

CS123A Final Exam Study Guide

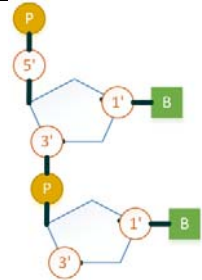
By: Zayd Hammoudeh

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

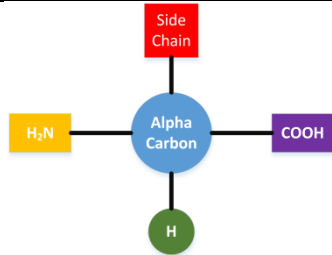
<p>Protein “Building blocks of life” “We are our proteins” They put into action the genetic information in DNA.</p> <p>Typical length: 300-500 Number of Amino Acids: 20</p> <p>Protein Function: Determine by its unique 3-dimensional structure.</p> <p>Four Structures of a Protein: Primary, Second, Tertiary,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

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

Sugar Phosphate Backbone

<p>Phosphodiester Backbone – 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 phosphodiester linkage.</p> <p>Because of the phosphodiester backbone DNA is laid out as 5' to 3' and complementary 3' to 5'. Hence backbones run in opposite directions.</p> <div style="text-align: center;"> <p>5' ----- 3'</p> <p> </p> <p>3' ----- 5'</p> </div>	
---	---

Amino Acids

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

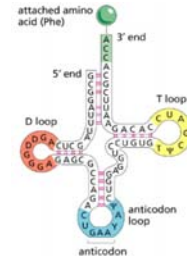
tRNA: Has a complementary 3' to 5' **anticodon** that binds to the mRNA codon and the ribosome in protein synthesis. **Clover leaf structure**.

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 (i.e. no splicing).

RNA Splicing

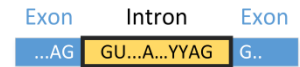
Exon – Coding portion of a gene.

Intron – Non-coding 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 discovery 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.

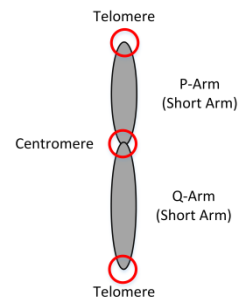
Large amount of biological data created in HGP required development of computer tools to:

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

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

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

Pairwise Alignment

<p>Pairwise Problem Definition Includes:</p> <ol style="list-style-type: none"> 1. Two (amino acid or nucleotide) sequences 2. A scoring system for match and mismatch 3. A gap penalty function <p>Optimal alignment/pairing – 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;">Global Alignment</p> <p>Pairwise alignment which attempts to maximize the total score at the expense of area with greatest local similarity.</p> <p>Perform global alignment when you want to determine whether the sequences are generally the same.</p>	<p style="text-align: center;">Local Alignment</p> <p>Finds the maximum scoring subsequences at the expense of the overall score.</p> <p>Perform when looking for high scoring local similarities. These local similarities may be due to common function or common substructure/subfunction.</p> <p>Optimal Local Alignment: Longest subsequence of high similarity between two sequences.</p> <p>Local Alignment is usually more meaningful than global alignment since it identifies conserved sequences.</p>	<p style="text-align: center;">Methods of Pairwise Alignment</p> <ol style="list-style-type: none"> 1. By Hand 2. Dot Matrix 3. Dynamic Programming 4. Heuristic Methods (e.g. BLAST, FASTA)
---	---	---	---

By Hand

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

Dot Matrix

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

Dynamic Programming

<p>Provides a reliable and optimal computational method for aligning DNA and protein sequences.</p> <p>Optimal alignments provide information that enables functional, structural, and evolutionary predictions of sequences.</p> <p>Requires use of a scoring system. Examples include PAM and BLOSUM.</p>	<p>Needleman Wunsch Algorithm: Variant of the longest common subsequence algorithm for global alignment. 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 conserved regions.</p>	<p style="text-align: center;">Advantages</p> <ol style="list-style-type: none"> 1. Provides exact answer on similarity. <p style="text-align: center;">Disadvantages</p> <ol style="list-style-type: none"> 1. Slow and takes significant computational resources.
---	---	---

Substitution Matrix/Scoring Matrix

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

Amino Acid Substitution Matrix Examples

PAM (Point Accepted Mutation)

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

Accepted Mutation: Any mutation that is not fatal to the organism nor does it 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 chance.

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

Genomic Imprinting: Methylating of a cytosine base to make a gene inactive.	Error Rate in DNA Replication: 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.	Hybridization: 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.	Control Regions: Surrounding regions of non-coding DNA that involved with whether a gene is expressed.
Gene Expression – Whether a gene is being transcribed into mRNA and proteins being transcribed from that mRNA.	RNA Microarray – Used to measure the level at which a gene is expressed.	Overlapping Genes: Where one or both strands in DNA encode to parts of different proteins. Most common in viruses but do occur in mammals and humans.	Degenerate: 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.
Methionine – Amino acid with codon AUG that is often removed from a newly synthesized protein.	Open Reading Frame – Amino acid sequence in mRNA that goes from the start codon to the stop codon inclusive of both. It may include introns.	Regulatory Elements – Elements that regulate transcription including promoter, response elements, enhancer, and repressor.	

Homework #2 Keywords

Promoter: 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.	Terminator: Sequence of DNA that tells RNA polymerase to stop transcription. More variable in prokaryotes than promoters.	Activator: Proteins that improve binding of RNA polymerase to the promoter sites. This increases gene expression .	Repressor: Proteins that inhibit the binding of RNA polymerase to the promoter region of DNA. They reduce gene expression .
TATA Box: DNA sequence found in the core promoter of most eukaryotes genes. Occurs about 25 bases upstream from TSS.	RNA Capping – Process of adding a modified guanosine molecule to the 5' end of mRNA in eukaryotes.	Polyadenylation – Adding of approximately 200 adenosine molecules to the 3' end of mRNA in eukaryotes.	RNA Splicing: Process of removing introns from the mRNA sequence and rejoining the exons.
Alternative Splicing: Different variants of RNA splicing where some exons or parts of exons are removed. 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.	Shine Dalgarno Sequence: In bacterial DNA, it is a short sequence at the 5' end of the mRNA that indicates the ribosome binding site . The consensus sequence is: AGGAGGU .	Operon – 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. Rarely found in eukaryotes.	Overlapping Genes: Multiple genes that overlap (i.e. are on top of each other). Common in viruses but rare in cellular genomes.
Viral Gene Replication: Involves inserting its DNA into the cellular DNA and hijacking the cell's replication (mRNA and DNA) mechanisms.	Mitochondria: "Powerhouse of the cell"	Plasmid: Extrachromosomal DNA that can be based from bacteria cell to bacteria cell. Commonly used mechanism by bacteria to achieve drug resistance.	Amino Acid Side Chain: 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.
Homologous: Proteins that share a common ancestor.	α-Helix: Secondary structure of protein. Formed by energetically favorable hydrogen bonds between atoms of the backbone of the amino acid chain.	β Strand – Extended strand of amino acids in a protein.	β-Sheet: Multiple β -strands that form a sheet by bonding with each other.

