Variation and Diversity

How Do Individuals of a Species Differ?

- *Physical Variation and Diversity*
- *Genetic Variation*

How Do Individuals of Species Differ?

- Genetic makeup of an individual is manifested in traits, which are caused by variations in genes

- While only the 0.1% of the 3 billion nucleotides in the human genome are different, small variations can have a large range of phenotypic expressions

- These traits make some more or less susceptible to disease, and the demystification of these mutations will hopefully reveal the truth behind several genetic diseases

The Diversity of Life

- Not only do different species have different genomes, but also different individuals of the same species have different genomes.
- No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
- Imagine the difficulty of biologists – sequencing and studying only one genome is not enough because every individual is genetically different!

Physical Traits and Variances

- Individual variation among a species occurs in populations of all sexually reproducing organisms.
- Individual variations range from hair and eye color to less subtle traits such as susceptibility to malaria.
- Physical variation is the reason we can pick out our friends in a crowd, however most physical traits and variation can only be seen at a cellular and molecular level.

Sources of Physical Variation

- Physical Variation and the manifestation of traits are caused by variations in the genes and differences in environmental influences.
- An example is height, which is dependent on genes as well as the nutrition of the individual.
- Not all variation is inheritable – only genetic variation can be passed to offspring.
- Biologists usually focus on genetic variation instead of physical variation because it is a better representation of the species.

Genetic Variation

- Despite the wide range of physical variation, genetic variation between individuals is quite small.
- Out of 3 billion nucleotides, only roughly 3 million base pairs (0.1%) are different between individual genomes of humans.
- Although there is a finite number of possible variations, the number is so high ($4^{3,000,000}$) that we can assume no two individual people have the same genome.
- What is the cause of this genetic variation?

Sources of Genetic Variation

- **Mutations** are rare errors in the DNA replication process that occur at random.
- When mutations occur, they affect the genetic sequence and create genetic variation between individuals.
- Most mutations do not create beneficial changes and actually kill the individual.
- Although mutations are the source of all new genes in a population, they are so rare that there must be another process at work to account for the large amount of diversity.

Sources of Genetic Variation

- **Recombination** is the shuffling of genes that occurs through sexual mating and is the main source of genetic variation.
- Recombination occurs via a process called **crossing over** in which genes switch positions with other genes during meiosis.
- Recombination means that new generations inherit random combinations of genes from both parents.
- The recombination of genes creates a seemingly endless supply of genetic variation within a species.

How Genetic Variation is Preserved

- **Diploid** organisms (which are most complex organisms) have two genes that code for one physical trait – which means that sometimes genes can be passed down to the next generation even if a parent does not physically express the gene.
- **Balanced Polymorphism** is the ability of natural selection to preserve genetic variation. For example, natural selection in one species of finch keeps beak sizes either large or small because a finch with a hybrid medium sized beak cannot survive.

Variation as a Source of Evolution

- Evolution is based on the idea that variation between individuals causes certain traits to be reproduced in future generations more than others through the process of Natural Selection.
- **Genetic Drift** is the idea that the prevalence of certain genes changes over time.
- If enough genes are changed through mutations or otherwise so that the new population cannot successfully mate with the original population, then a new species has been created.
- Do all variations affect the evolution of a species?

Neutral Variations

- Some variations are clearly beneficial to a species while others seem to make no visible difference.
- **Neutral Variations** are those variations that do not appear to affect reproduction, such as human fingerprints. Many such neutral variations appear to be molecular and cellular.
- However, it is unclear whether neutral variations have an effect on evolution because their effects are difficult, if not impossible to measure.
- There is no consensus among scientists as to how much variation is neutral or if variations can be considered neutral at all.

The Genome of a Species

- It is important to distinguish between the genome of a species and the genome of an individual.
- The genome of a species is a representation of all possible genomes that an individual might have since the basic sequence in all individuals is more or less the same.
- The genome of an individual is simply a specific instance of the genome of a species.
- Both types of genomes are important – we need the genome of a species to study a species as a whole, but we also need individual genomes to study genetic variation.

