Introduction to Biology
Origin of Life

Why do we need to know Biology?

- **To find solution to challenges faced by mankind**
- **Sustainability**
- **Design through mimicry – bio-mimicry**
- **Understanding human health and disease, improving the yield of crops, etc.**

Bio-mimicry

Ex-1: **surface of whales do not allow bacteria to stick and grow on them**

- **if we can have such a surface at micro-scale in hospitals, problem of spread of infections can be avoided.**

Ex-2: **beetles can capture water from the air for their requirement**

- **If a similar structure can be made, problems of Chennai can be solved.**

Ex-3: **We are all familiar with solar energy, which uses chemical based solar cells to capture light and generate electricity**

- **Photosynthesis does exactly the same thing, converting light energy into chemical energy, but in a very efficient manner**

Human Health

Second reason for studying biology is to study about us, what happens in the disease state

- **How can we improve our wellness, both physical and mental**
- **Be disease-free and mentally alert, or mentally at peace**
- **It should be possible through a proper understanding of biology, the cell, it's processes, the systems as a whole**

- **simple things such as obesity is still not understood properly - it is not a simple calorie counting thing**

Human Health

There are numerous examples of biological problems that are being solved by computer science engineers (or, electrical / mechanical / material engineers)

- **Developing artificial retina, which allows people with damaged retina to see**
- **Possibility of quadriplegics walking again just by using electrical signals in the brain – researchers working on human-computer interface**
- **Identifying rare disease variants by pattern search**
- **Organ transplant - In Jan 2022, the world's first pig-heart transplant was successfully achieved in a human patient**
- **etc.**

Scholarly View

The third reason for studying biology is

- **it is there and needs to be understood, e.g., Mendelian genetics, cell morphology, etc.**

- e.g., is a newborn child predisposed to certain disease, are bacterial cells different from plant or human cells, normal vs cancer cells, is there a difference between cells of different tissues, etc.

- **this may not have any immediate application, but may help make a better sustainable life for ourselves and our future generations**

- **The 21st century is called the century of Biology and major advances and breakthroughs are being made with the new biology of genome research – using the complete genetic blueprint of a species.**

Scholarly View

Compared to Physics wherein we find mostly 'Universal' laws,

- **In Biology the rules are generally not 'Universal', lot of exceptions are observed, the information available is not complete and understanding is rudimentary**
- **Not to mention, it's changing as we speak**

Reason for engineers to study biology

Today, Biology has become very quantitative because of large volumes of data generated

- **it's truly Big Data Analytics**

Developing algorithms and databases, addressing storage issues, data mining, applying ML/AI approaches, etc. are now routines tasks in biology

- a new discipline, Bioinformatics and Computational Biology, has emerged for the analysis of large volumes of biological data

In this half-course, we will try to understand some aspects of biology

- **Origin of Life - how life formed, how life evolved, and their relevance**
- **The basic biomolecules – DNA, RNA, proteins, metabolites, etc. and their analysis for knowledge extraction**

  - sequence alignment and sequence database search
- **Can we use molecular sequence data for finding the origin of SARS-CoV2 ?**

Origin of Life

First, let us understand what we mean by 'Life'

What are it's features that distinguishes it from non-living things?

- **the ability to replicate and reproduce**

In order to do that, to sustain its survival, and pass on its characters to its progeny

- an organism goes through a whole lot of processes that include its ability to utilize food, it's ability to synthesize food (if required), it's ability to digest food, it's ability to throw out the waste material and keep the machine going

Origin of Life

Life, or any biological system is a highly dynamical system, and its sole purpose is to survive and reproduce.

How did life evolve on the earth?

- **Based on studies done by geologist's, archaeologists, molecular biologists, there is enough evidence to speculate that life on earth evolved from non-living things**

⇒ we have essentially evolved from chemical reactions

Origin of Life

Radioactive dating estimates origin of earth ~4.8 billion yrs ago, and fossil records suggest earliest life forms ~ 3.6-3.8 billion yrs ago.

Origin of Life

Do any of the earliest life forms still exist on earth?

- **yes, bacteria, since over 3 Billion yrs**

Based on various evidences from geological excavations, chemical reactions and biotechniques, biologists divide the process of life and its existence into three phases:

- **Chemical evolution**
- **Acquisition of replicative ability**
- **Biological evolution**

Chemical Evolution

This phase corresponds to very early stages of earth's life, when the earth's crust was still very hot, the oceans and sea were still boiling, and the atmosphere was highly reducing.

Chemical Evolution

Any living organism –

~ 70% of its body wt is water,

major constituents of dry wt: C, H, O, N, P and S

- **Carbon is most abundant**
- **An interesting property of C is its ability to form covalent bonds with itself & other elements**
- **Another imp thing is highly reducing earth's atmosphere because of presence of H_2S, NH_3, CH_4, CO, and no atmospheric Oxygen**

Geologists, on analyzing early rocks on earth & meteorites, found to be rich in carbon compounds

⇒ the early building blocks were formed with carbon

Chemical Evolution

Hypothesis-1 (A. Oparin & J.B.S. Halden, 1929):

- the very first molecules/biomolecules would have arisen because of abiotic synthesis of small molecules.

- **due to initial reducing environment of earth, any high energy discharge, either in the form of ultra-violet lights, or in the form of lightning would have favoured spontaneous synthesis of simple molecules from existing geological molecules on earth's surface.**
- **In 1953, it was actually demonstrated to be true by Stanley Miller and Harold Urey.**

⇒ first set of sugars or amino acids were formed by a mere abiotic synthesis of these molecules

Chemical Evolution

Miller & Urey created the primordial soup in a laboratory setup

- **Collected water vapor under a very reduced environment of CH_4, NH_3 & H**
- **Energy supplied to mimic ultra-violet radiations or lightning through electrodes**
- **whatever was spontaneously generated was condensed and collected**

It consisted of amino-acids such as Alanine & Glycine, sugars, nucleic acid, and nucleic acid bases like Adenine and lipids.

Chemical Evolution

Hypothesis-2: Abiotic Polymerization of Small Molecules

How were macromolecules, such as cellulose (cell wall of plants) formed from small mols?

- **By spontaneous polymerization due to very high rate of dehydration, simply because of the presence of hot rock, clay or sand which would promote such a chemical reaction.**
- **due to presence of carbonaceous material in the earth's crust, the rate of synthesis was much higher than the rate of hydrolysis.**

Abiotic Polymerization of Small Molecules

- Hot rock, clay or sand
- Dehydration
- Rate of synthesis > Rate of hydrolysis
- Without enzymes

Proof by Sydney Fox (1950s)

- **dripping organic monomers on hot clay (iron pyrite) or sand, which do have these charged sites**
- **monomeric molecules joined together to form polymeric molecules, and in the process, the metal ions facilitate the condensation process.**

Chemical Evolution

Hypothesis-3: Abiotic Replication

When did the replicative ability begin?

- **we do not really have clear proof as to how this change happened from a non-replicative to the replicative form and life, but,**
- **we now have sufficient proof that the earliest genetic material, the ideal initiator of this process, must have been RNA.**

The flow of information in the living world:

$$\text{DNA} \longrightarrow \text{RNA} \longrightarrow \text{Protein}$$

Chemical Evolution

What makes RNA the probable candidate as the earliest molecule capable of acquiring the ability to replicate?

RNA is a very versatile molecule. It can

- **Polymerize itself**
- **Cleave itself**
- **Act as an enzyme and this catalytic activity is not dependent on proteins**
- **Involved in multiple steps of protein synthesis; even ribosomes are 2/3rd RNA**
- **Even DNA synthesis is dependent on RNA**

Sydney Altman et al, 1980

Chemical Evolution

How did cells come into existence?

Hypothesis-4: Protobionts

Aggregation of abiotically synthesized molecules for maintenance of internal environment

- **would have assembled as liposomes**
- **these are amphipathic, i.e., they have a polar head and non-polar tail**

These initial lipid liposomes would have enclosed abiotically synthesized self-replicating biological molecules to form the first and the earliest form of life, called protobionts

Liposomes

Hydrophilic pocket

A Liposome is a spherical vesicle made of phospholipids derived from natural or man-made materials

Phospholipids are the main components of biological membranes

Chemical Evolution

And somewhere down the line, the protobionts would have evolved into prokaryotes, time-scale – millions of years

The Oxygen Revolution

Almost for the first 2 ½ billion years, there was hardly any oxygen in the atmosphere.

The Oxygen Revolution

So what happened that led to the sudden burst from hardly any oxygen to a point where the atmosphere became highly oxidized.

It is believed that it's at this stage, in the earth's life that the plants or the photosynthetic organisms must have evolved.

By the time of 2 ½ billion years, resources must have reduced, the organism would have to find alternate sources of obtaining energy for survival.

One possible way would have been to use the reducing power using hydrogen sulphide.

The Oxygen Revolution

After exhausting H_2S, the organism must have developed a much smarter strategy of utilizing the unlimited pool of solar energy and the process of photosynthesis would have evolved.

- oxygen being a by-product, would have resulted in the atmosphere becoming highly oxidized.

New set of chemical reactions would have evolved under oxidizing conditions for breakdown of polymers into monomers and releasing energy

- somewhere at this point probably, the evolution of mitochondria would have happened.

To Summarize

As we understand today, the origin of life is divided into 3 stages:

- **First, the stage of chemical evolution - abiotic formation of organic molecules and its polymerization**
- **Second, self-organization, and the ability of these polymers to develop the property of replication and form the early protobionts.**
- **Third, the stage of metabolic evolution, i.e., evolutions in metabolic reactions that happen inside a living organism.**
  - **allow the organism to synthesize food, generate energy from food, and eventually transition from single-cell to multi-cell organism.**

Evolution

An intriguing point in the study of life is that despite the varied forms of life that we see today,

it is interesting to note that some of the fundamental metabolic reactions and the way genetic information is coded in our DNA are conserved right all the way from prokaryotes to humans.

Genetic and Biochemical similarity across the living world despite the diversity!

Evolution

Let us now see how from this early form of life, the present-day organisms evolved - a huge array of them.

So what is it that caused this variation, this distribution in different life forms, that we call evolution.

What is evolution?

- **Essentially it tries to explain how life must have diversified into multiple forms we see today.**
- **these present forms are believed to be the descendants of a common ancestor**

Evolution

During the course of history of earth's life, a lot of changes happened in organisms which were passed on to subsequent generations.

In other words, multiple heritable modifications from one generation to the next occurred, and it is this evolution that accounts for the unity and the diversity of life.

Let's consider an example - the basic skeletal architecture of the wings of Bat and compare it with the forelimbs of Humans/Cats and to the flipper of Whales.

Anatomical Homology

Forelimbs of mammals

Same basic structural plan is evident in all – one can "morph" one into another just by changing the sizes and proportions
- Genetic mutations and expression patterns of limb-development genes are responsible for these changes

Homology: Limb Structure

Different functionality despite having similar architecture:

throwing **walking** **flying** **swimming**

Evolution

Similarly, if we look at DNA, as far as the language which codes the information in DNA has been fairly conserved.

Yet, the DNA found in humans is slightly different from the DNA in earlier organisms, not in terms of its chemical entity, but in terms of further architectural development.

Evolution essentially talks about these changes - acquired by various life processes over the course of earth's history.

10-100M living species

~2M identified and classified

Evolution

One of the earliest scientist to bring in the concept of evolution was Lamarck.

- though his theories were disproved later.

According to him, it is a particular organ, or a particular feature in an organism that develops during the lifetime of the organism based on the usage in that organism.

He reasoned that the organs not being used by organisms eventually become useless, called vestigial organs, e.g., appendix in our body.

⇒ His theory based on use or disuse of a particular part was not the reason for evolution

Evolution

Remarkable observations made by Charles Darwin

Most interesting observations that he made were in the Galapagos Islands

- **in each island, there were unique birds, species and tortoises and no two islands had exactly the same kind of species, though similar to those found in the nearest island.**

Another observation he made during his voyage across the coastlines of Africa, Southern America, and Australia, was that the distribution of the species mirrored how the continental drift actually happened in the earth's history, e.g.,

- Giant flightless birds: Rhea (S. America), Ostrich (Africa), Emu (Australia)

Evolution

Darwin Qs:

- **Why are there so many species of living things?**
- **How do new species arise?**
- **How does the population of an organism evolve?**
- **How do organism adapt to their environment?**
- **How is one species related to another?**

He observed that, in a given population of a species, e.g., humans, there are individuals with different features (height, colour of hair/eyes/skin)

- these differences within a species, called variations, are heritable.

Evolution

The other observation he made in the Galapagos Islands was that

- **though the species were different in different islands, based on the environment which was being provided by the island, the species could adapt to that environment.**
- **e.g., in the islands where vegetations were found at a much lower level, the neck of the tortoises were smaller, while in those islands where vegetations were slightly at a higher level, the tortoises had a longer neck.**

⇒ the species can adapt to the changing environment.

Evolution

The third observation he made was that

- **as a population grows in size, eventually, the resources become limited.**
- **this would result in a competition among the individuals of a population for the resource**
- **Lead to 'the survival of the fittest'.**

⇒ Individuals with traits, or variations which allow them to best adapt to the environment are most likely to survive and reproduce, and also pass on these favourable characters to the next generation.

Evolution

To summarize Darwin's observations:

- **Individuals in a population exhibit variable traits: Variations**
- **Many traits are heritable**
- **Species adapt to their environment**
- **Limited resources**
- **Competition for survival**

Individuals with traits which allow them to best adapt to the environment will most likely survive and reproduce.

Natural Selection

In 1859 Darwin along with another naturalist, Alfred Russell Wallace, presented their ideas to the London Philosophical Society, and he published his most famous work, called 'The origin of species by means of natural selection'.

- **every species shows an evidence of descent from a previous species with modifications - they all have common ancestor**
- **the pressure which brings about these modifications, or variations, is the environment in which the organism lives.**
- **This pressure, provided by the environment is what is called the 'natural selection'.**

Natural Selection

A classic example of natural selection has been around the time of industrial revolution, giving rise to a species of peppered Moth (J.W. Tutt,)

1811 - most moths in England were light-colored

1848 - when industrial revolution was taking place in Manchester, a lot of peppered and black coloured moths were observed

- **by the end of 19th century, in a period of ~90yrs, light-coloured moths were totally outnumbered by dry, dark-coloured moths.**

1896 - J.W. Tutt presented it a case of natural selection

⇒ Change in the colour was a favourable trait, giving the moths survival advantage

Types of Natural Selection

How do organisms adapt to the environment?

- **Stabilizing**
- **Directional**
- **Diversifying**

Assume you have a population trait exhibiting normal distribution and that the environment is conducive and the conditions are fine,

- **In such a case there is no need for the outer extreme of the organisms to grow**
- **such kind of a selection where the central most acquired characters are retained is a stabilizing selection.**

Types of Natural Selection

If we have a situation where suddenly the environmental conditions force in such a fashion that it is this set of extreme outliers that have a survival advantage

- **this is an example of directional selection**

If a population, which because of a continental drift, or some other kind of catastrophe, gets split into two populations with different characteristics

- this is called as the diversifying selection.

Mendelian Genetics

Darwin did not know

- **what is the molecular mechanism which causes this variation.**
- **how are these variations passed on to successive generations?**
- **Why do certain species go extinct?**

Mendelian Genetics

How are the variations caused?

- **There are multiple ways, and the most classic way is the process of mutation.**
- **During DNA replication, an error gets incorporated in a gene and this mutation is favourable (or neutral) to the organism – it get's fixed in the population**
- **During sexual reproduction - we get a certain set of genes from our father, and a certain set of genes from our mother**
- **this has been one of the major reasons for variation and is one of the major reasons for success of survival in higher organisms.**

Mendelian Genetics

What are the causes of extinction?

- **we speculate that if a species fails to adapt itself to a changing environment, it would go extinct, or**
- **if it specializes to such a point that it cannot re-adapt itself, then also it can undergo extinction.**
- **Catastrophic event**

Mendelian Genetics

How do we try to understand evolution?

- **through fossil records,**
- **through geological excavations;**
- **by comparing the anatomical structures, or**
- **by comparative embryology**
- **Molecular phylogeny – look at similarities in DNA, RNA and proteins sequences/structures, e.g., DNA that codes for enzymes involved in basic metabolic reactions are highly similar across evolution.**

Embryonic Development of Vertebrates

Mendelian Genetics

Evolution is a continuous process!

the evolutionary changes accumulate over generations and then diversify into a newer species – but this happens over the scale of millions of years

- In molecular phylogeny we try to analyze these variations at the DNA level to understand evolution, and also to understand the function of various genes/proteins

Lecture-2
Biomolecules

Goal

Goal of molecular cell biology - to understand the physiology of living cells in terms of the information that is encoded in the genome of the cell

How computer science can help in achieving this goal?

⇒ To identify where are the genes in the genome, which gene is expressed when, where and how much, what factors affect its expression, what's its function, what happens in the disease state, what are the genetic variations and how to identify them, etc.

What kind of biological data analysis can help in combating the COVID-19 disease?

What kind of analysis can we carry out to know about the virus causing COVID-19?

- **How to identify if a person is infected with SAR-COV-2?**
- **Is it the only known human coronavirus?**
- **Comparing its genome with other viral genomes – to identify its closest relative**
- **What proteins aid in its transmission and infection?**
- **Identifying drug targets and develop vaccines**
- **What organs/tissues are affected by its infection?**
- **Its rate of propagation**
- **Is it mutating and becoming more virulent, or milder with time, etc.**

Cells

All living things are made of cells: small, membrane-enclosed units endowed with the ability to create copies of themselves by growing and dividing in two.

E. B. Wilson: "the key to every biological problem must be sought in the cell; for every living organism is, or at some time has been, a cell."

~ 10^{13} cells that form a human body, the whole organism has been generated by cell divisions from a single cell

Cells are fundamental units of life
- the vehicle for all the hereditary information that defines each species.

Study of the structure, function, and behavior of cells is called Cell Biology

© Encyclopædia Britannica, Inc.

Cells and Genomes

Do we carry cells of any other organism within us, apart from human cells?

How human are we?

- We have 10 trillion human cells and 100 trillion microbial cells: with respect to cell count we are just 10% human
- Our genome has 20-30K genes, our microbiome has 2-20M genes: with respect to genes we are 0.1-1% human
- Our microbiome weighs ~ 3 pounds, about the same weight as our brain, and maybe as important to our well being, if not more important
- We share 99.9% of our genome, but we share only 10% of our microbiome
- Microbiota include bacteria, archaea, viruses, eukaryota

Gut Microbiome

What does our microbiome do?

- It acts like an extra organ, helping us to digest molecules in our food that we couldn't break down ourselves, even can steal genes in order to help us digest exotic food.
- Paracetamol is toxic to our liver or not is determined by whether we have sulfite reducing microbes in our gut
- A diet works for you depends on the microbes you have before you started the diet
- The gut microbes train the immune system. It is now identified that our immune system plays in important role in response to COVID-19.
- They may influence your behaviour and affect your health

Cells and Chromosomes

- All cells store their hereditary information in the same linear chemical code: DNA (Deoxyribonucleic acid)
  - organized in coiled structures called chromosomes.
- Chromosomes are physically separate molecules that range in length from 50 - 250M base pairs.
- In most prokaryotes, each cell has a single chromosome.
- In eukaryotes, all cells contain same number of chromosomes, 8 in fruit flies, 46 in humans and bats, 84 in rhinoceros.
- In multi-cellular organisms, similar cells join together to form tissues.
- The complete DNA content of an organism is called genome.

Living organism divides into two major groups:

- **Prokaryotes - cells with no nucleus, e.g., bacteria and archaea**
- **Eukaryotes – cells with a nucleus, e.g., plants, animals and protozoas. These maybe unicellular (e.g., Yeasts) or multi-cellular (e.g., humans).**

Figure 1

Viruses

Viruses do not fall in either of the two divisions.

These are acellular entities with many of the properties of other life forms but are genetically and structurally too dissimilar to fit into the taxonomy scheme.

Viruses cannot be classified as living or non-living. They can replicate & carry out their functions only in a host – called obligate parasites.

May contain DNA or RNA as their nuclear material.

Viruses are acelluar. So, if they do not have a cell, how does their genetic material remain intact?

DNA

DNA (Deoxyribonucleic acid):

Composed of four basic units - called nucleotides

Each nucleotide contains - a sugar, a phosphate and one of the four bases:

Adenine (A), Thymine (T),

Guanine (G), Cytosine (C).

DNA

Bases: are ring-shaped and come in four types which fit together in pairs - this pairing forms the basis of information carrying capacity of DNA.

These are categorized as:

Pyrimidines **Purines**

Cytosine (C)

Thymine (T)

Guanine (G)

Adenine (A)

DNA

Backbone of a polynucleotide strand is made of linked sugar-phosphate-sugar-phosphate, with one base joined to each sugar molecule.

- built by creating a phosphodiester bond that links 3′ carbon on the sugar of growing chain with the phosphate attached to 5′ carbon of an incoming nucleotide.

DNA Strand

DNA

Base Pairing: **If two polynucleotide strands face each other, sugar-phosphate backbone runs down each side, and complementary pairs of bases in the middle spontaneously form hydrogen bonds:**

Triple bond between C and G

Double bond between A and T

Note: the two strands are not mirror images of each other - they run antiparallel

DNA

Double-Stranded DNA: **Nucleotides within each strand are linked by strong (covalent) chemical bonds; complementary nucleotides on opposite strands are held together more weakly, by hydrogen bonds:**

Why is DNA a helix?

Double-Stranded DNA: **If the sequence in the forward strand in 5′ to 3′ direction is:**

5′ CATTGCCAGT 3′

Then what would be the sequence on the reverse strand when read in 5′ to 3′ orientation?

DNA

If the sequence in the forward strand in 5′ to 3′ direction is:

5′ CATTGCCAGT 3′

Then what is the sequence on the reverse strand when read in 5′ to 3′ orientation?

First write its complement:

5′ CATTGCCAGT 3′

3′ GTAACGGTCA 5′

When read in 5′ to 3′ orientation, the sequence on the reverse strand is:

5′ ACTGGCAATG 3′

DNA Replication

In living cells DNA is not synthesized as a free strand in isolation, but on a template formed by a pre-existing DNA strand.

Copying of genetic information by DNA replication

templated polymerization

Central Dogma of Molecular Biology

All cells transcribe portions of their hereditary information into the same intermediary form: Ribonucleic acid (RNA).

DNA expresses its information, by letting the information guide synthesis of other molecules in the cell: RNAs and Proteins

Process begins with a templated polymerization called transcription, for the synthesis of RNA.

In the more complex process of translation, these RNA molecules direct the synthesis of proteins

Ribonucleic Acid (RNA):

It is single-stranded molecule

Composed of four basic units - called nucleotides:

Each nucleotide contains - a sugar (ribose), a phosphate and one of the four bases: Adenine (A), Uracil (U), Guanine (G), Cytosine (C)

RNA polynucleotide strand is built by creating a phosphodiester bond between nucleotides.

Intra-strand base pairing is a characteristic feature of RNA

Base Pairing – formed by weak H-bonds and follows the following complementarity rule:

A ⟷ U, G ⟷ C, and G ⟷ U

Ring Structure of Nucleic Acid bases

Adenine (A)

Guanine (G)

Purines

Cytosine (C)

Uracil (U)

Pyridimines

RNA

Nucleotide pairing between different regions of the RNA polymer chain causes the molecule to adopt a distinctive shape.

sugar-phosphate backbone

Base pairing: G-C, A-U, G-U

Distinct shapes of RNA molecules resulting from such internal associations enable it to recognize other molecules by binding to them selectively, or, in certain cases, catalyze chemical changes in the molecules that are bound.

RNA Synthesis:

RNA is also read in the 5′ to 3′ orientation.

RNA molecules that are copied from the genes (which ultimately direct the synthesis of proteins) are called messenger RNA (mRNA) molecules.

RNA Synthesis

1. For the following DNA sequence in the forward strand:

5′ CATTGCCAGT 3′

What will be the sequence of the RNA strand synthesized?

2. If the following DNA sequence is used as template for RNA synthesis:

5′ CATTGCCAGT 3′

Give the sequence of the RNA strand read in 5′ to 3′ orientation.

RNA Synthesis

1. If the DNA sequence in the forward strand is given as:

5′ CATTGCCAGT 3′

Template sequence used for RNA synthesis is its complement:

5′ CATTGCCAGT 3′

3′ GTAACGGTCA 5′ template

The synthesized RNA sequence is the reverse complement of the template:

3′ GTAACGGTCA 5′ template

5′ CAUUGCCAGU 3′ RNA

- i.e., synthesized RNA sequence is basically the DNA sequence in the forward strand with T replaced by U

RNA Synthesis

If the following DNA sequence is used as template for RNA synthesis:

5′ CATTGCCAGT 3′

First write its complement:

5′ CATTGCCAGT 3′

3′ GUAACGGUCA 5′ **complement**

Then the synthesized RNA sequence in 5′ to 3′ orientation is:

5′ ACUGGCAAUG 3′ **RNA**

- i.e., synthesized RNA sequence is basically the complement of the template DNA sequence with T replaced by U, when read in the 5′ to 3′ orientation

RNA Synthesis:

There are other RNA molecules also obtained from genes. The final product in such cases is RNA.