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

Hands On Exercise Keywords

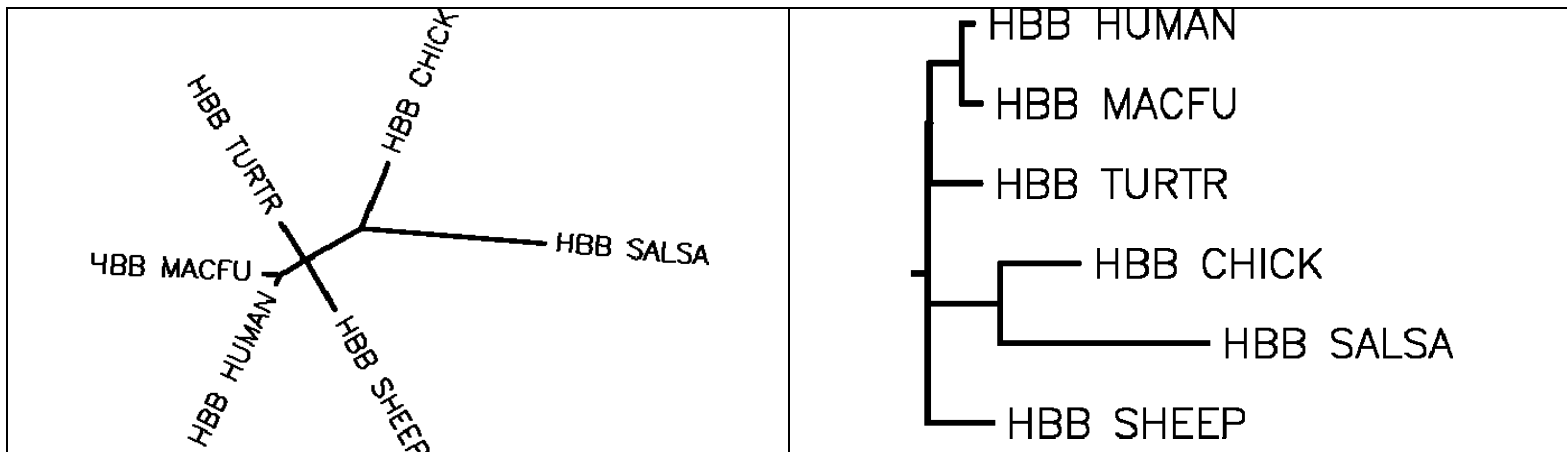
UTR – Untranslated region. It is found on either end of the unspliced mRNA molecule.	Synonymous SNP – A single nucleotide polymorphism that does not affect the resulting protein sequence.	Non-synonymous SNP – A single nucleotide polymorphism that changes the resulting protein sequence.	Phosphate Group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

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

TSS – Transcription Start Site	Silent Mutation: A mutation in a gene that has no effect on the resulting protein.	Coding Sequence (CDS): 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.	RNA Capping – Process of adding a modified guanosine molecule to the 5' end of mRNA in eukaryotes.
---------------------------------------	---	---	---

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

99.9% – Percentage of DNA similarity between individuals.	Phylogenetic Tree: 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.	cDNA – (Copied DNA) DNA that is reverse synthesized from mRNA. Hence it lacks mRNA and any control signals.	Expressed Sequence Tags (EST) – Partial cDNA sequence. It is a fragment of a gene.
--	--	--	---



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

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

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

Enzyme – Bind other molecules and catalyze their biochemical reactions.	Role of Protein as an Activator: Play a role in RNA transcription.	Communication role of proteins: Secreted by cells as chemical messengers to other cells	Receptor: Proteins on cell surface used to receive intercellular messages.
--	--	---	---

{A-Y} – {BJOUX} – Amino acid notation letters are between A and Y with the exception of letters B, J, O, U, and X.	Pairwise Sequence Alignment: Infer biological relationships from sequence similarity. Multiple Sequence Alignment: Infer sequence similarity from biological relationships.	1 in 10⁹ – DNA replication error rate.	IVS – Intervening sequence. Notation used to refer to introns.
---	--	--	---

Spliceosome – Consists of snRNA (small nuclear RNA) and proteins that contain the enzymatic activity to perform RNA splicing.	30% Identity Score – 90% chance the two genes are homologous.	UCSC Genome Browser – Tool used to visualize genome information.	Complement – Indication that the coding sequence is on the complementary strand in GenBank.
--	--	---	--

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.



The Big Jaw

<p>Big Jaw – Constraint that inhibited brain growth</p>	<p>Powerful masticatory muscles: Found in most primates including chimpanzees and gorillas</p> <p>Human Masticatory Muscles – Much smaller compared to other animals in the Homo genus.</p>	<p>Myosin Heavy Chain (MYH): Gene expressed in masticatory muscles. Inactivated in humans by a frameshift mutation after the lineage of humans and chimpanzees diverged.</p> <p>Mutation removed a barrier for the remodelling of the hominid cranium which consequently allowed for an increase in the size of the brain.</p>
--	---	--

Viruses

<p>Virus – Small living particles that can infect cells and change how the cells function. The effect on the cell's function depends on the type of virus and the cells that are infected. Surrounded by protein case.</p> <p>Retrovirus – Single stranded RNA virus that employs a double stranded DNA (dsDNA) intermediate for replication.</p>	<p>Pathogen: A disease product. It can include both infectious organisms (bacteria, fungi, etc.) as well as viruses.</p>	<p>Virulence: Ability of an infectious agent (i.e. pathogen) to cause a disease.</p> <p>Many viruses are virulent sometimes and asymptomatic at other times.</p>
--	---	---