Human Diversity Project

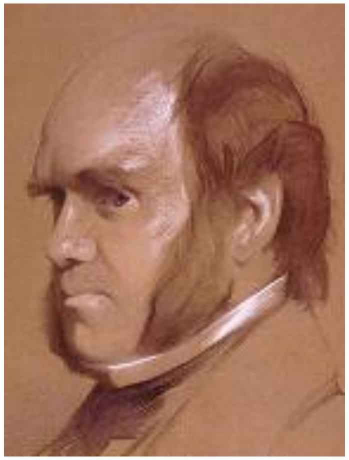
- The Human Diversity Project samples the genomes of different human populations and ethnicities to try and understand how the human genome varies.
- It is highly controversial both politically and scientifically because it involves genetic sampling of different human races.
- The goal is to figure out differences between individuals so that genetic diseases can be better understood and hopefully cured.

How Do Different Species Differ?

- *Section 10.1 – Molecular Evolution*
  - *What is Evolution*
  - *Molecular Clock*
  - *New Genes*
- *Section 10.2 – Comparative Genomics*
  - *Human and Mouse*
  - *Comparative Genomics*
  - *Gene Mapping*
  - *Cystic Fibrosis*
- *Section 10.3 – Genome Rearrangements*
  - *Gene Order*
  - *DNA Reversal*

The Biological Aspects of Molecular Evolution

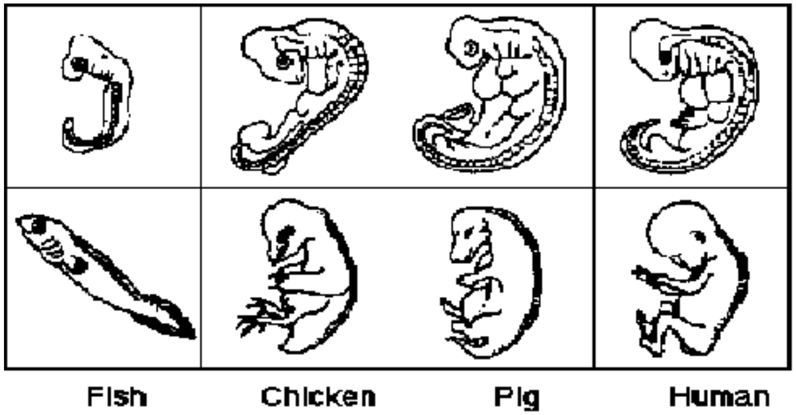
What is evolution?



- A process of change in a certain direction (*Merriam – Webster Online)*.
- In Biology: The process of biological and organic change in organisms by which descendants come to differ from their ancestor (*Mc GRAW –HILL Dictionary of Biological Science)*.
- Charles Darwin first developed the Evolution idea in detail in his well-known book *On the Origin of Species* published in 1859.

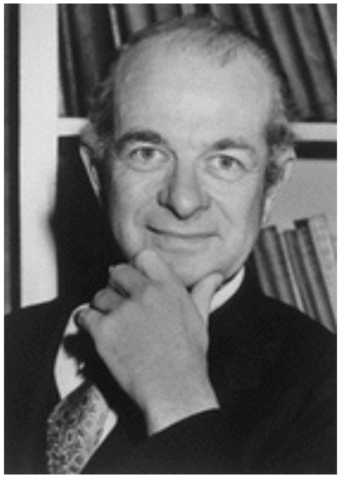
Some Conventional Tools For Evolutionary Studies

- Fossil Record: some of the biota found in a given stratum are the descendants of those in the previous stratum.
- Morphological Similarity: similar species are found to have some similar anatomical structure; For example: horses, donkeys and zebras.
- Embryology: embryos of related kinds of animals are astoundingly similar.



