

# CS123A Midterm #1 Study Guide

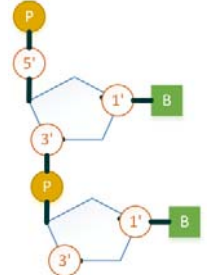
By: Zayd Hammoudeh

<p><b>Central Dogma of Biology</b>  <b>DNA → RNA → Proteins</b></p> <p><b>Single directional flow of information.</b> DNA is <b>replicated</b> into more DNA. DNA is <b>transcribed</b> into RNA. RNA is <b>translated</b> into proteins.</p>	<p><b>DNA – Deoxyribonucleic Acid</b>  <b>Double stranded</b> sequence of nucleotides with a sugar-phosphate backbone. Information Storage.</p> <p><b>Double helix structure</b> (Discovered by Watson and Crick).</p> <p><b>~3.2 billion</b> bases in human genome.</p>	<p><b>RNA – Ribonucleic Acid</b>  <b>Single stranded</b> sequence of nucleotides with a sugar-phosphate backbone.</p> <p>Less regular three dimensional structure than DNA due to hydrogen bonds in complementary sections of the strand.</p>
---	--	---

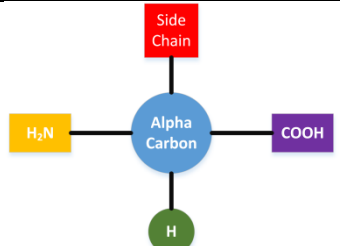
<p><b>Protein</b>  “Building blocks of life”  “We are our proteins”  They put into action the genetic information in DNA.</p> <p><b>Typical length: 300-500</b>  <b>Number of Amino Acids: 20</b></p> <p><b>Protein Function:</b> Determine by its unique 3-dimensional structure.</p> <p><b>Four Structures of a Protein: Primary, Second, Tertiary, Quaternary</b></p>	<p><b>Organism Types</b></p> <p><b>Prokaryote:</b> Single cell organism. No nucleus.  <b>Example:</b> Archea, bacteria</p> <p><b>Eukaryotes:</b> Higher level organism with one or more cells. <b>Have nuclei.</b>  <b>Example:</b> Plants, animals</p>	<p><b>Cell</b>  Contains all genetic instructions for an organism.</p> <p><b>Genome:</b> Entire set of genetic information for an organism.</p> <p><b>Genomics:</b> Field of genome studies.</p> <p><b>Chromosome:</b> Organized DNA clusters in a cell. Come in pairs. <b>Located in the nucleus of eukaryotes.</b></p> <p>Humans have <b>23 pairs of chromosomes.</b></p>	<p><b>Genes</b></p> <p><b>Gene:</b> A specific sequence of nucleotides in a chromosome that encodes a protein.</p> <p>Human genome has <b>about 23,500 genes.</b></p> <p><b>Gene Expression:</b> Gene has been transcribed in mRNA and protein synthesis is occurring.</p>
--	---	---	--

<p><b>Nucleotide</b>  <b>Nitrogen Base and sugar-phosphate molecule.</b></p> <p><b>Pentose:</b> Sugar in nucleotide molecule.</p> <p><b>DNA Nucleotides:</b> Adenine, Guanine, Cytosine, <b>Thymine</b></p> <p><b>RNA Nucleotides:</b> Adenine, Guanine, Cytosine, <b>Uracil</b></p>	<p><b>Nitrogen Base Pairings</b>  <b>Nitrogen base on sugar's 1' carbon.</b></p> <p>Cytosine and Guanine  Adenine and Thymine/Uracil</p> <p><b>Purine:</b> Adenine and Guanine  <b>Pyrimidine:</b> Cytosine and Thymine/Uracil</p> <p><b>Hydrogen bonds:</b> Type of bonds between nucleotides.  <b>3 Hydrogen bonds</b> between Cytosine and Guanine (<b>CG3</b>).  <b>2 Hydrogen bonds</b> between Adenine and Uracil/Thymine.</p>	<p><b>Nucleotide Sugars</b></p> <p><b>Pentose:</b> Classifier for sugar in nucleotide molecule.</p> <p><b>Deoxyribose –</b> Sugar in DNA. Deoxy means no oxygen. <b>Deoxyribose has no oxygen molecule on second carbon (2')</b> (just a hydrogen atom on 2' carbon).</p> <p><b>Ribose –</b> Sugar in RNA. On the 2' carbon, it has an –OH molecule.</p>
--	--	--

## Sugar Phosphate Backbone

<p><b>Phosphodiester Backbone –</b> Name of sugar phosphate backbone in DNA and RNA.</p> <p>The 3' carbon on the sugar bonds with the phosphate molecule on the 5' carbon of the next nucleotide. This bond is known as a <b>phosphodiester linkage</b>.</p> <p>Because of the phosphodiester backbone DNA is laid out as 5' to 3' and complementary 3' to 5'. Hence backbones run in <b>opposite directions</b>.</p> <div style="text-align: center;"> <p>5' ----- 3'</p> <p>     </p> <p>3' ----- 5'</p> </div>	
---	---

## Amino Acids

<p><b>Number of Amino Acids:</b> 20  Function of the amino acid is determined by the <b>side chain</b>.</p> <p><b>Classifications for Amino Acids:</b> Polar/nonpolar. Hydrophobic/hydrophilic. Acidic/basic. Size</p> <p><b>Two identification Schemes for Amino Acids:</b> Single letter capitalized or three letters with first letter capitalized only.</p> <p><b>Linkage Between Amino Acids in a Protein: Peptide bond</b></p> <p><b>Backbone of a Protein: Polypeptide Backbone</b></p>	
--	---

## Transcription

**mRNA** is the result of transcription of DNA. In eukaryotes, mRNA is **spliced** as it migrates from the **nucleus** to the **cytoplasm**.