<p>Immunodeficiency – The result when the immune system is unable to protect the host from disease causing agents or from malignant cells.</p>	<p>Acquired Immunodeficiency: Loss of immune function because the genetic or development deficiency was not acquired at birth. It results from exposure to various agents.</p>
---	---

<p>Virus – A single stranded RNA virus that employs a double stranded DNA (dsDNA) intermediate for replication.</p>	<p>Reverse Transcriptase: Turns viral RNA into DNA. It turns the RNA strand into DNA. It then uses the DNA to make it complementary strand.</p>	<p>cDNA – Complementary DNA made from mRNA by reverse transcriptase.</p>
---	--	---

<p>Capsid – Surrounds mRNA in virus particle. It's the outer protein shell.</p>	<p>Viral DNA is integrated into the DNA of the host cell.</p>	<p>Virion – Entire virus particle including the capsid (protein shell) and the inner core of nucleic acid.</p>
--	--	---

HIV

<p>HIV – Human Immunodeficiency Virus</p> <p>Type of retrovirus.</p> <p>Inherited from:</p> <ul style="list-style-type: none"> Chimpanzees Mangabys <p>Transmitted through bodily fluids (e.g. blood, semen) when the virus of an infected individual contacts the mucous membrane or enters the blood stream of an uninfected individual.</p>	<p>Cells Affected by HIV:</p> <ul style="list-style-type: none"> Macrophages - Large immune cells that devours invading pathogens and other intruders. Stimulates other immune cells by presenting them small pieces of the invaders. CD4+ T Cells (aka T-Helper Cells) – White blood cells that orchestrate the immune response. They signal other cells to perform their special functions. 	<p>Lentivirus – “Slow viruses” where the period between initial infection and the onset of serious symptoms is long.</p> <p>Other Lentiviruses:</p> <ul style="list-style-type: none"> FIV – Feline Immunodeficiency Virus SIV – Simian Immunodeficiency Virus (Infects monkeys and nonhuman primates)
--	---	---

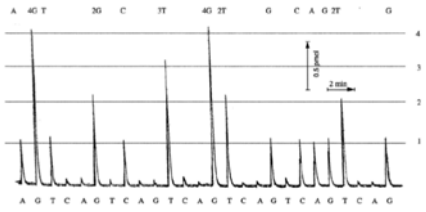
<p>Body's Immune Response to HIV</p> <p>Destroys the virions floating in the bloodstream before they can infect new cells.</p> <p>Destroys the infected CD4 helper T-cells depleting the body's ability to fight disease. This causes an immune system collapse. This leads to AIDS (Acquired immunodeficiency Syndrome).</p>	<p>Three Types of Proteins Involved in Viral (Virion) Replication</p> <ul style="list-style-type: none"> GAG – Encodes for core proteins and structural virion components. POL – Encodes for reverse transcriptase, integrase, and protease. Env – Encodes for the structural protein components that surrounds the virus. Needed for the virus to leave the cell. 	<p>Miscellaneous HIV Notes</p> <p>Many subgroups of HIV-1 exist.</p> <p>Within a single subtype and in a single infected person, the virus changes constantly.</p> <p>Transmission from chimp to humans happened multiple times in the past.</p>
---	--	---

Sanger Sequencing

<p>Developed by Frederick Sanger in 1977. Most widely used technology for ~25 years.</p> <p>Replaced by “Next Generation Sequencing” techniques.</p> <p>Still widely used for smaller scale projects and for long contiguous DNA sequences (>500 nucleotides).</p>	<p>Dideoxynucleotides – Nucleotides where the OH molecule on the 3' carbon of the sugar is modified to simply an –H making a subsequent phosphodiester linkage impossible.</p> <p>These are floating in the gel and sometimes DNA polymerase selects a normal nucleotide and other times the dideoxy analog which terminate the sequence. This sugar can fluoresce.</p>	<p>DNA polymerase makes a complement of a partial sequence within a DNA molecule. Synthesis is primed from a chemically synthesized fragment (i.e. primer) that is complementary to a part of the DNA sequence known from other studies.</p> <p>DNA polymerase builds strand from 5' to 3'.</p>
--	--	--

Pyrosequencing

<p>Developed in 1996.</p> <p>Based off the detecting of released pyrophosphate during DNA synthesis. This detection is through the detection of light.</p> <p>Sequences a single strand of DNA by synthesizing the strand's complement.</p>	<p>Benefits of Pyrosequencing:</p> <ul style="list-style-type: none"> • "Sequencing by synthesis" • Accurate • Simple and robust • No labels or gels • Real time results. <p>Nucleotides are dispensed sequentially and then removed from the reaction (this is done by apyrase). Light is only produced when the solution complements the first unpaired base of the template strand.</p> <p>A mini strand with a magnetic bead for DNA polymerase serves as a primer.</p>	<p>Example Pyrosequencing Instruments</p> <p>PSQ96</p> <ul style="list-style-type: none"> • 500 samples per hour • 4500 samples per day. • Includes CCD camera. <p>PSQHS96A</p> <ul style="list-style-type: none"> • 10,000 samples per day • 30,000 samples per day with triplex analysis. <p>Procedure:</p> <ol style="list-style-type: none"> 1. Prepare samples 2. Insert samples in PSQ96 3. Insert reagent cartridge 4. Start run. <p>96 represents the number of wells.</p>
--	--	--

<p style="text-align: center;">Reading a Pyrogram</p>  <p>Bases released sequentially. Depending on intensity of light you can determine the number of sequential bases.</p>	<p style="text-align: center;">Single Nucleotide Polymorphism</p> <ul style="list-style-type: none"> • Occurs every 500 to 1000 bases in DNA. • Most common cause of inter-individual variation. <p style="text-align: center;">HPV – Human Papillomavirus</p> <ul style="list-style-type: none"> • Sexually transmitted infection (STI) • Usually does not cause health problems but can cause cancer of the vulva, vagina, penis, and anus as well as in the back of throat. • Different primers used to detect the specific strain of HPV infection. 	<p>Wild Type – Original, non-mutated version of a gene. For bacteria, it would be the original non-drug resistant version.</p> <p>Pyrosequencing begins with a primer that binds to the DNA sequence. Primer has a "general primer site."</p>
--	--	---