Molecular Clock

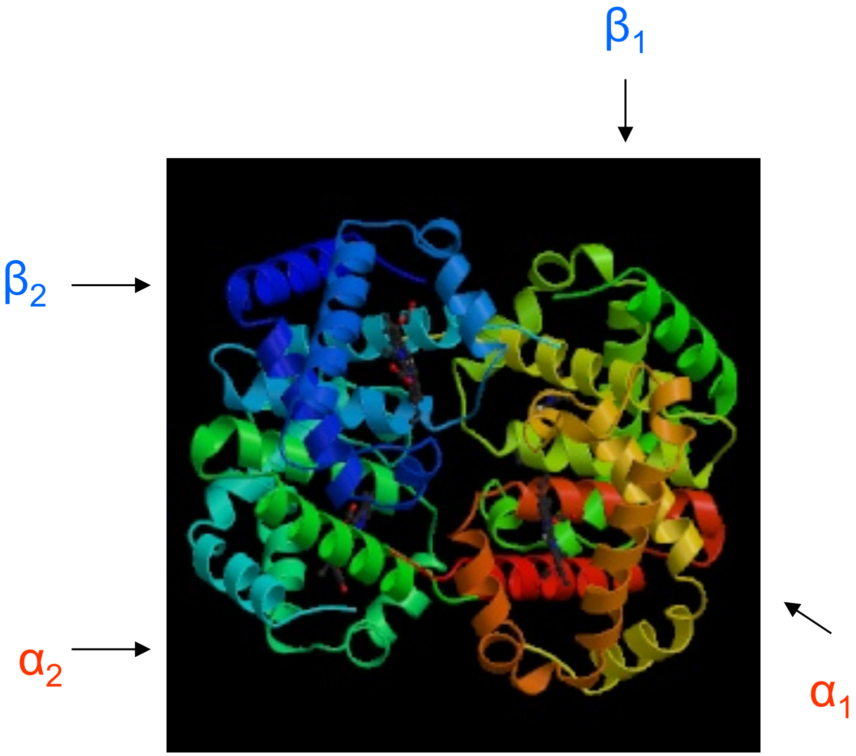
- Introduced by Linus Pauling and his collaborator Emile Zuckerkandl in 1965.
- They proposed that *the rate of evolution in a given protein (or later, DNA) molecule is approximately constant overtime and among evolutionary lineages.*



Linus Pauling

Molecular Clock Cont.

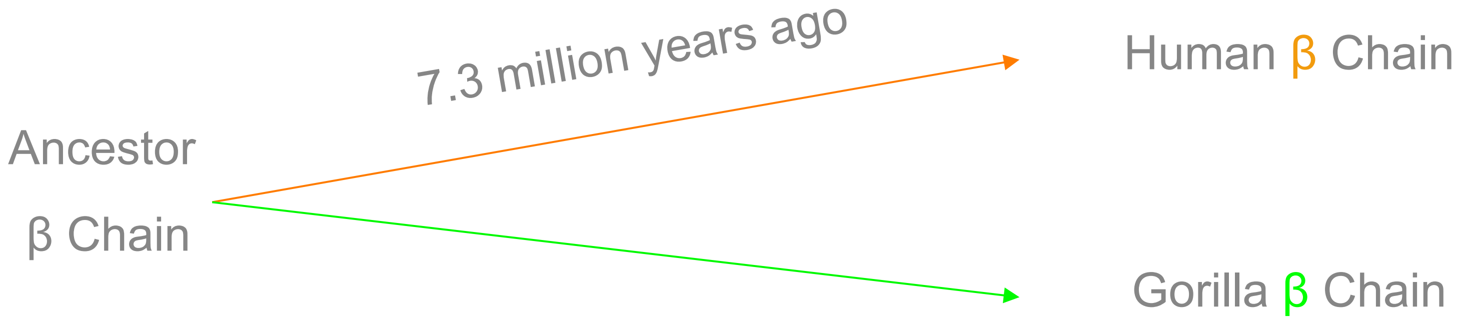
- Observing hemoglobin patterns of some primates, They found:
  - The gorilla, chimpanzee and human patterns are almost identical.
  - The further one gets away from the group of Primates, the primary structure that is shared with human hemoglobin *decreases*.
  - α and β chains of human hemoglobin are homologous, having a common ancestor.



Human Hemoglobin, A
2-α and 2-β tetramer.

Molecular Clock Cont.

- Linus and Pauling found that α-chains of human and gorilla differ by 2 residues, and β-chains by 1 residue.
- They then calculated the time of divergence between human and gorilla using evolutionary molecular clock.
- Gorilla and human α- and β-chains were found to diverge about 14.5 million and 7.3 million years ago, respectively.



7.3 million years ago

Ancestor

β Chain

Human β Chain

Gorilla β Chain

Molecular Evolution

- Pauling and Zuckerkandl research was one of the pioneering works in the emerging field of *Molecular Evolution*.
- *Molecular Evolution* is the study of evolution at molecular level, genes, proteins or the whole genomes.
- Researchers have discovered that as somatic structures evolves (*Morphological Evolution*), so does the genes. But the *Molecular Evolution* has its special characteristics.

