimating

their relative accuracy is fairly difficult. For example, neural networks are often trained on different datasets. The disorder prediction category is a part of biannual CASP experiment that is designed to test methods according accuracy in finding regions with missing 3D structure (marked in PDB files as REMARK465, missing electron densities in X-ray structures).

7.2.6 Disorder and disease

This section optional

Intrinsically unstructured proteins have been implicated in a number of diseases (47). Aggregation of misfolded proteins is the cause of many synucleinopathies and toxicity as those proteins start binding to each other randomly and can lead to cancer or cardiovascular diseases. Thereby, misfolding can happen spontaneously because millions of copies of proteins are made during the lifetime of an organism. The aggregation of the intrinsically unstructured protein -synuclein is thought to be responsible. The structural flexibility of this protein together with its susceptibility to modification in the cell leads to misfolding and aggregation. Genetics, oxidative and nitrative stress as well as mitochondrial impairment impact the structural flexibility of the unstructured -synuclein protein and associated disease mechanisms (48). Many key tumour suppressors have large intrinsically unstructured regions, for example p53 and BRCA1. These regions of the proteins are responsible for mediating many of their interactions. Taking the cell's native defense mechanisms as a model drugs can be developed, trying to block the place of noxious substrates and inhibiting them, and thus counteracting the disease (49).

References

1. Majorek K, Kozlowski L, Jakalski M, Bujnicki JM (December 18, 2008). "Chapter 2: First Steps of Protein Structure Prediction" (PDF). In Bujnicki J (ed.). Prediction of Protein Structures, Functions, and Interactions. John Wiley & Sons, Ltd. pp. 39–62. doi:10.1002/9780470741894.ch2. ISBN 9780470517673.
2. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001). "Intrinsically disordered protein". Journal of Molecular Graphics & Modelling. 19 (1): 26–59. CiteSeerX 10.1.1.113.556. doi:10.1016/s1093-3263(00)00138-8. PMID 11381529.
3. Dyson HJ, Wright PE (March 2005). "Intrinsically unstructured proteins and their functions". Nature Reviews Molecular Cell Biology. 6 (3): 197–208. doi:10.1038/nrm1589. PMID 15738986. S2CID 18068406.
4. Dunker AK, Silman I, Uversky VN, Sussman JL (December 2008). "Function and structure of inherently disordered proteins". Current Opinion in Structural Biology. 18 (6): 756–64. doi:10.1016/j.sbi.2008.10.002. PMID 18952168.

5. Andreeva A, Howorth D, Chothia C, Kulesha E, Murzin AG (January 2014). "SCOP2 prototype: a new approach to protein structure mining". Nucleic Acids Research. 42 (Database issue): D310–4. doi:10.1093/nar/gkt1242. PMC 3964979. PMID 24293656.
6. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014). "Classification of intrinsically disordered regions and proteins". Chemical Reviews. 114 (13): 6589–631. doi:10.1021/cr400525m. PMC 4095912. PMID 24773235.
7. Song J, Lee MS, Carlberg I, Vener AV, Markley JL (December 2006). "Micelle-induced folding of spinach thylakoid soluble phosphoprotein of 9 kDa and its functional implications". Biochemistry. 45 (51): 15633–43. doi:10.1021/bi062148m. PMC 2533273. PMID 17176085.
8. Anfinsen, Christian B. (20 July 1973). "Principles that Govern the Folding of Protein Chains". Science. 181 (4096): 223–230. doi:10.1126/science.181.4096.223. PMID 4124164.
9. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001-01-01). "Intrinsically disordered protein". Journal of Molecular Graphics & Modelling. 19 (1): 26–59. CiteSeerX 10.1.1.113.556. doi:10.1016/s1093-3263(00)00138-8. PMID 11381529.
10. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (March 2004). "Prediction and functional analysis of native disorder in proteins from the three kingdoms of life". Journal of Molecular Biology. 337 (3): 635–45. CiteSeerX 10.1.1.120.5605. doi:10.1016/j.jmb.2004.02.002. PMID 15019783.
11. Uversky VN, Oldfield CJ, Dunker AK (2008). "Intrinsically disordered proteins in human diseases: introducing the D2 concept". Annual Review of Biophysics. 37: 215–46. doi:10.1146/annurev.biophys.37.032807.125924. PMID 18573080.
12. Bu Z, Callaway DJ (2011). "Proteins move! Protein dynamics and long-range allostery in cell signaling". Protein Structure and Diseases. Advances in Protein Chemistry and Structural Biology. 83. pp. 163–221. doi:10.1016/B978-0-12-381262-9.00005-7. ISBN 9780123812629. PMID 21570668.
13. Kamerlin SC, Warshel A (May 2010). "At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?". Proteins. 78 (6): 1339–75. doi:10.1002/prot.22654. PMC 2841229. PMID 20099310.
14. Collins MO, Yu L, Campuzano I, Grant SG, Choudhary JS (July 2008). "Phosphoproteomic analysis of the mouse brain cytosol reveals a predominance of protein phosphorylation in regions of intrinsic sequence disorder" (PDF). Molecular & Cellular Proteomics. 7 (7): 1331–48. doi:10.1074/mcp.M700564-MCP200. PMID 18388127. S2CID 22193414.
15. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK (Oc-

tober 2002). "Intrinsic disorder in cell-signaling and cancer-associated proteins". Journal of Molecular Biology. 323 (3): 573–84. CiteSeerX 10.1.1.132.682. doi:10.1016/S0022-2836(02)00969-5. PMID 12381310.

16. Sandhu KS (2009). "Intrinsic disorder explains diverse nuclear roles of chromatin remodeling proteins". Journal of Molecular Recognition. 22 (1): 1–8. doi:10.1002/jmr.915. PMID 18802931. S2CID 33010897.
17. Wilson, Benjamin A.; Foy, Scott G.; Neme, Rafik; Masel, Joanna (24 April 2017). "Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth". Nature Ecology & Evolution. 1 (6): 0146–146. doi:10.1038/s41559-017-0146. PMC 5476217. PMID 28642936.
18. Willis, Sara; Masel, Joanna (September 2018). "Gene Birth Contributes to Structural Disorder Encoded by Overlapping Genes". Genetics. 210 (1): 303–313. doi:10.1534/genetics.118.301249. PMC 6116962. PMID 30026186. Lee SH, Kim DH, Han JJ, Cha EJ, Lim JE, Cho YJ, Lee C, Han KH (February 2012). "Understanding pre-structured motifs (PreSMos) in intrinsically unfolded proteins". Current Protein & Peptide Science. 13 (1): 34–54. doi:10.2174/138920312799277974. PMID 22044148.
19. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN (October 2006). "Analysis of molecular recognition features (MoRFs)". Journal of Molecular Biology. 362 (5): 1043–59. doi:10.1016/j.jmb.2006.07.087. PMID 16935303.
20. Gunasekaran K, Tsai CJ, Kumar S, Zanuy D, Nussinov R (February 2003). "Extended disordered proteins: targeting function with less scaffold". Trends in Biochemical Sciences. 28 (2): 81–5. doi:10.1016/S0968-0004(03)00003-3. PMID 12575995.
21. Sandhu KS, Dash D (July 2007). "Dynamic alpha-helices: conformations that do not conform". Proteins. 68 (1): 109–22. doi:10.1002/prot.21328. PMID 17407165. S2CID 96719019.
22. Tarakhovsky A, Prinjha RK (July 2018). "Drawing on disorder: How viruses use histone mimicry to their advantage". The Journal of Experimental Medicine. 215 (7): 1777–1787. doi:10.1084/jem.20180099. PMC 6028506. PMID 29934321.
23. Atkinson SC, Audsley MD, Lieu KG, Marsh GA, Thomas DR, Heaton SM, Paxman JJ, Wagstaff KM, Buckle AM, Moseley GW, Jans DA, Borg NA (January 2018). "Recognition by host nuclear transport proteins drives disorder-to-order transition in Hendra virus V". Scientific Reports. 8 (1): 358. Bibcode:2018NatSR...8..358A. doi:10.1038/s41598-017-18742-8. PMC 5762688. PMID 29321677.
24. Fuxreiter M (January 2012). "Fuzziness: linking regulation to protein dynamics". Molecular BioSystems. 8 (1): 168–77. doi:10.1039/c1mb05234a. PMID 21927770.
25. Fuxreiter M, Simon I, Bondos S (August 2011). "Dynamic protein-DNA recognition: beyond what can be seen". Trends in Biochemical Sciences. 36 (8): 415–23. doi:10.1016/j.tibs.2011.04.006. PMID 21620710.
26. Borgia A, Borgia MB, Bugge K, Kissling VM, Heidarsson PO, Fer-

nandes CB, Sottini A, Soranno A, Buholzer KJ, Nettels D, Kragelund BB, Best RB, Schuler B (March 2018). "Extreme disorder in an ultrahigh-affinity protein complex". Nature. 555 (7694): 61–66. Bibcode:2018Natur.555…61B. doi:10.1038/nature25762. PMC 6264893. PMID 29466338.

27. Feng H, Zhou BR, Bai Y (November 2018). "Binding Affinity and Function of the Extremely Disordered Protein Complex Containing Human Linker Histone H1.0 and Its Chaperone ProT ". Biochemistry. 57 (48): 6645–6648. doi:10.1021/acs.biochem.8b01075. PMC 7984725. PMID 30430826. Uversky VN (August 2011). "Intrinsically disordered proteins from A to Z". The International Journal of Biochemistry & Cell Biology. 43 (8): 1090–103. doi:10.1016/j.biocel.2011.04.001. PMID 21501695.
28. Oldfield, C. (2014). "Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions". Annual Review of Biochemistry. 83: 553–584. doi:10.1146/annurev-biochem-072711-164947. PMID 24606139.
29. Theillet FX, Binolfi A, Bekei B, Martorana A, Rose HM, Stuiver M, Verzini S, Lorenz D, van Rossum M, Goldfarb D, Selenko P (2016). "Structural disorder of monomeric -synuclein persists in mammalian cells". Nature. 530 (7588): 45–50. Bibcode:2016Natur.530…45T. doi:10.1038/nature16531. PMID 26808899. S2CID 4461465.
30. Minde DP, Ramakrishna M, Lilley KS (2018). "Biotinylation by proximity labelling favours unfolded proteins". bioRxiv. doi:10.1101/274761.
31. Minde DP, Ramakrishna M, Lilley KS (2020). "Biotin proximity tagging favours unfolded proteins and enables the study of intrinsically disordered regions". Communications Biology. 3 (1): 38. doi:10.1038/s42003-020-0758-y. PMC 6976632. PMID 31969649.
32. Minde DP, Maurice MM, Rüdiger SG (2012). Uversky VN (ed.). "Determining biophysical protein stability in lysates by a fast proteolysis assay, FASTpp". PLOS ONE. 7 (10): e46147. Bibcode:2012PLoSO…746147M. doi:10.1371/journal.pone.0046147. PMC 3463568. PMID 23056252.
33. Park C, Marqusee S (March 2005). "Pulse proteolysis: a simple method for quantitative determination of protein stability and ligand binding". Nature Methods. 2 (3): 207–12. doi:10.1038/nmeth740. PMID 15782190. S2CID 21364478.
34. Robaszkiewicz K, Ostrowska Z, Cyranka-Czaja A, Moraczewska J (May 2015). "Impaired tropomyosin-troponin interactions reduce activation of the actin thin filament". Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1854 (5): 381–90. doi:10.1016/j.bbapap.2015.01.004. PMID 25603119.
35. Minde DP, Radli M, Forneris F, Maurice MM, Rüdiger SG (2013). Buckle AM (ed.). "Large extent of disorder in Adenomatous Polyposis Coli offers a strategy to guard Wnt signalling against point mutations". PLOS ONE. 8 (10): e77257. Bibcode:2013PLoSO…877257M. doi:10.1371/journal.pone.0077257. PMC 3793970. PMID 24130866.
36. Brucale M, Schuler B, Samorì B (March 2014). "Single-molecule studies of intrinsically disordered proteins". Chemical Reviews. 114 (6): 3281–317.

doi:10.1021/cr400297g. PMID 24432838.

37. Neupane K, Solanki A, Sosova I, Belov M, Woodside MT (2014). "Diverse metastable structures formed by small oligomers of -synuclein probed by force spectroscopy". PLOS ONE. 9 (1): e86495. Bibcode:2014PLoSO…986495N. doi:10.1371/journal.pone.0086495. PMC 3901707. PMID 24475132.
38. Japrung D, Dogan J, Freedman KJ, Nadzeyka A, Bauerdick S, Albrecht T, Kim MJ, Jemth P, Edel JB (February 2013). "Single-molecule studies of intrinsically disordered proteins using solid-state nanopores". Analytical Chemistry. 85 (4): 2449–56. doi:10.1021/ac3035025. PMID 23327569.
39. Min D, Kim K, Hyeon C, Cho YH, Shin YK, Yoon TY (2013). "Mechanical unzipping and rezipping of a single SNARE complex reveals hysteresis as a force-generating mechanism". Nature Communications. 4 (4): 1705. Bibcode:2013NatCo…4.1705M. doi:10.1038/ncomms2692. PMC 3644077. PMID 23591872.
40. Miyagi A, Tsunaka Y, Uchihashi T, Mayanagi K, Hirose S, Morikawa K, Ando T (September 2008). "Visualization of intrinsically disordered regions of proteins by high-speed atomic force microscopy". ChemPhysChem. 9 (13): 1859–66. doi:10.1002/cphc.200800210. PMID 18698566.
41. Campen, Andrew; Williams, Ryan M.; Brown, Celeste J.; Meng, Jingwei; Uversky, Vladimir N.; Dunker, A. Keith (2008). "TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder". Protein and Peptide Letters. 15 (9): 956–963. doi:10.2174/092986608785849164. ISSN 0929-8665. PMC 2676888. PMID 18991772.
42. Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M, Rost B (June 2011). "Protein disorder–a breakthrough invention of evolution?". Current Opinion in Structural Biology. 21 (3): 412–8. doi:10.1016/j.sbi.2011.03.014. PMID 21514145.
43. Tompa, P. (2011). "Unstructural biology coming of age". Current Opinion in Structural Biology. 21 (3): 419–425. doi:10.1016/j.sbi.2011.03.012. PMID 21514142.
44. Ferron F, Longhi S, Canard B, Karlin D (October 2006). "A practical overview of protein disorder prediction methods". Proteins. 65 (1): 1–14. doi:10.1002/prot.21075. PMID 16856179. S2CID 30231497.
45. Uversky VN, Oldfield CJ, Dunker AK (2008). "Intrinsically disordered proteins in human diseases: introducing the D2 concept". Annual Review of Biophysics. 37: 215–46. doi:10.1146/annurev.biophys.37.032807.125924. PMID 18573080.
46. Wise-Scira O, Dunn A, Aloglu AK, Sakallioglu IT, Coskuner O (March 2013). "Structures of the E46K mutant-type -synuclein protein and impact of E46K mutation on the structures of the wild-type -synuclein protein". ACS Chemical Neuroscience. 4 (3): 498–508. doi:10.1021/cn3002027. PMC 3605821. PMID 23374074.
47. Dobson CM (December 2003). "Protein folding and misfolding". Nature. 426 (6968): 884–90. Bibcode:2003Natur.426..884D. doi:10.1038/nature02