**Enzyme that Creates mRNA: RNA Polymerase**

The **mRNA** (unspliced) is copy of the **coding/sense** DNA strand. The **template/anticoding/antisense** DNA strand is the complement of the mRNA strand.

**mRNA** is the “carrier of information”. It is what carries genetic information from the DNA in the nucleus to the cytoplasm.

## Translation

**Translation** is the process of converting mRNA genetic information into a protein.

**Codon:** Three nucleotide **triplet** that codes to an amino acid. Codon is 5' to 3'.

**Start Codon: AUG** (Methionine) Carried on **initiator tRNA**.

**Stop Codons: UAA, UAG, UGA. No tRNA for stop codon. A-site** of ribosome recognizes stop codon and release the two subunits of the ribosome.

**Cellular Structure where Translation Occurs:** Ribosome

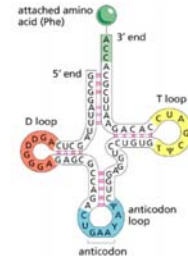
**tRNA:** Has a complementary 3' to 5' **anticodon** that binds to the mRNA codon and the ribosome in protein synthesis. **Clover leaf structure**.

**Anticodon:** Sequence in tRNA molecule that is the complement of the codon. In orientation 3' to 5'.

**Initiator tRNA:** Binds to P-Site of the ribosome to initiate protein synthesis. Usually codes to methionine.

**Ribosome** has **two subunits** (**large** (upper) and **small** (lower)) and **three binding locations** (**E-site**, **P-site**, and **A-site**). tRNA enters the ribosome at the A site and exits at the E-site.

**mRNA Binding Site:** Location in **lower subunit** of ribosome where mRNA binds.



## Protein Structure

**Primary Structure:** Sequence of amino acids that constitute the protein.

**Secondary Structure:** 3-Dimensional folding that is common to all proteins.

**Tertiary Structure:** Further folding and packing of the elements of secondary structure to produce the protein's final 3-D conformation.

**Quaternary Structure:** Multi-subunit protein formed of more than one protein chain.

Ability of a Protein to **interact** with other molecules **depends on its 3-D folding structure** which is dictated by its amino acid sequence.

A single strand of **RNA has 3 reading frames**. A double stranded **DNA has 6 possible reading frames**. A reading frame is from 5' to 3'.

**Eukaryote mRNA Construction Process (Done in Nucleus)**

1. **Transcription**
2. **RNA Capping**
3. **Splicing**
4. **Polyadenylation**

In prokaryotes, there is only transcription.

## RNA Splicing

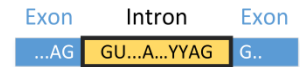
**Exon** – Coding portion of a gene.

**Intron** – Noncoding portion of DNA between exons.

**Typical Intron Cleaving Sequence:**

**AG|GU...AG|G**

**Lariat:** Circular structure of RNA formed by a spliced intron. The loop structure of the lariat binds at an **Adenine** molecule.



**Alternative Splicing:** Some exons or parts of exons are excluded from the RNA splicing.

## Bioinformatics

**Bioinformatics** is a field of study combining **biology**, **computer science**, and **information technology** into a single discipline.

**Goal of Bioinformatics** – Enable the discover of new biological insights and to create a global perspective from which unifying principles in biology can be discerned.

**Map:** Location of each gene on a chromosome.

### Three Major Bioinformatics Databases

1. **Genbank** (National Center for Biotechnology Information) Operated by the National Institutes of Health. Database of all public available DNA sequences.
2. **EBI** (European Bioinformatics Institute)
3. **DDBJ** (DNA Databank of Japan)

When using bioinformatics tools, it is not only **important to know how to use the tool** but understand the **results** and **errors** the tool can make.

## Human Genome Project

**Human Genome Project (HGP)** – Goal was to find all the genes in the human genome. More specific goals were:

1. **Identify** all 20,000-25,000 genes.
2. **Determine** the sequence of the 3.2 billion bases in the genome.
3. **Store** the information in databases.
4. **Improve** tools for data analysis.
5. **Address** ethical, legal, and social issues (ELSI) that may arrive from the project.

**Model Organism** – An organism that is extensively studied to understand particular biological phenomena.

Model organisms can provide insight into the inner workings of other organisms. This is because in evolution many fundamental biological principles are conserved.