Molecular Evolution Cont.

- Genes and their protein products evolve at different rates.

  For example, histones changes very slowly while fibrinopeptides very rapidly, revealing function conservation.

- Unlike physical traits which can evolved drastically, genes functions set severe limits on the amount of changes.

  Thought Humans and Chimpanzees lineages separated at least 6 million years ago, many genes of the two species highly resemble one another.

Beta globins:

- Beta globin chains of closely related species are highly similar:
- Observe simple alignments below:

Human β chain: MVHLT**PE**EK**S**AV**TA**LWGKV N**V**D**E**VGGEALGRLL

Mouse β chain: MVHLT**DA**EK**A**AV**NG**LWGKVN**P**D**D**VGGEALGRLL

Human β chain: VVYPWTQR**F**F**E**SFGDLS**TPD**A**V**MGNPKVKAHGKKV**LG**

Mouse β chain: VVYPWTQR**Y**F**D**SFGDLS**SAS**A**I**MGNPKVKAHGKK V**IN**

Human β chain: AF**S**DGL**A**HLDNLKGTFA**T**LSELHCDKLHVDPENFRLLGN

Mouse β chain: AF**N**DGL**K**HLDNLKGTFA**H**LSELHCDKLHVDPENFRLLGN

Human β chain: **VL**V**C**VL**A**HH**F**GKEFTP**PV**QAA**Y**QKVVAGVA**N**ALAHKYH

Mouse β chain: **MI** V**I** VL**G**HH**L**GKEFTP**CA**QAA**F**QKVVAGVA**S**ALAHKYH

There are a total of 27 mismatches, or (147 – 27) / 147 = 81.7 % identical

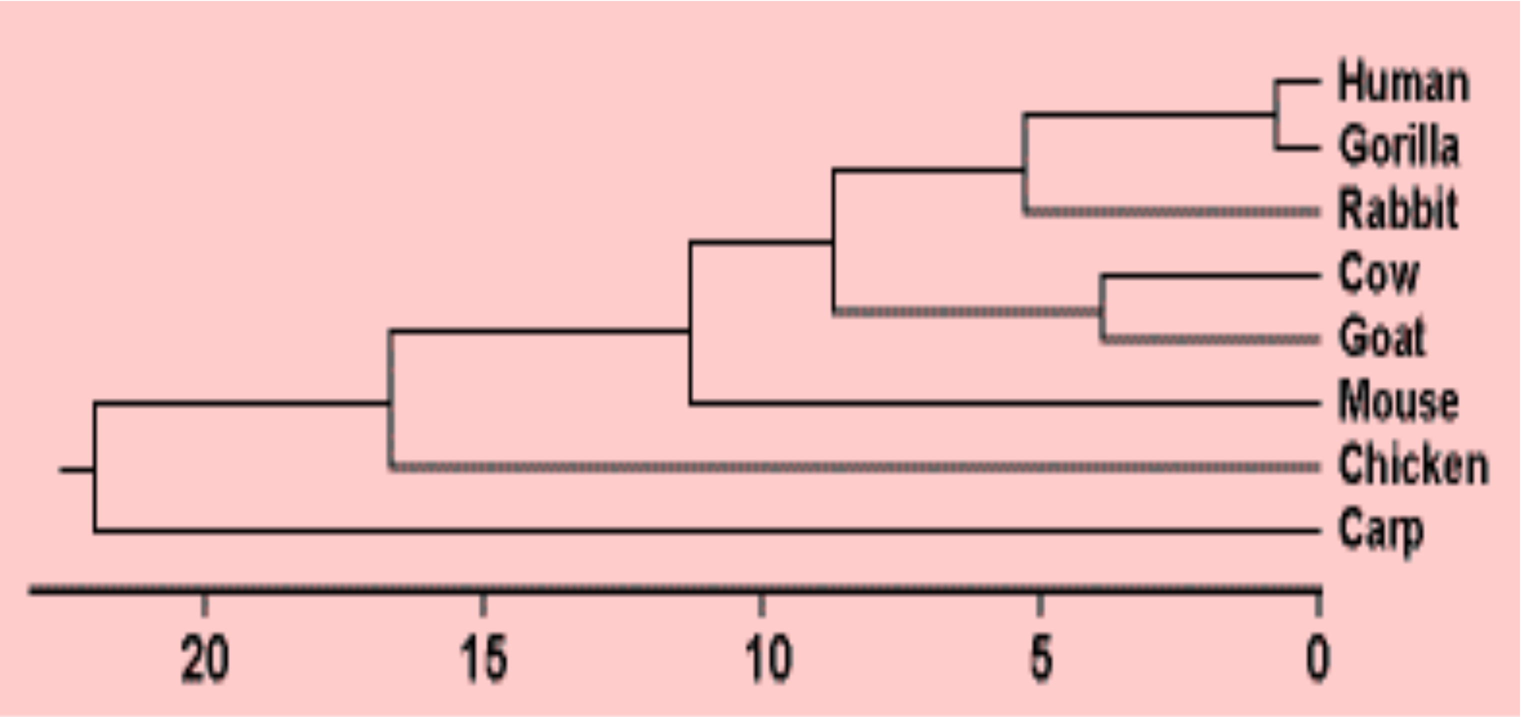
Beta globins: Cont.