These RNAs are known as noncoding RNAs because they do not code for protein.

e.g., in yeast *Saccharomyces cerevisiae*, over 1200 genes (~15%) produce RNA as their final product; Humans may produce on the order of 10,000 noncoding RNAs.

These RNAs, like proteins, serve as enzymatic, structural, and regulatory components for a wide variety of processes in the cell.

TABLE 6–1 Principal Types of RNAs Produced in Cells	
Type of RNA	Function
mRNAs	Messenger RNAs, code for proteins
rRNAs	Ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	Transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	Small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	Small nucleolar RNAs, help to process and chemically modify rRNAs
miRNAs	MicroRNAs, regulate gene expression by blocking translation of specific mRNAs and cause their degradation
siRNAs	Small interfering RNAs, turn off gene expression by directing the degradation of selective mRNAs and the establishment of compact chromatin structures
piRNAs	Piwi-interacting RNAs, bind to piwi proteins and protect the germ line from transposable elements
lncRNAs	Long noncoding RNAs, many of which serve as scaffolds; they regulate diverse cell processes, including X-chromosome inactivation

Note: rRNA, tRNA and snRNA play an important role in protein synthesis

Protein Synthesis

Proteins are synthesized from DNA in a two-step process:

Each chromosome has several genes that code for various traits in the body.

RNA molecules direct synthesis of proteins in a complex process called translation.

- information in mRNA is read out in groups of three nucleotides, called codons.

The Genetic Code

Second letter

First letter	U	C	A	G	Third letter
U	UUU, UUC } Phe UUA, UUG } Leu	UCU, UCC, UCA, UCG } Ser	UAU, UAC } Tyr **UAA Stop** **UAG Stop**	UGU, UGC } Cys **UGA Stop** UGG Trp	U C A G
C	CUU, CUC, CUA, CUG } Leu	CCU, CCC, CCA, CCG } Pro	CAU, CAC } His CAA, CAG } Gln	CGU, CGC, CGA, CGG } Arg	U C A G
A	AUU, AUC, AUA } Ile **AUG Met**	ACU, ACC, ACA, ACG } Thr	AAU, AAC } Asn AAA, AAG } Lys	AGU, AGC } Ser AGA, AGG } Arg	U C A G
G	GUU, GUC, GUA, GUG } Val	GCU, GCC, GCA, GCG } Ala	GAU, GAC } Asp GAA, GAG } Glu	GGU, GGC, GGA, GGG } Gly	U C A G

Start codon

The genetic code is degenerate

Protein Synthesis

Using the genetic code, obtain the amino acid sequence synthesized from the following mRNA sequence:

5′ ACU GGC AAU 3′

Thr Gly Asn

This genetic code is read out by a class of small RNA molecules, called transfer RNAs (tRNAs).

- **each type of tRNA attaches at one end a specific amino acid and at its other end has a specific sequence of 3 nucleotides - an anticodon that enables it to recognize, through base-pairing, a particular codon in the mRNA sequence.**
- **This process occurs on ribosome, a large multi-molecular machine composed of both proteins and ribosomal RNA.**

Protein synthesis

Proteins

Like DNA and RNA, Proteins carry information in linear sequence on a 20-letter alphabet, called amino acids:

ATRVGTCWPRA

Protein structure is divided in 4 hierarchical levels:

- **Primary structure - represented by AA sequences**
- **Secondary structure - α-helices & β-sheets**
- **Tertiary and Quaternary structures - represented by 3D structures**

Primary Structure: ATRVGTCWPRA

α-helix

Secondary Structures

β-sheets

Tertiary Structure

Quaternary Structure

Function of Proteins

- **Proteins make up much of the cellular structure – hair, skin, fingernails, etc.**
- **Enzymes – catalyze chemical reactions within the cell**
- **Transcription factors – regulate the manner in which genes direct production of other proteins**
- **Receptors – proteins on the surface of cells act as receptors for hormones and other signaling molecules**
- **Recognize and bind to Nucleic acids (DNA, RNA) and other Proteins – to carry out their functions in the cell**

Genes

Special sequences in the DNA code for genes:

- **Protein-coding genes, for which the final product is a protein.**
  - **Same gene may give rise to more than one protein (~ 6 per gene in humans).**
- **Non-coding RNA genes - for which the final product is RNA**

Genotype – An organism's genotype is the set of genes that it carries.

Phenotype – An organism's phenotype is all of its observable characteristics which are influenced both by its genotype and by the environment e.g., height, hair colour, levels of hormones, etc.

Differences in the genotypes can produce different phenotypes

Genes for ear form are different, causing one of the cats to have normal ears and the other to have curled ears

A change in environment also can affect the phenotype. Pinkness is <u>not</u> encoded in the genotype of flamingos – but the food they eat – brine shrimp and blue-green algae from which they obtain a natural pink dye, canthaxanthin.

The "Omics" Cascade
What can happen
GENOME
What appears to be happening
TRANSCRIPTOME
What makes it happen
PROTEOME
What has happened and is happening
METABOLOME
PHENOTYPE

Analyze Network of Networks

Types of interactions:

(1)enzyme catalysis,

(2)posttranscriptional control of gene expression by proteins

(3)effect of metabolite on gene transcription mediated by Protein

(4)protein-protein interaction

(5)effect of a downstream metabolite on transcription through binding by protein

(6)Feedback inhibition/activation of an enzyme by a metabolite

(7)exchange of a metabolite with outside of the system (eg ,cell)

Complexity and interdependency of biochemical networks

Systems Biology

The emphasis in biology is now shifting from identifying individual components and molecules to the study of vast networks that biological molecules create, which regulate and control life.

Aim of systems biology - to understand biological systems as a whole rather than by simply analyzing their parts

A "system" can be anything from a gene regulatory network to a cell, tissue, organ, or entire organism

- aim is to generate more accurate computer models to unravel the complexities of human physiology and the underlying mechanisms of disease.

Systems Biology - an integrated approach to the study of biology through experiment and use of computer models with both predictive and explanatory power

- exploits concepts & techniques from engineering, CS, and Maths to understand the complexity of living systems

Genes

The biological function of a gene is to preserve and express the genetic information encoded within it

Genes are normally very stable entities

Genetic stability is not absolute, however.

Genes may occasionally become altered; these changes called mutations create new alleles.

Mutant genes are also stable entities and are inherited in the same way as normal, wild-type genes.

Genes

Normal diploid cells such as somatic cells of humans contain two sets of genes – one set inherited from each parent.

- corresponding genes derived from each parent are called alleles.

Together the two alleles govern the phenotype of an organism.

What is the percentage of genes in a genome?

Genes

Gene-fraction varies from ~70% in prokaryotes to ~2 - 3% in humans

- **does that imply prokaryotes have more gene content than eukaryotes?**
- **Size of a prokaryotic genome? Eukaryotic genome?**

What is the function of the remaining ~97-98% of H. genome?

Major part of the genome is repetitive (~50%), and the remaining part consists of noncoding regions, whose functions may include providing chromosomal structural integrity and regulating where, when, and in what quantity proteins are made.

Mutations

Mutations - are local changes in the DNA content, caused by inexact replication and are of various kinds:

- **Substitution - a base is replaced by another - may or may not alter the protein sequence depending on the place it occurs.**
- **Insertion/Deletion – addition/removal of one or more bases – results in a frame-shift in coding regions.**
- **Rearrangement - a change in the order of complete segments along a chromosome.**

Mutations

Role of Mutations:

- **Mutations are the source of phenotypic variation on which natural selection acts, creating species & changing them.**

Without mutations there wouldn't be any evolution!

- **They are responsible for inherited disorders and diseases, which involve alterations in gene.**
- **Large scale rearrangements are responsible for creating new species, e.g., the human and mouse genome are very similar – major difference being the internal order of DNA segments.**

Chromosomal rearrangements occur both within and between chromosomes during evolution

The colors on the mouse chromosomes and the numbers alongside indicate the human chromosomes containing homologous segments.

Steps Leading from Gene to Protein

Prokaryotes

Eukaryotes

Gene Structure

UTR: Untranslated region

Eukaryote Gene Structure

Start codon: ATG **Stop Codon: TAA/TAG/TGA**

Transcription is initiated only at certain specific positions in the sequence, signaling the beginning of genes, called promoters.

Alternative Splicing

- **In many cases, the pattern of splicing can vary depending on the tissue in which the transcription occurs.**

  e.g., an exon maybe spliced in the gene transcribed in liver, but retained when transcribed in the brain.

- **This variation called alternative splicing, contributes to the overall protein diversity in the organism**

Alternative Initiation

- **Another type of variation that contributes to protein diversity is alternative initiation**

- **Alternative translation is an important mechanism of post-transcriptional gene regulation leading to the expression of different protein isoforms originating from the same mRNA**

Data Representation

DNA - a complex, dynamic, three-dimensional molecule represented as a string of alphabets

acgtcgtagttccagtc

- a perfect representation for computer analysis

Data Representation

For all computational purposes, a DNA sequence is considered to be a string on a 4-letter alphabet: A, T, G, C

ACGCTGAATAGC

Aim: to find grammar & syntax rules of DNA language based on this 4-letter alphabet

- similar to English Grammar to form meaningful sentences

Similarly, RNA sequence is represented as a string of 4 alphabets and protein sequence a string of 20 alphabets

Biological Sequence Analysis

Pattern Recognition:

Assumption in biological sequence analysis:

- strings carrying information will be different from random strings

If a hidden pattern can be identified in a string, it must be carrying some functional information

Biological Sequence Analysis

Order of occurrence of bases:

not completely <u>random</u>

- Different regions of the genome exhibit different patterns of the four bases, A, T, G, C

e.g., protein coding regions, regulatory regions, intron/exon boundaries, repeat regions, etc.

Aim: Identify various patterns to infer their functional roles

Example

❄♒♓♦ ♓♦ ♋ ●♏♍♦◆❒♏ □■
♘♓□♓■♐□❒○♋♦♓♍♦

♋♦♎♎ ●&♎♐● ♎♎♎♎ ♦♎♐♦♢♏
■❖♍❒□♦ ■⌘♍♎♎♒♦□◆

This is a lecture on bioinformatics

asjd lkjfl jdjd sjftye nvcrow nzcdjhspu

Frequency of letters

A.	7.3%	N.	7.8%
B.	0.9%	O.	7.4%
C.	3.0%	P.	2.7%
D.	4.4%	Q.	0.3%
E.	13.0%	R.	7.7%
F.	2.8%	S.	6.3%
G.	1.6%	T.	9.3%
H.	3.5%	U.	2.7%
I.	7.4%	V.	1.3%
J.	0.2%	W.	1.6%
K.	0.3%	X.	0.5%
L.	3.5%	Y.	1.9%
M.	2.5%	Z.	0.1%

Other statistics

Frequencies of the most common first letter of a word, last letter of a word, doublets, triplets, etc.

20 most used words in written English

- **the of to in and a for was is that on at he with by be it an as his**

20 most used words in spoken English

- **the and I to of a you that in it is yes was this but on well he have for**

Parallels in DNA language

ATGGTGGTCATGGCGCCCCGAACCCTCTTCCTGCTG
CTCTCGGGGGCCCTGACCCTGACCGAGACCTGGGCG
GGTGAGTGCGGGGTCAGGAGGGAAACAGCCCCTGC
GCGGAGGAGGGAGGGGCCGGCCCGGCGGG

GTCTCAACCCCTCCTCGCCCCCAGGCTCCCACTCCA
TGAGGTATTTCAGCGCCGCCGTGTCCCGGCCCGGCC
GCGGGGAGCCCCGCTTCATCGCCATGGGCTACGTGG
ACGACACGCAGTTCGTGCGGTTC

Parallels in DNA language

ATG GTG GTC ATG GCG CCC CGA ACC CTC TTC CTG CTG CTC TCG GGG GCC CTG ACC CTG ACC GAG ACC TGG GCG GGT GAG TGC GGG GTC AGG AGG GAA ACA GCC CCT GCG CGG AGG AGG GAG GGG CCG GCC CGG CGG…

GTC TCA ACC CCT CCT CGC CCC CAG GCT CCC ACT CCA TGA GGT ATT TCA GCG CCG CCG TGT CCC GGC CCG GCC GCG GGG AGC CCC GCT TCA TCG CCA TGG GCT ACG TGG ACG ACA CGC AGT TCG TGC GGT TC…

1st exon and 1st intron of Human HLA gene

This task needs to be automated because of the large genome sizes:

Smallest genome:

Mycoplasma genitalium 0.5×10^6 **bp**

Human genome: 3×10^9 **bp – not the largest!**

~ 10-100 times the Britannica Encyclopedia

Plant genomes are even larger.

DNA Sequence Analysis

- **Evolution has operated on every sequence that we see today - genes and sequences involved in gene regulation are conserved.**
- **these are transferred, like code modules, from one organism to another.**
- **Because of evolution, similar sequences have similar functions.**
- **Algorithms for comparing sequences and finding similar regions are at the heart of computational biology.**

Gene Technology
&
DNA Sequencing

Lecture-3

For all computational purposes, DNA is represented as a string of 4-letter alphabets - A, T, C, G:

attgctacgttacatcgctgca

How do we get this string representation from a dynamic double-stranded molecule?

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA

To carry out this task we need to be able to chop the DNA, store it, make copies of it.

To sequence a gene, we need to

- **Identify the region of interest**
- **Isolate it from the organism – DNA fragmentation**
- **move it to another easily manageable organism such as a bacterium for obtaining multiple copies – cloning**

Such manipulations are conducted by a toolkit of enzymes:

Restriction endonucleases - used as molecular scissors

DNA ligase - to bond pieces of DNA together

- a variety of additional enzymes that modify DNA are used to facilitate the process.

Restriction endonucleases are enzymes that make **site-specific** cuts in the DNA – chemical scissors

Ability to cut DNA into discrete fragments allows to understand

- how genetic material of an organism is **organized**
- how expression of genetic information is **controlled**
- how **alteration** of genetic information can give rise to genetically inherited disorders, etc.
- in **bulk production** of pharmaceutically important proteins

First restriction enzyme was isolated from H. influenzae in 1970 by Daniel Nathans and Kathleen Danna

- awarded the Nobel Prize for Medicine in 1978

Radioautogram of ^{14}C-labeled SV40 DNA cleaved with endonuclease R

SV40 DNA (a tumor virus) - after cutting, or "digesting" it with *H. influenzae* restriction enzyme, analyzed the pieces using a polyacrylamide gel electrophoresis.

- 11 distinct DNA bands were visible in the gel, indicating that the enzyme always cut SV40 DNA resulting in the same 11 pieces

Background

How were these restriction endonucleases identified?

Bacteria are under constant attack by viruses, e.g., bacteriophages

To protect themselves, bacteria have developed a method to chop up any foreign DNA, by an enzyme, called endonuclease

- it circulates in the bacterial cytoplasm, waiting for any attacking virus.

- also called restriction enzymes because they restrict the infection of bacteriophages.

Why do the restriction enzymes not chew up the genomic DNA of their host?

Background

A bacterium that makes a particular restriction endonuclease, also synthesizes a companion DNA methyltransferase,

- which methylates the DNA target sequence for that restriction enzyme, thereby protecting it from cleavage.

DNA from an attacking bacteriophage will not have these protective methyl groups and will be destroyed.

Methyl groups (attached to the cytosine in dinucleotide CG) block the binding of restriction enzymes, but do not block the normal reading and replication of the genomic information stored in the host DNA.

DNA Fragmentation

Different endonucleases present in different bacteria recognize different nucleotide sequences

Naming of restriction enzymes - after their host of origin, e.g.,

- **EcoRI -** ***Escherichia coli***
- **Hind II & Hind III -** ***Haemophilus influenzae***
- **XhoI -** ***Xanthomonas holcicola***

When cut with a restriction enzyme (RE), the ends of the cut DNA fragment can be cohesive or blunt-ended depending on the enzyme.

Generation of Cohesive & Blunt-ended Fragments

Cutting with Eco R I

5'... G ↓ AATTC... 3'
3'... CTTAA ↑ G ... 5'

5'... G AATTC....3'
3'... CTTAA G... 5'

Cohesive or "Sticky" Ends

Cutting with Pst I

5'... CTGCA ↓ G... 3'
...G ↑ ACGTC...'

5'... CTGCA G... 3'
3'... G ACGTC... 5'

Cohesive or "Sticky" Ends

(a)

Cutting with Sma I

↓

5'... CCC GGG... 3'
3'... GGG CCC... 5'

5'... CCC GGG... 3'
3'... GGG CCC... 5'

Blunt Ends

Restriction enzyme digestion of genomic DNA and plasmid vector DNA

When cut with same RE (*e.g., EcoRI*),
- overlapping ends of the vector and foreign DNA are compatible and complementary

A plasmid is a small, circular, double-stranded DNA molecule that is distinct from a cell's chromosomal DNA.

Features of Restriction Enzymes

- **Length of recognition sequence dictates how frequently the enzyme will cut a DNA sequence**

  Frequency of recognition sites of length, 4, 6, or 8?

- **Different REs can have the same recognition site and are called isoschizomers, e.g., *SacI* & *SstI* : GAGCTC**
- **Restriction recognitions sites can be unambiguous, e.g., *Hinf* I recognition site: GANTC – it's frequency of occurrence?**
- **Most recognition sequences are palindromes - they read the same forward and backward**

Can we use the property of palindrome sequence to identify restriction recognition sites?

Applications of Restriction Enzymes

- **To prepare a physical map of the genome**
- **In genetic engineering - to assemble customized genomes; create designer bacteria that make insulin, or growth hormones, or add genes for disease resistance to agricultural plants, etc.**
- **in DNA sequencing**

Restriction Map

Restriction map is a description of restriction endonuclease cleavage sites within a piece of DNA

- generating such a map is the first step in **characterizing** an unknown DNA

Multiple Complete Digest Mapping – creates a map by digesting DNA with multiple REs

- each recognizing a different specific short DNA sequence and producing a separate **fingerprint** for each clone

DNA to be restriction mapped is usually contained within a well-characterized plasmid or bacteriophage vector for which the sequence is known, which facilitates making the map.

Restriction Mapping

Ex: Consider a plasmid that contains a 3000 bp fragment of unknown DNA & unique recognition sites for enzymes Kpn I & BamH I.

Consider first separate digestions with Kpn I & BamH I:

Kpn I yields 2 fragments: 1000bp & "big"

BamH I yields 3 fragments: 600, 2200 & "big"

big – part of unknown DNA sequence + vector

⇒ one Kpn I site & two BamH I sites are present in the unknown DNA sequence

Restriction Mapping

One BamH I site is at 2800 bp. Trick to determine the location of 2nd BamH I site is to digest the plasmid with Kpn I & BamH I together

Restriction Mapping

Double digest yields fragments of 600, 1000 & 1200 bp (plus the "big" fragment).

Restriction Mapping

If the above process is conducted with a larger set of enzymes, a much more complete map would result

single digests - are used to determine which fragments are in the unknown DNA, and

multiple digests - to order and orient the fragments correctly.

Can we computationally generate a restriction map of DNA sequence?

Restriction Mapping

Using a Computer to Generate Restriction Maps

If the sequence is known, feed it to computer programs, which will search the sequence for various RE recognition sites and build a map.

- **Mapper - available as part of Molecular Toolkit**
  http://arbl.cvmbs.colostate.edu/molkit/mapper/
- **Webcutter**
  http://www.firstmarket.com/cutter/cut2.html
- **RebSite – as part of the REBASE Tools**
  http://tools.neb.com/REBsites/index.php3

REBASE - The Restriction Enzyme dataBASE

Back to...

REBsites

Program Guide

This tool will take a DNA sequence and digest it with one example of each of the known Type 2 restriction enzyme specificities.
The maximum size of the input file is 2 MByte, and the maximum sequence length is 200 KBases.

Local sequence file: Browse...

GenBank number: (*Browse GenBank*)

Name of sequence: (*optional*)

or Paste in your DNA sequence: (*plain or FASTA format*)

Standard sequences:
Lambda
pBR322
PhiX174
Ad2

Submit

The sequence is: Linear / Circular

Input sites: All specificities / Defined oligonucleotide sequences:

Clear the table below

Name	Oligonucleotide sequence

theoretical digest with all REBASE prototypes

REBsites

pBR322

Gel type: 0.7% agarose
Order by: enzyme name

zyme name for a list of fragments.
gel was generated by interpolating experimental data. See details.

Print

Fragment list

Close

pBR322 digested with AceIII

[Sites with flanks]

#	Coordinates	Length (bp)
1	3666-697	1393
2	698-1984	1287
3	2426-3665	1240
4	2126-2425	300
5	1985-2125	141

Type II Restriction Enzymes

Enzymes	Recognition Sequence	Isoschizomers	Suppliers
AaaI	C↓GGCCG	yes	-
AacI	GGATCC	yes	-
AaeI	GGATCC	yes	-
AagI	AT↓CGAT	yes	-
AamI	-	-	-
AaqI	GTGCAC	yes	-
AarI	CACCTGC (4/8)	-	y
AasI	GACNNNN↓NNGTC	yes	y
AatI	AGG↓CCT	yes	y
AatII	GACGT↓C	yes	y
AauI	T↓GTACA	yes	-
AbaI	T↓GATCA	yes	-
AbeI	CCTCAGC (-5/-2)	yes	-
AboORF2079P	AGGCCT	yes	-
AbrI	C↓TCGAG	yes	-
AcaI	TTCGAA	yes	-
AcaII	GGATCC	yes	-
AcaIII	TGCGCA	yes	-
AcaIV	GGCC	yes	-
AccI	GT↓MKAC	yes	y
AccII	CG↓CG	yes	y

AaaI

Type II restriction enzyme
subtype: P

Recognition Sequence:

C^GGCCG

5' .. C G G C C G 3' ..
3' .. G C C G G C 5' ..

REBASE Enz Num 1 entered Jan 1 1987 ... modified May 24 2004

Prototype: XmaIII
Org #: 1
Organism: *Acetobacter aceti ss aceti*
Organism source: M. Fukaya
Growth Temperature: 26 °

 # sites on
Adeno2: 19
Lambda: 2
pBR322: 1
PhiX174: 0
SV40: 0

Site frequency in sequenced genomes...

Related References...
sorted by date in new win
sorted by authors in new w
NOT commercially availa
Similar enzymes...

Assignment

- **Write a program to generate a restriction map for a specific RE and compare your results with Mapper.**
- **Write a program to identify restriction recognition sites in a given DNA sequence.**
- **Write to program to obtain the reverse strand in the forward strand of a DNA sequence in given**

Cloning

What is cloning?

The process of cloning involves the production of multiple copies of a DNA fragment of interest by amplification *in vivo*

- depends upon the ability of vectors to continue their life cycles in bacterial or yeast cells in spite of having foreign DNA inserted into them.

Cloning vector - a DNA molecule that carries foreign DNA into a host cell, replicates inside a bacterial (or yeast) cell and produces many copies of itself and the foreign DNA

Types of Vectors

Vector	Insert size (kb)
Plasmids	<10 kb
Bacteriophage	9 - 20 kb
Cosmids	33 - 47 kb
Bacterial artificial chromosomes (BACs)	75 - 125 kb
Yeast artificial chromosomes (YACS)	100-1000 kb

Types of Vectors

Plasmids - an extra-chromosomal double-stranded circular DNA molecules that replicates autonomously inside the bacterial cell

Plasmids are important as one can:

(i) isolate them in large quantities,

(ii) cut & splice them, add DNA of choice,

(iii) put them back into bacteria, where they replicate along with the bacteria's own DNA,

(iv) isolate them again to get billions of copies of inserted DNA

Limitation: size of DNA that can be introduced into the cell by transformation (~ 2 - 10kb)

Cloning into a plasmid

YAC - a functional self-replicating artificial chromosome. It includes <u>three</u> specific DNA sequences that enable it to propagate from one cell to its offspring:

- **TEL: The telomere which is located at each chromosome end, protects the linear DNA from degradation by nucleases**
- **CEN: The centromere which is the attachment site for mitotic spindle fibers, "pulls" one copy of each duplicated chromosome into each new daughter cell.**
- **ORI: Replication origin sequences, specific DNA sequences that allow the DNA replication machinery to assemble on the DNA and move at the replication forks**

It also contains few other specific sequences like:

- **A and B: selectable markers that allow easy isolation of yeast cells that have taken up the artificial chromosome.**
- **Recognition site for two REs: EcoRI & BamHI**

Cloning into a Yeast Artificial Chromosome (YAC

Why is it important to be able to clone large sequences?

To map the entire human genome ($3x10^9$ bps) would require more than 1000,000 plasmid clones (~10Kb limit).