Primers

<p>Primer Design: Required step before beginning pyrosequencing. This includes running PCR (polymerase chain reaction).</p> <p>General Primer – Will anneal with all alleles.</p> <p>Have a magnetic bead at 5' end.</p>	<p style="text-align: center;">Polymerase Chain Reaction</p> <p>Used to amplify a specific DNA sequence. Exponentially increases number of copies of a DNA sequence.</p> <p>Step #1: Denaturing – Heating the DNA sequence to render it single stranded (Double helix is also removed). Example time: 1 minute at 94C.</p> <p>Step #2: Annealing – Forward and reverse primers bind to the appropriate complementary strands.</p> <p>Step #3: Extension – DNA polymerase extends the primers.</p> <p>These three steps are repeated 30-40 times. First couple of PCR copying does not actually created double stranded DNA of the right gene. This is eventually achieved through the primer.</p>
--	--

Primer Characteristics

<ul style="list-style-type: none"> • Lack of secondary priming sites (uniqueness) • Absence of hairpin formation (bends in single strand). Caused by intermolecular interaction within the primer. 	<p>Uniqueness: There should be only one place the primer can bind in the template DNA. There should be no possible contaminant binding sites either (e.g. from other animals such as human, rat, mouse, etc.)</p>	<p>Length – Related to uniqueness and melting/annealing temperatures. The longer the primer, the more likely to be unique and the higher the melting/annealing temperatures.</p> <p>Minimum Length: 15 bases Ideal Length: 17-28 bases</p>	<p>Base Composition – Random base composition is best. Best to avoid long A/T and G/C chains.</p> <p>50-60% G+C content leads to the right annealing/melting temperatures.</p>
<p style="text-align: center;">Melting Temperature</p> <p>Temperature at which half the DNA strands are single stranded and half are double stranded.</p> <p>More G/C nucleotides in a strand means higher melting temperature since more hydrogen bonds.</p> <p>Notation: T_m Target: 52C to 65C</p>	<p style="text-align: center;">Annealing Temperature</p> <p>Temperature at the primer anneals (bonds) to the DNA stand.</p> <p>Calculated as: $T_{anneal} = T_{m_{primer}} - 4^{\circ}C$</p>	<p style="text-align: center;">Internal Structure</p> <p>Primers can anneal to themselves or to other primers.</p> <p>Hairpin: Primer bending back to bind to itself. Self Dimer: Primer bonding to another of the same primer Dimer: Primer binding to a complementary strand primer.</p> <p>Stability at 5' end of the primer is critical.</p>	<p>Primers work in pairs. Two types of primers:</p> <ul style="list-style-type: none"> • Forward Primer • Reverse Primer <p>Annealing temperatures of the two strands must be compatible (maximum 3 degrees) of each other.</p>

Primer Design

<p>Generally done best by computers.</p> <p>Example primer design tools: Primer3 (tool used in class from MIT), BioTools, GCG, Oligo</p>	<p>Adjustable Features in Primer Design Tools:</p> <ul style="list-style-type: none"> • Primer Length • Melting Temperature • (G+C)% 	<p>Forward and Reverse Primer</p> <p>Reverse (Right) Primer 3' <-----GGAA----- 5'</p> <p>Plus Strand 5'-----ATCG----- 3'-----TAGC-----</p> <p>Target 5'--ATCG----> 3'</p> <p>Minus Strand 5'-----ATCG----- 3'-----TAGC-----</p> <p>Forward (Left) Primer</p>
<p>Multiplex PCR</p> <p>Multiple primer pairs are added together in PCR. This allows for the amplifying of multiple sites.</p> <p>Challenges of this Approach:</p> <ul style="list-style-type: none"> • Different melting temperatures. • Ensuring no dimer formation 	<p>Most primers are designed to amplify a single product.</p> <p>Universal Primer: A single primer that can be used to amplify multiple products.</p> <p>Semi-universal Primer: A single primer can be used to amplify a subset of template sequences from a large group of similar sequences.</p>	<p>Guessmer</p> <p>In some cases, DNA sequences are unavailable or difficult to align. Example: Back-translated a protein which is degenerate so the nucleotide sequence cannot be known.</p> <p>Procedural Differences in Primer Design:</p> <ul style="list-style-type: none"> • Length: Primer should be longer than normal at about 30 bases to offset decreased hybridization. • Set higher annealing temperature to increase primer annealing stringency.

Homework #3 Keywords

Hydrophobic Amino Acid – Does not react with water. Tends to be buried within a protein surrounded by other hydrophobic amino acids.	Polar, hydrophilic Amino Acid – Typically surrounded by water molecules. Generally buried inside protein surrounded by other oppositely charged hydrophilic amino acids.	Protein Domain – Conserved part of protein where the protein has folded into these discrete structural units. Core of each domain is α -helixes, β -sheets, or a mixture of both.	Tetramer – Protein that consists of four monomer subunits . Example: Hemoglobin .
SQL – Structured Query Language HTML – Hypertext Markup Language XHTML – Extended Hypertext Markup Language	Flat File Database – Rely on basic text files to store information. Still in use due to their simplicity and relatively low cost.	Annotation – Non-sequence information in a database entry. It can include interpretation of data, relevant research citations, and related entries in other databases .	Non-redundant database – Database with no duplicate entries. GI Number – Unique identifier given to a sequence. If a sequence ever changes, a new GI is assigned. Accession Number – Unique number assigned to a database entry, and it never changes.
Ontology – Set of field-specific descriptors that enable the sharing of the same concepts and definitions for specific terms.	Gene Ontology – Collaborative project that provides a controlled vocabulary that describes genes and gene-associated information.	Hypothetical Gene – A gene identified in a protein sequence purely through computational methods. No experimental data supports this as a gene. May be correct or incorrect.	Swiss-Prot – Protein database that does not use computer based annotation. All curating is done manually by experts.