Human β chain:	MVH **L** T**P**EEK**SAV**T**A**LWGKVNV**D**E**V**G**G**EAL**G**RLL
Chicken β chain:	MVH**W**T**A**EEK**QL** **I** T**G**LWGKVNV**A**E**C**G**A**EAL**A**RLL
Human β chain:	**V**VYPWTQRFF**E**SFG**D**LS**T**P**D**A**VM**GNP**K**V**K**AHGKKVL**G**
Chicken β chain:	**I**VYPWTQRFF **A**SFG**N**LS**S**P**T**A **I** **L**GNP**M**V**R**AHGKKVL**T**
Human β chain:	**A**F**S**D**GLAH**LDNLK**G**TF**AT**LSELHCDKLHVDPENFRLLG**N**
Chicken β chain:	**S**F**G**D**AVKN**LDNIK **N**TF**SQ**LSELHCDKLHVDPENFRLLG**D**
Human β chain:	**V**L**VC**VLA**H**HF**G**K**E**FTP**PV**QAA**Y** QK**V**V**AG**VA**N**ALA**H**KYH
Mouse β chain:	**I** L **I** **I** VLA**A**HF**S**K**D**FTP**EC**QAA**W**QK**L**V**RV**VA**H**ALA**R**KYH

-There are a total of 44 mismatches, or (147 – 44) / 147 = 70.1 % identical

- As expected, mouse β chain is ‘*closer*’ to that of human than chicken’s.

Molecular evolution can be visualized with phylogenetic tree.



Phylogenetic tree of Beta globin (Aligned using Clustal, PAM250)

Origins of New Genes.

- All animals lineages traced back to a common ancestor, a protish about 700 million years ago.

Comparative Genomics

How Do Different Species Differ?

- As many as 99% of human genes are conserved across all mammals
- The functionality of many genes is virtually the same among many organisms
- It is highly unlikely that the same gene with the same function would spontaneously develop among all currently living species
- The theory of evolution suggests all living things evolved from incremental change over millions of years

Mouse and Human overview

- Mouse has 2.1 x10^9 base pairs versus 2.9 x 10^9 in human.
- About 97.5% of genetic material is shared.
- 99% of genes shared of about 30,000 total.
- The 300 genes that have no homologue in either species deal largely with immunity, detoxification, smell and sex*