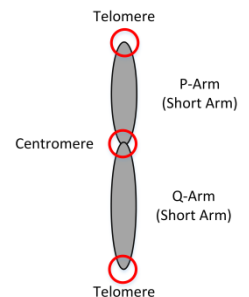
Large amount of biological data created in HGP required development of computer tools to:

1. **Collect** the data
2. **Organize** the data
3. **Maintain** the data
4. **Access** the data
5. **Analyze** the data

24% of DNA is introns. 15% of DNA is repeating DNA.

### Applications of Genome Research

Molecular Medicine, Microbial Genomics, Risk Assessment, Bio-archaeology, DNA Identification, Agriculture, Livestock Breeding, Bioprocessing. Genomes of other animals (e.g. chicken, rat, dolphin) have also been sequenced for comparison with the human genome.



**Chromosome (Cytogenetic) Map** – Descriptive diagram of gene locations on a chromosome.

## Applications of Genome Research

<p style="text-align: center;"><b>Molecular Medicine</b></p> <ol style="list-style-type: none"> <li>1. Improve diagnosis of disease</li> <li>2. Detect genetic predisposition</li> <li>3. Create drugs based on molecular information</li> <li>4. Use gene therapy as drugs</li> <li>5. Design custom drugs on individual genetic profiles.</li> </ol> <p><b>Personalized Medicine:</b> Genotype (i.e. genetic makeup) specific treatment of diseases.</p>	<p style="text-align: center;"><b>Microbial Genomics</b></p> <p><b>Pathogen:</b> Disease causing microbe.</p> <ol style="list-style-type: none"> <li>1. Swift detection and treatment of disease causing microbes and pathogens.</li> <li>2. Development of new energy sources through biofuels.</li> <li>3. Monitor the environment to detect biological warfare.</li> </ol>	<p style="text-align: center;"><b>DNA Identification</b></p> <ol style="list-style-type: none"> <li>1. Exonerate innocent suspects and incriminate guilty ones.</li> <li>2. Establish paternity and family relationships.</li> <li>3. Match organ donors with transplant recipients.</li> <li>4. Determine evolutionary history.</li> </ol>
--	---	---

**Locus** – Specific location of a gene, DNA sequence, or position on a chromosome.

**Allele** – Alternative forms of a gene.

**Ancestry Informative Marker (AIM)** – Set of polymorphisms for a locus with exhibits with substantially different frequencies in populations from different geographical regions.

**Single Nucleotide Polymorphism (SNP)** – Mutation in a single nucleotide locus that exists in more than 1% of the population. They make aligning DNA sequences more difficult. **Used in forensic identification.**

**Inverted Repeat:** A sequence of nucleotides followed downstream by its reverse complement.

**Conserved Sequence:** Similar or identical sequences that occur within nucleic or amino acid sequences across species.

Progress of Biology Research from *in vivo* (inside the body) to *in vitro* (inside the test tube) to *in silicon* (inside silicon, i.e. the computer).

## Sequence Alignment

<p><b>Sequence Alignment:</b> Procedure of comparing sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.</p> <p><b>Pairwise Sequence Alignment:</b> Comparing two sequences. It <b>infers</b> biological relationships <b>from</b> sequence similarity.</p> <p><b>Multiple Sequence Alignment:</b> Comparing more than two sequences. <b>From</b> biological relationships, it <b>infers</b> sequence similarity.</p>	<p style="text-align: center;"><b>Uses of Sequence Alignment</b></p> <p><b>Basic Principle of Sequence Alignment:</b> Similar DNA sequences produce similar proteins.</p> <p><b>Predicting a Protein's Structure (and in turn Function):</b> Compare the protein's <b>sequence/function</b> to that of other known sequences. <b>Same applies for DNA.</b></p> <p><b>Evolutionary Relationships:</b> Check whether two or more genes are proteins are evolutionarily related.</p>	<p><b>Quantifying Similarity:</b> To determine similarity, you must have a quantitative measure of alignment.</p> <p><b>Query Sequence:</b> Sequence that is searched for in a database.</p> <p><b>Aligned sequences may have:</b></p> <ol style="list-style-type: none"> <li>1. <b>A common ancestor</b></li> <li>2. <b>Same or similar structure/function</b></li> </ol> <p>Hence, one sequence can provide insight into another.</p>
---	---	---

<p style="text-align: center;"><b>Similarity</b></p> <p>Two or more similar DNA sequences <b>from different organisms</b> can be explained by the theory that all genetic material has one common ancestral DNA.</p> <p><b>Directly observable from alignment.</b></p> <p><b>Measurable quantity</b> which is dependent on the parameters used for comparison.</p> <p><b>Alignment in Protein:</b> A good alignment between two proteins has a high identity score and those amino acids that are different have similar physiochemical properties. There are also few gaps.</p>	<p style="text-align: center;"><b>Homology</b></p> <p>Two DNA sequences that share a common ancestor and are hence <b>evolutionarily related</b>. They derive from a <b>common ancestor gene</b>.</p> <p>Sequence and <b>usually structure</b> conservation.</p> <p>Can only be <b>inferred</b> and only reflect <b>probable</b> evolutionary history.</p>	<p style="text-align: center;"><b>Differences</b></p> <p>DNA mutations cause the differences between the families of contemporary species.</p> <p style="text-align: center;"><b>Types of Alignment Differences</b></p> <ol style="list-style-type: none"> <li>1. <b>Base/Amino Acid Substitution</b></li> <li>2. <b>Gap/Indel</b> (Insertion or deletion).</li> </ol> <p><b>Note:</b> Whether a change is an insertion or deletion depends on reference strand's perspective.</p>	<p style="text-align: center;"><b>Identity</b></p> <p>Simplest and most objective comparison metric.</p> <p>Quantified as the <b>percentage of identical characters</b>.</p> <p><b>Number of Identical matches divided by the length including gaps.</b></p>
--	--	--	--