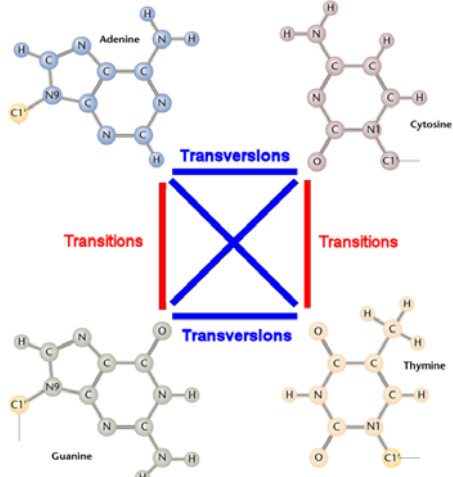
Three Types of Nucleotide Sequences in Databases

Raw Genomic Sequence – Represents chromosomal DNA and includes non-coding regions (e.g. introns, control regions, UTR) as well as coding regions (i.e. exons).	cDNA (Complementary DNA) – Result of reverse transcription from RNA to DNA . Does not include anything beyond the coding sequence.	Expressed Sequence Tag (EST) – A partial cDNA sequence that is around 300 bases in length.
---	--	--

Pseudogene – Sequences in genomic DNA that is similar to known coding genes but do not produce functional proteins . Mutate faster than normal genes since no longer under selection pressure. Possibly due to gene duplication. Up to 20,000 pseudogenes in human genome.	Window Size – Number of consecutive sequence objects used for comparison in a dot plot.	Stringency – Number of exact matches in a dot plot window that must be identical to be considered a match.	Protein versus Nucleotide Sequence Comparison <ul style="list-style-type: none">• Proteins have more 20 amino acids versus only 4 nucleotide bases. Hence, one character has more information in a protein.• Genetic code is redundant so insignificant variations are filtered out when looking at proteins.• Structure and function of protein is entirely dependent on amino acid sequence so amino acid sequences tend to change less over time.
--	--	---	--

RCSB Protein Database – Tool used to view protein structures.	Gene Locus Format: {Chromosome #}.{P/Q}{Chromosome Region} Example: 17.p21 is on the P-arm of chromosome 17 in region 21.	Nonsense Mutation – A point mutation in a DNA sequence that results in a premature stop codon (or nonsense codon). The mRNA is truncated , incomplete , and usually non-functional .	Missense Mutation – A point mutation in which a single nucleotide change results in a codon that codes for a different amino acid.
--	---	---	---

Homework #4 Keywords

<p>Purine: Adenine and Guanine – Two carbon rings (one pentagonal and one hexagonal)</p> <p>Pyrimidine: Cytosine and Thymine – Single hexagonal ring.</p> <p>Transition: Nucleotide substitution of:</p> <ul style="list-style-type: none"> Purine to Purine Pyrimidine to Pyrimidine <p>Transversion: Nucleotide substitution of:</p> <ul style="list-style-type: none"> Purine to Pyrimidine Pyrimidine to Purine <p>Transitions are more common than transversions since transitions tend to have less effect on the protein sequence.</p>	
--	--

Silent Mutation – Point mutation that does not affect the amino acid sequence of a protein. Can occur in both coding and non-coding regions .	Synonymous Mutation: Point mutation that does not affect the amino acid sequence of the gene. Limited to those substitutions in the coding region (i.e. exons).	“misc_feature” – Notation in a Genbank record to indicate a noteworthy feature in a sequence. Example: “ polymorphic (TAAA)n ” is a repeat sequence of “TAAA” multiple times.	Missense Mutation – A point mutation in which a single nucleotide change results in a codon that codes for a different amino acid.
---	---	--	---

Needleman-Wunsch – Global sequence alignment dynamic programming algorithm. Most rigorous and guaranteed to return the best alignment .	Smith-Waterman – Local sequence alignment dynamic programming algorithm. Most rigorous and guaranteed to return the best alignment . SSEARCH – Alignment program built off of Smith-Waterman.	Twilight Zone – Region between 20% and 30% identity in amino acid sequences. Homology may exist between the proteins but cannot be reliably assumed in the absence of other experimental data.	Midnight Zone – Region where there is less than 20% identity between two amino acid sequences. Very unlikely the two sequences are homologous.
--	--	--	--

Gap Penalty – Penalty that is subtracted from the alignment score anytime a gap is inserted in the alignment.	Gap Extension Penalty – Penalty to extend an existing gap. This penalty is generally less than when a new gap is inserted.	Gap Penalty Differences Based on Amino Acids – Some amino acids tend to be more important for a protein’s protein (e.g. Tryptophan). Hence, depending on the amino acid, the gap penalty may vary.	FASTA and BLAST: Heuristic based sequence alignment programs. Use heuristic to filter possible matches then runs dynamic program on those that passed the heuristic test . FASTA is a fast database-search method based on matching short identical sequences . BLAST based on finding very similar short segments .
--	---	---	--

Phenotype: A composite of an organism’s observable characteristics or traits. It is related to but not entirely dependent on the genotype as two organisms may have the same genotype but different physical characteristics due to the environmental factors.	Using Different Substitution Matrices – A single substitution score matrix is not ideal for all cases. Differences can include: <ul style="list-style-type: none"> Degree of similarity between the sequences Looking for closely related sequences or very distantly related evolution relationships 	<p>Blastp – Compares a query protein sequence against a protein database.</p> <p>Blastn – Compares a query nucleotide sequence against a nucleotide database.</p> <p>Blastx – Compares a nucleotide sequence translated into all six reading frames against a protein database.</p>	<p>E-Value – Expectation value. Statistical measure for estimating the significance of alignments.</p> <p>The smaller the E-value, the more likely that the two sequences are homologous.</p> <p>Closely related sequences have an E-value of less than 10⁻²⁰.</p>
---	--	--	--

<p>Low Complexity Region – Sequence segments (both nucleotide and amino acid) that have only a few types of bases or amino acids.</p> <p>Often removed from protein sequences before a database search as they can lead to misleading hits.</p>	<p>Motif – A conserved element of a sequence alignment.</p> <p>Constructed by the “consensus method” where multiple sequences are aligned and the most conserved regions are used to construct a pattern.</p>	<p>Logo – Visual representation of a set of aligned sequences. For each position in the sequence, a letter or set of letter is shown with larger letters indicating more conservation.</p>	
---	---	---	--

ClustalW and ClustalOmega

<p>Tools used for Multiple Sequence alignment.</p> <p>Can illustrate both transitions and transversions.</p> <p>Indels (insertions and deletions) are indicated with a “-”.</p>	<p>Example CLUSTAL OMEGA Output</p> <pre> gi 37222316 gb AY350716.1 ACAAGCTGATGACCACCTCCACAGCACTGCACCCCATTTTGTCCGCTGT gi 37222318 gb AY350717.1 ACAAGCTGATGACCACCTCCATAGCACCGCACCCATTTTGTCCGCTGT gi 37222320 gb AY350718.1 ACAAGCTGATGACCACCTCCATAGCACCGCACCCATTTTGTCCGCTGT gi 37222322 gb AY350719.1 ACAAGCTGATGACCACCTCCATAGCACCGCACCCATTTTGTCCGCTGT gi 37222324 gb AY350720.1 ACAAGCTGATGACCACCTCCATAGCACCGCACCCATTTTGTCCGCTGT gi 37222326 gb AY350721.1 ACAAGCTGATGACCACCTCCATAGCACCGCACCCATTTTGTCCGCTGT gi 37222328 gb AY350722.1 ACAAGCTGATGACCACCTCCATAGCACCGCACCCATTTTGTCCGCTGT gi 45752610:36298-36383 ACAAGCTGATGACCACCTCCATAG--CCGCACCCCATTTTGTCCGCTGT ***** </pre>
---	---