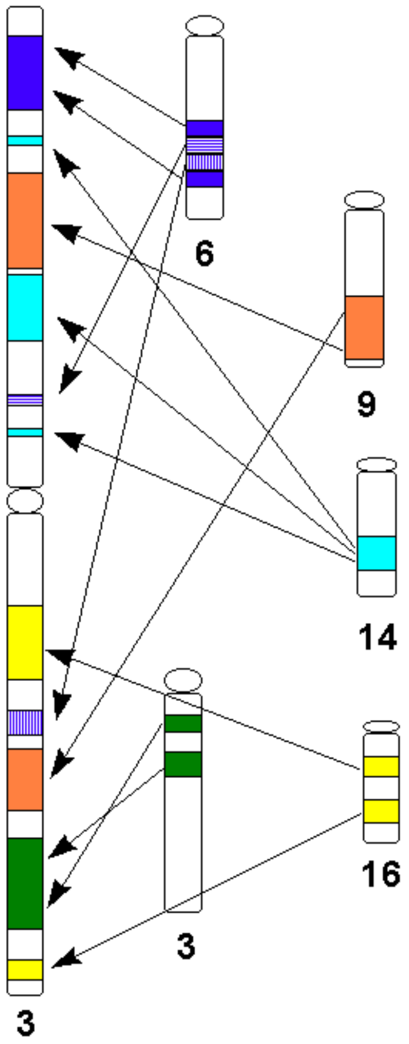
*Scientific American Dec. 5, 2002

Human and Mouse

Significant chromosomal rearranging occurred between the diverging point of humans and mice.

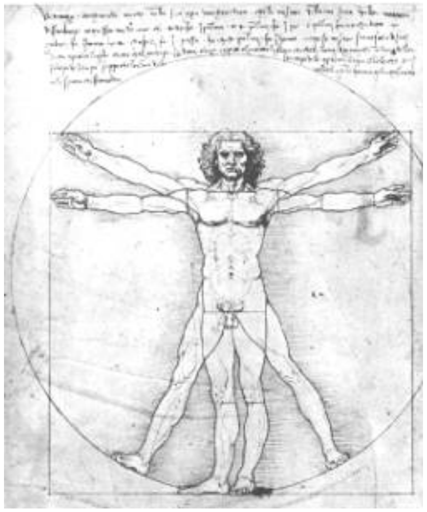
Here is a mapping of human chromosome 3.

It contains homologous sequences to at least 5 mouse chromosomes.



Comparative Genomics

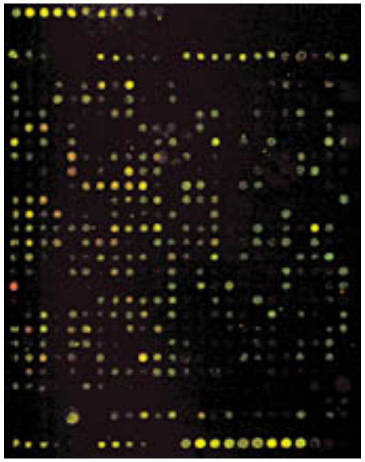
- What can be done with the full Human and Mouse Genome? One possibility is to create “knockout” mice – mice lacking one or more genes. Studying the phenotypes of these mice gives predictions about the function of that gene in both mice and humans.



Comparative Genomics

- By looking at the expression profiles of human and mouse (a recent technique using Gene Chips to detect mRNA as genes are being transcribed), the phenotypic differences can be attributed to genes and their expression.

A gene chip made by Affymetrix. The well can contain probes for thousands of genes.



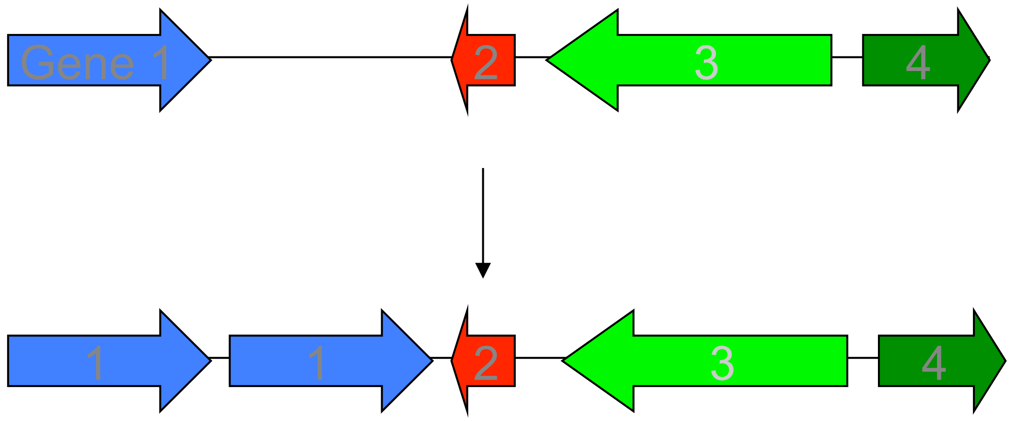
Imaging of a chip. The amount of fluorescence corresponds to the amount of a gene expressed.