## Pairwise Alignment

<p><b>Pairwise Problem Definition Includes:</b></p> <ol style="list-style-type: none"> <li>1. <b>Two</b> (amino acid or nucleotide) <b>sequences</b></li> <li>2. <b>A scoring system for match and mismatch</b></li> <li>3. <b>A gap penalty function</b></li> </ol> <p><b>Optimal alignment/pairing</b> – Pairwise alignment (including gaps) that achieves the optimal score. However, the optimal score is not necessarily the correct evolutionary pairing.</p>	<p style="text-align: center;"><b>Global Alignment</b></p> <p>Pairwise alignment which attempts to <b>maximize the total score</b> at the expense of area with greatest local similarity.</p> <p>Perform global alignment when you want to determine whether <b>the sequences are generally the same</b>.</p>	<p style="text-align: center;"><b>Local Alignment</b></p> <p>Finds the <b>maximum scoring subsequences</b> at the expense of the overall score.</p> <p>Perform when looking for high scoring <b>local similarities</b>. These local similarities may be due to <b>common function or common substructure/subfunction</b>.</p> <p><b>Optimal Local Alignment:</b> Longest subsequence of high similarity between two sequences.</p> <p><b>Local Alignment is usually more meaningful</b> than global alignment since it <b>identifies conserved sequences</b>.</p>	<p style="text-align: center;"><b>Methods of Pairwise Alignment</b></p> <ol style="list-style-type: none"> <li>1. <b>By Hand</b></li> <li>2. <b>Dot Matrix</b></li> <li>3. <b>Dynamic Programming</b></li> <li>4. <b>Heuristic Methods</b> (e.g. BLAST, FASTA)</li> </ol>
---	---	---	---

## By Hand

<p><b>Procedure:</b></p> <ol style="list-style-type: none"> <li>1. Write the two sequences on two rows.</li> <li>2. Place identical/similar characters in the same column.</li> <li>3. Place non-identical characters in either same column as a mismatch or with a gap.</li> </ol>
---

## Dot Matrix

<p>Simplest means of comparing two sequences.</p> <p>Provides a visual representation of areas of similarity between two sequences.</p> <p><b>Procedure:</b></p> <ol style="list-style-type: none"> <li>1. Create a 2-D grid and place one sequence along the top row and another along the leftmost column.</li> <li>2. When there two sequences match for a particular pair (set) of characters, place a dot in the corresponding X-Y grid location.</li> </ol>	<p style="text-align: center;"><b>Filtering in Dot Matrices</b></p> <p><b>Window Size:</b> Number of Characters to Compare at a Time.</p> <ul style="list-style-type: none"> <li>• DNA: 15 (Typical)</li> <li>• Protein: 2-3 (Typical)</li> <li>•</li> </ul> <p><b>Stringency:</b> Number of characters that must match exactly to be considered a match.</p> <ul style="list-style-type: none"> <li>• DNA: 10 (Typical)</li> <li>• Protein: 2 (Typical)</li> </ul> <p>Dot patterns along a diagonal in the dot matrix indicate matching sequences.</p>	<p style="text-align: center;"><b>Advantages</b></p> <ol style="list-style-type: none"> <li>1. All residues in the sequence are identified.</li> <li>2. Can reveal the presence of <b>insertions, deletions</b>, and <b>direct/inverted sequences</b>.</li> </ol> <p style="text-align: center;"><b>Disadvantages</b></p> <ol style="list-style-type: none"> <li>1. Relies on visual analysis to find trends.</li> <li>2. Not quantitative</li> <li>3. Hard to find optimal alignments.</li> <li>4. Do not allow gaps in the sequence</li> <li>5. Difficult to estimate significance of alignments.</li> </ol>
---	---	--

## Dynamic Programming

<p>Provides a <b>reliable</b> and <b>optimal</b> computational method for aligning DNA and protein sequences.</p> <p>Optimal alignments provide information that enables <b>functional, structural</b>, and <b>evolutionary predictions</b> of sequences.</p> <p>Requires use of a <b>scoring system</b>. <b>Examples include PAM and BLOSUM</b>.</p>	<p><b>Needleman Wunsch Algorithm:</b> Variant of the longest common subsequence algorithm for <b>global alignment</b>. However, it allows lateral/vertical moves in the case of a gap.</p> <p>Needleman Wunsch can be modified for local alignment. This is usually more useful since it emphasizes identifying <b>conserved regions</b>.</p>	<p style="text-align: center;"><b>Advantages</b></p> <ol style="list-style-type: none"> <li>1. Provides exact answer on similarity.</li> </ol> <p style="text-align: center;"><b>Disadvantages</b></p> <ol style="list-style-type: none"> <li>1. Slow and takes significant computational resources.</li> </ol>
---	---	---