In principle, the human genome could be represented in about 10,000 YAC clones (~1Mb limit)

DNA Sequencing

DNA Sequencing - determine the precise sequence of nucleotides in a sample of DNA – the order of A, T, G, C

Various types of sequencing:

- **Sequencing a region of interest, e.g., gene.**
- **Whole Genome/Exome Sequencing**
- **cDNA Sequencing – sequencing cDNA libraries of the expressed genes**
- **High-throughput sequencing – next-generation, 3rd & 4th generation sequencing - whole Genome/Exome/targetted**
- **Metagenome sequencing - sequencing of environmental samples**

- depending on the nature of analysis, type of sample, or type of sequencer used

Sequencing a Region of Interest

First requirement in sequencing a region of DNA is

- to have enough starting template for sequencing.

This is achieved by PCR - Polymerase Chain Reaction

- carried out in an automated cycler for 30 - 40 cycles.

Essential requirements for a PCR:

- **a mixture of 4 deoxy-nucleotides in ample quantities**
- **dATP, dGTP, dCTP, dTTP**
- **Taq DNA polymerase**
- **Primers ?**
- **Genomic DNA of interest**

What is the advantage of using PCR over traditional gene cloning?

Region of DNA to be amplified by PCR

to achieve geometric amplification of a DNA fragment, there must be two primers

Primers - short single-stranded oligonucleotides which anneal to the DNA template and serve as a starting point for DNA synthesis

Why are primers required?

The Cycling Reactions

Step-1: Denaturation at 94°C

- opens up double stranded DNA, all enzymatic reactions stop.

Step-2: Annealing at 54°C

- Primers jiggling around because of Brownian motion, binds to single stranded template once an exact match is found; the polymerase then attaches and start copying the template.

Step-3: Extension at 72°C

- ideal working temperature for the polymerase. Bases complementary to the template are coupled to the primer on 3' side (reading the template from 3' to 5' side)

Different Steps in PCR

PCR : Polymerase Chain Reaction

30 - 40 cycles of 3 steps :

Step 1 : denaturation

1 minut 94 °C

Step 2 : annealing

45 seconds 54 °C

forward and reverse primers !!!

Step 3 : extension

2 minutes 72 °C

only dNTP's

Different Steps in PCR

Exponential amplification of region of interest

<u>Both strands</u> are copied during PCR

- leading to an exponential increase of the number of copies of the region of interest.

Verification of PCR Product

Is the template copied during PCR and is it the right size?

Before the PCR product is used in further applications, it has to be checked if:

1. **A product is formed**
2. **The product is of the right size**
3. **Only one band is formed**

First 4 cycles of a PCR reaction

Verification of the PCR product

PCR Sequencing

For sequencing, we don't start from gDNA (like in PCR) but mostly from PCR fragments or cloned genes.

Amplified PCR product is supplied with

- **a mixture of all four <u>normal</u> (deoxy) nucleotides in ample quantities**
  - **dATP**
  - **dGTP**
  - **dCTP**
  - **dTTP**

- ***Taq*** **DNA polymerase**

PCR Sequencing

- a mixture of all four <u>dideoxynucleotides</u>, each present in limiting quantities and each labeled with a "tag" that fluoresces a different color:
  - ddATP
  - ddGTP
  - ddCTP
  - ddTTP

This method of DNA sequencing is called dideoxy method, or chain termination method, or Sanger's method.

PCR Sequencing

Dideoxy method: DNA is synthesized from four deoxynucleotide triphosphates.

Each new nucleotide is added to 3′ -OH group of the last nucleotide added.

When a dideoxynucleotide, **ddNTP is added** to the growing DNA strand, **chain elongation stops** because there is no 3′-OH for the next nucleotide to be attached to.

Steps in PCR Sequencing

I The sequencing reaction

- **Denaturation at 94°C**
- **Annealing at 50°C**
- **Extension at 60°C** ← **instead of 72°C**

II Separation of the fragments

III Detection on an automated sequencer

IV Assembling the sequenced parts

Different steps in Sequencing

Different steps in Sequencing

PCR Sequencing

Since only one primer is used, only <u>one strand</u> is copied during sequencing – results in linear increase of the No. of copies.

⇒ large amount of DNA in the starting mixture is required.

PCR Sequencing

II Separation of the molecules:

After the sequencing reactions, the mixture of strands of different lengths, all ending on a fluorescently labeled ddNTP are loaded on an acrylamide gel for separation

- gel electrophoresis.

During electrophoresis, a voltage is created across the gel making one end positive and the other negative.

DNA being –vely charged, migrates to the positive side;

- strands of different length migrate at different rates and thus are separated based on their size - the smallest strand travels the fastest.

Separation of molecules with electrophoresis

Very good resolution - a difference of even one nucleotide is enough to separate a strand from the next shorter or longer strand.

Four dideoxynucleotides fluoresces a different color when illuminated by a laser beam and an automatic scanner provides a printout of the sequence.

Separation of Molecules with Electrophoresis

Separation of the Molecules with Electrophoresis

PCR Sequencing

III Detection on an automated sequencer:

Fluorescently labeled fragments that migrate through the gel pass a laser beam at the bottom of the gel.

Scanning & Detection System on a Sequencer

PCR Sequencing

Plot of the colors detected in a 'lane' of the gel (one sample), scanned from smallest fragments to largest.

The computer interprets the colors by printing the nucleotide sequence across the top of the plot.

PCR Sequencing

IV Assembling the sequenced parts of a gene:

For publication, a gene sequence has to be confirmed in both directions using forward & reverse primers

Since it is only possible to sequence ~ 700-800 bases in one run, a gene of, say, 1800 bases, is sequenced with internal primers.

- the sequenced fragments are assembled using a computer program to obtain complete gene sequence.

Genome Sequencing

Genome Sequencing

By Sanger's method, we can sequence a fragment of DNA ~ 1000bp long.

But what about longer pieces?

Human genome is 3 billion bases long, arranged on 23 pairs of chromosomes.

Sequencing machine reads just a drop in the ocean!

Genome Sequencing

Solution: Break the entire genome into <u>manageable</u> pieces and sequence them.

Two approaches used for sequencing Human genome:

- **Publicly funded Human Genome Project (HGP) – clone-by-clone or hierarchical shotgun sequencing method**
- **Privately Funded Sequencing Project - Celera Genomics – whole genome shotgun sequencing method**

Genome Sequencing

Hierarchical shotgun sequencing approach:

- **genomic DNA is cut into pieces of about 150 Mb**
- **inserted into BAC cloning vectors,**
- **transformed into *E. coli* where they are replicated and stored.**

BAC inserts are isolated & mapped to determine the order of each cloned 150 Mb fragment

- **referred to as the Golden Tiling Path**

Begun formally in 1990, Human Genome Project was a 13-yr effort coordinated by the U.S. DAE and NIH.
- completed in 2003

Genome Sequencing

Each BAC fragment in the Golden Path is

- fragmented randomly into smaller pieces,

- each piece is cloned into a plasmid and sequenced on both strands.

These sequences are aligned so that identical regions overlap.

Contiguous pieces are then assembled into finished sequence once each strand has been sequenced about 5 times to produce 10× coverage of high-quality data.

Genome Sequencing

Whole genome shotgun sequencing (WGS)

- method developed and preferred by Celera Genomics

- skips the entire step of making libraries of BAC clones

Blast apart the entire human genome into fragments of 2 - 10 kb and sequence them.

Challenge is then to assemble these fragments into the whole genome sequence.

Human Genome Sequencing

Whole Genome Shotgun Method

What makes the task of assembling the genome fragments especially challenging

- repeats in the genome (~ 50% in human genome).

Because of the various ways a fragment could align with a repeat, and the different areas adjacent to the repeats in the original genome, assemblers need to be designed so as not to incorrectly join fragments

Whole Genome Shotgun Method

Adding to the challenge is the sheer computational complexity of the task.

Size of H. genome = 3 x 10^9 bp. Given length of read ~500 bps, for desired coverage of 10x, No. of reads required is:

RequiredReads = GenomeLength * DesiredCoverage / ReadLength

= 6 * 10^7

With 60M reads to assemble, we need algorithms that run in near linear time ($O(nlogn)$)

Whole Genome Shotgun Method

Which method is better?

Depends on the size and complexity of the genome

<u>Note</u>: Celera had access to the HGP data but the HGP did not have access to Celera data.

Which method is preferable for sequencing the genome of a novel coronavirus – SAR-CoV-2? Why?

High-throughput / Next-Generation Sequencing

13yrs, $3 billion **S.S.** **8days, $10,000** **15min, <$1,000**

DNA sequencing beating Moore's law

HTS/NGS Sequencing

High-throughput sequencing (HTS) technologies have revolutionized the way biologists acquire and analyze genomic data.

- massively parallel sequencing

	Roche GS FLX+	Illumina HiSeq 2000	SOLiD™ 4	Ion Torrent PGM
Bases per run	700Mb	600 Gb	100 GB	1 Gb
Time per run	23h	~11 days	~14 days	4.5 h
Reads per run	1 Million	6 Billion (paired-end) 3 Billion (single)	1.4 Billion	Millions
Read length	~700 bp	2 x 100 bases	2 x 50 bases	35–400 bases

- can generate tens of gigabases per week, at a cost 200-fold <u>less</u> than previous methods.

Sequencing Machines: Overview

	Roche GS FLX+	Illumina HiSeq 2000	SOLiD™ 4	Ion Torrent PGM
Bases per run	700Mb	600 Gb	100 GB	1 Gb
Time per run	23h	~11 days	~14 days	4.5 h
Reads per run	1 Million	6 Billion (paired-end) 3 Billion (single)	1.4 Billion	Millions
Read length	~700 bp	2 x 100 bases	2 x 50 bases	35–400 bases

Sequencing Machines: Overview

1. Pyrosequencing

Roche GS-FLX

3. Sequence by ligation

Life Technologies SOLiD

2. Sequence by Synthesis

Illumina HiSeq

4. Proton Detection

Life Technologies Ion Torrent

Basic workflow: Template Generation

Sequence library – convert starting material into a library of sequencing reaction templates.

Require common steps:

- **Fragmentation**
- **Size selection**
- **Adapter ligation**

by attachment to solid surfaces or beads

Amplification-based - "second-generation" sequencing technology

Single-molecule - "third-generation" sequencing technology

A library is either sequenced directly - Single-molecule templates
or, amplified then sequenced - Clonally amplified templates

Basic workflow: Detection & Data Analysis

Data Analysis

The scale and nature of data produced by all NGS platforms place substantial demands on IT at all stages of sequencing, including data tracking, storage, and quality control:

- **base calling - by proprietary software**
- **Quality check and filtering or reads**
- **aligning sequencing data to Reference genome. if available, or a *de novo* assembly is conducted.**

Once the sequence is aligned to a reference genome, the data needs to be analyzed in an experiment-specific fashion.

Sequence alignment & assembly is an active area of computational research

Third Generation Sequencing (TGS)

- **'Long read sequencing' – read length: ~ 10 – 60Kb**
- **Single molecule sequencing**
- **No PCR step involved**
- **Faster and portable**
- **Under active development**
- **e.g., PacBio Single molecule real time sequencing (SMRT) and Oxford Nanopore**

Oxford Nanopore - MinION

HTS Applications

One of the most prominent applications of NGS is re-sequencing:

- **whole genome resequencing**
- **target-region resequencing**
- **exome resequencing**

Any human individual's genome available in NCBI?

- **genome-wide analysis of single nucleotide variations and other structural variations, multiple individuals, or strains, cancer sequencing, population-based sampling of a species, migration patterns of a virus, e.g., SARS-CoV-2, etc.**

PCR Sequencing

How would you go about sequencing SARS-CoV-2 genome, 29903 bases long?

What technique is used for diagnostic testing of COVID-19?

While sequencing a novel genome for the first time, how are primers identified?

PCR Sequencing

Real time RT-PCR used for diagnostic testing of COVID-19

It is a laboratory technique combining reverse transcription of RNA into DNA (called complementary DNA or cDNA) and amplification of specific DNA targets using polymerase chain reaction (PCR).

It is primarily used to measure the amount of a specific RNA.

This is achieved by monitoring the amplification reaction using fluorescence, a technique called real-time PCR or quantitative PCR (qPCR).

- routinely used for analysis of gene expression and quantification of viral RNA in research and clinical settings.

Can we now answer these Qs:

- **How is the SARS-CoV-2 genome sequenced?**
- **How does one identify the coordinates of N gene on it? i.e., how to construct a physical map of a genome?**
- **How does one select which regions in this gene would give specificity for the presence of SARS-CoV-2?***
- **How is the specific probe regions extracted and amplified for detection?**
- **Is it possible to store the DNA sample for re-testing? How?**

References:

1. **Concepts in Biotechnology, ed. D. Balasubramanyam**
2. **Restriction Endonucleases and DNA Modifying Enzymes http://arbl.cvmbs.colostate.edu/hbooks/genetics/biotech/enzymes/index.html**
3. **REBASE: restriction enzymes and methyltransferases, Nucleic Acids Research, Vol. 31 (1), 418–420 (2003)**

Computational Genome Analysis

Lecture-4

Consider a DNA sequence:

ATGGTGGTCATGGCGCCCCGAACCCTCTTCCT
GCTGCTCTCGGGGGCCCTGACCCTGACCGAG
ACCTGGGCGGGTGAGTGCGGGGTCAGGAGGG
AAACAGCCCCTGCGCGGAGGAGGGAGGGGCC
GGCCCGGCGGGGTCTCAACCCCTCCTCGCCC
CCAGGCTCCCACTCCATGAGGTATTTCAGCGC
CGCCGTGTCCCGGCCCGGCCGCGGGGAGCCC
CGCTTCATCGCCATGGGCTACGTGGACGACAC
GCAGTTCGTGCGGTTC

Given a DNA sequence, there are a number of questions one might ask:

- **What sort of statistics should be used to describe this sequence?**
- **What sort of organism this sequence came from based on sequence content?**
- **Do parameters describing this sequence differ from those describing bulk DNA in that organism?**
- **What sort of sequence it might be: Protein coding? Centromere? Telomere? Transposable element? Control sequence?**

Let's attempt to address such Qs. by considering words:

- short strings of letters drawn from DNA alphabet, A, C, G, T.

What sort of information can we obtain from "*k*-tuple", or, *k*-mer analysis?

k = 1 (Base Composition)

***k*=1: Frequency of bases.**

DNA being a duplex,

No. of A's = No. of T's

No. of G's = No. of C's

What about on the same strand?

5'- GGATCGAAGCTAAGGGCT - 3'

3'- CCTAGCTTCGATTCCCGA - 5'

Top/Bottom Strand: No. of G = ?, C = ?

For the duplex molecule, G = ?, C = ?

$$N_A = N_T, \quad N_G = N_C, \quad N_A + N_G = N_T + N_C$$

- only for duplex molecule, not for a single strand.

Biological Words

For the duplex molecule,

$$fr(A+T) = 1 - fr(G+C)$$

$\Rightarrow$ only a <u>single parameter</u>, say fr(G+C), suffices in describing the base frequencies for duplex DNA.

i.e., there are 4 variables and 3 relations:

$$fr(A) = fr(T), \quad fr(G) = fr(C), \quad fr(A+T) = 1 - fr(G+C)$$

Base composition has been used as a descriptive statistic for genomes of various organisms since the early days of molecular biology.

Biological Words

If a genome is GC-rich, say, 60% GC, then find the fractions of A, T, G, and C.

The GC bond is stronger than the AT bond

- **higher temperatures are required for denaturation (opening the strands) if the genome is GC-rich**
- **Melting point of GC-rich genome is higher than a AT-rich genome**

So, what can you say about a bacterial species found in naturally occurring hot springs?

Base Compositions of Various Organisms

Organism	% G+C	Genome size (Mb)
Eubacteria		
Mycoplasma genitalium	31.6	0.585
Escherichia coli K-12	50.7	4.693
Pseudomonas aeruginosa PAO1	66.4	6.264
Archaebacteria		
Pyrococcus abyssi	44.6	1.765
Thermoplasma volcanium	39.9	1.585
Eukaryotes		
Caenorhabditis elegans (a nematode)	36	97
Arabidopsis thaliana (a flowering plant)	35	125
Homo sapiens (a bipedal tetrapod)	41	3080

Variation in GC content in different organisms – due to variation in selection, mutational bias & biased recombination-associated DNA repair

Biological Words

Distribution of individual bases within DNA molecule

- not uniform.

In prokaryotic genomes in particular, an excess of G over C is observed on the <u>leading</u> strand.

(strands whose 5' to 3' direction corresponds to the direction of replication fork movement).

This is described by "GC skew":

$$\text{GC skew} = (\#G - \#C)/ (\#G + \#C),$$

calculated in sliding "windows".

- It <u>changes sign</u> at positions of <u>replication origins</u> and termini in prokaryotic genomic sequences.

- another example of how a relatively simple statistic based on 1-tuple can be informative.

GC Skew

Third codon position GC skew (GCS), window size – 300Kb, step size – 10Kb.

Direction of skew switches at the genome's origin & terminus of replication, such that the leading strand in replication is always richer in G than C.

Biological Words

GC content is <u>not</u> uniformly constant throughout the genome - but is found to be markedly variable

- resulting in a mosaic-like formation with islet regions called isochores (GC-rich regions).

GC-rich isochores include many protein coding genes

- About 50% of humans are GC-rich

- another example of how a simple statistic based on *k* = 1 can be informative.

An isochore is a large region of DNA (> 300 KB) with a high degree of uniformity in G & C

k = 1 Analysis

To summarize:

- **G+C fraction gives the base composition of a genome**
- **Melting temperature of DNA sequence can be known**
- **GC skew on the leading strand can help in identifying the origin of replication in prokaryotes**
- **GC-rich regions – help in identifying protein-coding genes**
- **Since GC content of an organism is different, it is a useful measure in identifying horizontally transferred regions**

An interesting problem in genome analysis is in knowing whether certain patterns occur more often than by random chance in a given sequence;

- such sequences might be of biological significance

This can be addressed by using a <u>probabilistic model</u> of DNA sequence to identify if a pattern is over/under-represented in the genomic sequence compared to its expected frequency based on the iid model.

e.g., frequency of triplets is different in protein-coding regions compared to the genomic sequence.

<u>Probabilistic model</u>: a method for simulating observations from the model,

i.e., specify probabilistic rules to produce next letter in the simulated sequence, given the previous letters:

- **First base in the sequence is either an A, C, G, or T with probability p_A, p_C, p_G, p_T, respectively.**
- **Suppose the algorithm has generated r bases. To generate the base at position $r+1$, the algorithm pays no attention to what has been generated before and gives out A, C, G, or T with probabilities p_A, p_C, p_G, p_T.**

- similar to coin tossing experiment.

χ = {A, C, G, or T} for DNA seq

Output of simulation at every step, *X* takes any one character from set χ with probability p_A, p_C, p_G, p_T

$$p_A, p_C, p_G, p_T \geq 0, \qquad p_A + p_C + p_G + p_T = 1$$

p_A, p_C, p_G, p_T is the <u>probability distribution</u> of *X*.

i.e., first base L_1 in our model for a DNA sequence has probability distribution

$$P(L_1 = A) = p_A, \quad P(L_1 = C) = p_C,$$

$$P(L_1 = G) = p_G, \quad P(L_1 = T) = p_T. \qquad (1)$$

To study probability distribution of no. of times a given pattern occurs in a random DNA sequence $L_1, L_2, \ldots, L_n$,

Consider patterns of length one, $k = 1$. To address this question, define a new sequence $X_1, X_2, \ldots, X_n$ by

$$X_i = \begin{cases} 1, & \text{if} \quad L_i = \text{A} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

No. of times that A appears is:

$$N = X_1 + X_2 + \ldots + X_n \qquad (3)$$

What is the probability of getting runs of A's of certain length?

Is a homopolymer A region of any significance?

Polyadenylation signal – the site at which cleavage of 3'end of mature RNA takes place: AAUAAA

Typical structure of a mature eukaryotic mRNA

PolyA tail is important for nuclear export, translation and stability of mRNA

Alternative polyadenylation

Results of using different polyadenylation sites on the same gene

Starting from the initial probability distribution of L_i, the probability distribution of each X_i is:

$$P(X_i = 1) = P(L_i = A) = p_A,$$

$$P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T)$$

$$= p_C + p_G + p_T = 1 - p_A \qquad (4)$$

Different "runs" of the simulation would produce strings having different values of *N*, i.e., no.'s of A's.

- like in the case of a coin-tossing expt.

Then what is a "typical" value of *N*, i.e., its probability distribution?

To find the probability distribution of N is complicated because we need to know how individual outputs from our simulation are related to each other.

- it is not easy to find whether the random variables are independent,

Assume <u>independence</u> and then use the multiplicative rule to calculate the probabilities of different outcomes.

If the probability of a given pattern, assuming independence is similar to that observed, then the pattern is a <u>random independent</u> pattern, else the pattern is <u>unique</u>

Is this likely to be true in a DNA sequence?

Multiplicative Rule for Independent Events:

If events E & F are independent, probability of the event "E and F":

$$p(E \cap F) = p(E) \,.\, p(F)$$

For a DNA sequence model, L_i are independent, so probability of obtaining the sequence $l_1, l_2, \ldots, l_n$ is

$$\mathbb{P}\left(L_1 = l_1,\ L_2 = l_2,\ \ldots, L_n = l_n\right) = \mathbb{P}\left(L_1 = l_1\right)\mathbb{P}\left(L_2 = l_2\right)\mathbb{P}\left(L_n = l_n\right)$$

For S = ATTGCGTGAG, what is the probability of observing the pattern, P(ATTG) ?

$$P(ATTG) = p_A p_T p_T p_G$$

A distribution is defined by its <u>mean</u> & <u>variance</u>

Measures of location (central tendency) & spread for a sequence $X_1, X_2, \ldots, X_n$ given by

1100101000110110101000111011010

with probability distribution

$$P(X_i = 1) = p_A, \qquad X_i = \begin{cases} 1, & \text{if } L_i = \text{A} \\ 0, & \text{otherwise} \end{cases}$$

$$P(X_i = 0) = p_C + p_G + p_T = 1 - p_A$$

Define the <u>expected</u> value (or mean of expectation) of X

- to see if X is closely concentrated about its expected value, or is spread out, compute <u>variance</u>.

Mathematical Expectation

If X assumes discrete values $X_1, X_2, \ldots, X_k$ with respective probabilities $p_1, p_2, \ldots, p_k$, the expectation of X is defined as

$$E(X) = p_1X_1 + p_2X_2 + \cdots + p_kX_k = \sum_{i=1}^{k} p_iX_i = \sum pX$$

If probabilities p_i are replaced by relative frequencies

$$f_i/N, \quad N = \sum f_i$$

For large N

the expectation reduces to $\left(\sum fX\right)/N$

- arithmetic mean of a sample of size N

We interpret that $E(X)$ represents the mean of the population from which the sample is drawn.

Expected Values for Random Variables:

Using the fact that mean of the sum of X_i is the sum of the mean of the X_i:

$$\mathbb{E}\,(X_1 + X_2 + \ldots + X_n) = \mathbb{E}\,X_1 + \mathbb{E}\,X_2 + \ldots\ \mathbb{E}\,X_n$$

the expected No. of times we see an *A* in our *n* bp sequence is

$$\mathbb{E}\,N = \mathbb{E}\,X_1 + \mathbb{E}\,X_2 + \ldots\ \mathbb{E}\,X_n = n\,\mathbb{E}\,X_1 = n\,p_{\mathrm{A}}$$

where $N = X_1 + X_2 + \ldots + X_n$

– no. of times that A appears is the sum of X's, and

$$EX_i = 1 \times p_A + 0 \times (1-p_A) = p_A$$

$$X_i = \begin{cases} 1, & \text{if} \quad L_i = \mathrm{A} \\ 0, & \text{otherwise} \end{cases}$$

$P(X_i = 1) = P(L_i = A) = p_A,$

$P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T)$

$= p_C + p_G + p_T = 1 - p_A$

Expected Values for Random Variables:

Expected value of a random variable *X* gives a measure of its location - values of X tend to be scattered around this value,

$$\mu = \mathbb{E}\, X$$

So if we simulate this sequence a large no. of times, the no. of A's will take values around this mean value np_A.

The measure of spread is given by the variance.

Variance for Random Variables:

Measure of spread: Is *X* closely concentrated about its expected value, or is it spread out?

$$\mathrm{Var}\, X = \mathbb{E}\,(X - \mu)^2 = \sum_{i=1}^{J} (\boldsymbol{x}_j - \boldsymbol{\mu})^2\, \boldsymbol{p}_j$$

$$\mathrm{Var}\, X = \mathbb{E}\, X^2 - \mu^2 = \sum_{j=1}^{J} x_j^2\, p_j - \mu^2$$

For random variable X_i,

$$\mathrm{Var}\, X_i = [1^2 \text{ x } p_{\mathrm{A}} + 0^2 \text{ x } (1 - p_{\mathrm{A}})] - p_{\mathrm{A}}{}^2 = p_{\mathrm{A}}(1 - p_{\mathrm{A}})$$

+ve sq. root of variance is the standard deviation

Variance for Random Variables:

To calculate the variance of the no. *N* of A's in our simulated DNA sequences, we exploit the fact that

the variance of a sum of independent random variables is the sum of the individual variances,

$$\mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \ldots + \mathrm{Var}(X_n)$$

$$\mathrm{Var}\, N = \mathrm{n}\ \mathrm{Var}\, X_1 = n\, p_{\mathrm{A}}(1 - p_{\mathrm{A}})$$

Expected value & Variance of the no. of times letter A occurs in the given DNA sequence

- are <u>two</u> statistics that describe its probability distribution

<u>Ex</u>: Find the no. of times letter A occurs in a sequence of length n = 1000 if the probability of occurrence of A, p_A = 0.15.