<p>RNA Nucleotide: Note –OH molecule on 3' carbon in RNA not –H as in DNA.</p>	<p>2', 3'-dideoxy analog</p> <p>-OH on 3' carbon is changed to only a –H.</p>
---	---

<p>Hairpin</p> <p>Oligo, 3 bp (Loop=4), delta G = -0.1 kcal/m</p> <pre> 5' GGGAAA 3' TATCTAGGACCTTA </pre> <p>Oligo, 2 bp (Loop=3), delta G = 2.1 kcal/m</p> <pre> 5' GGGAA 3' TATCTAGGACCTTA </pre>	<p>Self-Dimer</p> <p>4 bp, delta G = -6.6 kcal/m (bad!) (worst= -36.6)</p> <pre> 5' GGGAAAATTCAGGATCTAT 3' 3' TATCTAGGACCTTAAAGGG 5' </pre> <p>4 bp, delta G = -5.4 kcal/m (bad!) (worst= -36.6)</p> <pre> 5' GGGAAAATTCAGGATCTAT 3' 3' TATCTAGGACCTTAAAGGG 5' </pre>	<p>Dimer</p> <p>forward primer</p> <pre> 5' TATCTAGGACCTTAAAGGG 3' 3' CATGGAAACGTAGGAGAC 5' </pre> <p>reverse primer</p>
---	--	--

Acceptable and Unacceptable Primer Bonding

<p>FASTA Format – File format for specifying sequence information.</p>	<p>Techniques to Reduce the Number of Search Hits</p> <ul style="list-style-type: none"> • Provide a maximum E-Value. • Search only newest elements in the database. 	<p>Rigorous Alignment Method – Example: Dynamic Programming. Guaranteed to find the optimal alignment.</p>
---	---	---

WARNING: Numbers in input sequence were deleted.

PRODUCT SIZE: 603, PAIR ANY_TH COMPL: 5.09, PAIR 3'_TH COMPL: 11.07
TARGETS (start len

1 aatggcacctgccctaaaatagcttcccatgtgagggctagagaaaggaaaagattagac

61 cctccctggatgagagagagaaagtgaaggaggcaggcgagccattg
>>>>>>>>>>>>>>>>>>>

121 agcgatotttgtcaagcatcccagaagactgcgccatggggctcagcgacgggggaatggc

181 agttcgtgctgaacgtotgggggaagggtggaggctgacatcccaggccatgggcaggaag

241 tctcatcaggctctttaagggtcacccagagactctggagaagtttgacaagttcaagc

301 acctgaagtcagaggacgagatgaaggcgtctgaggacttaaagaagcatgggtgccaccg

361 tgctcaccgccctgggtggcatccttaagaagaaggggcatcatgaggcagagattaagc

421 ccctggcacagtgcgatgccaccaagcacaagatccccgtgaagtacctggagttcatct

481 cggaatgcacatccagggttctgcagagcaagcatcccggggactttggtgctgatgcc

541 agggggccatgaacaaggccctggagctgttcgggaaggacatggcctccaactacaagg

601 agctggggttccagggctagggccctgccgctcccacccccaccatctggccccgggt

661 tcaacagagagcgggggtotgatctcgtgtagccatatagagtttgottctgagtgctgc

721 ttgttttagtagaggtgggcaggaggagctgaggggctggggctggggtgttgaagtgg

Created by MIT.

- >>> - Left primer (5' end of the sequence)
- <<< - Right primer (3' end of the sequence)
- *** Coding sequence (begins with AUG for methionine)
- tm – Melting point (Ideally 52C-65C)
- gc% - Percent content of Cytosine/Guanine (ideally between 50-60%)
- len – Primer length (i.e. number of nucleotides)

Specifying a range for primer design:

- Initial Number – Starting Nucleotide
- Second Number – Terminal Nucleotide

Genscan Codes:

```

Gn.Ex : gene number, exon number (for reference)
Type : 1st = Initial exon (ATG to 5' splice site)
      Intr = Internal exon (3' splice site to 5' splice site)
      Term = Terminal exon (3' splice site to stop codon)
      Sngl = Single-exon gene (ATG to stop)
      Prom = Promoter (TATA box / initiation site)
      PolyA = poly-A signal (consensus: AATAAA)
S : DNA strand (+ = input strand; - = opposite strand)
Begin : beginning of exon or signal (numbered on input strand)
End : end point of exon or signal (numbered on input strand)
Len : length of exon or signal (bp)
Fr : reading frame (a forward strand codon ending at x has frame x mod 3)
Ph : net phase of exon (exon length modulo 3)
1/Ac : initiation signal or 3' splice site score (tenth bit units)
Do/T : 5' splice site or termination signal score (tenth bit units)
CodRg : coding region score (tenth bit units)
P : probability of exon (sum over all parses containing exon)
Tscr : exon score (depends on length, 1/Ac, Do/T and CodRg scores)

```

Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRq P.... Tscr..

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
-----	-----	-	-----	-----	----	--	--	-----	-----	-----	-----	-----
1.01	Init	+	151	242	92	0	2	103	77	133	0.987	13.71
1.02	Intr	+	373	595	223	1	1	100	96	217	0.999	20.91
1.03	Term	+	1446	1574	129	2	0	116	43	119	0.969	7.40
1.04	PlyA	+	1682	1687	6							1.05

In mutation planning, the **first codon (i.e. methionine) is codon 0.**

In Expsy, an open reading frame (start codon to stop codon inclusive of introns) is shown in **red**.

bl2seq – Blasting two sequences against each other. It is a type of pairwise alignment.