261. PMID 14685248. S2CID 1036192.
48. Best RB, Zhu X, Shim J, Lopes PE, Mittal J, Feig M, Mackerell AD (September 2012). "Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone , and side-chain (1) and (2) dihedral angles". Journal of Chemical Theory and Computation. 8 (9): 3257–3273. doi:10.1021/ct300400x. PMC 3549273. PMID 23341755.
49. Best RB (February 2017). "Computational and theoretical advances in studies of intrinsically disordered proteins". Current Opinion in Structural Biology. 42: 147–154. doi:10.1016/j.sbi.2017.01.006. PMID 28259050.
50. Chong SH, Chatterjee P, Ham S (May 2017). "Computer Simulations of Intrinsically Disordered Proteins". Annual Review of Physical Chemistry. 68: 117–134. Bibcode:2017ARPC...68..117C. doi:10.1146/annurev-physchem-052516-050843. PMID 28226222.
51. Fox SJ, Kannan S (September 2017). "Probing the dynamics of disorder". Progress in Biophysics and Molecular Biology. 128: 57–62. doi:10.1016/j.pbiomolbio.2017.05.008. PMID 28554553.
52. Terakawa T, Takada S (September 2011). "Multiscale ensemble modeling of intrinsically disordered proteins: p53 N-terminal domain". Biophysical Journal. 101 (6): 1450–8. Bibcode:2011BpJ...101.1450T. doi:10.1016/j.bpj.2011.08.003. PMC 3177054. PMID 21943426.
53. Fisher CK, Stultz CM (June 2011). "Constructing ensembles for intrinsically disordered proteins". Current Opinion in Structural Biology. 21 (3): 426–31. doi:10.1016/j.sbi.2011.04.001. PMC 3112268. PMID 21530234.
54. Apicella A, Marascio M, Colangelo V, Soncini M, Gautieri A, Plummer CJ (June 2017). "Molecular dynamics simulations of the intrinsically disordered protein amelogenin". Journal of Biomolecular Structure & Dynamics. 35 (8): 1813–1823. doi:10.1080/07391102.2016.1196151. PMID 27366858. S2CID 205576649.
55. Zerze GH, Miller CM, Granata D, Mittal J (June 2015). "Free energy surface of an intrinsically disordered protein: comparison between temperature replica exchange molecular dynamics and bias-exchange metadynamics". Journal of Chemical Theory and Computation. 11 (6): 2776–82. doi:10.1021/acs.jctc.5b00047. PMID 26575570.
56. Granata D, Baftizadeh F, Habchi J, Galvagnion C, De Simone A, Camilloni C, Laio A, Vendruscolo M (October 2015). "The inverted free energy landscape of an intrinsically disordered peptide by simulations and experiments". Scientific Reports. 5: 15449. Bibcode:2015NatSR...515449G. doi:10.1038/srep15449. PMC 4620491. PMID 26498066.
57. Iida, Shinji; Kawabata, Takeshi; Kasahara, Kota; Nakamura, Haruki; Higo, Junichi (2019-03-22). "Multimodal Structural Distribution of the p53 C-Terminal Domain upon Binding to S100B via a Generalized Ensemble Method: From Disorder to Extradisorder". Journal of Chemical Theory and Computation. 15 (4): 2597–2607. doi:10.1021/acs.jctc.8b01042. ISSN 1549-9618. PMID 30855964.
58. Kurcinski M, Kolinski A, Kmiecik S (June 2014). "Mechanism of Folding

and Binding of an Intrinsically Disordered Protein As Revealed by ab Initio Simulations". Journal of Chemical Theory and Computation. 10 (6): 2224–31. doi:10.1021/ct500287c. PMID 26580746.

59. Ciemny, Maciej Pawel; Badaczewska-Dawid, Aleksandra Elzbieta; Pikuzinska, Monika; Kolinski, Andrzej; Kmiecik, Sebastian (2019). "Modeling of Disordered Protein Structures Using Monte Carlo Simulations and Knowledge-Based Statistical Force Fields". International Journal of Molecular Sciences. 20 (3): 606. doi:10.3390/ijms20030606. PMC 6386871. PMID 30708941.
60. Uversky VN (2013). "Digested disorder: Quarterly intrinsic disorder digest (January/February/March, 2013)". Intrinsically Disordered Proteins. 1 (1): e25496. doi:10.4161/idp.25496. PMC 5424799. PMID 28516015.
61. Costantini S, Sharma A, Raucci R, Costantini M, Autiero I, Colonna G (March 2013). "Genealogy of an ancient protein family: the Sirtuins, a family of disordered members". BMC Evolutionary Biology. 13: 60. doi: 10.1186/1471-2148-13-60. PMC 3599600. PMID 23497088.

Chapter 8

J

8.1 JoVe Videos

The JoVE Biotechnology education library provides high-level video summaries of key topics. A selection of videos are linked below. Additional videos are available at the JoVe site.

8.1.1 Biotechnology

- Genomics
- PCR
- Complementary DNA (cDNA)
- Recombinant DNA)
- Genetic engineering

8.1.2 Macromolecules

Macromolecules

- What are proteins
- Protein organization
- Protein folding
- What are nucleic acids
- Phosphodiester linkages

8.1.3 Cellular respiration

Cellcular respiration main page: https://www.jove.com/science-education-library/51/cellular-respiration

8.1.4 Population and community ecology

https://www.jove.com/science-education-library/74/population-and-community-ecology

- What are Populations and Communities?
- Life Histories
- Energy Budgets
- Population Growth
- Ecological Niches
- Ecological Succession
- Keystone Species
- Symbiosis
- Competition
- Predator-Prey Interactions
- Ecological Disturbance

8.1.5 Population genetics

https://www.jove.com/science-education-library/80/population-genetics

- What is population genetics
- Hardy-Weinberg Principle
- Mutation, Gene Flow, and Genetic Drift
- Genetic Drift
- Gene flow

Chapter 9

K

9.1 Talking glossary: Knockout (Genetic knockout, gene knockout)

Introduction: A knockout typically refers to an organism that has been genetically engineered to lack one or more specific genes. Scientists create knockouts (often in mice) so that they can study the impact of the missing genes and learn something about the genes' function.

Audio: https://www.genome.gov/genetics-glossary/Knockout

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/knockout.jpg

Transcript: Knockout does not refer to boxing, at least in this context. But it is a similar concept that what one is doing is inactivating a gene instead of inactivating your partner in boxing. This can be done in a variety of ways. Either one can insert a stop codon in the midst of a gene that one wants to eliminate the function of. Because, as we know, genes which have stock codons in the middle of them, when they occur naturally, they're frequently referred to as disease gene mutations, and they do that because they eliminate the function of the gene. So one can do this experimentally too, and when we do it experimentally on purpose–that's called a knockout. If one really wants to be extreme, one can knock out a gene by taking out the entire gene. One can go in and cut it at one end, cut it at the other, and take out the entire gene. That's the most extreme knockout. Then in that case, you're knocking the boxer out of the ring, so there's no possibility that he's going to come back. In the case of an induced stop mutation, it's a little bit like knocking out the boxer. It's likely he's not going to get up again, but it's always possible that he might. So you take a little bit of a chance. It's a little easier than knocking him out of the ring, but there's still a bit of a chance that your knockout may be incomplete.

Christopher P. Austin, M.D.

Chapter 10

L

10.1 Talking Glossary: Locus (1 min)

https://www.genome.gov/genetics-glossary/Locus

Abstract: "A locus is the specific physical location of a gene or other DNA sequence on a chromosome, like a genetic street address. The plural of locus is"loci"."

NOTE: Locus can refer to a single nucleotide in DNA or 3 nucleotides making up a codon in polypeptides.

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/locus.mp3

Transcript: "Locus" is a term that we use to tell us where on a chromosome a specific gene is. So it's really the physical location of a gene or of a DNA polymorphism on a chromosome. And it's sort of like a street address for people. And one of the things that we think about when we're thinking about genes and chromosomes is we may think of the chromosome as a country, and then a region of a chromosome would maybe be the city, and then we'll get down to a very specific area, which is the locus, and that would be equivalent to, say, a person's street address. And that's the street address of that gene. An important thing to remember is that the plural of "locus" is "loci", not "locuses".

Joan E. Bailey-Wilson, Ph.D.

Chapter 11

M

11.1 Manhattan Plot

This article is adapted Wikipedia, the free encyclopedia:https://en.wikipedia.org/wiki/Manhattan_plot

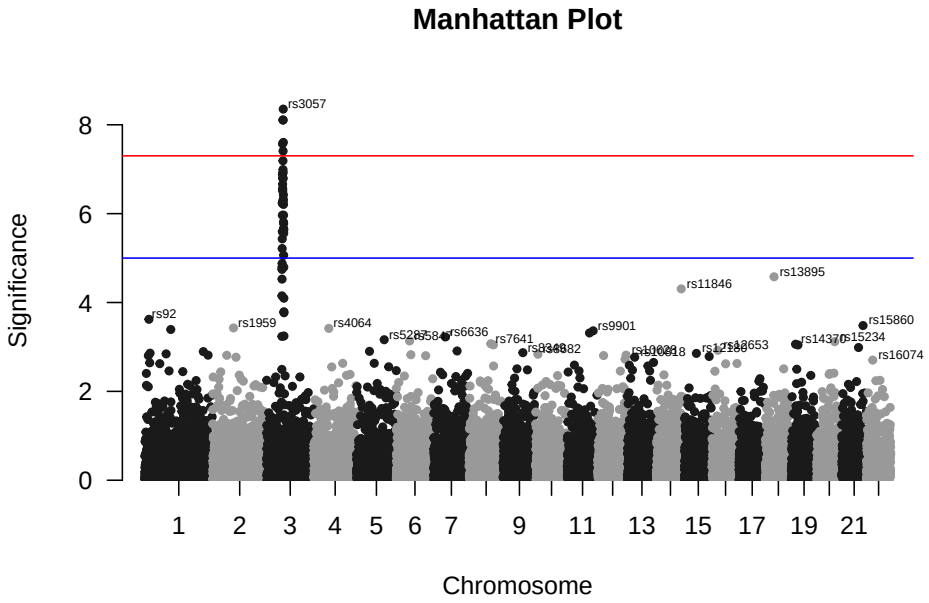
A **Manhattan plot** is a type of scatter plot commonly used in genome-wide association studies (GWAS) to display significant SNPs (single nucleotide polymorphisms). The data being plotted has some unique features because

1. All values are non-negative
2. Most values are low
3. Researchers are interested in only the highest values, which indicate SNPs which are near parts of a genome of potential biological importance.

```
## Load packages and data
#install.packages("qqman")
library(qqman)
data(gwasResults)

# ## Investigate results
# dim(gwasResults)
# names(gwasResults)
#
#
# ## Distributio of P values
# #hist(gwasResults$P)
#
# # Make Manhattan plot
# manhattan(gwasResults)
```

```
manhattan(gwasResults,
          annotatePval = 0.005,
          main = "Manhattan Plot",
          ylab = "Significance",
          annotateTop = TRUE)
```



Manhattan plots get their name from the similarity of such a plot to the Manhattan skyline: a profile of skyscrapers towering above the lower level "buildings" which vary around a lower height.

11.1.1 GWAS

In GWAS Manhattan plots, genomic coordinates are displayed along the X-axis, with each dot on the Manhattan plot signifies a SNP. The Y-axis represents the strength statistical significance of the SNP-phenotype association. The different colors of each block usually show the extent of each chromosome.

11.2 Talking Glossary: Microarray

https://www.genome.gov/genetics-glossary/Microarray-Technology

Microarray technology is a developing technology used to study the expression of many genes at once. It involves placing thousands of gene sequences in known locations on a glass slide called a gene chip. A sample containing DNA or RNA is placed in contact with the gene chip. Complementary base pairing between the sample and the gene sequences on the chip produces light that is measured.

Areas on the chip producing light identify genes that are expressed in the sample.

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/microarray_technology.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/microarray_technology.mp3

Microarrays are a technology that basically miniaturize processes that have been used in molecular genetics laboratories for years. They allow detection of DNA or RNA molecules by hybridizing or sticking to target DNA molecules or RNA molecules on a glass slide, with detection of that adherent DNA or RNA molecule by various labels or dyes that allow them to be seen under a microscope.

Leslie G. Biesecker, M.D.

11.3 Talking Glossary: Missense Mutation (1.25 min)

https://www.genome.gov/genetics-glossary/Missense-Mutation

Abstract: "A missense mutation is when the change of a single base pair causes the substitution of a different amino acid in the resulting protein. This amino acid substitution may have no effect, or it may render the protein nonfunctional."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/missense_mutation.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/missense_mutation.mp3

Transcript: "A missense mutation is a mistake in the DNA which results in the wrong amino acid being incorporated into a protein because of change, that single DNA sequence change, results in a different amino acid codon which the ribosome recognizes. Changes in amino acid can be very important in the function of a protein. But sometimes they make no difference at all, or very little difference. Sometimes missense mutations cause amino acids to be incorporated, which make the protein more effective in doing its job. More frequently, it causes the protein to be less effective in doing its job. But this is really the grist of evolution, when missense mutations happen, and therefore small changes, frequently small changes in proteins, happen, and it happens to be that it improves the function of a protein. That will sometimes give the organism that has it a competitive advantage over its colleagues and be maintained in the population."

Christopher P. Austin, M.D.

11.4 Talking Glossary: Mitochondrial DNA (1.5 min)

https://www.genome.gov/genetics-glossary/Mitochondrial-DNA

Abstract: "Mitochondrial DNA is the small circular chromosome found inside mitochondria. The mitochondria are organelles found in cells that are the sites of energy production. The mitochondria, and thus mitochondrial DNA, are passed [almost always - but there are exceptions!] from mother to offspring."

Note: Mitochondrial DNA is frequently used in population genetics and phylogenetics. Its almost always (like 99.999999% of the time) inherited maternally which has useful properties when constructing phylogenies and tracing the history of populations.

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/mitochondrial_dna.jpg Mitochondria and mitochondrial genome.

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/mitochondrial_dna.mp3

Animation: https://youtu.be/TiqTAfHhOCo

Transcript: "Inside the mitochondrion is a certain type of DNA. That's different in a way from the DNA that's in the nucleus. This DNA is small and circular. It has only 16,500 or so base pairs in it. And it encodes different proteins that are specific for the mitochondrial. Now, remember those pathways that are within the mitochondrion for producing energy. Some of the enzymes in those pathways, and some of the proteins that are needed to function in those pathways, are produced by the mitochondrial DNA. The mitochondrial DNA is critically important for many of the pathways that produce energy within the mitochondria. And if there's a defect in some of those mitochondrial DNA bases, that is to say a mutation, you will have a mitochondrial disease, which will involve the inability to produce sufficient energy in things like the muscle and the brain, and the kidney. Mitochondrial DNA, unlike nuclear DNA, is inherited from the mother [almost always - but there are exceptions!], while nuclear DNA is inherited from both parents. So this is very helpful sometimes in determining how a person has a certain disorder in the family. Sometimes a disease will be inherited through the mother's line, as opposed to both parents. You can tell from a pedigree or a group of family history whether or not this is a mitochondrial disease because of that."

William Gahl, M.D., Ph.D.