## Substitution Matrix/Scoring Matrix

<p>A matrix containing information on the <b>frequency of mutation</b> of one residue (e.g. nucleotide/amino acid) to another.</p> <p>Best substitution matrices are <b>derive from the analysis of numerous homologs of well-suited proteins from many difference species</b>.</p>	<p>It is a table showing the probability of a mutation from one residue to another.</p> <p>This will help determine whether a sequence matching is due to <b>random chance or is evolutionarily related</b>.</p>	<p style="text-align: center;"><b>Amino Acid Substitution Matrix Examples</b></p> <ol style="list-style-type: none"> <li>1. <b>PAM250</b> – Percent Accepted Mutation</li> <li>2. <b>BLOSUM62</b> – Block Substitution Matrix</li> </ol>
---	--	--

## Amino Acid Substitution Matrix Examples

### PAM (Percent Accepted Mutation)

Developed by Margaret Dayhoff. Also called **Dayhoff amino acid substitution matrices**.

**Accepted Mutation:** Any mutation that is not fatal to the organism or destroy the protein.

**Used to find optimal alignments of amino acid sequences in homologous proteins and to score that alignment.**

**One PAM (PAM1):** 1% or less of the amino acids have been changed.

For proteins that have less similarity, the One PAM matrix is multiplied against its self  $N$  times. Example:

$$PAM20 = (PAM1)^{20}$$

**The higher the PAM number, the less similar the expected similarity. These are useful for distantly related proteins:**

- PAM120: 40% similarity
- PAM80: 50% similarity
- PAM60: 60% similarity

#### Criticism of PAM:

1. Derives from a small number of closely related proteins and may not be indicative of all proteins.
2. Not much more useful for determining homology than simpler matrices (e.g. based on chemical grouping of amino acid side chains)

### BLOSUM (Block Substitution Matrix)

Derives from a much larger set of proteins than PAM1 matrix and is the **most widely used substitution matrix**.

Most famous BLOSUM matrix is **BLOSUM62**.

Used a larger set of conserved proteins (called **blocks**) when it was created than PAM

Appears able to capture more of the distant type of amino acid variations. More sensitive than PAM.

**BLOSUM80** is equivalent to PAM1 and is for the least divergent (1% divergence) amino acid sequences.

**BLOSUM62** is equivalent to PAM120 and is for moderately divergent (60% divergence) protein sequences.

**BLOSUM45** is equivalent to PAM250 and is for more divergent sequences.

**PAM numbering is from least divergent to most divergent while BLOSUM numbering is from most divergent to least divergent.**

## Heuristic Based Pairwise Alignment and BLAST

### BLAST (Basic Local Alignment Search Tool)

**Heuristic method for local alignment.**

Designed for quick searches of databases.

**Basic concept of BLAST:** "Good alignments contain short lengths of exact matches."

Managed by Genbank and the NCBI.

### Types of BLAST

**blastp:** Amino acid sequence local alignment

**blastn:** Nucleotide sequence local alignment

**blastx:** Six-frame dual strand query sequence against a protein database.

**tblastx:** Compares a six frame nucleotide sequences, converts them to an amino acid sequence and compares it against the six frame translation of a nucleotide database.

**tblastn:** Compares a protein sequence against all six reading frames in a nucleotide database.

**tblastn and blastx are the reverse flow of comparison.**

### E-Value

Quantifies the likelihood that similarity between two amino acid sequences is due to chance.

**The lower the E-value, the increased likelihood the proteins are homologs** and not due to change.

**FASTA** – Another heuristic based pairwise alignment method.

## Multiple Sequence Alignment

### Multiple Sequence Alignment Problem Definition:

1. Multiple amino or nucleic acid sequences
2. Match matrix
3. Gap penalty

**Result:** Alignment of sequences that returns an optimal score.

### Uses of Multiple Sequence Alignment

1. **Determine phylogenetic relationships and evolution.** (Example: root and unrooted phylogenetic trees)
2. **Structural analysis of proteins.**
3. Determine **relationships** between a group of sequences.
4. Determine **conserved regions**. **Conserved regions of proteins are usually the most important areas in the protein.**

### Types of Alignment

1. **Identical residue:** Amino acid matches exactly (i.e. identity)
2. **Conserved residue:** Amino acid belongs to the same amino acid partition.

### Exact algorithm

Traverses the entire search space and uses a performance measure to maximum quality.

**Advantage:** Returns the optimal solution.

**Disadvantage:** Computationally expensive. Impossible for large datasets (more than 7-8 sequences)

### Heuristic Algorithm

Returns an estimate of the exact solution. **Based off progressive pairwise alignment.**

**Based off:** Hidden Markov Models, Genetic Algorithms

**Examples:** ClustalW (**C**luster **A**lignment **W**eighted), MACAW

## Progressive Pairwise Alignment using ClustalW

Progressive alignment is a **greedy algorithm**.