Expected value = n x p_A = .15 x 1000 = 150,

Variance = n x p_A (1 - p_A) = 1000 x .15 x .85 = 127.5,

Std. deviation = 11.29

No. of times A would occur in a sequence of length 1000, if p_A = 0.15 will lie between

150 $\pm$ 11.29, i.e., 139 – 161

Applications: finding low-complexity regions, polyA tail, occurrence of GC boxes in the vicinity of a gene, etc.

<u>Ex</u>: Simulate observations having the binomial distribution with p = 0.25 and n = 1000.

Suppose that for a sequence of length n = 1000 bp and assuming that each base is equally likely, what is the probability of observing 280 A's or more in such a sequence?

There are 3 ways of doing this!

- **by using the Binomial distribution formula for the probability of observing j A's,**
- **by simulation (~ 10,000 runs),**
- **by Central Limit Theorem**

Binomial Distribution:

Since X_i are iid, the probability of the sequence that has j 1's in $x_1, x_2, \ldots, x_n$, is $p^j(1 - p)^{n-j}$

To compute the probability that the sequence contains j A's (i.e., $N = j$), we need to know

how many different realizations of the sequence $x_1, x_2, \ldots, x_n$ have j 1's (and $n - j$ 0's).

This is given by the binomial coefficient ${}^nC_j$, defined as

$$ {}^nC_j = \frac{n!}{j!(n-j)!} $$

Thus, the probability of observing j A's is

$$ \mathbb{P}(N = j) = {}^nC_j\, p^j (1-p)^{n-j}, \quad j = 0, 1, 2, \ldots, n $$

Simulating from Probability Distributions:

To understand the behaviour of random variables such as, the no. of A's (N), simulate no. of sequences and obtain the no. of A's: N_1, N_2, …, N_n, having the same probability distribution as N.

We can use them to study the properties of this distribution. e.g., the sample mean

$$\overline{N} = (N_1 + N_2 + \ldots + N_n)/n$$

can be used to estimate the expected value μ of N.

We can use the sample variance to estimate the variance σ^2 of N:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(N_i - \overline{N})^2$$

And use the histogram of the values of N_1, N_2, …, N_n to estimate the probability of different outcomes for N

Simulating from Probability Distributions:

The pseudo-random number generators can be used to produce a sequence of random nos. $U_1, U_2, \ldots, U_n$

We can thus simulate an observation with the distribution of X_1 given by

$$X_i = \begin{cases} 1, & \text{if} \quad L_1 = \mathrm{A} \\ 0, & \text{otherwise} \end{cases}$$

by taking a uniform random number U and setting $X_1 = 1$ if $U_1 \leq p \equiv p_A$ and 0 otherwise.

Repeating this procedure n times (with a new U each time) results in a sequence $X_1, X_2, \ldots, X_n$ from which N can be computed by adding up the X's.

Simulating from Probability Distributions:

We can simulate the sequence of bases $L_1, L_2, \ldots, L_n$, using random-number generator.

This is done by diving the interval (0, 1) into four intervals with endpoints at

$$p_A, \quad p_A + p_C, \quad p_A + p_C + p_G, \quad p_A + p_C + p_G + p_T = 1$$

If the random no. called lies in the leftmost interval, set L_1 = A;

if it is in the second interval, set L_1 = C;

if it is in the third interval, set L_1 = G;

otherwise set L_1 = T.

Repeating this procedure with a new series of random nos. each time to produce a no. of sequences and count the occurrence of A.

Using Binomial distribution:

$$P(N \geq 280) = \sum_{j=280 \text{ to } 1000} {}^{n}C_{j}\,(1/4)^{j}(1 - 1/4)^{n-j} = 0.016$$

By Simulation:

Simulate large no. of sequences, say, 10,000, and find out in how many cases, no. of A's is $\geq$ 280?

If 149 such instances observed, then

$$P(N \geq 280) = 149/10000 = 0.0149$$

By Central Limit Theorem:

$$\mu = Np = 250,\ \sigma = \sqrt{Npq} = 13.69,$$

$$z = (280-250)/13.69 \sim 2.1,$$

Probability corresponding to this z-score is

$$\sim 0.5 - 0.4857 = 0.0143$$

Biological Words: $k = 2$

A dinucleotide is $l_i l_{i+1}$, l_i is one of the 4 bases A, C, G, T

$\Rightarrow$ there are 16 different dinucleotides: AA, AC, AG, AT, …TG, TT; the sum of the dinucleotide frequencies is 1.

<u>Importance</u>:

- **Set of di-nucleotide relative abundance values constitutes a "genomic signature" of an organism**
- **Useful in identifying genomic islands (laterally transferred regions) devoid of coding.**

Biological Words: $k = 2$

Suppose we model the sequence $L_1, L_2, \ldots, L_n$ using iid model with base probabilities given by

$$\mathbb{P}(L_1 = \mathrm{A}) = p_{\mathrm{A}}, \quad \mathbb{P}(L_1 = \mathrm{C}) = p_{\mathrm{C}}, \quad \mathbb{P}(L_1 = \mathrm{G}) = p_{\mathrm{G}}, \quad \mathbb{P}(L_1 = \mathrm{T}) = p_{\mathrm{T}}$$

Assuming independence in the occurrence of bases, using multiplication rule, the probabilities of each dinucleotide $r_1 r_2$

$$\mathbb{P}(L_i = r_1, \; L_{i+1} = r_2) = p_{r1} p_{r2}$$

i.e., under the independence model, the chance of seeing dinucleotide AA is p_{A}^2, & chance of seeing CG is $p_{\mathrm{C}} p_{\mathrm{G}}$.

Biological Words: $k = 2$

To test if a given sequence has unusual dinucleotide frequencies compared with an iid model,

Compare the observed no. O of dinucleotide $r_1 r_2$

with the expected no. given by $E = (n - 1)\, p_{r1} p_{r2}$.

χ^2 – test:

$$\chi^2 = \frac{(O - E)^2}{E}$$

If the observed no. is close to the expected no. (the model is doing a good job of predicting the dinucleotide frequencies), χ^2 will be small.

Biological Words: *k* = 2

To determine which values of χ^2 are unlikely if in fact the model is true can be estimated:

(a)Calculate the number c given by

$$c = \begin{cases} 1 + 2p_{r1} - 3p_{r1}^2, & if \quad r_1 = r_2 \\ 1 - 3p_{r1}p_{r2}, & if \quad r_1 \neq r_2 \end{cases}$$

(b) Calculate the ratio χ^2 / c

(c) If this ratio is >3.84, the iid model is not a good fit.

If the base frequencies are <u>unknown</u>, the same approach works if these are estimated from the data.

For 1 degrees of freedom, and probability 0.05, chi-square value is 3.84.

Observed values of χ^2/c for the first 1000 bp of *E. coli* and *M. genitalium* genomes.

Dinucleotide	Observed of χ^2/c for	
	E. coli	*M. genitalium*
AA	**6.78**	0.15
AC	0.05	1.20
AG	**5.99**	0.18
AT	0.01	0.01
CA	2.64	0.01
CC	0.03	0.39
CG	0.85	**4.70**
CT	**4.70**	1.10
GA	2.15	0.34
GC	**10.04**	1.07
GG	0.01	0.09
GT	1.76	0.61
TA	**5.99**	1.93
TC	**9.06**	2.28
TG	3.63	0.05
TT	1.12	0.13

E. coli **base frequencies: (0.25, 0.25, 0.25, 0.25)**

M. genitalium **frequencies: (0.45, 0.09, 0.09, 0.37)**

E. coli **dinucl. frequencies not very well-described by the simple iid model**

Biological Words: *k* = 3

There are 61 codons that specify amino acids and three stop codons.

Since there are 20 AAs, most AAs are specified by more than one codon.

This has led to the use of a number of statistics to summarize "bias" in codon usage.

To see how these codon frequencies can vary, let's consider the specific example of the *E. coli* proteins.

Comparison of predicted & observed triplet frequencies in coding sequences for a subset of genes from *E. coli*.

	Codon	Predicted	Observed	
			Gene Class I (502)	Gene Class II (191)
Phe	TTT	0.493	0.551 ↔	0.291
	TTC	0.507	0.449 ↔	0.709
Ala	GCT	0.246	0.145	0.275
	GCC	0.254	0.276	0.164
	GCA	0.246	0.196	0.240
	GCG	0.254	0.382	0.323
Asn	AAT	0.493	0.409 ↔	0.172
	AAC	0.507	0.591 ↔	0.828

Class II genes ~ largely ribosomal proteins or translation factors
– genes expressed at high levels,
Class I genes ~ mostly those that are expressed at moderate levels

How are the predicted relative frequencies calculated?
Relative frequencies of two codons coding for the same A.A.?

Calculating the predicted relative frequencies:

For a sequence of independent bases L_1, L_2, …,L_n, the expected 3-tuple relative frequencies are computed as:

$$\mathbb{P}\,(L_i = r_1,\ L_{i+1} = r_2, L_{i+2} = r_3) = \mathbb{P}\,(L_i = r_1)\,\mathbb{P}\,(L_{i+1} = r_2)\,\mathbb{P}\,(L_{i+2} = r_3)$$

- provides the expected frequencies of particular codons

We first calculate the relative proportion of each of the codons making up a particular AA.

The relative proportion of each of the codons making up a particular AA.

$$₱\,(TTT) = 0.246 \text{ x } 0.246 \text{ x } 0.246 = 0.01489$$

$$₱\,(TTC) = 0.246 \text{ x } 0.246 \text{ x } 0.254 = 0.01537$$

using base frequencies. Among these codons making up the AA Phe, the expected proportion of TTT is

$$\frac{\mathbf{0.01489}}{\mathbf{0.01489 + 0.01537}} = \mathbf{0.492}$$

	Codon	Predicted	Observed	
			Gene Class I (502)	Gene Class II (191)
Phe	TTT	0.493	0.551	0.291
	TTC	0.507	0.449	0.709

Most widely used statistic for analyzing Codon usage bias in a protein is codon adaptation index (CAI)

- it compares the distribution of codons <u>actually used</u> in a particular protein with the preferred codons for highly expressed genes.

- **Highly expressed genes use mostly codons for tRNA species that are more abundant in the cell.**

⇒ CAI provides an indication of gene expression level under the assumption that there is translational selection to optimize gene sequences according to their expression levels

Consider a protein $X = x_1, x_2, \ldots, x_L$ with x_k representing the AA residue corresponding to codon k in the gene.

Here we are interested in comparing the actual codon usage with an alternative model:

- that the codons employed are the most probable codons for highly expressed genes.

p_k - probability that a particular codon k is used to code for the AA, and

q_k - probability for the <u>most</u> <u>frequently used</u> synonymous codon for that AA in highly expressed genes.

Codon Adaptation Index

CAI is defined as

$$\mathrm{CAI} = \left[\prod_{k=1}^{L} p_k / q_k \right]^{1/L}$$

i.e., the geometric mean of the ratios of the probabilities for the codons actually used to the probabilities for the codons most frequently used in highly expressed genes.

Alternatively,

$$\log(\mathrm{CAI}) = \frac{1}{L} \sum_{k=1}^{L} \log(p_k / q_k)$$

There is a <u>correlation</u> between CAI and mRNA levels - if CAI is large, the expression level is also large, i.e., it provides a first approx. of its expression level.

Larger Words

No. and distributions of *k*-tuples > 3 have practical & biological significance.

Some important *k*-tuples corresponding to *k* = 4, 5, 6 and 8:

- determine the distribution of restriction endonuclease digest fragments.

- useful for constructing the physical maps of genomes and identifying structural variations

Restriction sites for restriction endonucelases

Enzyme	Recognition Sequence
EcoRI	G↓AATTC
HindIII	A↓AGCTT
BamHI	G↓GATCC
BglI	GCCNNNN↓NGGC
PvuI	CGATC↓G
HaeIII	GG↓CC
MboI	GAT↓C

RFLP – difference in the lengths & number of fragments as a result of mutations in the recognition sites, VNTRs, other genetic disorders such as insertions, deletions, translocations & inversions.

Larger Words

Any other example of larger *k*-words ($k \geq 4$)?

- Modeling the no. of restriction sites in DNA

Some of the Qs that we can answer are:

- **If we were to digest the DNA with a RE such as EcoRI, approx. how many fragments would be obtained, and what would be their size distribution?**
- **Suppose we observed 761 occurrences of the sequence GCTGGTGG in a genome that is 50% G+C and 4.6Mb in size. How does this number compare with the expected no.? How would one find the expected no.? Expected according to what model?**

Larger Words

The pattern GCTGGTGG is known as Chi sequence:

5′-GCTGGTGG-3′ in *E. coli* (*k*=8)

- are significantly over-represented in particular genomes or on one or the other strand of the genome

e.g., Chi sequences occur 761 times in *E. coli* genome compared with approx. 70 instances predicted using base frequencies under iid model

Chi sequences are more abundant on the leading strand than on the lagging strand

- these observations relate in part to the involvement of Chi sequences in generalized recombination

Larger Words

Another example is the uptake sequences that function in bacterial transformation, e.g., (k=10)

5'-GCCGTCTGAA-3' in *Neisseria gonorrhoeae*

Some sequences may be under-represented, e.g.,

5'-CATG-3' occurs in the *E. coli* K-12 genome 1/20th of the expected frequency.

***k*-words ($k \geq 4$) are also useful for analyzing particular genomic subsequences, e.g., 6-word frequencies can be used to quantify the differences between *E. coli* promoter sequences and "average" genomic DNA.**

<u>Transformation</u> - genetic alteration of a cell resulting from direct uptake, incorporation & expression of exogenous genetic material from its surroundings and taken up through the cell membrane(s).

Summary and Applications

For word sizes, *k* = 1, 2, and 3, we saw that

- frequencies of words or statistics derived from them (*viz*., GC skew for *k* = 1) were not as predicted from independent, identically distributed (i.i.d.) base model.

This is not surprising – since genomes code for biological information, and we would therefore <u>not</u> expect the iid model to provide an accurate description for real genomes.

Summary and Applications

The frequencies of k-tuples have a number of applications:

- **GC skew - to predict locations of replication origins and termini in prokaryotes.**
- **k-mers ($k \geq 2$) – to identify regions having aberrant base compositions that may indicate genome segments acquired by lateral transfer.**

e.g., if there is gene transfer from organism 1 (GC content: 50%) to organism 2 (GC content: 70%), then a simple measure such as GC content can be computed in sliding windows to identify the laterally transferred regions.

Summary and Applications

Parametric Methods based on anomalous nucleotide compositions for identifying horizontally transferred regions:

- **GC content anomalies – defined as the ratio of GC content in sliding window by GC content of the whole genome**
- **Genomic signature - set of di-nucleotide relative abundance values constitutes a "genomic signature" of an organism**

$$\delta^*(f_w, g) = 1/16 \sum |\rho^*_{xy}(f_w) - \rho^*_{xy}(g)| \qquad \rho^*_{xy} = f^*_{xy} / f^*_x f^*_y$$

- ***k*-mer distribution - average *k*-mer difference is defined as**

$$\delta^*_k(w, g) = \frac{1}{n} \sum_{i=1}^{n} \left| f_i^w - f_i^g \right|$$

***n* = 4^k no. of distinct *k*-words (*k = 2 to 9*) are computed, both for the whole genome & the sliding window.**

Summary and Applications

Parametric Methods at the gene level are:

- **Codon Usage Bias** - i.e., unequal usage of synonymous codons
- **Amino Acid Usage Bias** - deviation in the frequency of usage of individual amino acids over the average usage of all 20 AAs
- **GC Content at Codon Positions** - frequency of occurrence of G & C at 3 codon positions, GC1, GC2 & GC3 for the set of genes compared to the whole genome/2nd set of genes

- All these are based on k-mer analysis

Summary and Applications

The frequencies of k-tuples have a number of applications:

- **For eukaryotes, gene regions, in general, have a different base composition than non-genic regions,**

  e.g., human genes are relatively GC-rich compared with the genome as a whole, triplet frequencies different in coding region compared to non-coding regions.

- **Different gene classes have different codon usage frequencies, e.g., highly expressed genes.**

- **Codon usage differs from organism to organism – useful in identifying horizontally transferred genes.**

Summary and Applications

- **Sometimes the observed frequencies of k-words can be used to make inferences about DNA sequences.**

  For e.g., suppose that we are given a sequence string that hypothetically could be a portion of a candidate exon or a prokaryotic gene:

GACGTTAGCTAGGCTTTAATCCGACTAAACCTTTGATGCATGCCTAGGCTG

  Simply by noting the frequency of stop codons in all the three reading frames, and knowing that a typical bacterial gene contains, on average, more than 300 codons, or that the typical human exon, contains around 50 codons, one can make a reasonable inference that this string does not code for a protein.

Summary and Applications

- ***k*-tuple (*k* > 3) frequencies can also assist in predicting whether an unannotated sequence is coding or noncoding.**

  Because coding sequences commonly specify AA strings that are functionally constraint, the distribution of *k*-tuple frequencies differ from that of noncoding sequences, e.g., intergenic or intronic sequences.

  e.g., consider in-frame hexamers (*k* = 6) – there are 4096 6-mers. From known polypeptide sequences we can easily predict that some 6-tuples will be infrequent, *viz*., the pair of residues Trp-Trp in succession is not common in protein sequences, implying that the corresponding dicodon, TTGTTG is likely to be relatively infrequent in coding sequences.

Summary and Applications

- **Predict gene expression:** using $k = 3$, compute CAI within ORFs to identify highly expressed genes (CAI ~1)
- ***k*-tuple frequencies and other content-based measures such as the presence of particular signals are among the statistical properties employed by computational gene finding tools.**

  Measures based on oligonucleotide counts:

  - Codon Usage
  - Amino Acid Usage
  - Codon Preference
  - Hexamer Usage

Summary and Applications

- ***k*-mer distributions are well-preserved among related strains/species - bacterial genomes can be clustered into natural groups according to *k*-mer distribution similarities.**

Reference

- **Computational Genome Analysis: An Introduction, R.C. Deonier, S. Tavare, M.C. Waterman (Chap-2).**

Mystery of the Chilean blob

- A 13-tonne blob containing no skin, bone or cells was washed ashore on a Chilean beach in July 2003
- Hypotheses ranged from remains of a giant squid, octopus, whale blubber, to some sea monster, alien
- DNA samples were extracted from the blob and sequenced (NADH2, control region)
- Search against the database unambiguously established the identity of the blob: sperm whale blubber

Pierce, S.K *et al.* 2004, *Biological Bulletinv*206, 125-133

Pairwise Sequence Alignment

Dynamic Programming Algorithm

A popular method for identification of conservation patterns between two genes/proteins is

- **pairwise sequence alignment**

Outcome of pairwise sequence alignment:

- **identifying regions of similarity**
- **a score which measures similarity between sequences - quantify**

Some Definitions

Alignment – process of lining up two or more sequences allowing for mismatches

HEAGAWGHEE
PAWHEAE

- for assessing the degree of similarity and the possibility of homology
- if the same letter occurs in both the sequences then this position has been **conserved** in evolution.
- if the letters **differ** it is assumed that the two derive from an **ancestral** letter (could be one of the two or neither)

Homology – having a common ancestral origin

- proteins with similar 3D-structures

Difference between similarity and homology:

- **Similarity is simply a measure of expression how alike two sequences are**
- **Homology means there is an evolutionary relationship between two sequences - there are no degrees of homology.**
- **Extending this to individual residues they are 'identical' or 'similar' residues - similar implies that they share certain physicochemical properties**
- **Homology cannot be observed, it is only an inference**

Difference between similarity and homology

Identical protein sequences result in identical 3-D structures - similar sequences may result in similar structures, and this is usually the case.

Converse is not true: identical 3-D structures do not necessarily indicate identical sequences. It is because of this that there is a distinction between "homology" and "similarity".

There are examples of proteins in the databases that have nearly identical 3-D structures, and are therefore homologous, but do not exhibit significant (or detectable) sequence similarity

Sequence identity and rmsd of Sperm Whale myoglobin

myoglobin
pig

rmsd = 0.5 Å
id = 86%

haemoglobin
pig

rmsd = 1.5 Å
id = 28%

globin-3
P. piclitum

rmsd = 2.2 Å
id = 18%

phycocyanin
F. diplosiphon

rmsd = 3.3 Å
id = 8%

Comparison of Sequences

The main objective of sequence alignment is

- to identify regions of similarity, i.e., *conserved regions*

- to find out if the two sequences are related or not

- this would enable us to extrapolate knowledge of the known sequence to the unknown query sequence

Any other reasons for Sequence Comparison?

Comparison of Sequences

Any other reasons for Sequence Comparison?

- **Identifying species** – as in the case of DNA barcoding
- **Phylogenetic analysis** – to find evolutionary relatedness between species
- Genome comparison **between individuals in a population** – for structural variation analysis
- Genome comparison **between diseased (e.g., cancer) and normal cells** – for identifying variations responsible for the disease
- Genome comparison **between species** – for understanding genome evolution

Inferring Function from Similarity

Biology

Zoology

Geology

Thanatology ?

Spectrology

Botany

Linguistics

Physics

Chemistry

Inferring Function from Similarity

Biology – study of life

Zoology – study of animals

Geology – study of earth

Thanatology – study of death

Spectrology – study of visible light/radiative energy as function of its wavelength/frequency

Botany - study of plants

Linguistics - study of languages

Physics – study of matter & its motion thru space & time

Chemistry - study of matter & energy & interactions between them

Basic task

Most basic sequence analysis task is - to find out if the two sequences are related or not, i.e.,

– to decide whether the alignment is more likely to have occurred because they are related, or just by chance ?

The other point to consider is – to use DNA (gene) or protein sequences?

Example

For the sequences gctgaacg and ctataatc:

An uninformative alignment:

```
- - -  - - -  - - g c t g a a c g
c t a t a a t c  - - - - - - - -
```

An alignment without gaps:

```
g c t g a a c g
c t a t a a t c
```

An alignment with gaps:

```
g c t g a - a - - c g
- - c t - a t a a t c
```

An alternative alignment:

```
g c t g - a a - c g
- c t a t a a t c -
```

with gaps

We also need to compute a score reflecting the quality of each alignment.

Key issues are:

- what sort of **alignment to be considered**
- the **scoring system** to rank alignments
- the **algorithm** to find optimal scoring alignments
- **statistical methods** to evaluate the significance of scores

Scoring scheme used is most crucial
- minor variations in scoring scheme may change the ranking of alignments, causing a different one to emerge as the best.