11.5 Talking Glossary: Mutagen (0.5 min)

Abstract: "A mutagen is a chemical or physical phenomenon, such as ionizing radiation, that promotes errors in DNA replication. Exposure to a mutagen can

produce DNA mutations that cause or contribute to diseases such as cancer."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/mutagen.mp3

Transcript: "A mutagen is a chemical or physical agent that has the ability to change our genetic code in a harmful way. The change in the genetic code is called a mutation, and throughout our lifetime we actually accumulate many mutations within our cells. And our body has the ability to recognize and repair these mutations. However, if some of these mutations escape repair, they can cause a normal cell to be transformed to become a tumor cell. Therefore, mutations are actually associated with the development of cancer."

Daphne W. Bell, Ph.D. Photo: https://i.irp.nih.gov/pi/0012727407.jpg

11.6 Talking Glossary: Mutation (0.5 min)

Abstract: "A mutation is a change in a DNA sequence. Mutations can result from DNA copying mistakes made during cell division, exposure to ionizing radiation, exposure to chemicals called mutagens, or infection by viruses. Germ line mutations occur in the eggs and sperm and can be passed on to offspring, while somatic mutations occur in body cells and are not passed on."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/mutation.mp3

Image: Sequence-scale mutations ("micro") versus chromosome-scale mutations ("macro") https://www.genome.gov/sites/default/files/tg/en/illustration/mutation.jpg

Transcript: "Mutation has been the source of many Hollywood movies, but it's really a simple process of a mistake made in a DNA sequence as it's being copied. Some of that's just the background noise that DNA copying is not perfect, and we should be glad of that or evolution couldn't operate. But mutation can also be induced by things like radiation or carcinogens in a way that can increase the risk of cancers or birth defects. But it's pretty simple; it's basically an induced misspelling of the DNA sequence. That's a mutation"

Francis S. Collins, M.D., Ph.D.

Chapter 12

N

12.1 Talking Glossary: Non-coding DNA (1 min)

Abstract: "Non-coding DNA sequences do not code for amino acids. Most non-coding DNA lies between genes on the chromosome and has no known function. Other non-coding DNA, called introns, is found within genes. Some non-coding DNA plays a role in the regulation of gene expression."

Note: Non-coding DNA (previously called junk DNA) makes up 98% of the genome. It is very useful for evolutionary and phylogenetic studies because it is not directly impacted by natural selection.

Transcript: "Non-coding DNA is just what it says; it's non-coding DNA. You can think of the genome as being split up into two parts. There's the stuff that codes for proteins. We call it coding DNA, and for a lack of a better term, the rest of genome is referred to as non-coding DNA. Some people will like to try and refer to this as junk DNA. But I would suggest otherwise, because this represents 98 percent of our genome sequence and it does all sorts of things, like regulate those genes to figure out where they should turn on, where they should turn off, how much we should turn on certain genes, how are we going to pack up the DNA into chromosomes, and so forth. And there are probably a whole host of functions that non-coding DNA does that we still don't know what it does yet."

Elliott Margulies, Ph.D.

12.2 Talking Glossary: Nonsene mutation

https://www.genome.gov/genetics-glossary/Nonsense-Mutation

Abstract: "A nonsense mutation is the substitution of a single base pair that leads to the appearance of a stop codon where previously there was a codon specifying an amino acid. The presence of this premature stop codon results in the production of a shortened, and likely nonfunctional, protein."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/nonsense_mutation.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/nonsense_mutation.jpg

Transcript

"A nonsense mutation, or its synonym, a stop mutation, is a change in DNA that causes a protein to terminate or end its translation earlier than expected. This is a common form of mutation in humans and in other animals that causes a shortened or nonfunctional protein to be expressed."

Leslie G. Biesecker, M.D.

12.3 Nucleic Acids - Overview

Authors: OpenStax / Libretext Formatted in RMarkdown by Nathan Brouwer under the Creative Commons Attribution License 4.0 license.

This chapter was adapted from LibreText General Biology, Chapter 3, Section 3.5: Nucleic Acids. The LibreText book is based on OpenStax Biology 2nd edition, Chapter 3, Section 3.5: Nucleic Acids.

Additional material taken from LibreText General Biology, Chapter 3, Section 3.4: Proteins. The LibreText book is based on OpenStax Biology 2nd edition, Chapter 3, Section 3.4: Nucleic Acids.

A full list of authors is found under the **Contributors and Attributions** section at the end of this document.

12.3.1 DNA and RNA

Nucleic acids are the most important molecules for the continuity of life. They carry the blueprint of a cell and carry instructions for the functioning of cell and organisms. There are two main types of nucleic acids: **deoxyribonucleic acid (DNA)** and **ribonucleic acid (RNA)**. DNA is the **genetic material** found in all living organisms, ranging from single-celled bacteria to **multicellular** mammals.

The entire genetic content of a cell is known as its **genome**, and the study of genomes is **genomics**. Many – but not all – genes contain the information to make proteins.

Proteins are one of the most abundant biological molecules in living systems and have a diverse range of functions. Each cell in a living system may contain thousands of proteins, each with a unique function. Their structures, like their functions, vary greatly. They are all, however, **polymers** of **amino acids**, arranged in a linear sequence. This sequence is determined by DNA.

DNA molecules don't directly code for proteins, but rather use an intermediary to communicate with the rest of the cell. This intermediary is **messenger RNA (mRNA)**.

DNA and RNA are made up of single subunits known as **nucleotides**. The nucleotides combine with each other to form a long chain of either DNA or RNA. The sequences of the nucleotides contains the information to make proteins. In DNA there are four nucleotides: adenine (A), guanine (G) cytosine (C), and thymine (T).

12.3.2 DNA Double-Helix Structure

DNA has a double-helix structure made up of two separate strands of nucleotides. The two strands of the helix run in opposite directions.

Only certain types of base pairing are allowed: A can pair with T, and G can pair with C. This is known as the **base complementary rule**. In other words, the DNA strands are complementary to each other. If the sequence of one strand is AATTGGCC, the complementary strand would have the sequence TTAACCGG. During DNA replication, each strand is copied, resulting in a new DNA double helix.

12.3.3 RNA

Ribonucleic acid, or RNA, is mainly involved in the process of **protein synthesis** . RNA is usually **single-stranded** and is made of ribonucleotides, which are slightly different than nucleotides. RNA can contain adenine, guanine, cytosine, or uracil ribonucleotides. Importantly there is no thymine ribonucleotide - uracil takes its place.

RMA molecules called **messenger RNA** (mRNA) carry the message from DNA. If a cell requires a certain protein to be synthesized, the gene for this product is turned "on" and the messenger RNA is made using the appropriate DNA as a template. The RNA base sequence is complementary to the coding sequence of the DNA from which it has been copied. However, as just noted, in RNA, the base T is absent and U is present instead. If the DNA strand has a sequence AATTGCGC, the sequence of the complementary RNA is UUAACGCG.

Once mRNA is made is can serve as a template for the synthesis of **proteins**. Proteins or the primary building blocks of cells and thereby tissues, organs and organisms. For example, muscles are made up of long filaments of large cells containing bundles of proteins.

mRNA is read in sets of three bases known as **codons**. Each codon codes for a single amino acid. Amino acids are the **monomers** (subunits) that make up proteins. In this way, the mRNA is read and the protein product is made.

Nucleotides and ribonucleotides are represented by a single letter: A, T, C, G and U. Similarly, amino acids are represented by a single upper case letter or a three-letter abbreviation. For example, valine is known by the letter V or the three-letter code val. The sequence and the number of amino acids ultimately determine a protein's shape, size, and function.

Information flow in an organism takes place from DNA to RNA to protein. DNA dictates the structure of mRNA in a process known as **transcription**, and RNA dictates the structure of protein in a process known as **translation**. This is known as the **"Central Dogma"** of molecular biology, (though there are important exceptions).

12.3.4 Protein function

The shape of a protein is critical to its function. For example, enzymes bind to other molecules at a site known as their **active site**. If this active site is altered because of the interference of a chemical such as a drug or toxin, the enzyme may be unable to bind to the substrate. Similarly, a mutation in the DNA that codes for the protein can change the shape of the active site and prevent it from functioning properly.

The unique sequence for every protein is therefore ultimately determined by the gene encoding the protein. A change in nucleotide sequence of the gene's coding region may lead to a different amino acid being added to the growing polypeptide chain, causing a change in protein structure and function. In condition sickle cell anemia, part of the protein hemoglobin has a single amino acid substitution, causing a change in protein structure and function. Hemoglobin has about 600 amino acids. The structural difference between a normal hemoglobin molecule and a sickle cell molecule – which dramatically decreases life expectancy - is a single amino acid of the 600. What is even more remarkable is that those 600 amino acids are encoded by three nucleotides each, and the mutation is caused by a single base change (point mutation), 1 in 1800 bases. Because of this change of one amino acid in the chain (and 1 nucleotide in the underlying DNA code), hemoglobin molecules form long fibers that distort the normal shape of red blood cells, turning them into a crescent or "sickle" shape, which clogs arteries. This can lead to myriad serious health problems such as breathlessness, dizziness, headaches, and abdominal pain for those affected by this disease.

12.3.5 Summary

Nucleic acids are molecules made up of **nucleotides** that direct cellular activities such as cell division and protein synthesis. There are two types of nucleic acids: DNA and RNA. DNA carries the genetic blueprint of the cell and is

passed on from parents to offspring (in the form of chromosomes). It has a double-helical structure with the two strands running in opposite directions. RNA is single-stranded. RNA provides the template for protein synthesis. Messenger RNA (mRNA) is copied from the DNA and contains information for the construction of proteins.

12.3.6 Glossary

amino acid: monomer of a protein; has a central carbon or alpha carbon to which an amino group, a carboxyl group, a hydrogen, and an R group or side **deoxyribonucleic acid (DNA)**: double-helical molecule that carries the hereditary information of the cell
messenger RNA (mRNA): DNA that carries information from DNA to allow protein synthesis
nucleic acid: biological molecule that carries the genetic blueprint of a cell and carries instructions for the functioning of the cell
nucleotide: monomer of nucleic acids; **protein**: biological macromolecule composed of one or more chains of amino acids
ribonucleic acid (RNA): single-stranded molecule that is involved in protein synthesis

transcription: process through which messenger RNA forms on a template of DNA

translation: process through which RNA directs the formation of protein

12.3.7 Contributors and Attributions

Connie Rye (East Mississippi Community College), Robert Wise (University of Wisconsin, Oshkosh), Vladimir Jurukovski (Suffolk County Community College), Jean DeSaix (University of North Carolina at Chapel Hill), Jung Choi (Georgia Institute of Technology), Yael Avissar (Rhode Island College) among other contributing authors. Original content by OpenStax (CC BY 4.0; Download for free at http://cnx.org/contents/185cbf87-c72…f21b5eabd@9.87).

Chapter 13

O

13.1 Talking Glossary : Open reading frame definition (3 min)

National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Open-Reading-Frame (Links to an external site.)

Abstract: "An open reading frame is a portion of a DNA molecule that, when translated into amino acids, contains no stop codons. The genetic code reads DNA sequences in groups of three base pairs, which means that a double-stranded DNA molecule can read in any of six possible reading frames–three in the forward direction and three in the reverse. A long open reading frame is likely part of a gene."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/open_reading_frame.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/open_reading_frame.jpg Note: This is a good figure, but doesn't contain all the details discussed in the glossary entry.

Transcript: "Open reading frame" is a terrible term that we're stuck with. What it refers to is a frame of reference, and what is being read, "reading", is the RNA code, and it is being read by the ribosomes in order to make a protein. And "open" means that the road is open to keep reading, and the ribosome will be able to keep reading the RNA code and add another amino acid one after another. Now, DNA, though it is a monotonous repetition of As, Cs, Ts, and Gs, has a language, which is transcribed, of course, into RNA and then translated into a protein. And when it's translated into a protein, the mRNA is not read one letter at a time, but it's read three letters at a time. And those three letters are called a codon, and each of those codons, whether it's an AAA or UUU or

an AUG, each of those codons is interpreted by the ribosome, the molecular machine, that's going to make the protein as a certain amino acid. So AUG codes for one amino acid, and UUU codes for another, and etc. So an open reading frame is the length of DNA, or RNA, which is transcribed into RNA, through which the ribosome can travel, adding one amino acid after another before it runs into a codon that doesn't code for any amino acid. And when that happens, it confuses the ribosome, and the ribosome stops. So you'll be pleased to hear that codons, which make that happen are called stop codons, and a stop codon ends an open reading frame. So an open reading frame is sometimes 300 amino acids long, and sometimes maybe it's 600, and sometimes it's longer. The longer an open reading frame is, the longer you get before you get to a stop codon, the more likely it is to be part of a gene which is coding for a protein. Now the finally confusing thing about an open reading frame is that because the codons are three nucleic acids long and DNA has two strands, the ribosome can read an RNA derived from one strand or another, and it can read it in 1-2-3s that are separated one from another so you can actually get three reading frames reading in one direction, three reading frames going in the other direction. So it's actually six different reading frames for every piece of DNA, which might give you an open reading frame."

Christopher P. Austin, M.D.

Chapter 14

P

14.1 Talking Glossary: PCR - The Polymerase Chain Reaction (0.5 min)

National Human Genome Research Institute

https://www.genome.gov/genetics-glossary/Polymerase-Chain-Reaction

Abstract: Polymerase chain reaction (PCR) is a laboratory technique used to amplify DNA sequences. The method involves using short DNA sequences called primers to select the portion of the genome to be amplified. The temperature of the sample is repeatedly raised and lowered to help a DNA replication enzyme copy the target DNA sequence. The technique can produce a billion copies of the target sequence in just a few hours.

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/polymerase_chain_reaction.mp3

Image: https://www.genome.gov/sites/default/files/media/images/2021-01/polymerase_chain_reaction.jpg

Transcript: "PCR, or the polymerase chain reaction, is a chemical reaction that molecular biologists use to amplify pieces of DNA. This reaction allows a single or a few copies of DNA to be replicated into millions or billions of copies. And by amplifying that DNA, it allows us to study that DNA molecule in detail in the laboratory."

Leslie G. Biesecker, M.D.

14.2 Protein Data Bank

Adapated from Wikipedia

The Protein Data Bank (PDB)[1] is a database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by **X-ray crystallography**, NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe,[2] PDBj,[3] RCSB,[4] and BMRB[5]).

The PDB is a key in areas of structural biology, such as structural genomics. Most major scientific journals and some funding agencies now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology.[6, 7]

Most structures are determined by X-ray diffraction, but about 10% of structures are determined by protein NMR. When using X-ray diffraction, approximations of the coordinates of the atoms of the protein are obtained, whereas using NMR, the distance between pairs of atoms of the protein is estimated. The final conformation of the protein is obtained from NMR by solving a distance geometry problem. After 2013, a growing number of proteins are determined by cryo-electron microscopy. Clicking on the numbers in the linked external table displays examples of structures determined by that method.