### Procedure of ClustalW Pairwise Alignment

1. Perform pairwise alignment on all sequence pairs. Create a distance matrix between all the sequence pairs. Distance is the number of exact matches excluding gaps.

2. Create a **Guide Tree** to determine what order the sequences are aligned. Note the Guide Tree or dendrogram has no phylogenetic meaning. It cannot be used to show evolutionary relationships.

3. Align the most closely related sequences first in a nearest neighbor clustering fashion.

### Example of ClustalW on Four Sequences

S1 and S2 are very closely aligned while S3 and S4 are very closely aligned (but not as closely aligned as S1 and S2). The way ClustalW would align these sequences in the following order:

a. Align S1 and S2. This results in an aligned sequence:  $S_{1,2}$ .

b. Align S3 and S4. This results in an aligned sequence:  $S_{3,4}$ .

c. Cluster  $S_{1,2}$  with  $S_{3,4}$ .

### Limitations of ClustalW

**Guide Tree Quality:** Insignificant for closely aligned sequences but it can matter for distantly aligned sequences.

**Local Minimum:** ClustalW progressively aligns sequences and/or sets of sequences. If initial clustering has an issue, it cannot be removed in later steps.

### Scalable Gap Penalties

Used in protein profile alignments. Provide variable weights for gap insertion. **Examples:**

a. The penalty for a gap varies depending on the types of amino acids that are adjacent to the gap. Example: A gap next to a hydrophobic amino acid may be weighted higher than a

b. A gap opening close to another gap may be penalized more than an isolated gap.

### When to Use ClustalW

Sets of amino or nucleic acid sequences that are similar over their entire lengths

Protein sequences that are entirely **co-linear** (i.e. share the same protein domains in the same order throughout the sequence).

### When Not to Use ClustalW

- a. Sequences do **not share common ancestors**.
- b. Sequences are **only partially related**.
- c. Sequences include **short non-overlapping segments**.

### Alternatives to ClustalW

#### Clustal Omega

**TCoffee** – Collection of tools for computing, evaluation, and manipulating multiple alignments of DNA, RNA, protein structures, and protein sequences.

**MUSCLE** – Multiple Sequence Comparison by Log Expression

#### Dialign

**MAFFT** – Multiple Alignment using Fast Fourier Transform. Good balance of accuracy and speed.

#### PRRN

### Issues in Multiple Sequence Alignment

Final results depend on the **order** the sequences were aligned.

Sequences of **different lengths** can cause issues.

**Gaps** can make alignment unrealistically long.

**Nonconserved regions can dilute conserved regions.**

## DNA versus Protein Alignment

The choice of whether to use DNA or protein alignment depends on the type of phenomenon you are investigating. Example:

**Protein Function:** Use protein alignment

**Genetic Changes:** Use DNA alignment

**Initial mutations take place in DNA while evolution pressure occurs with proteins.**

## Structural Alignment

**Goal of sequence alignment is to align sequences of similar structure (i.e. the areas that are evolutionarily conserved).**

Since the computer has no knowledge of structure behavior, manual adjusting of alignment results is often required.

Multiple Sequence Alignment Editor and Formatter Programs: GeneDoc, MACAW, CINEMA (**C**olor **I**nteractive **E**ditor for **M**ultiple **A**lignment), Boxshade, ClustalX



## Homework #1 Keywords

<b>Genomic Imprinting:</b> Methylating of a cytosine base to make a gene inactive.	<b>Error Rate in DNA Replication:</b> 1 in $10^9$ (one in a billion). It is this low because of the way DNA polymerase replicates the DNA and performs error checking.	<b>Hybridization:</b> DNA and RNA can pair a complementary sequence of nucleotides to form a DNA/RNA hybrid or double stranded RNA  Used in DNA microarrays, <i>in situ</i> hybridization to detect cell activity, and fluorescence in situ hybridization for locating genes in chromosomes.	<b>Control Regions:</b> Surrounding regions of non-coding DNA that involved with whether a gene is expressed.
<b>Gene Expression</b> – Whether a gene is being transcribed into mRNA and proteins being transcribed from that mRNA.	<b>RNA Microarray</b> – Used to measure the level at which a gene is expressed.	<b>Overlapping Genes:</b> Where one or both strands in DNA encode to parts of different proteins. Most common in viruses but do occur in mammals and humans.	<b>Degenerate:</b> Reference to the fact that multiple codons map to the same amino acid so the DNA sequence cannot be determined from an amino acid sequence.
<b>Methionine</b> – Amino acid with codon AUG that is often removed from a newly synthesized protein.	<b>Open Reading Frame</b> – Amino acid sequence in mRNA that goes from the start codon to the stop codon inclusive of both. It may include introns.	<b>Regulatory Elements</b> – Elements that regulate transcription including promoter, response elements, enhancer, and repressor.	