BRCA1 – Gene associated with an increased risk of contracting breast cancer. **It is a transcription factor and tumor suppressor.**

Steps in Cell Lifetime

1. G1 – Growth
2. S – DNA Synthesis
3. G2 – Growth and preparation for division
4. M – Mitosis

Reading a Clustal Omega Output

- ***** (**Asterisk**) - Identical alignment
- **:** (**Colon**) - Strongly conserved alignment
- **.** (**Period**)– Weakly conserved alignment
- **Blank** – No meaningful alignment

Homework #5 Keywords

Cryptic Splice Site – A splice site that is present in the genetic sequence but that is only activated when a variant disrupts the standard donor and acceptor splice sites.	NNSplice – Splice Site Predictor Program. Includes minimum donor site score and minimum acceptor site score for filtering the results.	Donor Splice Site – 5' end of intron. Usually beings with "GU".	Acceptor Splice Site – 3' end of intron. Usually beings with "AG".
BIC – Breast Cancer Information Core Database	Splice Site Prediction Tools – NNSplice, SSF, MaxEntScan, GeneSplice	Evolutionary/Genetic Distance – Metric for quantifying the difference between two or more sequences.	Phylogenetic Tree – Summarizes the key aspects of the reconstructed/hypothetical evolutionary history. Key Assumption of Phlogenetic Tree - Homology
Synonymous Mutation – Nucleotide mutations that do not change the amino acid sequence.	Nonsynonymous Mutation – Nucleotide mutation that alters the encoded amino acid.	Biased Mutation Pressure – Since many mutations in the third codon nucleotide are synonymous, more accepted mutations occur in that location than in the first two. Hence, it is often useful to ignore the third codon sites when performing evolutionary analysis.	Homoplasy – Genetic similarities that are not due to a homology/common ancestor. Example: Convergent Evolution.
Gene Loss – Loss of a gene in the chromosome. Example: Gene duplication in chromosome. One of the genes undergoes mutation such that it is no longer needed. Mutations can cause it to become a pseudogene . Duplicated gene continues to undergo mutation until it is no longer detectable. Can be caused by factors other than gene duplication.		Horizontal/Lateral Gene Transfer – Transfer of genetic information from one species to another. It is called "horizontal" to differentiate it from vertical gene transfer which is from parent to offspring. Most common in bacteria and archaea. However, it can occur in eukaryotes through viruses.	
Syntentic Region – Region of the genome that contains a series of genes in similar order to that found in a region of genome of another species. Not common as chromosomes often reorganize (shuffle/invert) and split during evolution.		Speciation Event – Point at which the population of a species divides into separate groups that subsequently diverge into different species.	small ribosome subunit rRNA – Used to determine that prokaryotes for two unique domains.
Two Assumptions in Phylogenetic Analysis 1. Rate of mutation is constant 2. Each location only when through mutation once	Taxa/Operational Taxonomic Unit (OTU) – Species in a phylogenetic tree.	Root in Phylogenetic Tree – Last Common ancestor	Leaf/External Nodes in Phylogenetic Tree – Existing species or extinct species that died out with no descent species.
Phylogenetic Tree – Representation to summarize key aspects of a reconstructed evolutionary history.	Branches/Edges in a Phylogenetic Tree – Evolutionary relationships between different species.	Bootstrap Analysis – Method of estimating support in sequence for the topological features of a phylogenetic tree where random selections of data are examined to determine the support for each split.	Condensed Tree – A tree where splits that were not well supported (e.g. occur in less than 60% of bootstrap trees) are removed.

Phylogenetic Trees

Phylogeny – History of descent of a group of organisms from a common ancestor.	True phylogeny cannot be known so it is inferred . Previously inference was done using morphological features. Now it is gone through sequence analysis.	Taxonomy – Science of classification of organisms.	Quagga – Extinct species that was a cross between zebras and horses.
Phylogenetic Analysis – Determination of how the set of organisms may have been derived during evolution.	Phylogenetic Tree – Diagram showing the inferred paths of species and genes.	Sequences that are next to each other in the tree are generally closely related while those further apart are less closely related.	Constructed from multiple sequence alignment . Allows us to understand the lineage of species and how functions evolved .
Leaves – Objects being compared. If the objects are species or classifications, then they are called taxa.	Internal Nodes – Extinct ancestral units or organisms.	Rooted Tree – Represents the evolutionary path from the for the set of species. Objects in a rooted tree should all share a common ancestor . Example: In a rooted tree, you CAN determine if yellow birds evolved from brown birds or brown birds evolved from yellow birds.	Unrooted Tree – Relationships among objects but not evolutionary paths. Example: In an unrooted tree, CANNOT determine if yellow birds evolved from brown birds or brown birds evolved from yellow birds.

Types of Phylogenetic Trees

Cladogram – Rooted tree in which the branch lengths have no meaning . Only tree topology is defined.	Additive Tree – Can be constructed from same data as a cladogram. Branch lengths represent a quantitative measure of evolution divergence . Can be rooted or unrooted.	Additive Tree with Outgroup – Same as an additive tree, but it includes a distantly related organism called an outgroup .	Ultrametric Tree – Same as an additive tree except that all branches in the tree have a common rate of mutation. This is referred to as the molecular clock . Always rooted . Present day at the bottom of the tree.

Types of Phylogenetic Trees 1. Distance Based (e.g. UPGMA) 2. Character Based – Uses morphological features.			
---	--	--	--

UPGMA

Candidate Root – Best location to start a phylogenetic tree. It is where the candidate group and outgroup diverged .	Distance Based Method – Given n sequences, build a diagonal matrix of the differences between pairs of sequences. Group sequences incrementally into pairs based off their distances.	UPGMA – Sequential clustering algorithm. Group pairs of sequences and when creating the pair, create a new amalgamated sequence.	UPGMA assumes the gene substitution rate is constant, that is mutations occur at the same rate at all points in the tree. Uniform Rate of Mutation entails that at any height in the tree, all branches have the same number of changes separating their base sequences.
--	--	---	---

Variations in Rates of Mutational Change

Molecular Clock – A measure of time for nucleotide substitutions per year.	Third site in a codon mutates faster than sites 1 and 2.	Introns mutate faster than exons.	Intergenic DNA (i.e. DNA between genes) mutates faster than intragenetic (i.e. open reading frame) DNA.	Transitions are more common than transversion.
---	---	--	--	---

Average Distance in UPGMA $d_{i,j} = \frac{1}{ C_i C_j } \sum_{p \in C_i, q \in C_j} d_{p,q}$ Distance ($d_{i,j}$) between clusters C_i and C_j where $ C_i $ and $ C_j $ is the size of C_i and C_j respectively.	Steps in UPGMA Algorithm		
	Step #1 Initialization – Assign sequence into its own cluster. Place these at height 0.	Step #2 Iteration – While there are more than two clusters, group two most similar clusters C_i and C_j into a new cluster C_k . Calculate distance ($d_{k,l}$) from C_k to all other remaining clusters C_l . Place C_k in the tree at height $\frac{d_{i,j}}{2}$. Replace clusters C_i and C_j with cluster C_k	Step #3 Termination – Group the last two remaining clusters C_i and C_j and place them in the tree at height $\frac{d_{i,j}}{2}$ in the tree.