Historically, the number of structures in the PDB has grown at an approximately exponential rate, with 100 registered structures in 1982, 1,000 structures in 1993, 10,000 in 1999, and 100,000 in 2014.[19, 20]

14.2.1 History

** OPTIONAL **

Two forces converged to initiate the PDB: a small but growing collection of sets of protein structure data determined by X-ray diffraction; and the newly available (1968) molecular graphics display, the Brookhaven RAster Display (BRAD), to visualize these protein structures in 3-D. In 1969, with the sponsorship of Walter Hamilton at the Brookhaven National Laboratory, Edgar Meyer (Texas A&M University) began to write software to store atomic coordinate files in a common format to make them available for geometric and graphical evaluation. By 1971, one of Meyer's programs, SEARCH, enabled researchers to remotely access information from the database to study protein structures offline.[8] SEARCH was instrumental in enabling networking, thus marking the functional beginning of the PDB.

The Protein Data Bank was announced in October 1971 in Nature New Biology[9] as a joint venture between Cambridge Crystallographic Data Centre, UK and Brookhaven National Laboratory, US.

Upon Hamilton's death in 1973, Tom Koeztle took over direction of the PDB for the subsequent 20 years. In January 1994, Joel Sussman of Israel's Weizmann

Institute of Science was appointed head of the PDB. In October 1998,[10] the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB);[11] the transfer was completed in June 1999. The new director was Helen M. Berman of Rutgers University (one of the managing institutions of the RCSB, the other being the San Diego Supercomputer Center at UC San Diego).[12] In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe),[2] RCSB (USA), and PDBj (Japan).[3] The BMRB[5] joined in 2006. Each of the four members of wwPDB can act as deposition, data processing and distribution centers for PDB data. The data processing refers to the fact that wwPDB staff review and annotate each submitted entry.[13] The data are then automatically checked for plausibility (the source code[14] for this validation software has been made available to the public at no charge).

For PDB structures determined by X-ray diffraction that have a structure factor file, their electron density map may be viewed. The data of such structures is stored on the "electron density server".[17, 18]

14.2.2 File format

OPTIONAL The file format initially used by the PDB was called the PDB file format. The original format was restricted by the width of computer punch cards to 80 characters per line. Around 1996, the "macromolecular Crystallographic Information file" format, mmCIF, which is an extension of the CIF format was phased in. mmCIF became the standard format for the PDB archive in 2014.[21] In 2019, the wwPDB announced that depositions for crystallographic methods would only be accepted in mmCIF format.[22]

An XML version of PDB, called PDBML, was described in 2005.[23] The structure files can be downloaded in any of these three formats, though an increasing number of structures do not fit the legacy PDB format. Individual files are easily downloaded into graphics packages from Internet URLs:

14.3 PFam

Adapted from Wikipedia.

Pfam is a database of **protein families** that includes their annotations and multiple sequence alignments generated using bioinformatics tools known as hidden Markov models (HMM; 1, 2, 3. The most recent version, Pfam 34.0, was released in March 2021 and contains 19,179 families (4).

14.3.1 Uses

The general purpose of the Pfam database is to provide a complete and accurate classification of protein families and domains.[5] Originally, the rationale behind creating the database was to have a semi-automated method of curating

information on known protein families to improve the efficiency of annotating genomes.[6] The Pfam classification of protein families has been widely adopted by biologists because of its wide coverage of proteins and sensible naming conventions.[7]

It is used by experimental biologists researching specific proteins, by structural biologists to identify new targets for structure determination, by computational biologists to organise sequences and by evolutionary biologists tracing the origins of proteins.[8] Early genome projects, such as human and fly used Pfam extensively for functional annotation of genomic data.[9, 10, 11]

14.4 Talking Glossary: Phenotype

https://www.genome.gov/genetics-glossary/Phenotype

Abstract: "A phenotype is an individual's observable traits, such as height, eye color, and blood type. The genetic contribution to the phenotype is called the genotype. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors."

Audio:

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/phenotype.jpg

Transcript: ""Phenotype" simply refers to an observable trait. "Pheno" simply means "observe" and comes from the same root as the word "phenomenon". And so it's an observable type of an organism, and it can refer to anything from a common trait, such as height or hair color, to presence or absence of a disease. Frequently, phenotypes are related and used–the term is used–to relate a difference in DNA sequence among individuals with a difference in trait, be it height or hair color, or disease, or what have you. But it's important to remember that phenotypes are equally, or even sometimes more greatly influenced by environmental effects than genetic effects. So a phenotype can be directly related to a genotype, but not necessarily. There's usually not a one-to-one correlation between a genotype and a phenotype. There are almost always environmental influences, such as what one eats, how much one exercises, how much one smokes, etc. All of those are environmental influences which will affect the phenotype as well."

Christopher P. Austin, M.D.

14.5 Phylogenetics vocab

Here are definitions of key terms related to phylogenetics.

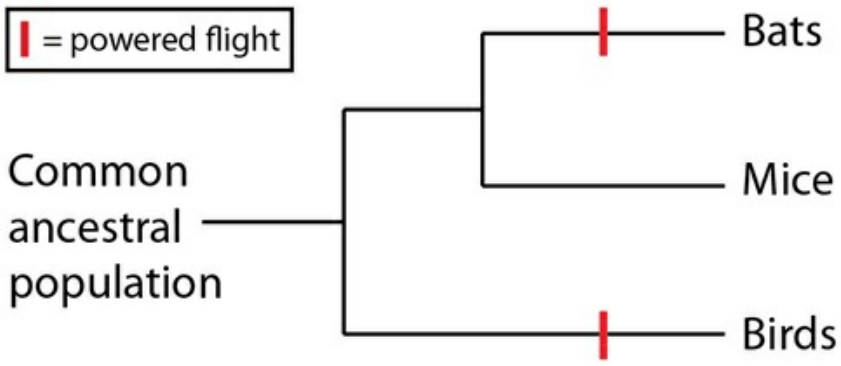
Apomorphy: A character present in one or more species in a clade that was not present in the clade's common ancestor; an evolutionary novelty. Also known

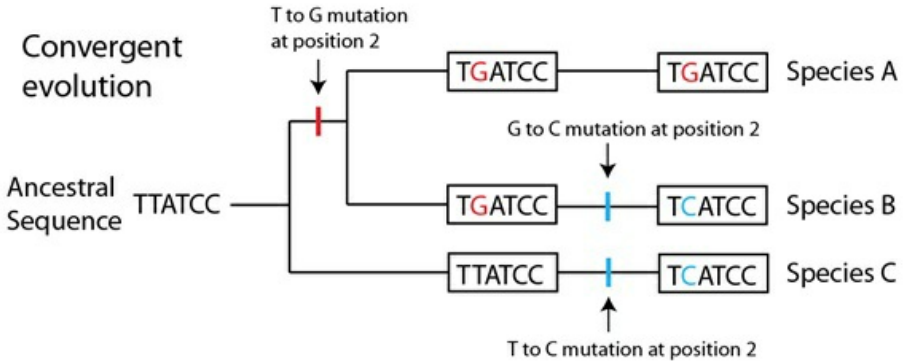
as a derived character. For a review of ancestral versus derived traits see this video . (Contrast with homplasy).

Autapomorphy: A derived character state that is restricted to one taxon in a particular data set. Apomorphies are often biologically interesting (like wings in birds) but actually have no value for building phylogenetic trees because they don't let you form clades. Stated another way, autapomorphis are unique to a single taxon and don't allow you to group taxa into a clade.

Convergence: The multiple and independent appearance in different lineages of similar evolutionary novelties (apomorphies). Often happens due to similar ecological conditions and evolutionary forces.

Convergent evolution: Similarity between species that is caused by a similar, but evolutionary independent, response to a common environmental problem (Figure 1). Convergence can also occur at the molecular level where the same mutation occurs independently in two different lineages (Figure 2).





Derived character: A character present in one or more species in a clade that

was not present in the clade's common ancestor; an evolutionary novelty; also known as an apomorphy. (Contrast with ancestral character)

Homologous: Describes characters derived from a common ancestor. (Contrast with analogous).

Homology: Similarity between species that results from inheritance of traits from a common ancestor. (Contrast with analogy). For more on homologies see Understanding Evolution: Homologies, Homologies: anatomy , Homologies: comparative anatomy, and Homologies: development .

Homoplasy: Technical term which indicates similarity in the characters/traits found in different species that are unrelated to common ancestry. Instead, they can be due to convergent evolution or reversal - not common descent. An analogous trait is a type of homoplasy. For an in-depth discussion of homology versus homoplasy see this video (Contrast with homology).

Lineage: A group of ancestral and descendant populations, species, or other taxa that are descended from a common ancestor. Synonymous with clade.

Outgroup: A taxonomic group that diverged prior to the rest of the focal taxa (focal clade) in a phylogenetic analysis.

Parsimony: A criterion for selecting among alternative patterns or explanations based on minimizing the total amount of change or complexity. Can be approximate as "simpler answers are to be preferred."

Reversal / evolutionary reversal: An event that results in the reversion of a derived trait (apomorphy) to the ancestral form. This happens frequently at the DNA level, where a mutation can change a base to something different, then later another mutation changes it back. This can be called a back mutation. Reversals can also occur at the protein or morphological level.

Trait matrix (aka character matrix, character state matrix, data matrix): A table showing the state of each trait occurring in each taxon. Generally, row represents taxa, columns represent characters, and numbers (usually just 0 or 1) represent character states.

14.6 Talking Glossary: Plasmid (1 min)

National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Plasmid

Intro: "A plasmid is a small, often circular DNA molecule found in bacteria and other cells. Plasmids are separate from the bacterial chromosome and replicate independently of it. They generally carry only a small number of genes, notably some associated with antibiotic resistance. Plasmids may be passed between different bacterial cells."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/plasmid.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/plasmid.mp3

Transcription: "Electrophoresis is a laboratory technique used to separate DNA, RNA, or protein molecules based on their size and electrical charge. An electric current is used to move molecules to be separated through a gel. Pores in the gel work like a sieve, allowing smaller molecules to move faster than larger molecules. The conditions used during electrophoresis can be adjusted to separate molecules in a desired size range."

14.7 Talking Glossary: Point mutation (0.5 min)

Point mutation

https://www.genome.gov/genetics-glossary/Point-Mutation

Abstract: "A point mutation is when a single base pair is altered. Point mutations can have one of three effects. First, the base substitution can be a silent mutation where the altered codon corresponds to the same amino acid. Second, the base substitution can be a missense mutation where the altered codon corresponds to a different amino acid. Or third, the base substitution can be a nonsense mutation where the altered codon corresponds to a stop signal."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/point_mutation.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/point_mutation.mp3

NOTE: Sources vary a little bit in how they define point mutations. Like this glossary entry, I include indels (insertions / deletions) as a type of point mutation.

Transcript: "Point mutations are a large category of mutations that describe a change in single nucleotide of DNA, such that that nucleotide is switched for another nucleotide, or that nucleotide is deleted, or a single nucleotide is inserted into the DNA that causes that DNA to be different from the normal or wild type gene sequence.

Leslie G. Biesecker, M.D."

14.8 Talking Glossary: Polymorphism (1 min)

Abstract: "Polymorphism involves one of two or more variants of a particular DNA sequence [and/or protein sequence]. The most common type of polymorphism involves variation at a single base pair [or amino acid]. Polymorphisms

can also be much larger in size and involve long stretches of DNA. Called a single nucleotide polymorphism, or SNP (pronounced snip), scientists are studying how SNPs in the human genome correlate with disease, drug response, and other phenotypes."

Transcript: "Polymorphism, by strict definitions which hardly anybody pays attention to anymore, is a place in the DNA sequence where there is variation, and the less common variant is present in at least one percent of the people of who you test. That is to distinguish, therefore, polymorphism from a rare variant that might occur in only one in 1,000 people. A polymorphism, it has to occur in at least one in 100 people. Polymorphisms could be not just single-letter changes like a C instead of T. They could also be something more elaborate, like a whole stretch of DNA, that is either present or absent. You might call that a copy number variant; those are all polymorphisms. But this is basically a general term to talk about diversity in genomes in a species."

Francis S. Collins, M.D., Ph.D.

14.9 Talking Glossary U03: Primer (0.5 min)

National Human Genome Research Institute

Introduction: "A primer is a short, single-stranded DNA sequence used in the polymerase chain reaction (PCR) technique. In the PCR method, a pair of primers is used to hybridize with the sample DNA and define the region of the DNA that will be amplified. Primers are also referred to as oligonucleotides."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/primer.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/primer.mp3

Transcript "Primer refers to a small set of nucleotides of DNA, typically 18 to 24 base pairs in length. And a primer can be used for a multitude of other experimental processes. You can use primer in PCR to target a locus to allow for amplification for further analysis. You'd use a primer for sequencing a sequencing reaction where you want to target a very specific region and then do analysis in the extension of that DNA molecule.

Stacie Loftus, Ph.D.

14.10 Talking Glossary: Promoter (1.5 min)

Abstract: "A promoter is a sequence of DNA needed to turn a gene on or off. The process of transcription is initiated at the promoter. Usually found near the beginning of a gene, the promoter has a binding site for the enzyme used to make a messenger RNA (mRNA) molecule."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/promoter.mp3

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/promoter.jpg

Transcript: "The promoter region is the sequence typically referred to that's right upstream or right next to where a gene is about to be transcribed. It's the region where certain regulatory elements will bind; these are proteins that will bind to help RNA get transcribed. Now,"promoter", the term "promoter", can actually be a little bit of a nebulous term because it's not very exact. There are specific sequences that are generally found within a promoter region, but sometimes people refer to even extended promoter region that might include sequences that are farther upstream of the gene that might help enhance or repress the particular gene that's about to be transcribed in certain cell types. In general, if you think of the promoter as that piece of DNA that's just upstream of the transcription start site of a gene, that's pretty much what we refer to as promoters."

Elliott Margulies, Ph.D.

Chapter 15

Q

Chapter 16

R

16.1 Talking Glossary: Recessive

https://www.genome.gov/genetics-glossary/Recessive

Abstract: "Recessive is a quality found in the relationship between two versions of a gene. Individuals receive one version of a gene, called an allele, from each parent. If the alleles are different, the dominant allele will be expressed, while the effect of the other allele, called recessive, is masked. In the case of a recessive genetic disorder, an individual must inherit two copies of the mutated allele in order for the disease to be present."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/recessive.jpg

Transcript: Recessive refers to a type of allele which will not be manifested in an individual unless both of the individual's copies of that gene have that particular genotype. It's usually referred to in conjunction with a Punnett square, other types of Mendelian genetics, and frequently contrasted with a dominant pattern of inheritance wherein if one has one copy of the gene, regardless of what the other copy is, that dominant allele will show itself. In the case of a recessive allele, the individual will show the trait which corresponds to that genotype only if both alleles are the same and have that particular recessive characteristic. Now, that recessive characteristic can be one of no functional consequence. This results in differences between individuals such as in eye color or hair color, but it can also refer to a disease. For instance, in cystic fibrosis, which is a very common Mendelian disorder, that disease exists only when there's a malfunction of both genes that correspond to cystic fibrosis. If there is only one mutation, then that recessive mutation can be compensated for by the normal allele. However, when the function of both are lost, then the disease manifests itself as a recessive disease where there is a loss of function and

therefore observable disease.

Christopher P. Austin, M.D.