## Homework #2 Keywords

<b>Promoter:</b> Control region of DNA that which RNA polymerase binds to initiate transcription. RNA polymerase binds more closely to the promoter than to other regions of DNA.  In bacteria, the promoter typically occurs right before the TSS.	<b>Terminator:</b> Sequence of DNA that tells RNA polymerase to stop transcription.  More variable in prokaryotes than promoters.	<b>Activator:</b> Proteins that <b>improve</b> binding of RNA polymerase to the promoter sites. This <b>increases gene expression</b> .	<b>Repressor:</b> Proteins that <b>inhibit</b> the binding of RNA polymerase to the promoter region of DNA. They <b>reduce gene expression</b> .
<b>TATA Box:</b> DNA sequence found in the core promoter of most eukaryotes genes. Occurs about 25 bases upstream from TSS.	<b>RNA Capping</b> – Process of adding a modified guanosine molecule to the 5' end of mRNA in eukaryotes.	<b>Polyadenylation</b> – Adding of approximately 200 adenosine molecules to the 3' end of mRNA in eukaryotes.	<b>RNA Splicing:</b> Process of removing introns from the mRNA sequence and rejoining the exons.
<b>Alternative Splicing:</b> Different variants of RNA splicing where some exons or parts of exons are removed. <b>This allows different versions of a protein to be made by the same gene. It allows for the larger number of proteins than there are genes.</b>	<b>Shine Dalgarno Sequence:</b> In bacterial DNA, it is a short sequence at the 5' end of the mRNA that indicates the <b>ribosome binding site</b> . The consensus sequence is: <b>AGGAGGU</b> .	<b>Operon</b> – Functionally related proteins that are clustered together in DNA. An operon is transcribed as one long mRNA and multiple different proteins are made some the same mRNA molecule. <b>Rarely found in eukaryotes.</b>	<b>Overlapping Genes:</b> Multiple genes that overlap (i.e. are on top of each other). <b>Common in viruses but rare in cellular genomes.</b>
<b>Viral Gene Replication:</b> Involves inserting its DNA into the cellular DNA and hijacking the cell's replication (mRNA and DNA) mechanisms.	<b>Mitochondria:</b> "Powerhouse of the cell"	<b>Plasmid:</b> Extrachromosomal DNA that can be based from bacteria cell to bacteria cell. Commonly used mechanism by bacteria to achieve drug resistance.	<b>Amino Acid Side Chain:</b> Molecule connected to the alpha carbon of the amino acid that defines the properties of the amino acid. Side chains are divided into different groups including size, polar/nonpolar, hydrophilic/hydrophobic, etc.
<b>Homologous:</b> Proteins that share a common ancestor.	<b><math>\alpha</math>-Helix:</b> Secondary structure of protein. Formed by energetically favorable hydrogen bonds between atoms of the backbone of the amino acid chain.	<b><math>\beta</math> Strand</b> – Extended strand of amino acids in a protein.	<b><math>\beta</math>-Sheet:</b> Multiple $\beta$ -strands that form a sheet by bonding with each other.

<b>Globular Protein:</b> Roughly spherical in shape. Can be composed of multiple globular proteins or a single protein.	<b>Fibrous Protein:</b> Rod or wire like proteins. Examples include hair, wool, and the silk protein.	<b>Translation and Transcription in Prokaryotes:</b> Occurs in the cytoplasm. Since no splicing in prokaryotes, translation can begin while transcription is in progress.	
---	---	---	--

### Hands On Exercise Keywords

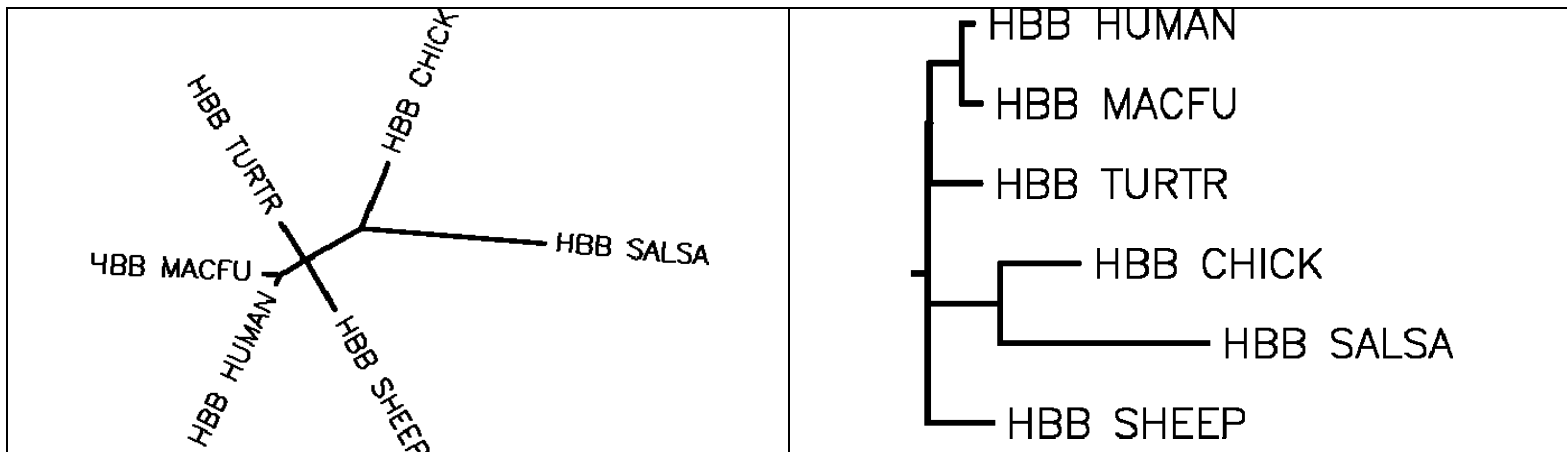
<b>UTR</b> – Untranslated region. It is found on either end of the unspliced mRNA molecule.	<b>Synonymous SNP</b> – A single nucleotide polymorphism that does not affect the resulting protein sequence.	<b>Non-synonymous SNP</b> – A single nucleotide polymorphism that changes the resulting protein sequence.	<b>Phosphate Group</b> – Molecule attached to the 5' carbon of the sugar in DNA and RNA.
---	---	---	--