Complexity of the Problem

Consider three pairwise alignments, all to same region of human alpha globin protein sequence (HBA_Human):

- **human beta globin (HBB_Human)**
- **leghaemoglobin from yellow lupin (LGB2 LUPLU)**
- **nematode glutathione S-transferase (F11G11.2)**

Complexity of Problem

(a) **I=18, S=17, G=0**

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
            G+ +VK+HGKKV  A+++++AH+D++ +++++LS+LH  KL
HBB_HUMAN   GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

both hydrophilic

(b) **I=8, S=17, G=4**

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
            ++ ++++H+ KV   + +A  ++          +L+ L+++H+ K
LGB2_LUPLU  NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

(c) **I=13, S=12, G=6**

```
HBA_HUMAN   GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
            GS+ + G +   +D L  ++ H+ D+  A +AL D    ++AH+
F11G11.2    GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Complexity of the Problem

Challenge:

How to distinguish cases like (b) from those like (c)?

One needs to carefully choose the scoring system to evaluate alignments.

Even then it may not always be possible to distinguish true alignments from spurious ones.

e.g., it is extremely difficult to find significant similarity between lupin leghaemoglobin and human alpha globin (Alignment (b)) by pairwise alignments

The Scoring Model

When comparing two sequences one is looking for evidence that they have diverged from a common ancestor by a process of mutation or natural selection

Basic mutational processes are:

substitution – which change residues in a sequence

insertions and deletions – which add or remove residues, together referred as gaps or indels

Further apart two sequences are from each other, more frequent these changes are expected to occur.

The Scoring Model

Mutations potentially **affect the function** of the gene, which can either be beneficial, or lead to reduction in functionality and adaptability of the protein.

Natural selection comes into play – allowing mutations that are either **evolutionarily advantageous** or, **occur in non-functional regions of the sequence**

Natural selection has the effect of **screening the mutations** – some changes are seen more often than others

- results in a **mosaic pattern** of conserved & unconserved regions, the functional regions being conserved across evolutions

⇒ Some identities, Ala-Ala score **lower** than others, Trp-Trp

The Scoring Model

Total score of an alignment

- sum of the terms for **each aligned pair** of residues, plus terms for **each gap**

Identities and conservative substitutions
- contribute a positive score, while

non-conservative changes and gaps
- accumulate negative scores

Similarity-based scoring scheme

The Scoring Model

Additive scoring scheme corresponds to an assumption that - mutations at different sites in a sequence **occurred independently** (treating gap of arbitrary length as a single mutation)

- a reasonable approximation for DNA and protein sequences, though **interactions between residues** play a critical role in determining protein structure.

This assumption is **inaccurate** for structural **RNAs** where base pairing introduces long-range dependencies

Substitution Matrices

The scoring system is composed of a substitution matrix (4×4 for DNA, 20×20 for Proteins) and gap penalties

- **entries in substitution matrix reflect the likelihood of a substitution from one nucleic acid (amino acid) base to another during the course of evolution,**

e.g., how often is Ala replaced by Val, Gly, etc.

i.e., the frequency of Ala-Val substitution at a given evolutionary distance

The Scoring Model

What is the probability that the alignment observed is a <u>true alignment</u> and not by chance?

In probabilistic interpretation – this corresponds to the logarithm of the relative likelihood that the sequences are <u>related</u>, compared to being unrelated

- assign a probability to the alignment in each of the two cases and then consider the ratio of the two probabilities

Substitution Matrix

Random or unrelated model *R*

- assumes that symbol ***a*** occurs independently with some frequency q_a, hence the probability of two sequences is just the product of the probabilities of each nucleic/amino acid:

$$P(x, y \mid R) = \prod_i q_{x_i} \prod_j q_{y_j}$$

using the multiplication rule of probabilities

Substitution Matrix

Match model *M* - aligned pairs occur with a joint probability p_{ab}. Thus, the probability for the whole alignment:

$$P(x, y \mid M) = \prod_i p_{x_i y_i}$$

Odds-ratio - the ratio of these two likelihoods :

$$\frac{P(x, y \mid M)}{P(x, y \mid R)} = \frac{\prod_i p_{x_i y_i}}{\prod_i q_{x_i} \prod_i q_{y_i}} = \prod_i \frac{p_{x_i y_i}}{q_{x_i} q_{y_i}}$$

Substitution Matrix

To have an additive scoring system, take logarithm of this ratio, to obtain log-odds ratio:

$$S = \sum_i s(x_i, y_i) \qquad s(a,b) = log\frac{p_{ab}}{q_a q_b}$$

***s*(*a*,*b*) is the log likelihood ratio of the residue pair (*a*,*b*) occurring as an <u>aligned</u> pair, as opposed to an <u>unaligned</u> pair.**

– these individual scores for each pair of residues are arranged in a matrix, called *scoring matrix* or *substitution matrix*

Gap Penalties

Gaps need to be penalized

The standard cost associated with a gap of length *g* is given either by a linear score

$$\gamma(g) = -gd$$

or, an affine score

$$\gamma(g) = -d - (g-1)e$$

$\gamma(g) = ?$
for $d = 8$, $e = 2$,
$g = 10$

d – gap-open penalty, *e* – gap-extension penalty;

$e < d$ - allows long insertions and deletions to be penalized less than they would be by the linear gap cost

Why?

Gap Penalties

An affine score assumes that consecutive deletions/insertions are a single mutation event as opposed to multiple insertions/deletions and hence should be penalized less.

Why not assign a long gap the penalty of a single gap?

Dynamic Programming Algorithm

Aim: Given a scoring system, to have an algorithm for finding an optimal alignment for a pair of sequences.

For two sequences of length n , there are

$$\begin{bmatrix} 2n \\ n \end{bmatrix} = \frac{(2n)!}{(n!\ )^2} \approx \frac{2^{2n}}{\sqrt{\pi n}}$$

possible global alignments!

Not computationally feasible

For a sequence of length 100, $2^{2n} \approx 10^{30}$, for n=1000, it is 10^{300}

Dynamic Programming Algorithm

Algorithm for finding optimal alignments given an additive alignment score is called the dynamic programming

The scoring scheme being introduced as a log-odds ratio

- *better* alignments will have *higher* score

So the aim is to *maximize* the score

When scores are assigned as *costs* or *edit distances*, the aim is to *minimize* the score

Dynamic programming algorithms apply to either case

- the differences are trivial changes of 'min' for 'max'

Dynamic Programming Algorithm

Consider the following two amino acid sequences:

HEAGAWGHEE and PAWHEAE

To score the alignments, consider

Substitution scoring matrix: BLOSUM50

Gap penalty, d = 8

BLOSUM50 Matrix

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	**5**	−2	−1	−2	−1	−1	−1	0	−2	−1	−2	−1	−1	−3	−1	1	0	−3	−2	0
R	−2	**7**	−1	−2	−4	1	0	−3	0	−4	−3	3	−2	−3	−3	−1	−1	−3	−1	−3
N	−1	−1	**7**	2	−2	0	0	0	1	−3	−4	0	−2	−4	−2	1	0	−4	−2	−3
D	−2	−2	2	**8**	−4	0	2	−1	−1	−4	−4	−1	−4	−5	−1	0	−1	−5	−3	−4
C	−1	−4	−2	−4	**13**	−3	−3	−3	−3	−2	−2	−3	−2	−2	−4	−1	−1	−5	−3	−1
Q	−1	1	0	0	−3	**7**	2	−2	1	−3	−2	2	0	−4	−1	0	−1	−1	−1	−3
E	−1	0	0	2	−3	2	**6**	−3	0	−4	−3	1	−2	−3	−1	−1	−1	−3	−2	−3
G	0	−3	0	−1	−3	−2	−3	**8**	−2	−4	−4	−2	−3	−4	−2	0	−2	−3	−3	−4
H	−2	0	1	−1	−3	1	0	−2	**10**	−4	−3	0	−1	−1	−2	−1	−2	−3	2	−4
I	−1	−4	−3	−4	−2	−3	−4	−4	−4	**5**	2	−3	2	0	−3	−3	−1	−3	−1	4
L	−2	−3	−4	−4	−2	−2	−3	−4	−3	2	**5**	−3	3	1	−4	−3	−1	−2	−1	1
K	−1	3	0	−1	−3	2	1	−2	0	−3	−3	**6**	−2	−4	−1	0	−1	−3	−2	−3
M	−1	−2	−2	−4	−2	0	−2	−3	−1	2	3	−2	**7**	0	−3	−2	−1	−1	0	1
F	−3	−3	−4	−5	−2	−4	−3	−4	−1	0	1	−4	0	**8**	−4	−3	−2	1	4	−1
P	−1	−3	−2	−1	−4	−1	−1	−2	−2	−3	−4	−1	−3	−4	**10**	−1	−1	−4	−3	−3
S	1	−1	1	0	−1	0	−1	0	−1	−3	−3	0	−2	−3	−1	**5**	2	−4	−2	−2
T	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	2	**5**	−3	−2	0
W	−3	−3	−4	−5	−5	−1	−3	−3	−3	−3	−2	−3	−1	1	−4	−4	−3	**15**	2	−3
Y	−2	−1	−2	−3	−3	−1	−2	−3	2	−1	−1	−2	0	4	−3	−2	−2	2	**8**	−1
V	0	−3	−3	−4	−1	−3	−3	−4	−4	4	1	−3	1	−1	−3	−2	0	−3	−1	**5**

Global Alignment

It's end-to-end alignment of two sequences, allowing <u>gaps</u>, e.g., consider two sequences:

HEAGAWGHEE and PAWHEAE

Alignment:

HEAGAWGHE – E

– – P – AW – HEAE

is called a **global alignment**

The dynamic programming algorithm for obtaining global alignment of biological sequences is called the **Needleman-Wunsch algorithm**

Initialize: $F(0, 0)=0$

Boundary conditions:

$F(i, 0) = F(0, i) = -id,\ d = 8$

Score of the pair (P, H) = ?

If $F(i\text{-}1, j\text{-}1)$, $F(i\text{-}1, j)$ and $F(i, j\text{-}1)$ are <u>known</u>, it is possible to calculate $F(i, j)$

The idea is to build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences

Global Alignment: Needleman-Wunsch Algorithm

I G A x_i	A I G A x_i	G A x_i – –
L G V y_j	G V y_j – –	S L G V y_j

- to align a letter from the horizontal sequence, x_i, with a letter from the vertical sequence, y_j:

$$F(i, j) = F(i\text{-}1, j\text{-}1) + s(x_i, y_j)$$

Global Alignment: Needleman-Wunsch Algorithm

I G A x_i	A I G A x_i	G A x_i – –
L G V y_j	G V y_j – –	S L G V y_j

- to align a letter from the horizontal sequence, x_i, against a gap in the vertical sequence; in which case

$$F(i, j) = F(i\text{-}1, j) - d$$

- to align a gap from the horizontal sequence against a letter in the vertical sequence, y_j, in which case

$$F(i, j) = F(i, j\text{-}1) - d$$

Global Alignment: Needleman-Wunsch Algorithm

The best score up to (i, j) will be the largest of these three options:

$$F(i,j) = max\begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

This equation is applied repeatedly to fill in the matrix of $F(i, j)$ values

A <u>pointer</u> is kept in each cell back to the cell from which its $F(i, j)$ was derived.

<u>Boundary conditions</u>: $F(i, 0) = -id$, $F(0, j) = -jd$

Score of pair (P, H) = -2

$F(i, j) = -2$, and the direction is diagonal

$F(i, 0) = F(0, j) = -8$

Global Alignment: Needleman-Wunsch Algorithm

		H	E	A	G	A	W	G	H	E	E
	0	**–8**	**–16**	–24	–32	–40	–48	–56	–64	–72	–80
P	–8	–2	–9	**–17**	**–25**	–33	–41	–49	–57	–65	–73
A	–16	–10	–3	–4	–12	**–20**	–28	–36	–44	–52	–60
W	–24	–18	–11	–6	–7	–15	**–5**	**–13**	–21	–29	–37
H	–32	–14	–18	–13	–8	–9	–13	–7	**–3**	–11	–19
E	–40	–22	–8	–16	–16	–9	–12	–15	–7	**3**	–5
A	–48	–30	–16	–3	–11	–11	–12	–12	–15	**–5**	2
E	–56	–38	–24	–11	–6	–12	–14	–15	–12	–9	**1**

HEAGAWGHE-E

--P-AW-HEAE

Score

Global Alignment: Needleman-Wunsch Algorithm

Final cell of the matrix, $F(n, m)$ gives the best score for global alignment of $x_{1...n}$ to $y_{1...m}$

To find the alignment itself, we must find the <u>path of choices</u> that led to this final value. The procedure for doing this is known as traceback.

It works by building the alignment in reverse, starting from the final cell, and following the pointers stored when building the matrix.

Global Alignment: Needleman-Wunsch Algorithm

Algorithmic complexity:

NW algorithm takes $O(nm)$ time & $O(nm)$ memory, where n & m are lengths of the two sequences

i.e., the computer time and memory storage required scales as the product of sequence lengths

Can this algorithm be used for comparing genomes?

Local Alignment

A common situation is where one is looking for the best alignment between subsequences of *x* and *y*.

This arises for e.g., when it is suspected that two protein sequences may share a common domain, or when comparing extended sections of genomic DNA.

It is also the most sensitive way to detect similarity when comparing two very highly diverged sequences sharing common evolutionary origin, or,

when one has no knowledge about divergence, as in database search.

Local Alignment

The highest scoring alignment of subsequences of *x* and *y* is called the best *local* alignment,

e.g., consider the sequences:

HEAGAWGHEE and PAWHEAE

Local Alignment: AWGHE

AW- HE

Dynamic programming algorithm for finding optimal local alignment in biological sequences is the Smith-Waterman algorithm

Local Alignment: Smith-Waterman Algorithm

There are two differences for finding optimal local alignment, compared to global alignment

First, in each cell, an extra possibility is added, allowing $F(i, j)$ to take the value 0 if all other options have value less than 0:

$$F(i, j) = max \begin{cases} 0 \\ F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) - d \\ F(i, j-1) - d \end{cases}$$

Option 0 corresponds to starting a new alignment anywhere in the sequence.

Local Alignment: Smith-Waterman Algorithm

A consequence of 0 is that the <u>boundary conditions</u> are:

$$F(i, 0) = 0, \; F(0, j) = 0$$

instead of $-id$ and $-jd$ as for the global alignment.

<u>Second</u> change is that now an alignment can end anywhere in the matrix

For this, look for the highest value of $F(i, j)$ in the whole matrix to start the traceback, instead of the (n, m) cell.

Traceback ends when a cell with value '0' is encountered.

Local Alignment: Smith-Waterman Algorithm

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

AWGHE
AW-HE

Score

Tools - EMBOSS

Global Alignment:

- **needle** – uses Needleman-Wunsch global alignment algorithm (including gaps).

Local Alignment:

- **water** – uses Smith-Waterman algorithm (modified for speed enhancements) to calculate local alignment.
- **matcher** - compares two sequences looking for local sequence similarities using a rigorous algorithm, based on Bill Pearson's 'lalign'

Suboptimal/Repeated Matches

Smith-Waterman algorithm - gives the best local match between two sequences

If one or both the sequences are long, it is possible to have many different local alignments with a significant score, and one may be interested in all of these

Example - many copies of a repeated domain or motif in a protein, multi-domain proteins, distantly related sequences can have more than one conserved regions

Local Alignment: Smith-Waterman Algorithm

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0
A	0	0	0	5	0	**5**	0	0	0	0	0
W	0	0	0	0	2	0	**20**	**12**	4	0	0
H	0	10	2	0	0	0	12	18	**22**	14	6
E	0	2	16	8	0	0	4	10	18	**28**	20
A	0	0	8	21	13	5	0	4	10	20	27
E	0	0	6	13	18	12	4	0	4	16	26

HEAG
HEAE

AWGHE
AW-HE

AWGHEE
AW - HEA

AWGHE - E
AW - HEAE

Suboptimal Matches

Suboptimal matches are obtained by taking a traceback not only from the maximum scoring cell, but from other **high-scoring cells ($\geq T$)**

All other conditions for local alignment hold true in this case.

Some of the high-scoring alignments may be overlapping – one may output only non-overlapping ones

Suboptimal Matches

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	**1**	1	1	1	1	3	**9**
P	0	0	0	0	**1**	1	1	1	1	3	9
A	0	0	0	5	1	**6**	1	1	1	3	9
W	**0**	0	0	0	2	1	**21**	**13**	5	3	9
H	0	**10**	2	0	1	1	13	19	**23**	15	9
E	0	2	**16**	8	1	1	5	11	19	**29**	21
A	0	0	8	**21**	13	6	1	5	11	21	28
E	0	0	6	13	18	12	4	1	5	17	27

HEAGAWGHEE
HEA.AW-HE.

T = 20

There are two separate match regions, with scores 1 and 8
Dots are used to indicate unmatched regions.

Suboptimal Matches

Initial Conditions: $F(0,0) = 0$

Boundary Conditions: $F(0,j) = 0$

$$F(i,0) = max\begin{cases} F(i-1,0) \\ F(i-1,j) - T \qquad j = 1,\ldots m \end{cases} \qquad (i)$$

- handles **unmatched regions** and **ends of matches**, only allowing matches to end when they have score at least **T**

Recurrence Relations:

$$F(i,j) = max\begin{cases} F(i,0) \\ F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases} \qquad (ii)$$

- handles **starts of matches** and **extensions**

Suboptimal Matches

Total score of all the matches is obtained by adding an extra cell to the matrix, ***F(n*+1,0),*** using eq. (i)

This score will have ***T*** subtracted from each match; if there were no matches of score greater than ***T*** it will be **0** [repeated application of 1st option in (i)]

$$F(i,0) = max\begin{cases} F(i-1,0) \\ F(i-1,j)-T \qquad j=1,\ldots m \end{cases} \qquad \text{(i)}$$

Suboptimal Matches

EMBOSS Programs:

- **matcher:** Rigorous Smith-Waterman alignment
- **supermatcher:** Finds a match of a large sequence against one or more sequences
- **wordmatch:** Finds all exact matches of a given size between two sequences

Overlap Matches

Another important situation commonly encountered with biological sequences is -

- one sequence contained in the other
- have common overlapping regions.

e.g., when comparing fragments of genomic DNA sequence to each other, or to large chromosomal sequences, in sequence assembly

Several different types of configurations can occur

Overlap Matches

Start: F(0, *j*), or F(*i*, 0)

End: F(*i*, *m*), or F(*n*, *j*)

i*: 1, … *n*, *j*: 1, … *m

Overlap Matches

This is a special a case of global alignment that does not penalize overhanging ends – also called semi-global alignment.

i.e., an algorithm for a match that may start on the left / top border of the matrix, and end on the right / bottom border.

Overlap Matches

Initialization equations: (same as for <u>local</u> alignment)

$$F(0,0) = 0$$

$$F(i,0) = 0 \qquad i = 1,\ldots n$$

$$F(0,j) = 0 \qquad j = 1,\ldots m$$

Recurrence relations: (same as for <u>global</u> alignment)

$$F(i,j) = max\begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

Overlap Alignment

		H	E	A	G	A	W	G	H	E	E
	0	0	0	0	0	0	0	0	0	0	0
P	0	–2	–1	–1	**–2**	–1	–4	–2	–2	–1	–1
A	0	–2	–3	4	–1	**3**	–4	–4	–4	–3	–2
W	0	–3	–5	–4	1	–4	**18**	**10**	2	–6	–6
H	0	10	2	–6	–6	–1	10	16	**20**	12	4
E	0	2	16	8	0	–7	2	8	16	**26**	18
A	0	–2	8	21	13	5	–3	2	8	18	**25**
E	0	0	4	13	18	12	4	–4	2	14	24

GAWGHEE
PAW-HEA

Note the boundary conditions

Overlap Matches

F_{max}- the maximum value on the **bottom** border (i, m), $i = 1, \ldots, n$, or the **right** border (n, j), $j = 1, \ldots, m$.

Traceback starts from F_{max} and continues until the **top** $(i, 0)$ **or left** $(0, j)$ edge is reached.

Overlap Matches

EMBOSS programs:

est2genome: Align EST & genomic DNA sequences

merger: Merges two overlapping sequences

megamerger: Same as merger for large sequences

Tandem Repeats

Consider a situation where a repetitive sequence *y* is found in **tandem copies** not separated by gaps, then the **recurrence relations** for filling the path matrix are:

Initial & Boundary conditions:

$$F(0,0) = 0$$

$$F(i,0) = 0$$
$$F(0,j) = 0$$

$$F(i,1) = max\begin{cases} F(i-1,0) + s(i,1) \\ F(i-1,m) + s(i,1) \\ F(i-1,1) - d \\ F(i,0) - d \end{cases}$$

?

$$F(i,j) = max\begin{cases} F(i-1,j) + s(i,j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

Similar to global alignment

Allows a bypass of $-T$ penalty so the threshold applies **only once** to each tandem cluster of repeats, not once to each repeat

Tandem Repeats

EMBOSS:

etandem: Finds tandem repeats in nucleotide sequence

equicktandem: Finds tandem repeats up to a specified size, the results can be used to run etandem on the candidate repeat lengths

einverted: Finds DNA inverted repeats using DP

palindrome: Finds DNA inverted repeats

When comparing sequences we should ideally always consider

- **what types of match we are looking for**
- **choose the appropriate algorithm accordingly**
- **choose an appropriate scoring matrix & gap penalties** ←

Alignment with Affine Gap Scores

Affine gap cost is given by

$$\gamma(g) = -d - (g-1)e$$

While using this gap cost we need to keep track of multiple values for each pair of residue coefficients (i, j) in place of single value $F(i, j)$

I G A x_i	A I G A x_i	G A x_i - -
L G V y_j	G V y_j - -	S L G V y_j

$M(i, j)$ - score up to (i, j), x_i aligned to y_j (L)

$I_x(i, j)$ - best score given that x_i aligned to a gap in y (C)

$I_y(i, j)$ - best score given that y_j aligned to a gap in x (R)

Alignment with Affine Gap Scores

x_i aligned to consecutive gaps in y

x_i aligned to y_j

y_j aligned to consecutive gaps in x

a finite state automaton model representing relationships between the three states used for affine gap alignment

Alignment with Affine Gap Scores

Recurrence relations (global alignment):

$$
\begin{aligned}
M(i,j) &= max\begin{cases} M(i-1,j-1)+s(x_i,y_j) \\ I_x(i-1,j-1)+s(x_i,y_j) \\ I_y(i-1,j-1)+s(x_i,y_j) \end{cases} \\
I_x(i,j) &= max\begin{cases} M(i-1,j)-d \\ I_x(i-1,j)-e \end{cases} \\
I_y(i,j) &= max\begin{cases} M(i,j-1)-d \\ I_y(i,j-1)-e \end{cases}
\end{aligned}
$$

Assumption: a deletion will not be followed directly by an insertion (true for the optimal path if – d – e is less than the lowest mismatch score)

Affine gap versions provide the most sensitive sequence matching methods

State Assignments for an Alignment with Affine Gap Scores

Complexity of Dynamic Program

It is of order – O(nm), where n, m are the length of the two sequences.

Not feasible for comparing complete genomes or chromosomes ~ a few Mbs long

- Space complexity needs to be addressed

In database search, a query sequence of length n is searched a database of size ~ few Gbs

- Time complexity is an issue in this case

References

(1) Biological Sequence Analysis, R. Durbin,S. Eddy, A.Krogh, G.Mitchison, Cambridge University Press

Chap-2: Pairwise Alignment

Pairwise Sequence Alignment

contd….

Heuristic Alignment Algorithms

Dynamic programming algorithms though very sensitive, are not the fastest available sequence alignment methods

- time complexity ~ $O(nm)$, and in many cases speed is the issue, e.g., database searches

Protein database contain ~100M residues, this requires ~10^{11} matrix cells to be evaluated to search complete database for a query of length 1000; DNA Dbs are even larger

At 10M matrix cells per second this would be 10^4 secs, or ~ 3 hrs for a single search

Heuristic Alignment Algorithms

For homology-based gene identification and annotation purposes one routinely screens genes against a database of proteins.

- need for algorithms faster than pure DP

Approx. methods can detect close relationships well and quickly but fail to identify very distant relationships.

Heuristic Alignment Algorithms

Goal of such methods - to search as small a fraction as possible of the cells in the DP matrix, while still looking at all high scoring alignments.

A typical approximation approach - take a small integer *k*, and determine all instances of each *k*-tuple of residues in the probe sequence that occur in any sequence in the database

A candidate sequence in the database would contain many matching *k*-tuples, with equivalent spacing in probe and candidate sequences.

- perform DP only for probable homologies.

Heuristic Alignment Algorithms

For selected candidate sequences, approximate optimal alignment calculations are carried out, with the "time" and "space" saving restriction

- paths through the matrix considered are restricted to bands around the diagonals containing the matching *k*-tuples.

e.g. of Heuristic approaches: BLAST, FASTA

Another computational resource that can limit dynamic programming alignment is memory usage

- memory requirement for $F(i, j) \sim O(nm)$, product of sequence lengths

For two protein sequences (few hundred residues long) - manageable on a desktop computers

But if one or both the sequences are genomic DNA sequences (hundreds of thousands of bases long), required memory for the full matrix can exceed machine's physical capacity.

Assignment

Find out the size of protein database, UniProt, and nucleotide database, GenBank.

Compute No. of matrix cells to be computed using DP for:

(1) searching protein database, UniProt, and nucleotide database, GenBank and the time required assuming query sequence of length 1000 bases.

(2) Comparing Human Chr 1 ~249Mbp with a query sequence of 1000 bases using DP, and comparing it with Chr 1 of Mouse (~195Mbp)? What is the space requirement in the two cases?

Use computation time as 10M matrix cells per second (or operation time of your machine)

Linear Space Alignments

Fortunately, situation is better with memory than speed:

- there are techniques that give the optimal alignment in limited memory, of order $n + m$ rather than nm, but comes with a cost of doubling of time.

These methods are commonly referred to as linear space alignment methods.

Linear Space Alignments

If only the maximal score is needed, the problem is very simple:

Recurrence relation for $F(i, j)$ depends only on entries one row back - one can throw away rows that are further than one back from the current point.

For local alignment, the maximum score in the whole matrix is required - easy to keep track of the maximum value as the matrix is being built.

Linear Space Alignments

While this gives us the score, it will not find the Alignment!

If we throw away rows to avoid $O(nm)$ storage, then we also loose the traceback pointers.

An approach used to obtain the alignment uses the principle of divide and conquer.

Snapshot of Execution of Hirschberg's algorithm

Linear Space Alignments

Steps:

- Let $u = m/2$ (integer part)
- Identify a v such that the cell (u, v) is on the optimal alignment, i.e., v is the column where the alignment crosses the $u = m/2$ row of the matrix

  - this splits the DP problem into two subproblems: from $(0, 0)$ to (u, v), & from (u, v) to (m, n)

  - full alignment will be concatenation of the optimal alignments for these two separate submatrices

Linear Space Alignments

- This is done **recursively**, by successively halving each region, until sequences of zero length are being aligned, or alternatively, sequences are short enough & standard $O(nm)$ alignment and traceback method can be used.

But how do we find v?

- By combining the results of "forward" and "backward" DP passes at row u, for <u>each point</u> along the middle row, i.e.,

- optimal score from (0, 0) to each point in row u and optimal score from that point to (m, n).

Linear Space Alignments

- Adding these numbers gives the optimal score over all paths from (0, 0) to (m, n) that pass through all the points in row u.
- A sweep along the middle row, checking these sums, determines a point ($m/2$, v) where an optimal path crosses the middle row.
- This is then done recursively.

Linear Space Alignments

EMBOSS Programs:

matcher: linear-space version of local alignment algorithm by M. S. Waterman & M. Eggert

stretcher: global alignment algorithm using linear space

Database Search - Need for Heuristics

- Dynamic prog'g algorithms are computer intensive & time consuming ~ O(nm)
- Searching large databases using these algorithms is not feasible

  e.g., if the time to compare two 1kb sequences is ~ milliseconds, comparison of 1kb sequence with the human genome will take 3 hours!

Database Search - Need for Heuristics

- **Solution 1:** implement the algorithm in hardware
  – expensive

  SW algorithm has been implemented on Hybrid Core machines

- **Solution 2:** distribute the job to several processors to do in parallel mode

  – gets expensive as no. of processors ~ 1000

- **Solution 3:** use heuristics – gives approx. solutions

Heuristic Algorithms

Basic idea: first locate high-scoring short stretches and then extend them

Two well known programs:

~ 50 times faster than DP

- FASTA: Fast Alignment Tool
- BLAST: Basic Local Alignment Search Tool

Both find high scoring local alignments between a query sequence and a target database

Heuristic algorithms aim at speeding up the database search at the price of possibly missing the best scoring alignment

FASTA

- Proposed by Pearson & Lipman, 1988
- Compares sequences pairwise
- Heuristics: A good alignment will have exact matching subsequences
- Gain on speed but loss of sensitivity
- Best suited for global alignment
- Works better with DNA sequences

http://www.ebi.ac.uk/Tools/sss/fasta

FASTA: The algorithm

- Based on the logic of the dot matrix method
- View sequences as sequences of short words (*k*-tuple)

Motivation:

- Good alignments should contain many exact matches
- Hashing can find exact matches in O(n) time
- Diagonals can be formed from exact matches quickly and sorted
- Apply more precise alignment to small search space at the end

FASTA: The algorithm

- Look for all *k*-tuple matches between query & database sequences
  - For DNA *k* = 2-6 (default 6)
  - For proteins, *k* =1,2 (default 2)
- This is done by a lookup table, or hash
- For each *k*-tuple the database is pre-processed to identify its positions
- Query is scanned – for each *k*-tuple in the query, look-up the table/hash to identify matches in the database

Lookup method for finding an alignment

Position	1	2	3	4	5	6	7	8	9	10	11
Seq A	n	c	s	p	t	a	.	.	.	.	.
Seq B	.	.	.	.	.	.	a	c	s	p	r	k

Amino acid	Position in		Offset pos A – pos B
	Protein A	Protein B	
a	6	6	0
c	2	7	-5
k	-	11	
n	1	-	
p	4	9	-5
r	-	10	
s	3	8	-5
t	5	-	

Dot matrix method

	n	c	s	p	t	a				
a						•				
c		•								
s			•							
p				•						
r										
k										

~ O(n)

Possible alignment:

```
n c s p t a
  | | |
a c s p r k
```

FASTA: The algorithm

- Look for adjacent hot spots (substrings of exact matches) on the same diagonal
  - join them to form larger segments
  - space between hot spots gets –ve score
- Find 10 highest scoring diagonals
- Find the best diagonal run and filter out low scoring runs – by recomputing the score for each diagonal using scoring matrix and threshold score
- Combine close diagonal runs, including indels
- Compute alternative local alignments, using a band around the best diagonal run
- Finally use DP to align the query against best ranking resulting sequences

- Identify regions of identity (hot spots),
- Scan regions using a scoring matrix & save the best initial regions, regions with score < threshold are discarded, * marks best region
- Optimally join initial regions with score > threshold,
- Recalculate an optimised alignment around the highest scoring region

BLAST

- Basic Local Alignment Search Tool

  - proposed by Altshul *et al*, 1990, developed by NCBI

- The algorithm has been designed to balance speed and sensitivity in aligning sequences

- Designed to work best for local ungapped alignments – now incorporates gaps

- Website: http://www.ncbi.nlm.nih.gov/blast/

BLAST: Fastest alignment tool

Motivation:

- Good alignments should contain many <u>close matches</u>
- Statistics can determine which matches are significant
  - Much more sensitive than % identity
- Extending matches in both directions finds alignment
  - Yields high-scoring/maximum segment pairs (HSP/MSP)

BLAST: Fastest alignment tool

- View sequences as sequence of short words (*k*-tuples)
  - DNA: 7, 11, 15 (default 11),
  - protein: 2, 3, 6 (default 6)
- Create hash table of <u>neighborhood</u> (closely-matching) words
- Use statistics to set threshold for "closeness"

BLAST: Fastest alignment tool

- First look for short matching segments
- Choose matching segments above a threshold score, hits
- Overlapping hits form a larger segment
- Large segment extended in either direction if its score is above a threshold
- Extension of alignment continues until the score falls below the drop-off threshold from maximum value

PQG - 18
PEG - 15
PRG - 14
PSG - 13
PQA - 12

BLAST: Example

- Sequence: ASTNC
- Word ASTN has score of 9 for exact match using PAM250 scoring matrix
- If threshold score is 17, ASTN will not be used for querying the database
- STNC has score 19 for exact match using PAM250, and will be used

Note: not all exact matches are used if the score is below the threshold.

BLAST: The Algorithm

- Look for high scoring segments pairs (HSPs)
  - looks for similar instead of identical pairs
  - uses scoring matrix to score aligned pairs
  - only those pairs which score above a threshold are considered for extension
  - the extension is without gaps

```
L P   P Q G   L L   query
M P   P E G   L L   Db seq
      <word>
←             →
Ext. to Left Ext. to Rt
```

HSP score
= 9 + 15 + 8
= 32

BLAST: The Algorithm

- Ungapped extension of HSPs with scores > *T* (threshold) - identifies maximal segment pairs (MSPs)
- Extension continues until the score drops below a threshold drop-off from the maximum score encountered
- Highest scoring segment pair, MSP (Maximal scoring Segment Pair) identified
- For gapped alignment, BLAST uses the same strategy as FASTA of joining segments on different diagonals.

BLAST: Algorithm

1) Convert 1st sequence into words (using all frames for given word size)

2) Calculate for each word list of "neighborhood" words (scoring threshold T) and enter in dictionary

3) Scan 2nd sequence, find matching words in dictionary, store locations

4) For each match, extend alignment in both directions while score above threshold S, merge segments

5) Align best segments using dynamic programming, report statistically significant matches

Selecting a BLAST program

BLAST is suite of programs, the most popular being

- blastn: compares nucleotide sequence with nucleotide database
- blastp: compares protein sequence with a protein sequence database

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

End of updates for BLAST+ version 4 databases (dbV4)

Start moving to the new version 5 databases!

Fri, 27 Sep 2019 16:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide

blastx
translated nucleotide ▶ protein

tblastn
protein ▶ translated nucleotide

Protein BLAST
protein ▶ protein

BLAST Genomes

Enter organism common name, scientific name, or tax id | Search

Human | Mouse | Rat | Microbes

BLAST® Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

Clear

Query subrange

From

To

Enter coordinates for a **subrange** of the query sequence. The BLAST search will apply only to the residues in the range. Sequence coordinates are from 1 to the sequence length.The range includes the residue at the **To** coordinate. more...

Or, upload file

Browse...

Job Title

Enter a descriptive title for your BLAST search

This title appears on all BLAST results and saved searches.

Align two or more sequences

Choose Search Set

Database

Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Genomic plus Transcript

- Human genomic plus transcript (Human G+T)
- Mouse genomic plus transcript (Mouse G+T)

Other Databases

- Nucleotide collection (nr/nt)
- Reference mRNA sequences (refseq_rna)
- Reference genomic sequences (refseq_genomic)
- NCBI Genomes (chromosome)
- Expressed sequence tags (est)
- Non-human, non-mouse ESTs (est_others)
- Genomic survey sequences (gss)
- High throughput genomic sequences (HTGS)
- Patent sequences(pat)
- Protein Data Bank (pdb)
- Human ALU repeat elements (alu_repeats)
- Sequence tagged sites (dbsts)
- Whole-genome shotgun reads (wgs)
- Environmental samples (env_nt)

Organism
Optional

Enter organism name or id--completions will be suggested

Exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude
Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Show results in a new window

- Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.
- Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.
- BlastN is slow, but allows a word-size down to seven bases.

Algorithm parameters

Algorithm parameters

General Parameters

Max target sequences 100

Select the maximum number of aligned sequences to display

Maximum number of aligned sequences to display (the actual number of alignments may be greater than this).

Short queries ☑ Automatically adjust parameters for short input sequences

Automatically adjust word size and other parameters to improve results for short queries.

Expect threshold 10 ← **statistical significance threshold**

Expected number of chance matches in a random model. more... YouTube Expect value tutorial

Word size 28

The length of the seed that initiates an alignment. more...

Max matches in a query range 0 ←

Limit the number of matches to a query range. This option is useful if many strong matches to one part of a query may prevent BLAST from presenting weaker matches to another part of the query. The algorithm is based upon http://www.ncbi.nlm.nih.gov/pubmed/10890403

Scoring Parameters

Match/Mismatch Scores 1,-2

1,-2
1,-3
1,-4
2,-3
4,-5
1,-1

Gap Costs Linear

Existence: 5 Extension: 2
Existence: 2 Extension: 2
Existence: 1 Extension: 2
Existence: 0 Extension: 2
Existence: 3 Extension: 1
Existence: 2 Extension: 1
Existence: 1 Extension: 1

Filters and Masking

Applied only to query seq

Filter ☑ Low complexity regions

Mask regions of low compositional complexity that may cause spurious or misleading results. more...

☐ Species-specific repeats for: Homo sapiens (Human)

Mask ☑ Mask for lookup table only

Mask query while producing seeds used to scan database, but not for extensions. more...

☐ Mask lower case letters

BLAST — Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

My NCBI [Sign In] [Register]

NCBI/ BLAST/ blastp suite

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Reset page | Bookmark

Enter Query Sequence

Enter accession number, gi, or FASTA sequence — Clear

Query subrange: From ____ To ____

Or, upload file — Browse...

Job Title

Enter a descriptive title for your BLAST search

☐ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

- Non-redundant protein sequences (nr)
- Reference proteins (refseq_protein)
- Swissprot protein sequences(swissprot)
- Patented protein sequences(pat)
- Protein Data Bank proteins(pdb)
- Environmental samples(env_nr)

Organism (Optional): Enter organism name or id--completions will be suggested

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Entrez Query (Optional)

Enter an Entrez query to limit search

Program Selection

Algorithm

- (•) blastp (protein-protein BLAST)
- ( ) PSI-BLAST (Position-Specific Iterated BLAST)
- ( ) PHI-BLAST (Pattern Hit Initiated BLAST)

Choose a BLAST algorithm

blastp - compares a protein query to a protein database

PSI-BLAST - allows the user to build a PSSM using the results of the first BlastP run

PHI-BLAST - performs the search but limits alignments to those that match a pattern in the query

BLAST — Search database nr using Blastp (protein-protein BLAST)

☐ Show results in a new window

▶ Algorithm parameters

Algorithm parameters

General Parameters

Max target sequences 100

Select the maximum number of aligned sequences to display

Short queries ☑ Automatically adjust parameters for short input sequences

Expect threshold 10

Word size 3

PAM30
PAM70
BLOSUM80
BLOSUM62
BLOSUM45

Scoring Parameters

Matrix BLOSUM62

Gap Costs Existence: 11 Extension: 1

Existence: 9 Extension: 2
Existence: 8 Extension: 2
Existence: 7 Extension: 2
Existence: 12 Extension: 1
Existence: 11 Extension: 1
Existence: 10 Extension: 1

Compositional adjustments Conditional compositional score matrix adjustment ?

to compensate for the compositions of the two sequences being compared

Filters and Masking

Filter ☐ Low complexity regions

Mask ☐ Mask for lookup table only
☐ Mask lower case letters

Difference between the two options?

BLAST

Search database nr using Blastp (protein-protein BLAST)

Specialised BLAST: PSI-BLAST

- Position Specific Iterated (PSI) BLAST
- Designed to find remote homologues (15 – 25% identity levels)
- Construct scoring matrices by multiple alignment of hits obtained
- Search the database with the new scoring matrix for every iteration
- Iterate until convergence is reached

Specialised BLAST: PSI-BLAST

- The idea of constructing a scoring matrix from the hits is that the new scoring matrix is tailor-made to find sequences similar to the query.

- allows detection of homologues in the range of 15%-25% sequence identity levels.

Specialised BLAST: PHI-BLAST

- Pattern Hit Iterated (PHI) BLAST
- Designed to find motifs
- Given a protein sequence and a motif/pattern within it, the program finds all proteins which carry that pattern and have similar surrounding residues
- Combines regular expression with local alignment around the matching region

BLAST: Statistics

Since BLAST uses a heuristic approach to accelerate aligning of two sequences, it becomes essential to compute the statistical significance of the alignment.

- Compute the probability that the alignment is obtained by "chance".

Strategy adopted by BLAST - compute distribution of alignment scores for random sequences.

- very little known about the random distribution of optimal global alignment scores

Statistics for ungapped local alignment score is quite well understood

BLAST: Statistics

- Model random sequences generated: for proteins choose amino acids randomly with specific probabilities
- Align these random sequences using the Smith-Waterman algorithm and compute the score for the optimal alignment
- It is known that the distribution of the maximum of independent identically distributed (i.i.d.) random variables is an extreme value distribution
- the scores of optimal alignment of random sequences falls in this category

For 10000 pairs of 1000-length protein sequences, 10000 optimal local alignments were computed using Smith-Waterman. From these scores the values for *K* and λ were derived: 0.035 and 0.252

***K* and *λ* are parameters related to the position of maximum and width of the distribution**

BLAST: Statistics

- In the limit of large sequence lengths m and n, expected number of HSPs with score at least S is:

$$E = K(mn)e^{-\lambda S}$$

Dependent on database size

- K and λ are empirical parameters derived from the distribution
- E-value depends on the size of query sequence as well as that of the database.

⇒ for same aligned pair of sequences, E-value may be different depending on the database searched

Rule of thumb – search a larger database whenever possible

Significance of a hit: example

Search against a database of 10,000 sequences.

An extreme-value distribution (blue) is fitted to the distribution of all scores.

It is found that 99.9% of the blue distribution has a score below 112.

This means that when searching a database of 10,000 sequences you'd expect to get 0.1% * 10,000 = 10 hits with a score of 112 or better for random reasons

10 is the E-value of a hit with score 112. You want E-values well below 1!

BLAST: Bit Scores

- Raw scores have little meaning without detailed knowledge of the scoring system used, or more simply its statistical parameters *K* and λ

```
Query:  1  SGLKSLVGKTALLSGTSSKL  20
           SGLKSLVGKTALLSGTSSKL
Sbjct:  1  SGLKSLVGKTALLSGTSSKL  20
```

```
Score =  91
```

```
Query:  1  CQHMWYQWMIQCIWMYHCMQ  20
           CQHMWYQWMIQCIWMYHCMQ
Sbjct:  1  CQHMWYQWMIQCIWMYHCMQ  20
```

```
Score = 138
```

Based on scores $\Rightarrow$ alignment 2 is better than alignment 1, but both the alignments are of the same length and have 100% identity

BLAST: Bit Scores

- Raw score, S, - sum of the alignment's pair-wise scores using a specific scoring matrix.
- By normalizing a raw score using the formula

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

E-value corresponding to a given bit score is

$$E = mn\, 2^{-S'}$$

Note: λ and K are parameters dependent upon the scoring system employed.

BLAST: Bit Scores

- Bit score is calculated from the raw score by normalizing with the statistical variables that define a given scoring system.
- effect of normalization is to change the score distribution into standard normal distribution
- So, bit scores from different alignments, even those employing different scoring matrices can be compared.
- Higher the score, better the alignment, but the significance of an alignment can not be deduced from the score alone.

Using BLAST: Inferring Homology

How do you know when a certain level of similarity implies homology?

Does doing multiple searches help?

Using BLAST: Inferring Homology

Rules of thumb expressed in terms of percent identity in the optimal alignment:

- ~ 45% identity - the proteins will have very similar structures and are very likely to have a common or at least a similar function.
- ~ 25% identity - they are likely to have a similar general folding pattern.
- 18–25% identity - defined as 'twilight zone' - lower degree of sequence similarity cannot rule out homology.

Using BLAST: Inferring Homology

- In general, two sequences significantly similar over the entire length are likely to be homologous
- 50% similarity over a short sequence often occurs by chance
- Low complexity regions can be highly similar without being homologous
- Homologous sequences are not always highly similar
- Suggested BLAST cutoffs
  - For nucleotide-based searches, look for hits with e-value $\leq 10^{-6}$ and sequence identity $\geq 70\%$
  - For protein-based searches, look for hits with e-value $\leq 10^{-3}$ and sequence identity $\geq 25\%$

Substitution Matrices

Importance of scoring matrices

- **Scoring matrices appear in all analysis involving sequence comparison.**
- **Choice of matrix can <u>strongly</u> influence the outcome of the analysis.**
- **Scoring matrices <u>implicitly</u> represent a particular theory of evolution.**
- **Understanding theories underlying a given scoring matrix can aid in making a proper choice**

Substitution Matrices for Nucleotides

BLAST Matrix: In general, different substitution matrices are tailored to detecting similarities among sequences that are diverged by differing degrees, e.g.

	A	T	C	G
A	2	-3	-3	-3
T	-3	2	-3	-3
C	-3	-3	2	-3
G	-3	-3	-3	2

default – blastn

	A	T	C	G
A	1	-3	-3	-3
T	-3	1	-3	-3
C	-3	-3	1	-3
G	-3	-3	-3	1

default - megablast

- Default scoring scheme for blastn target sequences that are **90% identical,** while the default scoring matrix for megablast is appropriate for sequences that are **99% identical,**

Substitution Matrices for Nucleotides

Distance matrix:

Another measure, called edit distance, or cost is also used, e.g., Transition/Transversion Matrix:

	A	T	C	G
A	0	5	5	1
T	5	0	1	5
C	5	1	0	5
G	1	5	5	0

Using a Transition/Transversion matrix reduces noise in comparisons of distantly related sequences

Substitution Matrices For Nucleotides

ATGCCGTGATAGTCGAT
ACGGCTCGATCTACTAC

Identity Matrix: Score: 8 1's = **8**

Blast Matrix: Score: 8 x 1 – 3 x 9 = 8 – 27 = **- 19**

Transition/Transversion Matrix: (distance/cost)

Score: 8 x 0 + 3 x 1 + 6 x 5 = **33**

Which is better?

It is clear that raw alignment scores are meaningless without specific knowledge of the scoring matrix used.

PAM Matrices

PAM units & PAM matrices - developed by Margaret Dayhoff and coworkers.

- examined 1572 accepted mutations between 71 families of closely related sequences of proteins and noticed that the substitutions were not random.

⇒ evolutionarily related proteins need not have same AA at every position: can have a comparable one.

PAM stands for "Point Accepted Mutations" or "Percent of Accepted Mutations" ("accepted" refers to mutations that have become fixed in the population)

PAM matrices refer to various degrees of sensitivity depending on the evolutionary distance between sequence pairs

PAM Matrices

PAM units measure the amount of evolutionary distance between two protein sequences.

One PAM of evolution means that the total number of substitutions is 1% of the sequence length

After 100 PAMs of evolution, not every position would have changed, because some positions will have mutated several times, perhaps returning to their original state

In fact, even after 250 PAMs, proteins are still sufficiently similar that sequence homology can frequently be detected.

PAM Matrices

- **If changes were purely random**
  - **Frequency of each possible substitution would be proportional to background frequencies**
- **In related proteins:**
  - **Observed substitution frequencies called the target (replacement) frequencies are biased toward those that do not disrupt the protein's function**
  - **These point mutations are "accepted" during evolution**
- **Log-odds approach:**
  - **Scores proportional to the natural log of the ratio of target frequencies to background frequencies**

PAM Matrices

Score matrix entry for time *t* given by:

$$s(a,b|t) = \log \frac{P(a|b,t)}{q_a q_b}$$

Conditional probability that *b* is substituted by *a* in *t*

Frequency of AAs *a* & *b*

$$P(a|b) = P(a \cap b)/P(b)$$

PAM Matrices Construction

- Based on the hypothesis that proteins diverge by accumulating uncorrelated mutations
- Align closely related sequences (> 85% identity)
- considering very similar sequences allow the correct alignments to be determined with high certainty
- Observe the probability of AA changes & compute the log-odds ratio
- Normalize the matrix (relative frequencies of various mutations multiplied by a carefully chosen constant) to give an average change of 1% of all positions to obtain PAM-1 matrix

PAM Matrices Construction

How to derive scoring matrices for distantly related sequences from data about closely related sequences?

- **Scoring matrices to <u>any</u> PAM distance can be determined by extrapolating from PAM 1 – by successive iteration of a reference mutation matrix:**

$$M_n = (M_1)^n$$

e.g., for PAM 250, multiply PAM-1 250 times with itself.

Assumption in this evolutionary model is that AA substitutions observed over short periods of evolutionary history can be extrapolated to longer distances.

PAM Matrices Construction

$$M_n = (M_1)^n$$

M_1 - matrix reflecting 99% sequence conservation and one accepted point mutation (PAM 1) per 100 residues

M_n - substitution probabilities after *n* PAMs

PAM250 – most frequently used matrix that is geared to very distant, but still detectable homologies

PAM matrices are derived from global alignments of closely related sequences

Table 1 – The log odds matrix for 250 PAMs (multiplied by 10)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	2	-2	0	0	-4	1	-1	-1	-1	-2	-1	0	1	0	-2	1	1	0	-6	-3
C		12	-5	-5	-4	-3	-3	-2	-5	-6	-5	-4	-3	-5	-4	0	-2	-2	-8	0
D			4	3	-6	1	1	-2	0	-4	-3	2	-1	2	-1	0	0	-2	-7	-4
E				4	-5	0	1	-2	0	-3	-2	1	-1	2	-1	0	0	-2	-7	-4
F					9	-5	-2	1	-5	2	0	-4	-5	-5	-4	-3	-3	-1	0	7
G						5	-2	-3	-2	-4	-3	0	-1	-1	-3	1	0	-1	-7	-5
H							6	-2	0	-2	-2	2	0	3	2	-1	-1	-2	-3	0
I								5	-2	2	2	-2	-2	-2	-2	-1	0	4	-5	-1
K									5	-3	0	1	-1	1	3	0	0	-2	-3	-4
L										6	4	-3	-3	-2	-3	-3	-2	2	-2	-1
M											6	-2	-2	-1	0	-2	-1	2	-4	-2
N												2	-1	1	0	1	0	-2	-4	-2
P													6	0	0	1	0	-1	-6	-5
Q														4	1	-1	-1	-2	-5	-4
R															6	0	-1	-2	2	-4
S																2	1	-1	-2	-3
T																	3	0	-5	-3
V																		4	-6	-2
W																			17	0
Y																				10

Most / Least Mutable AAs?

Elements of matrix are multiplied by 10

Non-diagonal +ve values represent evolutionarily conservative replacements

PAM Matrices

Salient Points:

- Matrices for greater evolutionary distances are **extrapolated** from those for lesser ones
- **Number** with the matrix (PAM40, PAM100, PAM250) refers to the **evolutionary distance**; larger numbers represent greater distances
- **No clear correspondence** between PAM distance and evolutionary time, since different protein families evolve at different rates
- Does not take into account **different evolutionary rates** between conserved & non-conserved regions.

BLOSUM – BLocks SUbstitution Matrix

BLOSUM matrix – given by Henikoff and Henikoff (1992)

- **Uses an alternative approach to determine a family of scoring matrices**
- **Uses structurally conserved protein domains from the BLOCKS database, which contains ungapped multiple alignments, called blocks, of core regions from hundreds of proteins**
- **Directly tabulates frequencies $p(x,y)$ for distantly related proteins, instead of having a need to extrapolate from observation**

BLOSUM Matrices

Each matrix is tailored to a particular evolutionary distance, *viz*., for BLOSUM62 matrix, the default matric for BLAST:

In each block, sequences sharing at least 62% AA identity are clustered together

Then frequencies of aligned pairs $p(x, y)$ counted

Sequences more identical than 62% are represented by a single sequence in the alignment so as to avoid over-weighting closely related family members.

BLOSUM50 Estimation

- For each alignment in BLOCKS db, the sequences are grouped into clusters with at least 50% identical residues
- All pairs of sequences are compared, observed pair frequencies noted (e.g., A aligned with A makes up 1.5% of all pairs, A aligned with C makes up 0.01% of all pairs, etc.)
- Expected pair frequencies are computed from single AA frequencies, e.g,

  $f_{A,C}=f_A \times f_C=7\% \times 3\% = 0.21\%$
- For each AA pair substitution scores are computed as:

$$\log \frac{\text{Pair_freq (obs)}}{\text{Pair_freq (expected)}}$$

- $S_{i,j}$ are scaled to give an integer value