16.2 Talking Glossary: Recombinant DNA (0.5 min)

National Human Genome Research Institute https://www.genome.gov/genetics-glossary/Recombinant-DNA

Introduction: "Recombinant DNA (rDNA) is a technology that uses enzymes to cut and paste together DNA sequences of interest. The recombined DNA sequences can be placed into vehicles called vectors that ferry the DNA into a suitable host cell where it can be copied or expressed."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/recombinant_DNA.mp3

Transcript: "Pieces of DNA, such as human DNA, can be engineered in a fashion that allows them to be copied, or replicated, in bacteria or yeast. This involves attaching appropriate elements to a piece of DNA and then transferring into a bacterial or yeast cell, with those elements then instructing the bacterial or yeast cell to copy this DNA along with its own. This process is known as DNA cloning, with the resulting cloned DNA often referred to as recombinant DNA."

Eric D. Green, M.D., Ph.D.

16.3 Talking Glossary: Repressor (1 min)

Abstract: "A repressor is a protein that turns off the expression of one or more genes. The repressor protein works by binding to the gene's promoter region, preventing the production of messenger RNA (mRNA)."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/repressor.mp3

Transcript: "A repressor is a protein that has a negative effect on gene expression. So these usually are proteins that bind to DNA, and they either prevent the RNA transcription machinery from getting in there and transcribing that DNA, or they just slow it down. So repressors are present in cells where you don't want a particular gene expressed. So if the repressor cell recognizes a sequence in that gene, it will travel to there and keep that gene off in that cell. And this is how you prevent hemoglobin from being expressed in neurons, and how you would prevent liver enzymes from being expressed in red blood cells. Repressors are very difficult to study because it's much easier to study things that give you more of what you're looking for. But I think as we go along we're

going to find they play as important a role in gene regulation as the activating proteins."

David M. Bodine, Ph.D.

16.4 Talking Glossary: Restriction Enzyme (0.5 min)

National Human Genome Research Institute

https://www.genome.gov/genetics-glossary/Restriction-Enzyme

Intro: "A restriction enzyme is an enzyme isolated from bacteria that cuts DNA molecules at specific sequences. The isolation of these enzymes was critical to the development of recombinant DNA (rDNA) technology and genetic engineering."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/restriction_enzymes.mp3

Text "Restriction enzymes are proteins that bind to DNA in a very specific manner. So they actually recognize the base pairs within the DNA. And typically they will bind to a palindromic sequence, for instance, a sequence that is a mirror copy of itself–AGCCGA."

Stacie Loftus, Ph.D.

Chapter 17

S

17.1 Structural Classification of Proteins database (SCOP)

Adapted from Wikipedia https://en.wikipedia.org/wiki/Structural_Classification_of_Proteins_database

The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein **structural domains** based on similarities of their structures and amino acid sequences. A motivation for this classification is to determine the **evolutionary relationship** between proteins. Proteins with the same shapes but having little sequence or functional similarity are placed in different superfamilies, and are assumed to have only a very distant common ancestor. Proteins having the same shape and some similarity of sequence and/or function are placed in "families", and are assumed to have a closer common ancestor.

Similar to **CATH** and **Pfam** databases, SCOP provides a classification of individual structural domains of proteins, rather than a classification of the entire proteins which may include a significant number of different domains.

The SCOP database is freely accessible on the internet. SCOP was created in 1994 in the Centre for Protein Engineering and the Laboratory of Molecular Biology (3). It was maintained by Alexey G. Murzin and his colleagues in the Centre for Protein Engineering until its closure in 2010 and subsequently at the Laboratory of Molecular Biology in Cambridge, England (4, 5, 6, 1).

17.1.1 Hierarchical organisation

The source of protein structures is the Protein Data Bank. The unit of classification of structure in SCOP is the protein domain. What the SCOP authors

mean by "domain" is suggested by their statement that small proteins and most medium-sized ones have just one domain (8), and by the observation that human hemoglobin (9), which has an 2 2 structure, is assigned two SCOP domains, one for the and one for the subunit.

The shapes of domains are called "**folds**" in SCOP. Domains belonging to the same fold have the same major secondary structures in the same arrangement with the same topological connections. 1195 folds are given in SCOP version 1.75.

17.1.2 Superfamilies

Domains within a fold are classified into **superfamilies**. This is a largest grouping of proteins for which structural similarity is sufficient to indicate evolutionary relatedness and therefore share a common ancestor. However, this ancestor is presumed to be distant, because the different members of a superfamily have low sequence identities. For example, the two superfamilies of the "Globin-like" fold are: the Globin superfamily and alpha-helical ferredoxin superfamily.

17.1.3 Families

Protein families are more closely related than superfamilies. Domains are placed in the same family if that have either:

- less than 30% sequence identity
- some sequence identity (e.g., 15%) AND perform the same function

The similarity in sequence and structure is evidence that these proteins have a closer evolutionary relationship than do proteins in the same superfamily. Sequence tools, such as BLAST, are used to assist in placing domains into superfamilies and families.

17.1.4 PDB entry domains

OPTIONAL

A "TaxId" is the taxonomy ID number and links to the NCBI taxonomy browser, which provides more information about the species to which the protein belongs. Clicking on a species or isoform brings up a list of domains. For example, the "Hemoglobin, alpha-chain from Human (Homo sapiens)" protein has >190 solved protein structures, such as 2dn3 (complexed with cmo), and 2dn1 (complexed with hem, mbn, oxy). Clicking on the PDB numbers is supposed to display the structure of the molecule, but the links are currently broken (links work in pre-SCOP).

17.2 Sequence alignment

Adapated from Wikipedia https://en.wikipedia.org/wiki/Sequence_alignment

In bioinformatics, a sequence alignment is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences (1). Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a grid or matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.

17.2.1 Interpretation

If two sequences in an alignment share a common ancestor, mismatches can be interpreted as point mutations and gaps as indels (that is, insertion or deletion mutations) introduced in one or both lineages in the time since they diverged from one another. In sequence alignments of proteins, the degree of similarity between amino acids occupying a particular position in the sequence can be interpreted as a rough measure of how conserved a particular region or sequence motif is among lineages. The absence of substitutions, or the presence of only very conservative substitutions (that is, the substitution of amino acids whose side chains have similar biochemical properties) in a particular region of the sequence, suggest (3) that this region has structural or functional importance. Although DNA and RNA nucleotide bases are more similar to each other than are amino acids, the conservation of base pairs can indicate a similar functional or structural role.

17.2.2 Alignment methods

Very short or very similar sequences can be aligned by hand. However, most interesting problems require the alignment of lengthy, highly variable or extremely numerous sequences that cannot be aligned solely by human effort. Instead, human knowledge is applied in constructing algorithms to produce high-quality sequence alignments, and occasionally in adjusting the final results to reflect patterns that are difficult to represent algorithmically (especially in the case of nucleotide sequences).

Computational approaches to sequence alignment generally fall into two categories: **global alignments** and **local alignments**. Calculating a global alignment is a form of global optimization that "forces" the alignment to span the *entire* length of all sequences. By contrast, local alignments identify regions of similarity within long sequences that are often widely divergent overall. Local alignments are often preferable, but can be more difficult to calculate because of the additional challenge of identifying isolated regions of high similarity (4). BLAST searches are an example of local alignment.

17.2.3 Representations

Alignments are commonly represented both graphically and in text format. In almost all sequence alignment representations, sequences are written in rows arranged so that aligned residues appear in successive columns.

```
library(ggmsa)
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```

```
fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
ggmsa(fasta, 164, 213, color="Chemistry_AA", none_bg = T,
      consensus_views = T,
      use_dot = T)
```

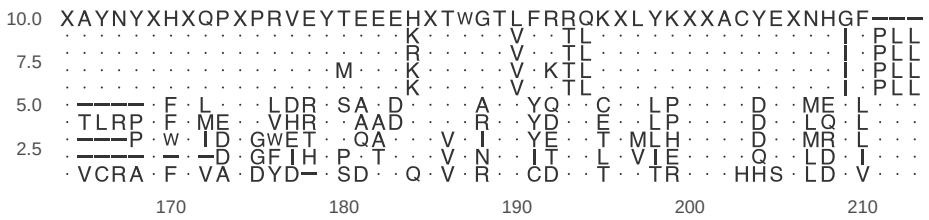


Figure 1: Dots * for concensus

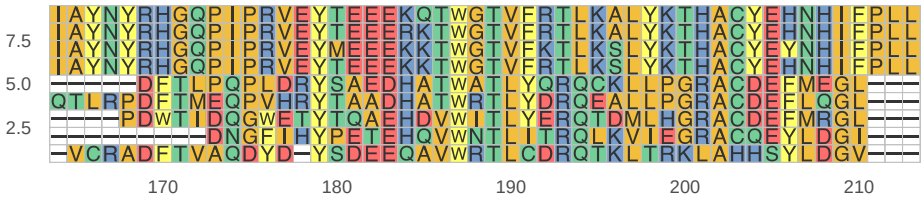
```
library(ggmsa)
fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
ggmsa(fasta, 164, 213, color="Chemistry_AA", none_bg = T,
      consensus_views = T,
      use_dot = F)
```

Figure 2: Pipes (|) for consensus

In text formats, aligned columns containing identical or similar characters are indicated with a system of conservation symbols. In the image above, an asterisk (*) is used to show identity between two columns. Some programs use the pipe symbol (|). Other symbols that may be used include a colon for **conservative substitutions** and a period for **semiconservative substitutions**.

```
library(ggmsa)
# fasta <- system.file("extdata", "sample.fasta", package = "ggmsa")
ggmsa(fasta, 164, 213, color="Chemistry_AA"
      #none_bg = T
```

```
)
```



Many sequence visualization programs also use color to display information about the properties of the individual sequence elements; in DNA and RNA sequences, this equates to assigning each nucleotide its own color. In protein alignments, such as the one in the image below, color is often used to indicate amino acid properties to aid in judging the conservation of a given amino acid substitution. For multiple sequences the last row in each column is often the consensus sequence determined by the alignment; the consensus sequence is also often represented in graphical format with a **sequence logo** in which the size of each nucleotide or amino acid letter corresponds to its degree of conservation (5).

```
library(ggseqlogo)
data(ggseqlogo_sample)
ggseqlogo( seqs_dna[[1]] )
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```

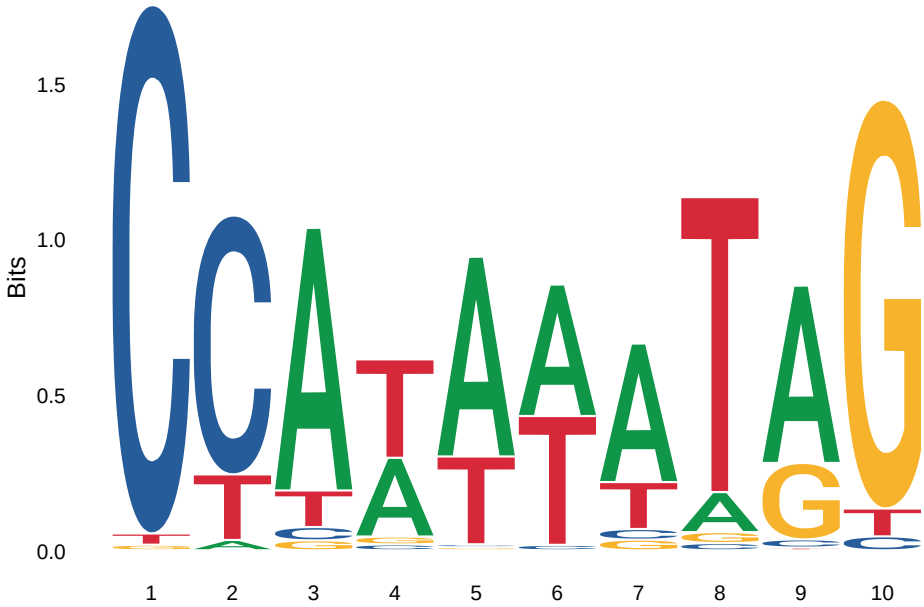


Figure 3: Sequence log

This section is optional:

Sequence alignments can be stored in a wide variety of text-based file formats, many of which were originally developed in conjunction with a specific alignment program or implementation. Most web-based tools allow a limited number of input and output formats, such as FASTA format and GenBank format and the output is not easily editable. Several conversion programs that provide graphical and/or command line interfaces are available, such as READSEQ and EMBOSS. There are also several programming packages which provide this conversion functionality, such as BioPython, BioRuby and BioPerl. The SAM/BAM files use the CIGAR (Compact Idiosyncratic Gapped Alignment Report) string format to represent an alignment of a sequence to a reference by encoding a sequence of events (e.g. match/mismatch, insertions, deletions).[6]

17.2.4 Global and local alignments

Global alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. (This does not mean global alignments cannot start and/or end in gaps.) A general global alignment technique is the **Needleman–Wunsch algorithm**, which is based on an approach known as dynamic programming. Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The **Smith–Waterman algorithm** is a general local alignment method based on the same dynamic programming scheme but with additional choices to start and end at any place (4).

17.2.5 Pairwise alignment

Pairwise sequence alignment methods are used to find the best-matching piecewise (local or global) alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high similarity to a query).

The three primary methods of producing pairwise alignments are (1)

1. graphic dot-matrix methods
2. dynamic programming,
3. and word methods

Multiple sequence alignment techniques can also begin by aligning pairs of sequences. Although each method has its individual strengths and weaknesses, all three pairwise methods have difficulty with highly repetitive sequences of low information content - especially where the number of repetitions differ in the two sequences to be aligned.

17.2.6 Dot-matrix methods

The dot-matrix approach, which implicitly produces a family of alignments for individual sequence regions, is qualitative and conceptually simple, though time-consuming to analyze on a large scale. In the absence of noise, it can be easy to visually identify certain sequence features—such as insertions, deletions, repeats, or inverted repeats—from a dot-matrix plot. To construct a dot-matrix plot, the two sequences are written along the top row and leftmost column of a two-dimensional matrix and a dot is placed at any point where the characters in the appropriate columns match—this is a typical recurrence plot. Some implementations vary the size or intensity of the dot depending on the degree of similarity of the two characters, to accommodate conservative substitutions. The dot plots of very closely related sequences will appear as a single line along the matrix's main diagonal.

Problems with dot plots as an information display technique include: noise, lack of clarity, non-intuitiveness, difficulty extracting match summary statistics and match positions on the two sequences. There is also much wasted space where the match data is inherently duplicated across the diagonal and most of the actual area of the plot is taken up by either empty space or noise, and, finally, dot-plots are limited to two sequences. None of these limitations apply to Miropeats alignment diagrams but they have their own particular flaws.

Dot plots can also be used to assess repetitiveness in a single sequence. A sequence can be plotted against itself and regions that share significant similarities will appear as lines off the main diagonal. This effect can occur when a protein consists of multiple similar structural domains.

17.2.7 Dynamic programming

The technique of "dynamic programming" can be applied to produce global alignments via the **Needleman-Wunsch algorithm**, and local alignments via the **Smith-Waterman algorithm**. In typical usage, protein alignments use a **substitution matrix** to assign scores to amino-acid **matches** or **mismatches**, and a **gap penalty** for matching an amino acid in one sequence to a gap in the other.