Motifs and Logos

Motifs – Nucleotide sequence patterns of functional significance. Examples: Core promoter TATA box . CAT box at the start of transcription.	TSS – Transcription Start Site	5' UTR – 5' Untranslated Region	3' UTR – 5' Untranslated Region	Coding Sequence – Start codon to stop codon with introns removed.
--	---------------------------------------	--	--	--

Transcription – Converting of DNA information to mRNA.	Translation – Conversion of mRNA to a protein.	Three Conserved Sequences in Splicing		
		Donor Site: 5' Splice Site	Branch Site: Where Lariat Joins	Acceptor Site: 3' Splice Site

snRNP – Small nuclear ribonucleic proteins which are complexes of proteins and snRNA (small nuclear RNA)	Spliceosome – Team of snRNP molecules.	Probability of an Adenine or Thymine: 56%/2 (0.56/2)	Probability of an Cytosine or Guanine: 44%/2 (0.44/2)
---	---	---	--

Calculating Log Odds $\text{LogOdds} = \lg \left(\frac{\text{NumSeqsWithBaseX}}{\text{TotalNumSeqs} * \text{BaseFreq}} \right)$	Logs Odds Greater than 0 (Base Ratio Greater than 1.0) – Base is more frequent in that position than the average	Logs Odds Less than 0 (Base Ratio Less than 1.0) – Base is less frequent in that position than the average	Position Weight Matrix – Formed by the log odds for each base in each position of the sequence.	Determining if a Sequence is a Motif – Using the position weight matrix, add all the base weights of a sequence of equal length and if the score is above some threshold, it is a match.
--	--	--	--	---

Logos

Logo – Visual representation of a set of aligned sequences (e.g. motif) that indicate position preference in information theory .	Size of characters in a logo is proportional to that character's frequency in the sequence (i.e. its information content).	Information Theory – Quantifies the amount of information	Information at Position j in the Sequence: $I_j = \lg(4) - H_j$ $I_j = 2 + \sum_x P_{x,j} \lg(P_{x,j})$ x is one of the four nucleotide bases.
---	--	--	---

Hands On Exercise #09 and #10

Tool Names 1. NCBI Dot Plot Viewer 2. NCBI Genome Browser	FASTA – Format for storing sequence information.	"/protein_id" – Used to indicate accession number for protein sequence in a nucleotide entry. It is with the "translation".	Wild Type – Strain that prevails in natural conditions.
--	---	--	--

Four Stages in Cell Lifecycle 1. G1 – Growth 2. S – DNA Synthesis 3. G2 – Growth and preparation for cell division 4. M – Mitosis	Fishbones in a gene browser indicate the 5' end.	Degree of similarity – Same as identity score. Number of identical matches.	Mutant Type – Mutated version of the wild type.
--	--	--	--

Hands On Exercise #5 - Disease

Red line in UCSC Genome Browser shows the gene locus.	Synonymous SNPs are shown in green .	Nonsynonymous SNPs are shown in red .	Strand + or Strand – indicates strand with the mutation. Strand + is shown in UCSC genome browser. Strand – is the complementary strand.
---	--	---	--

When displaying CDS information, common notation is to make exon nucleotides in CAPITAL letters and intron nucleotides in lowercase letters .			
---	--	--	--

Hands On Exercise #6 – Sequence Comparison

SDSC – San Diego Supercomputer Center Biology Workbench – Tool that can do phylogenetic trees. Performs Multiple Sequence Alignment.	Dendrogram 	BOXSHADE – Used to create phylogenetic tree. 	DRAWGRAM – Builds Rooted Tree. Uses PHYLIP
---	-----------------------	---	--

Hands On Exercise #7 - Thalassemia

--	--	--	--

Hands On Exercise #11 – Detecting Motifs

JASPAR – Tool for visualizing logos.	In Jaspas, "REVERSE COMPLEMENT" shows the complementary strand. It swaps the strand order where first base in old strand is now last base and swaps A<->T and C<->G.	When using the log odds table, to calculate the log odds score for the reverse complement of a strand, you must swap its order (first base last and last base first and take strand's complement) similar to what JASPAR did.	
---	--	--	--

Hands On Exercise #12 – Constructing Position Weight Matrices

WebLogo – Tool from the University of California at Berkley to make logos.	Laplace Rule for Pseudocounts: Prevents overflow due to $\lg(0)$. Involves adding 1 to the numerator of base frequency and add to the denominator the number of possible characters in a given location (e.g. add 4 for a nucleotide and 20 for an amino acid).	Laplace Rule for Pseudocount Examples:		
		Nucleotide: $\frac{1}{7} \rightarrow \frac{2}{11}$	Nucleotide: $\frac{0}{7} \rightarrow \frac{1}{11}$	Amino Acid: $\frac{5}{10} \rightarrow \frac{9}{30}$

Hands On Exercise #13 – Origins of HIV

ClustalOmega – Tool used for multiple sequence alignment.	In ClustalOmega – Select "PHYLIP" to get output in format for Phylogenetic trees.	In the top row of the PHYLIP output, first number (e.g. 12) indicates the number of aligned sequences .	In the top row of the PHYLIP output, second number (e.g. 950) indicates the number of columns in the alignment .
--	--	--	---

In Pasteur tool, select	In Pasteur, we selected "Distance Based"	In Pasteur, you can enable bootstrap	Replicates – Number of bootstrap
-------------------------	---	---	---

“ Phylogeny ” to construct a phylogenetic tree.	trees (as opposed to “ Character Based ”).	analysis to verify support for features in the true.	replicates in bootstrap analysis.
--	---	---	-----------------------------------

Pasteur Phylogenetic analysis is used to run the “ neighbor ” analysis. To run it, you click on “ further analysis. ”	Pasteur refers to a “ condensed tree ” as a “ consensus tree ” where links without support are removed.	Tree out of Pasteur is unrooted.	Tool Name for Phylogenetic Tree: Mobyle @ Pasteur
---	---	---	--