```
ID   FIBRONECTIN_2; BLOCK
COG9_CANFA  GNSAGEPCVFPFIFLGKQYSTCTREGRGDGHLWCATT
COG9_RABIT  GNADGAPCHFPFTFEGRSYTACTTDGRSDGMAWCSTT
FA12_HUMAN  LTVTGEPCHFPFQYHRQLYHKCTHKGRPGPQPWCATT
HGFA_HUMAN  LTEDGRPCRFPFRYGGRMLHACTSEGSAHRKWCATTH
MANR_HUMAN  GNANGATCAFPFKFENKWYADCTSAGRSDGWLWCGTT
MPRI_MOUSE  ETDDGEPCVFPFIYKGKSYDECVLEGRAKLWCSKTAN
PB1_PIG     AITSDDKCVFPFIYKGNLYFDCTLHDSTYYWCSVTTY
SFP1_BOVIN  ELPEDEECVFPFVYRNRKHFDCTVHGSLFPWCSLDAD
SFP3_BOVIN  AETKDNKCVFPFIYGNKKYFDCTLHGSLFLWCSLDAD
SFP4_BOVIN  AVFEGPACAFPFTYKGKKYYMCTRKNSVLLWCSLDTE
SP1_HORSE   AATDYAKCAFPFVYRGQTYDRCTTDGSLFRISWCSVT
COG2_CHICK  GNSEGAPCVFPFIFLGNKYDSCTSAGRNDGKLWCAST
COG2_HUMAN  GNSEGAPCVFPFTFLGNKYESCTSAGRSDGKMWCATT
COG2_MOUSE  GNSEGAPCVFPFTFLGNKYESCTSAGRNDGKVWCATT
COG2_RABIT  GNSEGAPCVFPFTFLGNKYESCTSAGRSDGKMWCATS
COG2_RAT    GNSEGAPCVFPFTFLGNKYESCTSAGRNDGKVWCATT
COG9_BOVIN  GNADGKPCVFPFTFQGRTYSACTSDGRSDGYRWCATT
COG9_HUMAN  GNADGKPCQFPFIFQGQSYSACTTDGRSDGYRWCATT
COG9_MOUSE  GNGEGKPCVFPFIFEGRSYSACTTKGRSDGYRWCATT
COG9_RAT    GNGDGKPCVFPFIFEGHSYSACTTKGRSDGYRWCATT
FINC_BOVIN  GNSNGALCHFPFLYNNHNYTDCTSEGRRDNMKWCGTT
FINC_HUMAN  GNSNGALCHFPFLYNNHNYTDCTSEGRRDNMKWCGTT
FINC_RAT    GNSNGALCHFPFLYSNRNYSDCTSEGRRDNMKWCGTT
MPRI_BOVIN  ETEDGEPCVFPFVFNGKSYEECVVESRARLWCATTAN
MPRI_HUMAN  ETDDGVPCVFPFIFNGKSYEECIIESRAKLWCSTTAD
PA2R_BOVIN  GNAHGTPCMFPFQYNQQWHHECTREGREDNLLWCATT
PA2R_RABIT  GNAHGTPCMFPFQYNHQWHHECTREGRQDDSLWCATT
```

$$S_{A,C} = \log \frac{0.01}{0.21} = -1.3$$

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	5																			
R	-2	7																		
N	-1	-1	7																	
D	-2	-2	2	8																
C	-1	-4	-2	-4	13															
Q	-1	1	0	0	-3	7														
E	-1	0	0	2	-3	2	6													
G	0	-3	0	-1	-3	-2	-3	8												
H	-2	0	1	-1	-3	1	0	-2	10											
I	-1	-4	-3	-4	-2	-3	-4	-4	-4	5										
L	-2	-3	-4	-4	-2	-2	-3	-4	-3	2	5									
K	-1	3	0	-1	-3	2	1	-2	0	-3	-3	6								
M	-1	-2	-2	-4	-2	0	-2	-3	-1	2	3	-2	7							
F	-3	-3	-4	-5	-2	-4	-3	-4	-1	0	1	-4	0	8						
P	-1	-3	-2	-1	-4	-1	-1	-2	-2	-3	-4	-1	-3	-4	10					
S	1	-1	1	0	-1	0	-1	0	-1	-3	-3	0	-2	-3	-1	5				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	2	5			
W	-3	-3	-4	-5	-5	-1	-3	-3	-3	-3	-2	-3	-1	1	-4	-4	-3	15		
Y	-2	-1	-2	-3	-3	-1	-2	-3	2	-1	-1	-2	0	4	-3	-2	-2	2	8	
V	0	-3	-3	-4	-1	-3	-3	-4	-4	4	1	-3	1	-1	-3	-2	0	-3	-1	5

BLOSUM50 matrix:

- **Positive scores on diagonal (identities)**
- **Similar residues get positive scores (marked in red)**
- **Dissimilar residues get smaller (negative) scores**

BLOSUM 62 Matrix

Table 2 - The log odds matrix for BLOSUM 62

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	4	0	-2	-1	-2	0	-2	-1	-1	-1	-1	-2	-1	-1	-1	1	0	0	-3	-2
C		9	-3	-4	-2	-3	-3	-1	-3	-1	-1	-3	-3	-3	-3	-1	-1	-1	-2	-2
D			6	2	-3	-1	-1	-3	-1	-4	-3	1	-1	0	-2	0	-1	-3	-4	-3
E				5	-3	-2	0	-3	1	-3	-2	0	-1	2	0	0	-1	-2	-3	-2
F					6	-3	-1	0	-3	0	0	-3	-4	-3	-3	-2	-2	-1	1	3
G						6	-2	-4	-2	-4	-3	0	-2	-2	-2	0	-2	-3	-2	-3
H							8	-3	-1	-3	-2	1	-2	0	0	-1	-2	-3	-2	2
I								4	-3	2	1	-3	-3	-3	-3	-2	-1	3	-3	-1
K									5	-2	-1	0	-1	1	2	0	-1	-2	-3	-2
L										4	2	-3	-3	-2	-2	-2	-1	1	-2	-1
M											5	-2	-2	0	-1	-1	-1	1	-1	-1
N												6	-2	0	0	1	0	-3	-4	-2
P													7	-1	-2	-1	-1	-2	-4	-3
Q														5	1	0	-1	-2	-2	-1
R															5	-1	-1	-3	-3	-2
S																4	1	-2	-3	-2
T																	5	0	-2	-2
V																		4	-3	-1
W																			11	2
Y																				7

Default matrix in BLAST

BLOSUM Matrices

Salient Points:

- Derived from local, ungapped alignments of distantly related sequences
- Uses blocks of protein sequence fragments from different families (the BLOCKS database)
- Blocks represent structurally conserved regions
- Amino acid pair frequencies calculated by summing over all possible pairs in block
- Different evolutionary distances are incorporated into this scheme with a clustering procedure (identity over particular threshold = same cluster)

BLOSUM Matrices

- All matrices are directly calculated; no extrapolations are used – no explicit model
- Number after the matrix (BLOSUM62) refers to the minimum percent identity of the blocks used to construct the matrix; greater numbers represent lesser distances.
- BLOSUM 62 is the default matrix in BLAST
- For database searches, Blosum matrices have often been the better performers

- the reason being these matrices are based on the replacement patterns found in more highly conserved regions of the sequences.

BLAST: Nucleotide Scoring Matrix

- **Scoring Matrices for match (M)/mismatch (N):**
  - **1, -2**
  - **1, -3**
  - **1, -4**
  - **2, -3**
  - **4, -5** ⟵ default
  - **1, -1**

Relative magnitudes of M & N determines the No. of nucleic acid PAMs (point accepted mutations per 100 residues) for which they are most sensitive at finding homologs.

BLAST: Nucleotide Scoring Matrix

A ratio of 0.33 (1/-3) is appropriate for sequences that are about 99% conserved

A ratio of 0.5 (1/-2) is best for sequences that are 95% conserved

A ratio of one (1/-1) is best for sequences that are 75% conserved

- the (absolute) reward/penalty ratio should be increased as one looks at more divergent sequences

Multiple Sequence Alignment (MSA)

Multiple Sequence Alignment

Most important contribution of MB to evolutionary analysis is the discovery that DNA sequences of different organisms are often related.

i.e., genes are conserved across widely divergent species, often performing a similar or even identical function, and at other times, mutating or rearranging to perform an altered function.

Through simultaneous alignment of gene sequences, sequence patterns that have been subject to alteration may be analyzed.

Multiple Sequence Alignment

Aligning more than two sequences

A C - - B C D B

- C A D B - D -

A C A - B C D -

In an MSA, homologous residues among a set of sequences are aligned together in columns.

'Homologous' is meant in both the structural and evolutionary sense.

Motivation for MSA

- MSA helps identify conserved regions and those allowed to vary – regions resistant to change are functionally most important to the molecule.
- Carries more information than mere pair-wise alignment
- Multiple sequence similarity suggests common structure for the protein, a common function or evolutionary origin
- MSA requirements are different in the various applications

Motivation for MSA

Multiple alignments can improve pairwise alignments:

```
(A) p110α        TFILGIGDRHNSNIMVKDDG-QLFHIDFGHFLDHKKKKFGYKRERVPFVLT--QDFLIVI 142
    cAMP-kinase  QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPE 179

(B) p110β        SYVLGIG----------DRHSDNINVKKTGQLFHIDFGHILGNFKSKFGIKRERVPFILT 136
    p110δ        TYVLGIG----------DRHSDNIMIRESGQLFHIDFGHFLGNFKTKFGINRERVPFILT 136
    p110α        TFILGIG----------DRHNSNIMVKDDGQLFHIDFGHFLDHKKKKFGYKRERVPFVLT 135
    p110γ        TFVLGIG----------DRHNDNIMITETGNLFHIDFGHILGNYKSFLGINKERVPFVLT 135
    p110_dicti   TYVLGIG----------DRHNDNLMVTKGGRLFHIDFGHFLGNYKKKFGFKRERAPFVFT 135
    cAMP-kinase  QIVLTFEYLHSLDLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCG--TPEYLA 177
```

Catalytic domains of 5 P13-kinases and cAMP-dependent protein kinase

MSA for DNA Sequences

In DNA sequences MSA is used in

- Genome sequence assembly - shotgun sequencing
- Discovering new regulatory elements
- Inferring evolutionary relationships
- DNA barcoding
- SNP identification
- Develop primers & probes - use conserved regions to develop
  - Primers for PCR
  - Probes for DNA microarrays

In which of these applications do we look for similarity/differences?

MSA for Protein Sequences

In protein sequences, MSA is used in

- Homology modeling of proteins
- Building phylogenetic tree
- Constructing scoring matrices – PAM, BLOSUM
- Predicting secondary & tertiary structures of new sequences
- Identifying conserved patterns, motifs, blocks in protein sequences - to characterize protein families
- Identify related proteins in database searches, e.g., Profiles, PSI-BLAST, HMMs

MSA

Visual alignment of MSA tables use different colours for displaying AAs of different physico-chemical type - aids in identifying conserved patterns:

Colour	Residue Type	Amino acids
Yellow	Small nonpolar	Gly, Ala, Ser, Thr
Green	Hydrophobic	Cys, Val, Ile, Leu, Pro, Phe, Tyr, Met, Trp
Magenta	Polar	Asn, Gln, His
Red	Negatively charged	Asp, Glu
Blue	Positively charged	Lys, Arg

Structure prediction tools also give more reliable results when based on MSAs than on single sequences.

16
19
Synechocystis_sp._PCC_6803/1-40
Halothece_sp._PCC_7418/1-42
Cyanothece_sp._PCC_8802/1-42
Microcystis_aeruginosa_DIANCHI905/1-42
Microcystis_aeruginosa_NIES-843/1-56
Cyanothece_sp._ATCC_51472/1-42
cyanobacterium_UCYN-A/1-42
Crocosphaera_watsonii_WH_8501/1-42
Microcoleus_sp._PCC_7113/1-42
Cyanothece_sp._PCC_7822/1-39
Chamaesiphon_minutus_PCC_6605/1-41
Thermosynechococcus_elongatus_BP-1/1-41
Moorea_producens_3L/1-41
Synechococcus_sp._PCC_6312/1-43
Oscillatoria_nigro-viridis_PCC_7112/1-42
Synechococcus_sp._PCC_7002/1-37
Leptolyngbya_sp._PCC_7376/1-39
Trichodesmium_erythraeum_IMS101/1-41
Leptolyngbya_sp._PCC_7375/1-41
Synechococcus_sp._PCC_7335/1-41
Pseudanabaena_sp._PCC_7367/1-42
Cyanobacterium_stanieri_PCC_7202/1-42
Cyanobacterium_aponinum_PCC_10605/1-42
Dactylococcopsis_salina_PCC_830/1-42
Gloeocapsa_sp._PCC_73106/1-42
Nodularia_spumigena_CCY9414/1-50
Oscillatoriales_cyanobacterium_JSC-12/1-43
Anabaena_sp._90/1-48
Coleofasciculus_chthonoplastes_PCC_7420/1-42
Calothrix_sp._PCC_7507/1-49
Synechococcus_sp._CB0101/1-38
Calothrix_sp._PCC_6303/1-52
Synechococcus_elongatus_PCC_7942/1-44
Leptolyngbya_sp._PCC_6406/1-40
Synechococcus_sp._CB0205/1-38
Synechococcus_sp._CC9311/1-39
Nostoc_sp._PCC_7107/1-49
Nostoc_sp._PCC_7120/1-49
'Nostoc_azollae'_0708/1-49
Nostoc_sp._PCC_7524/1-49
Nostoc_punctiforme_PCC_73102/1-49
Cylindrospermum_stagnale_PCC_7417/1-49
Synechococcus_sp._BL107/1-38
Lyngbya_sp._PCC_8106/1-41
Synechococcus_sp._RCC307/1-38
Synechococcus_sp._WH_7805/1-39
Synechococcus_sp._RS9917/1-39
Synechococcus_sp._WH_7803/1-37
Synechococcus_sp._RS9916/1-38

MSA

To be informative a MSA should

- contain a distribution of closely- and distantly-related sequences.

If all closely-related - information contained is largely redundant

⇒ few inferences can be drawn.

If all very distantly-related - difficult to construct an accurate alignment

⇒ quality of results & inferences might be questionable

Ideally, one should have a complete range of similarities, including distantly-related examples linked through chains of close relationships

Inferences from MSA

Some examples:

- Identify highly conserved regions - likely to be essential sites for structure/function, e.g. active site
- Regions rich in insertions/deletions – may correspond to loops/turns in proteins
- Build gene/protein families – use conserved regions to guide search
- Basis for phylogenetic analysis – infer evolutionary relationships between genes

Inferences from MSA

Some examples:

- Secondary structure prediction:

- Conserved pattern of hydrophobicity with spacing 2 with intervening residues more variable and including hydrophilic residues – suggests β-strand

- Conserved pattern of hydrophobicity with spacing 4 suggests a helix

12 glutamyl tRNA reductase sequences
13 - 16
E. coli
Axc
Synts
Cypar
Oleg
Horvu
Arab
Cupep
B. subt.
Chvib
Cljoj
PHD Sec. Pred.
PHD Acc. Pred.
SOPMA Sec. Pred.
SSPRED Sec. Pred.
Conservation
Consensus
β-strand
hydrophobic
Non-hydrophobic (active sites)
JPRED server
38 – 45
α-helix
i, i+3,
i+4, i+7

Relationship between MSA & Phylogenetic Analysis

SeqA N – F L S

SeqB N – F – S

SeqC N KY L S

SeqD N – Y L S

A hypothetical evolutionary tree that could have generated these sequence changes.

MSA of 8 fragments of immunoglobin sequences

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCSSSGFIFSS--YAMYWVRQAPG
LSLTCTVSGTSFDD--YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

Conserved residues, regions, patterns

alternating hydrophobicity pattern, conserved Cys, W

Multiple Alignment: Evaluation

```
VTISCTGSSSNIG-AGNHVKWYQQLPG
VTISCTGTSSNIG--SITVNWYQQLPG
LRLSCSSSGFIFS--SYAMYWVRQAPG
LSLTCTVSGTSFD--DYYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNW--YVDG
ATLVCLISDFYPG--AVTVAW--KADS
AALGCLVKDYFPE--PVTVSW--NS-G
VSLTCLVKGFYPS--DIAVEW--ESNG
```

```
VTISCTGSSSNIGAG-NHVKWYQQLPG
VTISCTGTSSNIGS--ITVNWYQQLPG
LRLSCS-SSGFIFSS-YAMYWVRQAPG
LSLTCT-VSGTSFDD-YYSTWVRQPPG
PEVTCVVVDVSHEDPQVKFNWYVDG--
ATLVCLISDFYPGA--VTVAWKADS--
AALGCLVKDYFPEP--VTVSWNSG---
VSLTCLVKGFYPSD--IAVEWESNG--
```

It is not enough to just get a multiple alignment; we need to score the alignment

Multiple Alignment: Evaluation

A simple way to evaluate a multiple alignment is to evaluate the cost column by column

Sum of Pairs (SP)

$$= \sum_{i<j} D(S_i, S_j)$$

Using unit cost:
mismatch costs 1,
match 0, and
indel costs 1

$$ColumnCost\begin{pmatrix} L \\ L \\ A \\ P \\ G \\ S \\ - \\ G \end{pmatrix} = ?$$

Summing the scores of all possible combinations of AA pairs in a column of MSA

Multiple Alignment: Evaluation

$$ColumnCost\begin{pmatrix} \mathtt{L} \\ \mathtt{L} \\ \mathtt{A} \\ \mathtt{P} \\ \mathtt{G} \\ \mathtt{S} \\ \mathtt{-} \\ \mathtt{G} \end{pmatrix} = 26.$$

Assumes a model for evolutionary change in which any of the sequence could be the ancestor of others

SP Scoring Method

- There are problems with SP scoring system as illustrated in the example:

Sequence	Col. A	Col. B	Col. C
1	…N…	…N…	…N…
2	…N…	…N…	…N…
3	…N…	…N…	…N…
4	…N…	…N…	…C…
5	…N…	…C…	…C…
Score	60	24	9

(Using Blosum62):
N-N: 6, N-C: - 3, C-C: 9

What's the problem?
Score for N = 10 seq

Multiple Alignment: DP

- Pair-wise alignment: DP involves a L x L (L^2) matrix
- Multiple alignment: DP involves an L^N matrix, an N-dimensional hyper-lattice, *N* - no. of sequences of length *L* each, to align simultaneously
- Computationally not feasible: for 5 sequences, each ~100 bp long, 10^{10} matrix elements need to be computed

  - equivalent to a pairwise alignment of two 100,000 bp sequences

The Recursive Relation

For 3 sequences, to assign a value to a node (i,j,k), we need to consider 7 values; for 2 seqs, we needed only 3!

Multiple Sequence Alignment

$\alpha_{i_1,i_2,\ldots,i_N}$ - maximum score of an alignment up to the subsequences ending with $x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}$

Recursive relation for multiple sequences:

$$\alpha_{i_1,i_2,\ldots,i_N} = max \begin{cases} \alpha_{i_1-1,i_2-1,\ldots,i_N-1} + S(x^1_{i_1}, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1,i_2-1,\ldots,i_N-1} + S(-, x^2_{i_2}, \ldots, x^N_{i_N}), \\ \alpha_{i_1-1,i_2,i_3-1,\ldots,i_N-1} + S(x^1_{i_1}, -, \ldots, x^N_{i_N}), \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{i_1-1,i_2-1,\ldots,i_N} + S(x^1_{i_1}, x^2_{i_2}, \ldots, -), \\ \alpha_{i_1,i_2,i_3-1,\ldots,i_N-1} + S(-,-,\ldots,x^N_{i_N}), \\ \cdot \\ \cdot \\ \cdot \\ \alpha_{i_1,i_2-1,\ldots,i_{N-1}-1,,i_N} + S(-, x^2_{i_2}, \ldots, -), \\ \cdot \\ \cdot \\ \cdot \end{cases}$$

Multiple Sequence Alignment

To calculate each entry, need to maximize over all $2^N - 1$ combinations of gaps in a column, excluding the case where all Δ_k are zero.

Introducing the notation Δ_i which is 0 or 1 and define the 'product'

$$\Delta_i \cdot x = \begin{cases} x & if \quad \Delta_i = 1, \\ - & if \quad \Delta_i = 0. \end{cases}$$

Recursion relation can now be written as

$$\alpha_{i_1, i_2, \ldots, i_N} = \max_{\Delta_1 + \ldots + \Delta_N > 0} \left\{ \alpha_{i_1 - \Delta_1, i_2, -\Delta_2 \ldots, i_N - \Delta_N} + S(\Delta_1 \cdot x^1_{i_1}, \Delta_2 \cdot x^2_{i_2}, \ldots, \Delta_N \cdot x^N_{i_N} \right\}$$

For N = 3, 4, 5, and 10, $2^N - 1$ = ?

DP Hyperlattice: Example

Overall score $S(m)$ for an alignment is defined as a sum of scores $S(m_i)$ for each column i: $S(m) = \Sigma_i S(m_i)$

Time and Space complexity

Assuming all sequences of roughly same length L, memory complexity of the multi-dimensional DP algorithm is $O(L^N)$ and time complexity is $O(2^N L^N)$

- impractical for more than a few sequences.

For 6 sequences, each 100bp long, time taken will be $2^6 \times 100^6 \times 10^{-9}$ = 64000 seconds (~ 18 hrs)

Add 2 more sequences of same length and the no. is 2.56×10^9 seconds (over 81 yrs)

Even worse is memory space requirement - 10^{12} for 6 sequences!

Carrilo-Lipman/MSA Algorithm

- Carrillo and Lipman proposed a heuristic method for accelerating the search, implemented in the program called MSA.
- It is based on the property that if the strings are relatively similar, the alignment path would be close to the main diagonal.

⇒ not all values in the multi-dimensional cube need to be calculated.

- Their central idea is - every multiple alignment imposes a pw-alignment on each pair of sequences
- For each pair, it sets the upper bound equal to the cost of the imposed alignment.

MSA can be projected onto the sides of cube, defining no. of positions within the cube to be evaluated
sequence C
sequence B
sequence A
Pairwise alignment
B-C
A-B
A-C

Bounds within which an optimal alignment will be found by MSA
Coloured areas on each cube surface - 2d projections of the gray volume
it is only necessary to calculate optimal alignment scores within the gray volume

Multiple alignment DP: Heuristics

- Optimal alignment path will lie around the main diagonal (MSA, by Carillo-Lipman, 1988)

While visiting a node and looking for the minimum along all the incoming edges - ignore the edges "coming from outside polyhedron". Edges 1, 2, 3, 6, 7 are coming from inside and edges 4 and 5 (not shown) can be ignored

Multiple alignment DP: Heuristics

- Optimal alignment path will lie around the main diagonal (MSA, by Carillo-Lipman, 1988)
- It uses a heuristic approach to choose upper bounds for the pairs.
- Works for very similar sequences
- MSA can find optimal multiple alignments for ~ 6-8 sequences, 200-300 length

Alignment by Consensus

- Aim of alignment by Consensus: to minimize the consensus error
- Align sequence 1 & 2 and obtain a consensus sequence
- Building consensus: an example

```
A V – C – D
A V C C – D
A V C D E –

A V C S E D
```

- As is clear from above, the consensus need not be similar to any of the aligned sequences

Consecutive consensus

- Align sequence 3 to the consensus of first two sequences to obtain a new consensus
- Continue until the consensus alignment converges to a global consensus
- Problem - Depends on the order in which the sequences are introduced

Solution ?

Consecutive consensus

- Align sequence 3 to the consensus of first two sequences to obtain a new consensus
- Continue until the consensus alignment converges to a global consensus
- Problem - Depends on the order in which the sequences are introduced
- Solution - one may use a guide tree obtained by p-w alignment scores to build the consensus

Progressive approach

- Align each sequence to every other pair-wise
- Compute distances between each aligned pair (*e.g.* no. of mismatches)
- Construct a phylogenetic tree
- Cluster closely related sequences
- Align closely related sequences first
- Gaps inserted in closely related sequences are propagated throughout

Progressive alignment - involves constructing a succession of pairwise alignments.

Tools: ClustalW,T-Coffee, MUSCLE

Pairwise alignments