Substitution matrices are symmetric matrices that assign high scores to amino acids that are more chemically similar or which substitutions are most common. Low values are assigned to chemically dissimilar amino acids / unlikely substitutions.

```
##    A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V  B  J
## A  5 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## R -2  7 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## N -1  0  6 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## D -2 -1  2  7 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## C -1 -3 -2 -3 12 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
```

```
## Q -1  1  0  0 -3  6 NA NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## E -1  0  0  2 -3  2  6 NA NA NA NA NA NA NA NA NA NA NA NA NA NA
## G  0 -2  0 -1 -3 -2 -2  7 NA NA NA NA NA NA NA NA NA NA NA NA NA
## H -2  0  1  0 -3  1  0 -2 10 NA NA NA NA NA NA NA NA NA NA NA NA
## I -1 -3 -2 -4 -3 -2 -3 -4 -3  5 NA NA NA NA NA NA NA NA NA NA NA
## L -1 -2 -3 -3 -2 -2 -2 -3 -2  2  5 NA NA NA NA NA NA NA NA NA NA
## K -1  3  0  0 -3  1  1 -2 -1 -3 -3  5 NA NA NA NA NA NA NA NA NA
## M -1 -1 -2 -3 -2  0 -2 -2  0  2  2 -1  6 NA NA NA NA NA NA NA NA
## F -2 -2 -2 -4 -2 -4 -3 -3 -2  0  1 -3  0  8 NA NA NA NA NA NA NA
## P -1 -2 -2 -1 -4 -1  0 -2 -2 -2 -3 -1 -2 -3  9 NA NA NA NA NA NA
## S  1 -1  1  0 -1  0  0  0 -1 -2 -3 -1 -2 -2 -1  4 NA NA NA NA NA
## T  0 -1  0 -1 -1 -1 -1 -2 -2 -1 -1 -1 -1 -1 -1  2  5 NA NA NA NA
## W -2 -2 -4 -4 -5 -2 -3 -2 -3 -2 -2 -2 -2  1 -3 -4 -3 15 NA NA NA
## Y -2 -1 -2 -2 -3 -1 -2 -3  2  0  0 -1  0  3 -3 -2 -1  3  8 NA NA
## V  0 -2 -3 -3 -1 -3 -3 -3 -3  3  1 -2  1  0 -3 -1  0 -3 -1  5 NA
## B -1 -1  5  6 -2  0  1 -1  0 -3 -3  0 -2 -3 -2  0  0 -4 -2 -3  5 NA
## J -1 -3 -3 -3 -2 -2 -3 -4 -2  4  4 -3  2  1 -3 -2 -1 -2  0  2 -3  4
```

Figure 1: The BLOSUM45 substitution matrix

```
nucleotideSubstitutionMatrix(match = 1, mismatch = -1, baseOnly = T, type = "DNA")
```

```
##    A  C  G  T
## A  1 -1 -1 -1
## C -1  1 -1 -1
## G -1 -1  1 -1
## T -1 -1 -1  1
```

DNA and RNA alignments may use a **scoring matrix**, but in practice often simply assign a positive match score when both sequences have the same base, a negative mismatch score when they are different, and a negative gap penalty.

A common extension to standard linear gap costs, is the usage of two different gap penalties for **opening** a gap and for **extending** a gap. Typically the former is much larger than the latter, e.g. -10 for gap open and -2 for gap extension. Thus, the number of gaps in an alignment is usually reduced and residues and gaps are kept together, which typically makes more biological sense.

17.2.8 Word methods

OPTIONAL

Word methods are approximate methods that are not guaranteed to find an optimal alignment solution, but are significantly more efficient than dynamic programming. These methods are especially useful in large-scale **database searches** where it is understood that a large proportion of the candidate sequences will have essentially no significant match with the query sequence.

Word methods are best known for their implementation in the database search

tools FASTA and the **BLAST** family (1). Word methods identify a series of short, nonoverlapping **subsequences** ("words") in the query sequence that are then matched to candidate database sequences. The relative positions of the word in the two sequences being compared are subtracted to obtain an offset; this will indicate a region of alignment if multiple distinct words produce the same offset. Only if this region is detected do these methods apply more sensitive alignment criteria; thus, many unnecessary comparisons with sequences of no appreciable similarity are eliminated.

BLAST uses a word search of length k and evaluates only the most significant word matches. Most BLAST implementations use a fixed default word length that is optimized for the query and database type, and that is changed only under special circumstances, such as when searching with repetitive or very short query sequences. Implementations can be found via a number of web portals, such as EMBL FASTA and NCBI BLAST.

17.2.9 Multiple sequence alignment

Multiple sequence alignment (MSA) is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set. Multiple alignments are often used in identifying **conserved** sequence regions across a group of sequences hypothesized to be evolutionarily related. Such conserved sequence **motifs** can be used in conjunction with structural and mechanistic information to locate the catalytic active sites of enzymes. MSA therefore occur regularly in molecular biology and biochemistry publications.

Alignments are also used to aid in establishing evolutionary relationships by constructing phylogenetic trees. Multiple sequence alignments are computationally difficult to produce (most formulations of the problem lead to NP-complete combinatorial optimization problems; 10, 11). Nevertheless, the utility of these alignments in bioinformatics has led to the development of a variety of methods suitable for aligning three or more sequences.

17.2.10 Phylogenetic analysis

Phylogenetics and sequence alignment are closely related fields due to the shared necessity of evaluating sequence relatedness (25). The field of phylogenetics makes extensive use of sequence alignments in the construction and interpretation of phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The degree to which sequences in a query set differ is qualitatively related to the sequences' **evolutionary distance** from one another. Roughly speaking, high **sequence identity** suggests that the sequences in question have a comparatively young **most recent common ancestor**, while low identity suggests that the divergence is more ancient.

OPTIONAL

This approximation, which reflects the "molecular clock" hypothesis that a roughly constant rate of evolutionary change can be used to extrapolate the elapsed time since two genes first diverged (that is, the coalescence time), assumes that the effects of mutation and selection are constant across sequence lineages. Therefore, it does not account for possible difference species in the possible functional conservation of specific regions in a sequence. (In the case of nucleotide sequences, the molecular clock hypothesis in its most basic form also discounts the difference in acceptance rates between silent mutations that do not alter the meaning of a given codon and other mutations that result in a different amino acid being incorporated into the protein). More statistically accurate methods allow the evolutionary rate on each branch of the phylogenetic tree to vary, thus producing better estimates of coalescence times for genes.

Progressive multiple alignment techniques produce a phylogenetic tree by necessity because they incorporate sequences into the growing alignment in order of relatedness. Other techniques that assemble multiple sequence alignments and phylogenetic trees score and sort trees first and calculate a multiple sequence alignment from the highest-scoring tree. Commonly used methods of phylogenetic tree construction are mainly heuristic because the problem of selecting the optimal tree, like the problem of selecting the optimal multiple sequence alignment, is NP-hard (26).

17.2.11 Assessment of significance

Sequence alignments are useful in bioinformatics for identifying sequence similarity, producing phylogenetic trees, and developing homology models of protein structures. However, the biological relevance of sequence alignments is not always clear. Alignments are often assumed to reflect a degree of evolutionary change between sequences descended from a common ancestor; however, it is formally possible that **convergent evolution** can occur to produce apparent similarity between proteins that are evolutionarily unrelated but perform similar functions and have similar structures.

In database searches such as BLAST, statistical methods can determine the likelihood of a particular alignment between sequences or sequence regions arising by chance given the size and composition of the database being searched. These values can vary significantly depending on the **search space**. In particular, the likelihood of finding a given alignment by chance increases if the database consists only of sequences from the same organism as the query sequence. Repetitive sequences in the database or query can also distort both the search results and the assessment of statistical significance; BLAST automatically filters such repetitive sequences in the query to avoid apparent hits that are statistical artifacts.

17.2.11.1 Assessment of credibility

Statistical significance indicates the probability that an alignment of a given quality could arise by chance, but does not indicate how much superior a given alignment is to alternative alignments of the same sequences. Measures of alignment credibility indicate the extent to which the best scoring alignments for a given pair of sequences are substantially similar.

17.2.11.2 Scoring functions

The choice of a scoring function that reflects biological or statistical observations about known sequences is important to producing good alignments. Protein sequences are frequently aligned using substitution matrices that reflect the probabilities of given character-to-character substitutions. A series of matrices called **PAM matrices** (Point Accepted Mutation matrices, originally defined by Margaret Dayhoff and sometimes referred to as "Dayhoff matrices") explicitly encode evolutionary approximations regarding the rates and probabilities of particular amino acid mutations. Another common series of scoring matrices, known as **BLOSUM** (Blocks Substitution Matrix), encodes empirically derived substitution probabilities. Variants of both types of matrices are used to detect sequences with differing levels of divergence, thus allowing users of BLAST or FASTA to restrict searches to more closely related matches or expand to detect more divergent sequences. Gap penalties account for the introduction of a gap - on the evolutionary model, an insertion or deletion mutation - in both nucleotide and protein sequences, and therefore the penalty values should be proportional to the expected rate of such mutations. The quality of the alignments produced therefore depends on the quality of the scoring function.

It can be very useful and instructive to try the same alignment several times with different choices for scoring matrix and/or gap penalty values and compare the results. Regions where the solution is weak or non-unique can often be identified by observing which regions of the alignment are robust to variations in alignment parameters.

17.2.12 Other biological uses

Sequenced RNA, such as expressed sequence tags and full-length mRNAs, can be aligned to a sequenced genome to find where there are genes and get information about alternative splicing (35) and RNA editing (36). Sequence alignment is also a part of genome assembly, where sequences are aligned to find overlap so that contigs (long stretches of sequence) can be formed (37). Another use is SNP analysis, where sequences from different individuals are aligned to find single basepairs that are often different in a population (38).

17.2.13 References

1. Mount DM. (2004). Bioinformatics: Sequence and Genome Analysis (2nd ed.). Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 978-0-87969-608-5.
2. "Clustal FAQ Symbols". Clustal. Archived from the original on 24 October 2016. Retrieved 8 December 2014.
3. Ng PC; Henikoff S (May 2001). "Predicting deleterious amino acid substitutions". Genome Res. 11 (5): 863–74. doi:10.1101/gr.176601. PMC 311071. PMID 11337480.
4. Polyanovsky, V. O.; Roytberg, M. A.; Tumanyan, V. G. (2011). "Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences". Algorithms for Molecular Biology. 6 (1): 25. doi:10.1186/1748-7188-6-25. PMC 3223492. PMID 22032267. S2CID 2658261.
5. Schneider TD; Stephens RM (1990). "Sequence logos: a new way to display consensus sequences". Nucleic Acids Res. 18 (20): 6097–6100. doi:10.1093/nar/18.20.6097. PMC 332411. PMID 2172928.
6. "Sequence Alignment/Map Format Specification" (PDF).
7. Brudno M; Malde S; Poliakov A; Do CB; Couronne O; Dubchak I; Batzoglou S (2003). "Glocal alignment: finding rearrangements during alignment". Bioinformatics. 19. Suppl 1 (90001): i54–62. doi:10.1093/bioinformatics/btg1005. PMID 12855437.
8. Delcher, A. L.; Kasif, S.; Fleishmann, R.D.; Peterson, J.; White, O.; Salzberg, S.L. (1999). "Alignment of whole genomes". Nucleic Acids Research. 27 (11): 2369–2376. doi:10.1093/nar/30.11.2478. PMC 148804. PMID 10325427.
9. Wing-Kin, Sung (2010). Algorithms in Bioinformatics: A Practical Introduction (First ed.). Boca Raton: Chapman & Hall/CRC Press. ISBN 978-1420070330.
10. Wang L; Jiang T. (1994). "On the complexity of multiple sequence alignment". J Comput Biol. 1 (4): 337–48. CiteSeerX 10.1.1.408.894. doi:10.1089/cmb.1994.1.337. PMID 8790475.
11. Elias, Isaac (2006). "Settling the intractability of multiple alignment". J Comput Biol. 13 (7): 1323–1339. CiteSeerX 10.1.1.6.256. doi:10.1089/cmb.2006.13.1323. PMID 17037961.
12. Lipman DJ; Altschul SF; Kececioglu JD (1989). "A tool for multiple sequence alignment". Proc Natl Acad Sci USA. 86 (12): 4412–5. Bibcode:1989PNAS...86.4412L. doi:10.1073/pnas.86.12.4412. PMC 287279. PMID 2734293.
13. Higgins DG, Sharp PM (1988). "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer". Gene. 73 (1): 237–44. doi:10.1016/0378-1119(88)90330-7. PMID 3243435.
14. Thompson JD; Higgins DG; Gibson TJ. (1994). "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix

choice". Nucleic Acids Res. 22 (22): 4673–80. doi:10.1093/nar/22.22.4673. PMC 308517. PMID 7984417.
15. Chenna R; Sugawara H; Koike T; Lopez R; Gibson TJ; Higgins DG; Thompson JD. (2003). "Multiple sequence alignment with the Clustal series of programs". Nucleic Acids Res. 31 (13): 3497–500. doi:10.1093/nar/gkg500. PMC 168907. PMID 12824352.
16. Notredame C; Higgins DG; Heringa J. (2000). "T-Coffee: A novel method for fast and accurate multiple sequence alignment". J Mol Biol. 302 (1): 205–17. doi:10.1006/jmbi.2000.4042. PMID 10964570. S2CID 10189971.
17. Hirosawa M; Totoki Y; Hoshida M; Ishikawa M. (1995). "Comprehensive study on iterative algorithms of multiple sequence alignment". Comput Appl Biosci. 11 (1): 13–8. doi:10.1093/bioinformatics/11.1.13. PMID 7796270.
18. Karplus K; Barrett C; Hughey R. (1998). "Hidden Markov models for detecting remote protein homologies". Bioinformatics. 14 (10): 846–856. doi:10.1093/bioinformatics/14.10.846. PMID 9927713.
19. Chothia C; Lesk AM. (April 1986). "The relation between the divergence of sequence and structure in proteins". EMBO J. 5 (4): 823–6. doi:10.1002/j.1460-2075.1986.tb04288.x. PMC 1166865. PMID 3709526.
20. Zhang Y; Skolnick J. (2005). "The protein structure prediction problem could be solved using the current PDB library". Proc Natl Acad Sci USA. 102 (4): 1029–34. Bibcode:2005PNAS..102.1029Z. doi:10.1073/pnas.0407152101. PMC 545829. PMID 15653774.
21. Holm L; Sander C (1996). "Mapping the protein universe". Science. 273 (5275): 595–603. Bibcode:1996Sci…273..595H. doi:10.1126/science.273.5275.595. PMID 8662544. S2CID 7509134.
22. Taylor WR; Flores TP; Orengo CA. (1994). "Multiple protein structure alignment". Protein Sci. 3 (10): 1858–70. doi:10.1002/pro.5560031025. PMC 2142613. PMID 7849601.
23. Orengo CA; Michie AD; Jones S; Jones DT; Swindells MB; Thornton JM (1997). "CATH–a hierarchic classification of protein domain structures". Structure. 5 (8): 1093–108. doi:10.1016/S0969-2126(97)00260-8. PMID 9309224.
24. Shindyalov IN; Bourne PE. (1998). "Protein structure alignment by incremental combinatorial extension (CE) of the optimal path". Protein Eng. 11 (9): 739–47. doi:10.1093/protein/11.9.739. PMID 9796821.
25. Ortet P; Bastien O (2010). "Where Does the Alignment Score Distribution Shape Come from?". Evolutionary Bioinformatics. 6: 159–187. doi:10.4137/EBO.S5875. PMC 3023300. PMID 21258650.
26. Felsenstein J. (2004). Inferring Phylogenies. Sinauer Associates: Sunderland, MA. ISBN 978-0-87893-177-4.
27. Altschul SF; Gish W (1996). Local Alignment Statistics. Meth.Enz. Methods in Enzymology. 266. pp. 460–480. doi:10.1016/S0076-6879(96)66029-7. ISBN 9780121821678. PMID 8743700.
28. Hartmann AK (2002). "Sampling rare events: statistics of local sequence alignments". Phys. Rev. E. 65 (5): 056102. arXiv:cond-mat/0108201.