<b>Hydroxyl</b> – OH Molecule on the 2' carbon of an RNA ribose molecule.	<b>Four Functional Groups</b> Attached to the Alpha Carbon on an Amino Acid.	<b>Amino (N) Terminal</b> – A reactive amino group that is located at what is the 5' end of the mRNA sequence.	<b>Carboxyl (C) Terminal</b> – A reactive carboxyl group that is located at what is the 3' end of the mRNA sequence.
---	--	--	--

<b>TSS</b> – Transcription Start Site	<b>Silent Mutation:</b> A mutation in a gene that has no effect on the resulting protein.	<b>Coding Sequence (CDS):</b> Part of a gene that codes for exons. It is bounded by the 5' and 3' untranslated regions (UTR). It is from the start codon to the stop codon.	<b>RNA Capping</b> – Process of adding a modified guanosine molecule to the 5' end of mRNA in eukaryotes.
---------------------------------------	---	---	---

<b>Hydrophobic</b> – Not attracted to water (i.e. some say repel)	<b>Hydrophobic:</b> Attracted to water molecules.	<b>Dystrophin:</b> Longest gene in the human body at 2.4 million bases.	<b>Annotation</b> – Additional information in a database entry.
---	---	---	---

<b>99.9%</b> – Percentage of DNA similarity between individuals.	<b>Phylogenetic Tree:</b> Can be rooted or unrooted. Graphical means to depict evolutionary relationships between organisms. Branch length indicates time and an inner node represent a common ancestor.	<b>cDNA</b> – (Copied DNA) DNA that is reverse synthesized from mRNA. Hence it lacks mRNA and any control signals.	<b>Expressed Sequence Tags (EST)</b> – Partial cDNA sequence. It is a fragment of a gene.
--	--	--	---



<b>Primary Data</b> – Raw experimental results. It is the initial experiment interpretation.	<b>Secondary Data</b> – Derived from the primary sources. Less reliable than primary data. Must be rederived regularly.	<b>Two Methods for Checking the Accuracy of a Database</b> – Automated computerized analysis and manual curating.	<b>Nonredundant Database</b> – Each gene (or splice-form of a gene) has a unique entry in the database.
--	---	---	---

<b>Convergent evolution</b> is when organs, proteins, and DNA sequence that unrelated in their evolutionary origin acquire the same structure or function.	<b>Divergent evolution</b> produces different structures or sequences from a common ancestor	<b>Pseudogene</b> - Sequences in genomic DNA that are similar to known coding-genes but do not produce a functional protein.	<b>Frameshift Mutation</b> – An indel (insertion/deletion) where the number of effected bases is not divisible by 3. It can cause a change in the reading frame.
--	--	--	--

<b>Missense</b> – Mutation in which a single nucleotide change which results in a codon that codes for an amino acid.	<b>Nonsense Mutation:</b> A point (i.e. single nucleotide) mutation that results in a premature stop codon.	<b>Protein Domain:</b> A discrete structural unit in a protein.	<b>Denatured Protein:</b> An unfolded protein.
---	---	---	--



<b>Enzyme</b> – Bind other molecules and catalyze their biochemical reactions.	<b>Role of Protein as an Activator:</b> Play a role in RNA transcription.	<b>Communication role of proteins:</b> Secreted by cells as chemical messengers to other cells	<b>Receptor:</b> Proteins on cell surface used to receive intercellular messages.
<b>{A-Y} – {BJOUX}</b> – Amino acid notation letters are between A and Y with the exception of letters B, J, O, U, and X.	<b>Pairwise Sequence Alignment:</b> Infer biological relationships from sequence similarity. <b>Multiple Sequence Alignment:</b> Infer sequence similarity from biological relationships.	<b>1 in 10<sup>9</sup></b> – DNA replication error rate.	<b>IVS</b> – Intervening sequence. Notation used to refer to introns.
<b>Spliceosome</b> – Consists of snRNA (small nuclear RNA) and proteins that contain the enzymatic activity to perform RNA splicing.	<b>30% Identity Score</b> – 90% chance the two genes are homologous.		

**Swissprot** – Protein sequence and functional information database run by the Swiss. Curated by hand.

**Genscan** – Exon, intron, and coding sequence predictor. From MIT.

Biology Workbench – San Diego Supercomputer Center – Used to run ClustalW to generate protein sequence alignment and dendograms.