```
Sheep    STCVLSAYWKDLNNYH
Cattle   STCVLSAYWKDLNNYH

Sheep    STCVLSAYWK-DLNNYH
Pig      STCVLSAYWRNELNNFH

Sheep    STCVLSAYWKDLNNYH
Human    STCMLGTY-QDFNKFH

Sheep    STCVLSAYWKDLNNYH
Rat      STCMLGTY-QDLNKFH

Sheep    STCVLSAYWKD-LNNYH
Salmon   STCVLGKL-SQELHKLQ

Pig      STCVLSAYWRNELNNFH
Human    STCMLGTY-QD-FNKFH

Pig      STCVLSAYWRNELNNFH
Rat      STCMLGTY-QD-LNKFH

Pig      STCVLSAYWRNELNNFH
Salmon   STCVLGKL-SQELHKLQ

Human    STCMLGTYQDFNKFH
Rat      STCMLGTYQDLNKFH

Human    STCMLGTY-QDFNKFH
Salmon   STCVLGKLSQELHKLQ

Rat      STCMLGTY-QDLNKFH
Salmon   STCVLGKLSQELHKLQ
```

Hierarchy of Addition

Sheep-Cattle	0	Pig-Rat	8
Sheep-Pig	4	Pig-Salmon	10
Sheep-Human	8	Human-Rat	1
Sheep-Rat	7	Human-Salmon	9
Sheep-Salmon	11	Rat-Salmon	8
Pig-Human	9		

- Align Sheep and Cattle first
- Align Human and Rat
- Align Pig to Sheep and Cattle
- Align these two clusters to each other
- …
- Align Salmon to large alignment last

Progressive alignment

```
step 1   Sheep    STCVLSAYWKDLNNYH
         Cattle   STCVLSAYWKDLNNYH

step 2   Human    STCMLGTYQDFNKFH
         Rat      STCMLGTYQDLNKFH

step 3   Sheep    STCVLSAYWK-DLNNYH
         Cattle   STCVLSAYWK-DLNNYH
         Pig      STCVLSAYWRNELNNFH

...      Salmon   STCVLGKLSQE-LHKLQ
```

Aligning Alignments

All possible cases that arise in progressive alignment approach:

- Align two sequences to each other
- Align a sequence to an existing alignment
- Align two alignments to each other

Note - computationally this is always a PW

Aligning Alignments

- Pairwise alignment of alignments is also called <u>profile</u> alignment

Again, we can use DP

$$S(i,j) = \max[S(i-1,j-1)+m(i,j),\ S(i-1,j)+g,\ S(i,j-1)+g]$$

$m(i,j)$ – similarity score averaged over characters at that position, here i and j no longer stands for single sequences, but an average of several, when aligning alignments

g – gap penalty

	S	T	C	V	L	S	A	YWKD-LNNYH	Sheep
	S	T	C	V	L	S	A	YWKD-LNNYH	Cattle
	S	T	C	V	L	S	A	YWRNELNNFH	Pig
SS									
TT									
CC									
MM									
LL									
GG									
TT									
YY									
QQ									
DD									
FL									
NN									

Aligning Alignments

```
Alignment 1:   ATA
               CCA
               ↓
Alignment 2:  TCAFE
              TAT-E
              TATF-
              AGTFD
```

Score 1st column of 1st alignment against 2nd column in the other alignments using:

= 1/8 (score(A,C) + score(A,A) + score(A,A) + score(A,G) + score(C,C) + score(C,A) + score(C,A) + score(C,G))

Aligning Alignments

- Once sequences are aligned & gaps introduced, these are not altered – the alignment method is <u>hierarchical</u>
- ClustalW finds a local optimum as early alignment decisions are "locked in" by the "greedy" algorithm
- Early errors will be propagated and cause the final alignment to be worse

```
Sheep    STCVLSAYWK-DLNNYH
Cattle   STCVLSAYWK-DLNNYH
Pig      STCVLSAYWRNELNNFH

Sheep    STCVLSA--YWK-DLNNYH
Cattle   STCVLSA--YWK-DLNNYH
Pig      STCVLSA--YWRNELNNFH
```

ClustalX/ClustalW

- Performs pair-wise alignments of all the sequences
  - k-tuple based alignment (fast/approx.), full DP (slow/accurate)
- Uses alignment scores to produce a phylogenetic tree
  - Genetic distance: no. of mismatches/no. of matches (positions against gaps not considered)
- Aligns sequences sequentially, guided by the phylogenetic relationships indicated by tree.
  - Sequence contributions are 'weighted' by their relationship on the predicted tree, weights based on the distance of each sequence from root

Alignment scores between two positions in msa calculated using resulting weights as multiplication factors

Weights are normalized so that largest weight is 1

A. Calculation of sequence weights

B. Use of sequence weights

Column in alignment 1

Sequence A (weight a) .....K.....

Sequence B (weight b) .....I.....

Column in alignment 2

Sequence C (weight c) .....L.....

Sequence D (weight d) .....V.....

Score for matching these two column in an msa =

[a x c x score (K,L) +
a x d x score (K,V) +
b x c x score (I,L) +
b x d x score (I,V)] / 4

Basic idea in Progressive Heuristic Approach:

- compute pairwise alignments and merge alignments consistently

Consider alignment of 3 sequecens:

acg, cga, gac

Get optimal pairwise alignments:

a c g -	- a c g	c g a -
- c g a	g a c -	- g a c

Order of merging matters !
Note once a gap, always a gap …

Example

```
C   PADKTNVKAAWGKVGAHAGEYGA
D   AADKTNVKAAWSKVGGHAGEYGA

A   PEEKSAVTALWGKVNVDEYGG
B   GEEKAAVLALWDKVNEEEYGG

C   PADKTNVKAAWG_KVGAHAGEYGA
D   AADKTNVKAAWS_KVGGHAGEYGA
E   AA__TNVKTAWSSKVGGHAPA__A

A   PEEKSAV_TALWG_KVN__VDEYGG
B   GEEKAAV_LALWD_KVN__EEEYGG
C   PADKTNVKAA_WG_KVGAHAGEYGA
D   AADKTNVKAA_WS_KVGGHAGEYGA
E   AA__TNVKTA_WSSKVGGHAPA__A
```

Once a gap, always a gap

Steps in Progressive Alignment

(B) Multiple alignment following the tree from A

ClustalW: Refinements

- Different BLOSUM matrices used based on similarity of sequences – to reflect evolutionary changes better
- Recall: BLOSUM80/ BLOSUM62 are based on sequences that are 80% / 62% identical, *i.e.* lower numbers for more distant sequences
- ClustalW calculates gaps in a way to place them between conserved domains, based on observation that structurally related proteins align in a way to have gaps between secondary structural elements
- Frequency of gaps next to each amino acid computed in these regions is used (Pascarella and Argos)

MSA of 7 globins by CLUSTALW

		1	2	3	4	5	6
Hbb_Human	1	-					
Hbb_Horse	2	.17	-				
Hba_Human	3	.59	.60	-			
Hba_Horse	4	.59	.59	.13	-		
Myg_Phyca	5	.77	.77	.75	.75	-	
Glb5_Petma	6	.81	.82	.73	.74	.80	-
Lgb2_Luplu	7	.87	.86	.86	.88	.93	.90

Pairwise alignment: Calculate distance matrix

Rooted Neighbor Joining tree (guide tree)

Progressive alignment: Align following the guide tree

```
--------VHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLST
--------VQLSGEEKAAVLALWDKVN--EEEVGGEALGRLLVVYPWTQRFFDSFGDLSN
---------VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS--
---------VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLS--
---------VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKT
PIVDTGSVAPLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLTT
--------GALTESQAALVKSSWEEFNANIPKHTHRFFILVLEIAPAAKDLFSFLKGTSE
         *           *                       *     *

PDAVMGNPKVKAHGKKVLGAFSDGLAHLD-----NLKGTFATLSELHCDKLHVDPENFRL
PGAVMGNPKVKAHGKKVLHSFGEGVHHLD-----NLKGTFAALSELHCDKLHVDPENFRL
----HGSAQVKGHGKKVADALTNAVAHVD-----DMPNALSALSDLHAHKLRVDPVNFKL
----HGSAQVKAHGKKVGDALTLAVGHLD-----DLPGALSNLSDLHAHKLRVDPVNFKL
EAEMKASEDLKKHGVTVLTALGAILKKKG-----HHEAELKPLAQSHATKHKIPIKYLEF
ADQLKKSADVRWHAERIINAVNDAVASMDDT--EKMSMKLRDLSGKHAKSFQVDPQYFKV
VP--QNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-VADAHFPV
          *                                   *    *

LGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
LGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH------
LSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR------
ISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
LAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
VKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
```

Known locations of 7 α-helices in the structure of this group shown in boxes

Problems with Progressive Alignment

Gaps at the ends are penalized less, so CAT is aligned with FAT in sequence 2

The greedy approach results in efficiency of the algorithm at the cost of accuracy

```
CLUSTALW (Score=20, Gop=-1, Gep=0, M=1)

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST CA-T ---
SeqC GARFIELD THE VERY FAST CAT
SeqD -------- THE ---- FA-T CAT
```

```
CORRECT (Score=24)

SeqA GARFIELD THE LAST FA-T CAT
SeqB GARFIELD THE FAST ---- CAT
SeqC GARFIELD THE VERY FAST CAT
SeqD -------- THE ---- FA-T CAT
```

Problems with Progressive Alignment

- Depends on the very first closely related sequences used for constructing the multiple alignment
- If these sequences align well, there will be few errors
- More distantly related these sequences are, more errors will get propagated through the alignment
- Using Bayesian methods such as Hidden Markov models (HMMs) may be useful for aligning more distantly related sequences.

Solution: Stochastic or Iterative Methods

ClustalW Misapplied

ClustalW and other algorithms that include an initial pair-wise comparison step should not be used to align sequences that do not all share a common block.

Not suitable for sequence assembly!

ClustalW / ClustalΩ online

- **EBI:**
  https://www.ebi.ac.uk/Tools/msa/

The latest version is called Clustal Omega - uses seeded guide trees and HMM profile-profile techniques to generate alignments

- one of the fastest multiple sequence alignment tools

- **DDBJ:**
  http://clustalw.ddbj.nig.ac.jp/

Iterative Methods of MSA

- To correct for errors introduced by initial alignment, use iterative methods: re-align sub-groups of sequences and then align these sub-groups into a global alignment
- Objective – to improve overall alignment score
- Selection of groups may be based on phylogenetic tree, separation of one or two sequences from the rest, or a random selection of the groups.
- Programs using iterative methods – MultiAlin, PRRP and DIALIGN

Other Methods

Genetic Algorithm: It is a machine learning approach, implemented in the program SAGA, and produces alignments by attempted simulation of the evolutionary changes in sequences.

Basic idea behind this method is to try to generate many different MSAs by rearrangements that simulate gap insertion & recombination events during replication in order to generate a higher and higher score for the MSA.

Hidden Markov Models (HMM): HMMs are trained to recognize a specific family of proteins; require different HMM for each protein family

Examine alignments by eye

Some artifacts observed in the output of algorithms:

Always examine your alignment manually to see if it can be improved.

Edit alignments manually

Multiple sequence alignment tools

- Viewers
  - ClustalX, Jalview, Cinema, Sequence logos
- Editors / annotation
  - SeqVu, MACAW, BioEdit
- BioEdit available at:
  https://bioedit.software.informer.com/7.2/

Work with proteins

- Twenty symbols to match as against four for DNA
- No noise resulting from the degeneracy of genetic code
- More sensitive scoring matrices
- Requires less of manual editing

Choose genes judiciously

- When inferring phylogeny choose genes carefully
- For closely related organisms choose genes which mutate fast
- For distantly related species choose slowly mutating genes
- Compare orthologous genes between species and paralogous ones within an organism

Summary

- Treat the output of multiple alignment programs as a first alignment
- Examine it by eye and edit it manually to improve it
- Get rid of low confidence or highly divergent regions
- Ensure you have started with a sensible evolutionary hypothesis

Summary

How does one perform an MSA?

- By hand: too hard!
- Automated alignment: Fast, but doesn't necessarily produce the "correct" alignment

Best approach = Automated alignment with manual editing

References

- David W. Mount, Bioinformatics: Sequence and Genome Analysis, CBS Publishers & distributors, New Delhi, and references therein.
- Thompson, J.D., Higgins, D.G. and Gibson. T.J. "CLUSTALW: improving the sensitivity of multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice" *Nuc. Acids Res.* 22, 4673-80 (1994).
- Carrillo, H. and Lipman, D. "The multiple sequence alignment problem in biology" *SIAM J. Appl. Math.* 48, 1073-82 (1988).

Molecular Phylogeny

Ernst Haeckel, who coined the word ‘phylogeny’ illustrated evolutionary relationships using trees

Tree of Life, Ernst Haeckel, 1866

Phylogeny by Anatomy

- Before the birth of molecular biology, phylogeny was derived by looking for homologous anatomical structures or pattern of embryonic development

Anatomical Homology

Embryonic Development of Vertebrates

Reconstructing a tree

Molecular homology

Number of mutations between the human haemoglobin alpha chain (142 amino acids long) and

Human beta chain	0
Gorilla	1
Gibbon	2
Rhesus monkey	8
Dog	15
Horse, cow	25
Mouse	27
Gray kangaroo	38
Chicken	45
Frog	67
Lamprey	125
Sea slug (a mollusk)	127
Soybean (leghemoglobin)	124

Now phylogeny is based on proteins & DNA sequences

Molecular Phylogeny

- Mol. Phylogeny is the study of evolution of biological sequences (nucleic acid / protein)
- An MSA is the first step for phylogenetic analysis
- Aligned sequences are represented as outer branches on an evolutionary tree
- In a gene family phylogenetic relationships
  - one can determine, family membership and how the family might have evolved over time
  - helps in identifying genes with equivalent function, (can then be tested by genetic experiments)

Molecular Phylogeny

- Phylogenetic analysis may also be used to follow changes occurring in a rapidly changing species *e.g.* influenza
  - Study rapidly changing genes
  - Next year's strain can be predicted
  - Flu vaccination can be developed
- Analysis of the types of changes (synonymous / nonsynonymous) within a population can reveal, for example, whether a particular gene is under selection

Concept of Evolutionary Tree

- 2-dimensional graph showing evolutionary relationship between organisms, or genes from organisms
- Each sequence referred to as taxon, defined as phylogenetically distinct unit
- Taxa lie as outer branches (leaves) of tree, while internal nodes and branches represent relationship between taxa;

  - closely related taxa are found on neighbouring branches

Concept of Evolutionary Tree

- Branch length indicates the No. of mutations (a measure of evolutionary time) before the next level of separation
- Internal nodes represent the splitting of evolutionary path of the gene into two different species that are reproductively isolated
- Molecular clock hypothesis assumes uniform rate of mutation, suitable for closely related species

  – represented by equal branch lengths from the common ancestor

Concept of Evolutionary Tree

- Trees can be rooted or unrooted
- To construct a rooted tree include a taxon that we are reasonably sure branched off earlier than others, also termed as <u>outgroup</u>
- The sum of all branch lengths is termed as the <u>length</u> of the tree
- Trees are usually binary with two branches bifurcating at each node; may have more branches if taxa are too close to be resolved

Example: Rooted & Unrooted

Rooted

Unrooted

Example: Scaled & UnScaled

Concept of Evolutionary Tree

Ultrametricity

All tips are at equal distance from the root.

$$a = b + c + d + e$$

Additivity

Distance between any two tips equals the total branch length between them.

$$\overline{XY} = a + b + c + d + e$$

Tree reconstruction

- Tree reconstruction becomes difficult if all possible trees have to be scored before deciding on the optimal tree
- For three species there are only three possible trees

Tree reconstruction

- With 4 species the No. of rooted trees becomes 15, 105 for 5 species and over 13 billion for 12 species
- For *n* species, the No. of rooted trees is given by

$$(2n - 3)!/(2^{(n-2)}[n - 2]!)$$

- For unrooted trees, the No. is significantly smaller (1 for n=3, 3 for n=4, 15 for n=5, and over half a billion for n=12):

$$(2n - 5)!/(2^{(n-3)}[n - 3]!)$$

- This problem has been shown to be NP-hard

Number of possible evolutionary trees to consider as a function of number of sequences

Taxa or sequence no.	No. of rooted trees	No. of unrooted trees
3	3	1
4	15	3
5	105	15
—	—	—
7	10,395	954

Terminology

- **Tree topology: ordering of species independent of branch lengths**

Same topology

Different topology

METHODS

Choose set of related sequences.[1]
Obtain multiple sequence alignment (Chapter 4).[2]
Is there strong sequence similarity?[3]
Yes
Maximum parsimony methods
No
Is there clearly recognizable sequence similarity?[4]
Yes
Distance methods
No
Maximum likelihood methods[5]
Analyze how well data support prediction.[6]

Methods for Phylogeny

- Maximum parsimony approach is used for closely related sequences, and the variation is similar among all pairs of sequences
- Distance methods are used when more variation and intermediate level of similarity is present
- Maximum likelihood methods work for any set of sequences, but they are useful when sequences are very variable

Maximum Parsimony Method

- Predicts trees that minimize the No. of steps to generate observed variation
- For each aligned position, trees that require smallest no. of evolutionary steps to explain the observed changes are identified
- Finally the trees that produce the smallest no. of changes overall for all aligned positions are identified
- Works best for a small no. of closely related sequences

parsimony score = 4

Maximum Parsimony Method

- Maximum Parsimony method does not use any model of molecular evolution.
- It is based on an implicit assumption that mutation is rare, and the best explanation of evolutionary history is the one that requires the least mutation.

Phylip Programs: Parsimony

- **dnapars**: treats gaps as 5th state
- **dnapenny**: uses branch and bound method
- **dnacomp**: based on compatibility criteria; finds tree that supports largest number of sites
- **dnamove**: performs parsimony and compatibility interactively
- **protpars**: based on no. of mutations to change a codon for aa1 to codon for aa2 for non-synonymous changes only

Distance Methods

- Uses No. of changes between pairs of sequences in a group to construct a tree
- Sequences with fewest changes are neighbours, *i.e.* they share a node to which they are joined by a branch
- Aim is to position neighbours correctly and to compute branch lengths that best fit the data
- Methods for tree construction: UPGMA, Fitch-Margoliash, Neighbour-Joining, etc., based on distance matrix that gives substitution rates.

Distance Methods

- Distance is computed based on conditional probability of observing a change.
- A simplest model, called Jukes-Cantor assumes all substitutions have same rate of mutation, while more complex one define different rates for each type of substitution.

$$M = \begin{pmatrix} p_{A|A} & p_{A|G} & p_{A|C} & p_{A|T} \\ p_{G|A} & p_{G|G} & p_{G|C} & p_{G|T} \\ p_{C|A} & p_{C|G} & p_{C|C} & p_{C|T} \\ p_{T|A} & p_{T|G} & p_{T|C} & p_{T|T} \end{pmatrix}$$

Ancestral base

Descendent base

Jukes-Cantor model

Transition matrix for Jukes-Cantor model:

$$M = \begin{pmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{pmatrix}$$

$$d_{\mathrm{JC}}(\mathrm{S}_0, \mathrm{S}_1) = t\alpha \approx -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Value of α depends on the time step we use and features of the particular DNA sequence being modeled.

α is a probability, but can be interpreted as a rate at which observable base substitutions occur over one time step and is measured in units of

α = (substitutions per site) / time step

Kimura Models

Kimura 2-parameter model allows for different rates for transitions (β) and transversions (γ), while substitution rates to be different:

$$M = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix} \qquad M = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix}$$

with diagonal entries being 1-β-2γ, and 1-β-γ-δ, respectively.

$$d_{K2} = -\frac{1}{2}\ln(1-2p_1-p_2) - \frac{1}{4}\ln(1-2p_2)$$

$$d_{K3} = -\frac{1}{4}(\ln(1-2\beta-2\gamma)+\ln(1-2\beta-2\delta)+\ln(1-2\gamma-2\delta))$$

Tree Construction: UPGMA

Unweighted Pair-Group Method with Arithmetic Means (UPGMA), also called the average distance method.

- assumes that rate of change along the branches of a tree is a **constant**, i.e., it assumes a molecular clock.

	S_1	S_2	S_3	S_4
S_1		.45	.27	.53
S_2			.40	.50
S_3				.62

Distances Between Taxa

Step -1: pick the two closest taxa, S1 and S3, .27 distance apart.

Draw the edges, each of length 0.27/2 = 0.135, equidistant from the common ancestor.

UPGMA

Combine S1 & S3 into a group

	S_1	S_2	S_3	S_4
S_1		.45	.27	.53
S_2			.40	.50
S_3				.62

e.g., distance between S1-S3 and S2 is

(.45 + .40)/2 = .425

The table then collapses to:

	S1-S3	S2	S4
S1-S3		.425	.575
S2			.50

UPGMA

Step-2: Repeat the process, using the collapsed distance table. Because the closest taxa and/or groups in the new table are S1-S3 and S2, .425 apart:

Edge to S2 will have length .425/2 = .2125, while the other new edge will be (.425/2) - .135 = .0775

	S1-S3	S2	S4
S1-S3		.425	.575
S2			.50

UPGMA

Step-3: Again combining taxa, we form a group S1-S2-S3, and compute its distance from S4 by averaging the original distances from S4 to each of S1, S2, and S3: (.53 + .5 + .62)/3 = .55

	S1-S2-S3
S4	.55

Final tree is drawn by estimating S4 as .55/2 = .275 from the root.

The other edge has length 0.275 – 0.2125 = .0625, since that places all other taxa .275 from the root as well.

UPGMA

	S_1	S_2	S_3	S_4
S_1		.45	.27	.53
S_2			.40	.50
S_3				.62

Does the constructed tree exactly fit the data?

Distance on the tree from S3 to S4 =.55, while according to the original data, it is.62!

To an approximation, the tree distances are reasonably close to the distances given by the data

UPGMA

Note: molecular clock assumption is implicit in UPGMA.

In this example, when we placed S1 & S3 at the ends of equal length branches, we assumed that the amount of mutation each underwent from their common ancestor was equal.

UPGMA always places all the taxa at the same distance from the root, so that the amount of mutation from the root to any taxon is identical.

UPGMA always produces a rooted tree.

Phylip Programs: Distance Methods

- **dnadist** is the Phylip program for computing distance among DNA sequences
- **protdist** is the program for computing distances for protein sequences
- **fitch** estimates branch length assuming additivity of branch lengths using Fitch-Margoliash method; molecular clock not assumed
- **kitsch** same as fitch but assumes molecular clock
- **neighbor** estimates phylogeny using the neighbour-joining or UPGMA method

Maximum Likelihood

- Maximum likelihood is similar to maximum parsimony, in that analysis is performed for each column of the alignment, all possible trees are considered, and trees with fewest changes are usually more likely
- ML allows corrections for variations in the mutation rates by considering explicit evolutionary models
- Method can be used to explore relationships among more diverse sequences

Maximum Likelihood

The basic approach of maximum likelihood is:

- Specify a particular model of molecular evolution (e.g., Jukes-Cantor (all substitutions are equally likely), Kimura (different rates for transitions and transversions).
- Consider a specific tree for relating our taxa. Assuming the model of evolution and specific tree are correct, compute the probability that the DNA sequence in our data could have been produced.

- This is the likelihood of the tree, given our data.

Maximum Likelihood

- Compute the likelihood for each tree
- Probability of each tree is the product of mutation rates in each branch
- Likelihoods given by each column are multiplied to give the likelihood of the tree
- Choose the tree with the greatest likelihood as the tree best fitting the data.
- Phylip programs dnaml and dnamlk (same as dnaml except that is assumes a molecular clock) and proml and promlk (for proteins)

Bootstrapping

- Once a tree has been chosen by some method, it would be desirable to quantify how confident one is of it.
- This is given by the statistical technique - bootstrapping.
- In this procedure, the true data sequences are used to create a set of new pseudo-replicate sequences of the same length.
- Bases at a particular site in the new sequences are chosen to be the bases appearing in a randomly chosen site in the original sequences.

Bootstrapping

Initial alignment and tree:

In bootstrap analysis:

- **All sites are considered independent as if freely available in a 'hat' to pick**
- **Available sites are picked up randomly to reconstruct a new alignment of the <u>original size</u> and a new phylogeny**

Bootstrapping

- Process is repeated many times to ascertain the strength of clustering. New “alignment” may contain several sites multiple times while some other sites may be absent - sampling with replacement

Bootstrapping

- Phylogenies are compared to calculate values [Bootstrap value] that signify the number of times a given branch/cluster occurred in the Multiple bootstrap trees
- Higher the value - higher the confidence of phylogenetic inference
- In general values < 50% provide very poor support

Which method is the Best?

One of the difficulties of picking a method is that one can find good arguments for and against them all.

Cautious approach - always use a number of different methods on the data.

Rather than trusting a single method to give an accurate tree, check to see if different methods give roughly the same results.

They often do, and if they do not, it is worth investigating why they don't.

References

- Bioinformatics Sequence & Genome Analysis, David W. Mount
- Biological Sequence Analysis, Probabilistic Models of Proteins and Nucleic Acids, R. Durbin, S.R. Eddy, A. Keoghs and G. Mitchison