Bibcode:2002PhRvE..65e6102H. doi:10.1103/PhysRevE.65.056102. PMID 12059642. S2CID 193085.

29. Newberg LA (2008). "Significance of gapped sequence alignments". J Comput Biol. 15 (9): 1187–1194. doi:10.1089/cmb.2008.0125. PMC 2737730. PMID 18973434.
30. Eddy SR; Rost, Burkhard (2008). Rost, Burkhard (ed.). "A probabilistic model of local sequence alignment that simplifies statistical significance estimation". PLOS Comput Biol. 4 (5): e1000069. Bibcode:2008PLSCB...4E0069E. doi:10.1371/journal.pcbi.1000069. PMC 2396288. PMID 18516236. S2CID 15640896.
31. Bastien O; Aude JC; Roy S; Marechal E (2004). "Fundamentals of massive automatic pairwise alignments of protein sequences: theoretical significance of Z-value statistics". Bioinformatics. 20 (4): 534–537. doi:10.1093/bioinformatics/btg440. PMID 14990449.
32. Agrawal A; Huang X (2011). "Pairwise Statistical Significance of Local Sequence Alignment Using Sequence-Specific and Position-Specific Substitution Matrices". IEEE/ACM Transactions on Computational Biology and Bioinformatics. 8 (1): 194–205. doi:10.1109/TCBB.2009.69. PMID 21071807. S2CID 6559731.
33. Agrawal A; Brendel VP; Huang X (2008). "Pairwise statistical significance and empirical determination of effective gap opening penalties for protein local sequence alignment". International Journal of Computational Biology and Drug Design. 1 (4): 347–367. doi:10.1504/IJCBDD.2008.022207. PMID 20063463. Archived from the original on 28 January 2013.
34. Newberg LA; Lawrence CE (2009). "Exact Calculation of Distributions on Integers, with Application to Sequence Alignment". J Comput Biol. 16 (1): 1–18. doi:10.1089/cmb.2008.0137. PMC 2858568. PMID 19119992.
35. Kim N; Lee C (2008). Bioinformatics detection of alternative splicing. Methods in Molecular Biology. 452. pp. 179–97. doi:10.1007/978-1-60327-159-2_9. ISBN 978-1-58829-707-5. PMID 18566765.
36. Li JB, Levanon EY, Yoon JK, et al. (May 2009). "Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing". Science. 324 (5931): 1210–3. Bibcode:2009Sci...324.1210L. doi:10.1126/science.1170995. PMID 19478186. S2CID 31148824.
37. Blazewicz J, Bryja M, Figlerowicz M, et al. (June 2009). "Whole genome assembly from 454 sequencing output via modified DNA graph concept". Comput Biol Chem. 33 (3): 224–30. doi:10.1016/j.compbiolchem.2009.04.005. PMID 19477687.
38. Duran C; Appleby N; Vardy M; Imelfort M; Edwards D; Batley J (May 2009). "Single nucleotide polymorphism discovery in barley using autoSNPdb". Plant Biotechnol. J. 7 (4): 326–33. doi:10.1111/j.1467-7652.2009.00407.x. PMID 19386041.
39. Abbott A.; Tsay A. (2000). "Sequence Analysis and Optimal Matching Methods in Sociology, Review and Prospect". Sociological Methods and Research. 29 (1): 3–33. doi:10.1177/0049124100029001001. S2CID 121097811.

40. Barzilay R; Lee L. (2002). "Bootstrapping Lexical Choice via Multiple-Sequence Alignment" (PDF). Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). 10: 164–171. arXiv:cs/0205065. Bibcode:2002cs……5065B. doi:10.3115/1118693.1118715. S2CID 7521453.
41. Kondrak, Grzegorz (2002). "Algorithms for Language Reconstruction" (PDF). University of Toronto, Ontario. Archived from the original (PDF) on 17 December 2008. Retrieved 21 January 2007.
42. Prinzie A.; D. Van den Poel (2006). "Incorporating sequential information into traditional classification models by using an element/position-sensitive SAM". Decision Support Systems. 42 (2): 508–526. doi:10.1016/j.dss.2005.02.004. See also Prinzie and Van den Poel's paper Prinzie, A; Vandenpoel, D (2007). "Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modeling sequential information in NPTB models". Decision Support Systems. 44 (1): 28–45. doi:10.1016/j.dss.2007.02.008.
43. EMBL-EBI. "ClustalW2 < Multiple Sequence Alignment < EMBL-EBI". www.EBI.ac.uk. Retrieved 12 June 2017.
44. T-coffee
45. "BLAST: Basic Local Alignment Search Tool". blast.ncbi.nlm.NIH.gov. Retrieved 12 June 2017.
46. "UVA FASTA Server". fasta.bioch.Virginia.edu. Retrieved 12 June 2017.
47. Thompson JD; Plewniak F; Poch O (1999). "BAliBASE: a benchmark alignment database for the evaluation of multiple alignment programs". Bioinformatics. 15 (1): 87–8. doi:10.1093/bioinformatics/15.1.87. PMID 10068696.
48. BAliBASE
49. Thompson JD; Plewniak F; Poch O. (1999). "A comprehensive comparison of multiple sequence alignment programs". Nucleic Acids Res. 27 (13): 2682–90. doi:10.1093/nar/27.13.2682. PMC 148477. PMID 10373585.
50. "Multiple sequence alignment: Strap". 3d-alignment.eu. Retrieved 12 June 2017.

17.3 Talking Glossary U03: Shotgun sequencing (1 min)

Now just called "sequencing"

Introduction: "Shotgun sequencing is a laboratory technique for determining the DNA sequence of an organism's genome. The method involves breaking the genome into a collection of small DNA fragments that are sequenced individually. A computer program looks for overlaps in the DNA sequences and uses them to place the individual fragments in their correct order to reconstitute the genome."

Transcript: "The most efficient way to sequence a large piece of DNA involves

a process known as shotgun sequencing. For this, the starting DNA is broken up randomly into many smaller pieces, sort of in a shotgun fashion, with each of those pieces then sequenced individually. The resulting sequence reads generated from the different pieces are then analyzed by a computer program, looking for stretches of sequence from different reads that are identical with one another. When identical regions are identified, they are overlapped with one another, allowing the two sequence reads to be stitched together. This computer process is repeated over and over and over again, eventually yielding the complete sequence of the starting piece of DNA. The initial random fragmenting and reading of the DNA gave this approach the name"shotgun sequencing". "

Eric D. Green, M.D., Ph.D.

17.4 Talking Glossary: Sickle Cell Disease (4 min)

https://www.genome.gov/genetics-glossary/Sickle-Cell-Disease

Abstract: "Sickle cell disease is a hereditary disease seen most often among people of African ancestry. Caused by mutations in one of the genes that encode the hemoglobin protein, the disease is inherited as an autosomal recessive trait. The mutation causes the red blood cells to take on an unusual sickle shape. Individuals affected by sickle cell disease are chronically anemic and experience significant damage to their heart, lungs, and kidneys."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/sickle_cell_disease.jpg

Transcript: "Sickle cell disease is the first human inherited disease that was understood at the protein and the DNA level. Sickle cell is a disease that's primarily seen in people of African descent. In studying the genomics of people from Africa, it's very clear now that three different times during the history of the human race the mutation in the beta-globin gene that changes an amino acid at the sixth position of the protein, the same mutation has happened. And this has been expanded throughout history in Africa in three different groups and then migrated all over the world. Now, the mutation does not really affect the ability of beta-globin to participate with alpha-globin and make hemoglobin, and it doesn't affect the ability to carry oxygen. In matter of fact, people who inherit a hemoglobin S gene from their mother and a hemoglobin S gene from their father, so their homozygous for hemoglobin S, have perfectly normal oxygen-carrying capacity of hemoglobin. The problems happen when the red cell containing hemoglobin S gets into the muscles or into your brain and discharges the oxygen. And what happens is that hemoglobin S has a tendency when the oxygen goes off to stick to each other, and they form these polymers. Now, when they get oxygen back on, they go back into solution. The difference between hemoglobin S and wild-type hemoglobin is that the wild-type

hemoglobin is soluble when it's got oxygen, and it's also soluble when it doesn't have oxygen. Now, if everything is working well, these polymers in people with sickle cell disease don't get very big. But that can be worsened when you have extreme oxygen stress, and that's what causes the characteristic sickle cells. The polymers get very, very long, and they stretch the cell out of shape. Now your spleen is very good at raking these cells out of the peripheral blood as they come through. And that's what causes the anemia, and the old name for sickle cell was sickle cell anemia. However, though, short polymers are very, very dangerous. If you think of a red blood cell going through your veins and arteries and capillaries as a water balloon, you'll see how it can squish itself down into a long cylinder to get through a capillary and flatten out like a pancake to get through areas in the spleen that kind of rake out the bad-shape red blood cells, and then to form right back to a regular water balloon shape to go into the arteries and veins. But if you have a red blood cell full of these small hemoglobin S polymers, it's like having a water balloon full of ice chips, and as that goes through the capillaries and the small veins, it tears up the lining of these things, just as if you had a cut, and it activates your clotting response and micro-clots will form. And sometimes these get to be bigger and bigger clots, and so the real lethality in sickle cell disease is not from the anemia, it's from this vascular disease. And so it's very unfortunate that about one-third of people who are homozygous for hemoglobin S will have one or more strokes before they're 10 years old. And in those fortunate few that are able to get to be adolescents, these clots can build up in the lungs and give a very severe disease called acute chest syndrome, which is basically emphysema or lung destruction from these obstructions. About one-third of the patients will live healthily to adulthood, but they have many problems with iron-overloaded organs, and their life span is significantly shorter. And it is simply amazing to me that this all comes from one very small change, changes in amino acid, that really isn't having any effect on the normal function of the protein, but has big effects on all of the other systems in the body that the red cells pass through."

David M. Bodine, Ph.D.

17.5 Medline: SNPs (single nucleotide polymorphism)

Adapted from: Medline: "What are single nucleotide polymorphisms (SNPs)?" https://medlineplus.gov/genetics/understanding/genomicresearch/snp/

Single nucleotide polymorphisms, frequently called SNPs (pronounced "snips"), are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA.

SNPs occur normally throughout a person's DNA. They occur almost once in

every 1,000 nucleotides on average, which means there are roughly 4 to 5 million SNPs in a person's genome. These variations may be unique or occur in many individuals; scientists have found more than 100 million SNPs in populations around the world. Most commonly, these variations are found in the DNA between genes. They can act as biological markers, helping scientists locate genes that are associated with disease. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function.

Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families. Future studies will work to identify SNPs associated with complex diseases such as heart disease, diabetes, and cancer.

17.6 Talking glossary: Somatic cells (0.75 min)

https://www.genome.gov/genetics-glossary/Somatic-Cells

Abstract: "A somatic cell is any cell of the body except sperm and egg cells. Somatic cells are diploid, meaning that they contain two sets of chromosomes, one inherited from each parent. Mutations in somatic cells can affect the individual, but they are not passed on to offspring."

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/somatic_cells.jpg

Transcript: "Somatic cells is a fairly general term which refers to essentially all the cells of the body except for the germ line; the germ line being the cells in the sexual organs that produce sperm and eggs. So anything that doesn't have the job of producing sperm or eggs is a somatic cell. It is very important, of course, for every living organism to be alive, but it contributes nothing in terms of inheritance through genetics, inheritance to the next generation. So it is only of use to the living organism and has no relation to anything that happens in the next generation of that organism."

Shawn Burgess, Ph.D.

17.7 Sequence annotation

Adapted from Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/DNA_annotation

Note: Wikipedia titles this "DNA annotation." In my experience this phrase is not used; rather we refer to sequence annotation, genome annotation, or just annotation.

Genome annotation is the process of predicting the locations of genes and all of the coding regions in a genome and predicitng what those genes do. An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.[1] For DNA annotation, a previously unknown sequence representation of genetic material is enriched with information of intron -exon boundaries, regulatory sequences , gene names and protein products. This annotation is stored in genomic databases.

17.7.1 Process

Genome annotation consists of two main steps:

1. predicting coding elements on the genome , a process called gene prediction
2. attaching biological information to these elements

Structural genomic annotation consists of the identification of genomic elements.

1. ORFs [open reading frames] and their localization
2. predicted coding regions
3. predicted location of regulatory motifs (e.g. promoters)
4. Functional annotation consists of attaching biological information to genomic elements.

17.8 Talking Glossary: Substitution mutation (1.75 min)

https://www.genome.gov/genetics-glossary/Substitution

Abstract: "Substitution is a type of mutation where one base pair is replaced by a different base pair. The term also refers to the replacement of one amino acid in a protein with a different amino acid."

Transcript: "Subsititution refers to the replacement of one amino acid with another amino acid in a protein or the replacement of one nucleotide with another in DNA or RNA. Substitutions generally give rise to–or they always give rise to–either a polymorphism, that is, a difference between one person, one individual, in a population or another, or a special kind of polymorphism that we call a mutation. In either case, all individuals in the population originally had the same sequence of a gene. There were substitution events that resulted in a change in DNA sequence, which resulted in a change in RNA sequence, which then could result in a change in amino acid sequence. When that happens, that change in DNA sequence or amino acid sequence, or both, could have no effect on the protein, in which case the substitution is benign and has no functional

effect. In other cases, if it changes the function of the protein, then it will be observed as either a functional polymorphism, something which increases the effectiveness of the protein product, and therefore would be evolutionarily selected for, or is bad–deleterious–in which case the person might die early and get evolutionarily selected against. Substitutions which lead to mutations, which lead to a deleterious outcome, that is the organism having difficulty with living or dying early, those we call mutations, but they're the result of a certain kind of a substitution."

Christopher P. Austin, M.D.

Chapter 18

T

18.1 Taxa, Taxon, and clades: A Brief Primer

By Nathan Brouwer

The following reading discusses what biologists mean when we use the terms taxa, taxon and clade. A few advanced topics are discussed that will be further detailed in future lessons. Names of specific species, families, orders etc. are used as examples; you are not expected to memorize them. A google-doc version of this file can be found here (Links to an external site.).

Key Vocab: * Taxa * Taxon * Clade * Order * Taxonomic group * Primates

Advanced topics:** * Monophyletic group * Non-monophyletic group

18.1.1 What is a "taxon"

Taxon and **clade** refer to different kinds of biological groups. **Taxa** is the plural of **taxon**. A taxon (or **taxonomic group**) can refer to any species *or* group of species. Humans, *Homo sapiens*, is a taxon. We are the genus *Homo*, a larger taxon composed of humans and our close relatives, including Neanderthals (genus: *Homo*, species: *Homo neanderthalensis*). We are part of the **family** Hominidae, a larger taxon which includes us, chimps and bonobos (genus *Pan*), gorillas (genus *Gorilla*), and orangutans (genus *Pongo*). All primates are part of an **order**, which includes all the taxa mentioned previously as well as taxa such as monkeys, gibbons, and lemurs. We can say all primates considered together are taxon (a group of species), which is composed of many taxa (many individual species).

The upshot: each level of the hierarchy is a taxon which contains multiple taxa:

- Order: Primates
- Family: Hominidae

- Genus: Homo
- Species: *Homo sapiens*

The word "taxa" is often used as a generic stand-in for species, family, or order. I can say about my own research "I studied two taxa of plants" – that is, I studied two plant species, though it could also mean I studied two plant families, two plant orders, etc. Someone who studies the phenomena of punctuated equilibrium can say "the number of taxa on earth increased dramatically during the Cambrian explosion." A key way to remember this: Each of the branch tips on a phylogenetic tree is a taxon. Often the tips of a tree are species, but not always.

When using "taxa" you don't necessarily have to refer to the current evolutionary or phylogenetic understanding of the species you are talking about. Most taxonomic groups were thought to reflect evolutionary relationships when they were first proposed, but subsequent information has indicated that the group isn't necessarily coherent. For historical reasons and convenience, these taxonomic groups are maintained.

For example, the taxonomic group **reptiles** includes lizards, snakes, crocodiles, and turtles. These were once thought to be a cohesive group. We now know that birds and crocodiles are **sister taxa** and that the group we call "reptiles" should include birds if it were to be evolutionarily consistent (see here for a summary). The field known as herpetology focuses on the biology of reptiles, as well as amphibians (frogs and salamanders). This is a group of taxa that excludes birds and evolutionarily isn't coherent. Similarly, the taxonomic group of "fish" is problematic from an evolutionary perspective (The precise reasons for this are beyond the scope of this particular reading).

One key point is that "taxa" and "taxon" do not have to refer to the species level - it could refer to subspecies, populations, or other levels of the hierarchy of biological classification. It is most commonly used in reference to the level of species, but this isn't necessary.

18.1.2 What is a clade?

In contrast to taxon and taxa, **clade** has a very specific meaning - a clade is **all** of the taxa that descended from a common ancestor, *plus* the common ancestor. Clade is synonymous with **monophyletic group**. If you exclude one of the taxa that descended from a common ancestor you end up with what is called a **non-monophyletic group**. "Reptiles" and "fish" are taxonomic groups (taxa) that are not clades because they don't contain all of the descendants of their common ancestor and therefore are not evolutionarily coherent.

Glossary:

- Clade – a group composed of an ancestor and all of its descendants.
- Monophyletic group – a group composed of an ancestor and all of its descendants (=clade). Example: Reptilia is a monophyletic group if (and

only if) it includes birds
- Non-monophyletic group – a group composed of an ancestor and only some of its descendants, where the missing ones have been placed in another group. Example: "Reptilia" is a paraphyletic group if birds are excluded from it. The names of paraphyletic groups are often placed in quotation marks by convention.
- Taxon – a named group and its constituent members. Normally a taxon will be a named clade. Phylogenetic definition (of a group) – a definition for a group that is based on common ancestry. Example 1: Tetrapoda is the group composed of the last common ancestor of living amphibians and amniotes and all taxa more closely related to that clade than to lungfish.

18.1.3 Further Reading:

For more on clades: http://en.wikipedia.org/wiki/Clade For more on taxa: https://en.wikipedia.org/wiki/Taxon

"In biology, a taxon (plural taxa), from taxonomy, is a group of one or more populations of an organism or organisms seen by taxonomists to form a unit." ICZN (1999) International Code of Zoological Nomenclature.

"A taxonomic unit, whether named or not: i.e. a population, or group of populations of organisms which are usually inferred to be phylogenetically related and which have characters in common which differentiate (q.v.) the unit (e.g. a geographic population, a genus, a family, an order) from other such units."

For information on non-monophyletic groups and why fish aren't a clade, see these sites

- http://ess17.blogspot.com/2008/06/q-can-you-please-explain-question-3-on.html
- https://en.wikipedia.org/wiki/Paraphyly
- https://en.wikipedia.org/wiki/Fish

18.2 Talking Glossary: Trait (0.75 min)

Abstract: "A trait is a specific characteristic of an organism. Traits can be determined by genes or the environment, or more commonly by interactions between them. The genetic contribution to a trait is called the genotype. The outward expression of the genotype is called the phenotype."

Comments: In ecology and evolutionary biology, we often use the terms trait and character interchangeably. In phylogenetics the term character is common.

In phylogenetics, we organize our data in a grid called a trait matrix, or a character matrix. We can consider phenotypic traits/phenotypic characters or sequences (genetic traits/genetic characters). Often we refer to phenotypic

traits as morphological traits if they have to do with the size, shape, structure, anatomy etc. of an organism.

The definition here in the Talking Glossary highlights the fact that phenotypic traits are due to both genes and the environment. When we do phylogenetics with phenotypic traits we strive to use consistent morphological differences that are due to genetics. Because phenotypic traits like height, weight, color etc can vary so much due to the environment we use traits that can take on distinct states, like tails: monkeys have tails while humans don't; dogs long have tails but bears have very short ones. We therefore draw a distinction between characters (e.g. tails) and character states (tail present, tail absent).

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/trait.mp3

Transcript "Trait is a specific characteristic of an individual. For example, their hair color or their blood type. Traits are determined by genes, and also they are determined by the interaction with the environment with genes. And remember that genes are the messages in our DNA that define individual characteristics. So the trait is the manifestation of the product of a gene that is coded for by the DNA. The word"phenotype" is sometimes used interchangeably with the word trait, and "phenotype" may also define a whole compendium of traits together."

Donna Krasnewich, M.D., Ph.D. Program director in the Division of Genetics and Molecular, Cellular, and Developmental Biology, National Institutes of Health

Photograph: https://www.nigms.nih.gov/about/PublishingImages/headshots/headshot-donna-krasnewich-small.jpg

Chapter 19

U

19.1 Uniprot

Adapated from Wikipedia, the free encyclopedia

UniProt (www.uniprot.org/) is a freely accessible database of protein sequence and functional information, many entries being derived from genome sequencing projects. It contains a large amount of information about the biological function of proteins derived from the research literature. It is maintained by the UniProt consortium, which consists of several European bioinformatics organisations and a foundation from Washington, DC, United States.

19.1.1 The UniProt consortium

OPTIONAL

The UniProt consortium comprises the European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR). EBI, located at the Wellcome Trust Genome Campus in Hinxton, UK, hosts a large resource of bioinformatics databases and services. SIB, located in Geneva, Switzerland, maintains the ExPASy (Expert Protein Analysis System) servers that are a central resource for proteomics tools and databases. PIR, hosted by the National Biomedical Research Foundation (NBRF) at the Georgetown University Medical Center in Washington, DC, US, is heir to the oldest protein sequence database, Margaret Dayhoff's Atlas of Protein Sequence and Structure, first published in 1965.[2] In 2002, EBI, SIB, and PIR joined forces as the UniProt consortium.[3]

19.1.2 The roots of UniProt databases

OPTIONAL

Each consortium member is heavily involved in protein database maintenance and annotation. Until recently, EBI and SIB together produced the Swiss-Prot and TrEMBL databases, while PIR produced the Protein Sequence Database (PIR-PSD).[4][5][6] These databases coexisted with differing protein sequence coverage and annotation priorities.

Swiss-Prot was created in 1986 by Amos Bairoch during his PhD and developed by the Swiss Institute of Bioinformatics and subsequently developed by Rolf Apweiler at the European Bioinformatics Institute.[7][8][9] Swiss-Prot aimed to provide reliable protein sequences associated with a high level of annotation (such as the description of the function of a protein, its domain structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases. Recognizing that sequence data were being generated at a pace exceeding Swiss-Prot's ability to keep up, TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was created to provide automated annotations for those proteins not in Swiss-Prot. Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families.

The consortium members pooled their overlapping resources and expertise, and launched UniProt in December 2003.[10]

Chapter 20

V

20.1 Talking Glossary: Vector (2 min)

National Human Genome Research Institute

Note: Usually refers to a plasmid, though in this entry they mention how it can also apply to viruses.

Introduction: "A vector is any vehicle, often a virus or a plasmid [usually a plasmid in practice, and that's what we focus on in this class] that is used to ferry a desired DNA sequence into a host cell as part of a molecular cloning procedure. Depending on the purpose of the cloning procedure, the vector may assist in multiplying, isolating, or expressing the foreign DNA insert."

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/vector.mp3

Transcript "A vector is a way to take a sequence of DNA, usually, and introduce it into another place. So what vectors do is allow you to propagate the DNA you're interested in, in the organism you've chosen to propagate it in. So the simplest one is the origins of recombinant DNA technology: They made copies of RNAs, and they were able to insert these into what is known as plasmids. Now, plasmids are kind of mini-bacterial chromosomes. They have a way to replicate themselves, and what makes it work is they also carry one or two genes on them that make them resistant to specific antibiotics. So if you can insert the gene you're interested in into this plasmid, you can select for the bacteria that have picked up that plasmid by growing them on an antibiotic that, if they haven't picked it up, would kill them. So that plasmid is a vector for taking a particular DNA sequence into a bacteria. And then you can isolate one colony of bacteria and clone that, grow that clone up, and that's how you would propagate that. There are other vectors that are larger and will have multiple sites of origins of replication, and these are known as bacterial artificial chromosomes, and they

can handle much larger pieces of DNA. There are yeast artificial chromosomes that allow very large fragments of DNA to be grown in yeast cells. And recently, human artificial chromosome have been developed that allow enormous pieces of DNA to be introduced and propagated into human cells. So vector is really just a means to take a piece of DNA that you're interested in and insert it, and select for it, and identify it in the organism that you want to propagate it in."

David M. Bodine, Ph.D

Chapter 21

W

21.1 Talking Glossary: X-chromosome (1.25 min)

Length: 1.25 min

https://www.genome.gov/genetics-glossary/X-Chromosome

Abstract: The X chromosome is one of two sex chromosomes. Humans and most mammals have two sex chromosomes, the X and Y. Some individuals have two X chromosomes in their cells, while others have X and Y chromosomes in their cells. Egg cells all contain an X chromosome, while sperm cells contain an X or a Y chromosome. This arrangement means that during fertilization, it is the XY parent that determines the whether the offspring will be XY or XX.

Image: https://www.genome.gov/sites/default/files/tg/en/illustration/x_chromosome.jpg

Audio: https://www.genome.gov/sites/default/files/tg/en/narration/x-chromosome.mp3

Transcript "So this, because I'm a female, is truly one of my favorite chromosomes. As you know, females have two X chromosomes. They're quite large in comparison to the male chromosomes. They are carried by the egg, and so consequently you pass on–if you have an egg–you can only pass on an X chromosome to your offspring. What's also quite interesting is the number of genes that are found on the X chromosome. It is predicted that there are approximately 155 million base pairs, which translates to about 900 to 1,400 genes on the X chromosome. Meaning that it carries about five percent of the total DNA in the entire cell. Which is quite in contrast to the Y chromosome, which is considerably smaller. Again, if you look at the X chromosomes in the genes that it carries, often times you'll see that sex-linked disorders are carried on the

X chromosome. Which is why, as I stated before, they're more predominant in male, because there's not a protective mechanism against having a mutation on one of those genes because we don't have the normal copy of that gene on the X chromosome."

Carla Easter, Ph.D.

Image: https://www.genome.gov/sites/default/files/genome-old/images/content/easter_carla.jpg

Biography

Carla Easter, Ph.D., is chief of the Education and Community Involvement Branch at the National Human Genome Research Institute (NHGRI). She played a major role in the development of the NHGRI/Smithsonian exhibition Genome: Unlocking Life's Code, and its accompanying website, and serves as a liaison to the K-12 and university community as a speaker on genomic science and career preparation and pathways. Dr. Easter also serves as an adjunct faculty member at the University of the District of Columbia Department of Biology, Chemistry and Physics.

From 2003-2006, Dr. Easter was director of outreach for Washington University School of Medicine's Genome Sequencing Center. Before assuming her role as outreach director, Dr. Easter was a research associate in the Department of Education at Washington University (2001-2003) where she explored the notions of science among secondary students. She served as pre-college coordinator for the NASA Summer High School Apprenticeship Research Plus Program and project associate for the Quality Education for Minorities Network. From 1997-2000, Dr. Easter conducted post-doctoral research at Washington University School of Medicine on the virulence factors associated with Streptococcus pyogenes.

Dr. Easter earned her bachelor's degree in microbiology from the University of California, Los Angeles and her doctoral in biology with an emphasis on molecular genetics from the University of California, San Diego.

Video: https://youtu.be/S1-Y8cO-sMc Dr. Easter is a laboratory scientist who now specializes in science communication and outreach. The video below is a discussion between Dr. Easter and a highschool student sponsored by the Children's Science Center Lab

The YouTube video below contains an audio-only interview with Dr. Easter https://youtu.be/X_CX7FWSfoM

21.2 Talking Glossary: X-inactivation ("Lyonization"; 1.45 min))

https://www.genome.gov/genetics-glossary/Lyonization (Links to an external site.)

Abstract: "Lyonization is commonly known as X-inactivation. In mammals, males receive one copy of the X chromosome while females receive two copies. To prevent female cells from having twice as many gene products from the X chromosomes as males, one copy of the X chromosome in each female cell is inactivated. In placental mammals, the choice of which X chromosome is inactivated is random, whereas in marsupials it is always the paternal copy that is inactivated."

Audio

Transcript: "Lyonization is named after Mary Lyon, who was a geneticist who first figured out that in females who have two copies of the X chromosome, that one copy of each gene is turned off permanently in one chromosome or another. So that females, who have two copies of the X chromosome, and males, who have one copy of the X chromosome, are both human, and they can both operate fairly normally. So this process of turning off one copy of one gene or another on the X chromosome is called lyonization, and it happens from a series of essentially irreversible chemical modifications to one copy of the gene. The fascinating thing about this is that in so-called X-linked diseases, if the female inherits a gene responsible for an X-linked disease, and has one copy that's abnormal and one copy that's normal, the abnormal gene is almost always the one that's turned off. And the normal gene is almost always allowed to stay on. In X-linked diseases in males, of course, X-linked diseases manifest because they only have one X chromosome, and so those mutated genes have to show up. But the fascinating part about it, which we really don't understand, is how it is that the body knows if the female inherits one copy of a gene that is abnormal and would otherwise cause an X-linked disease if the normal copy was lyonized and turned off. That doesn't happen. The abnormal copy gets turned off, leaving the normal copy to function, and saving the female from having the disease."

Christopher P. Austin, M.D.

History:

English geneticist Mary Lyons (https://en.wikipedia.org/wiki/Mary_F._Lyon) first characterized X-inactivation. You can read a short article about her life here (https://www.nature.com/articles/518036a).

Chapter 22

X

Chapter 23

V

23.1 Additional Glossaries

https://www.ncbi.nlm.nih.gov/books/NBK5191/

Chapter 24

Z