Comparative Genome Sizes

- The genome of a protist Plasmodium falciparum, which causes malaria, is 23 Mb long.
- Human genome is approximately 150 times larger, mouse > 100 times, and fruit fly > 5 times larger.
- Question: How genomes of old ancestors get bigger during evolution?

Mechanisms:

- Gene duplications or insertions



Comparative Genomics

- Knowing the full sequence of human and mouse genomes also gives information about gene regulation. Because the promoter regions tend to remain conserved through evolution, looking for similar DNA upstream of a known gene can help identify regulatory sites. This technique gets more powerful the more genomes can be compared.

Gene Mapping

- Mapping human genes is critically important
  - Insight into the evolutionary relationship of human to other vertebrate species
  - Mapping disease gene create an opportunity for researchers to isolate the gene and understand how it causes a disease.

Genomics: the sub discipline of genetics devoted to the mapping, sequencing, and functional analysis of genomes

Gene Mapping

- The procedure for mapping chromosomes was invented by Alfred H.Sturterant.
  - Analysis of experiment data from Drosophilia
- Experimental data demonstrated that genes on the same chromosome could be separated as they went through meiosis and new **combination** of genes is formed.
- Genes that are tightly linked seldom recombine, whereas genes that are loosely linked recombine

Gene Mapping

- Genetic maps of chromosomes are based on recombination frequencies between markers.
- Cytogenetic maps are based on the location of markers within cytological features such as chromosome banding patterns observed by microscope.
- Physical maps of chromosomes are determined by the molecular distances in base pairs, kilobase pairs, or mega base pairs separating markers.
- High-density maps that integrate the genetic, cytological and physical maps of chromosomes have been constructed for all of human chromosomes and for many other organisms

Gene Mapping

- Recombinant DNA techniques have revolutionized the search for defective genes that cause human disease.
- Numerous major “disease genes” have already been identified by positional cloning.
  - Huntington’s disease (HD gene)
  - Cystic fibrosis (CF gene)
  - Cancer

Cystic fibrosis

- Symptoms:
  - excessively salty sweat
  - The lungs, pancreas, and liver become clogged with thick mucus, which results in chronic infections and eventual malfunction
  - Mucus often builds up in the digestive tract, causing malnourishment
  - Patients often die from infections of the respiratory system.

Cystic Fibrosis

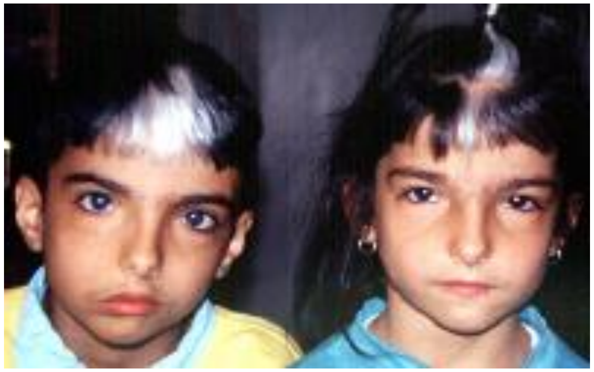
- In 1989, Francis Collins and Lap-Chee Tsui
  - identified the CF gene
  - characterized some of the mutation that cause this disease.
- A cDNA (complimentary DNA) library was prepared from mRNA isolated from sweat gland cells growing in culture and screened by colony hybridization
- CF gene product is similar to several ion channels protein,
  - which form pores between cells through which ions pass.
- Mutant CFTR protein does not function properly
  - salt accumulates in epithelial cells and mucus builds up on the surfaces of the cells.

Cystic Fibrosis

- Chromosome walking and jumping and complementary DNA hybridization were used to isolate DNA sequences, encompassing more than 500,000 base pairs, from the cystic fibrosis region on the long arm of human chromosome 7.
- neither gene therapy nor any other kind of treatment exists
- doctors can only ease the symptoms of CF
    1. antibiotic therapy combined with treatments to clear the thick mucus from the lungs.
    2. For patients whose disease is very advanced, lung transplantation may be an option.

Waardenburg’s syndrome

- Genetic disorder
- Characterized by loss of hearing and pigmentary dysphasia
- Found on human chromosome 2



Waardenburg’s syndrome

- A certain breed of mice (with splotch gene) that had similar symptoms caused by the same type of gene in humans
- Mice and Human genomes very similar → but easier to study mice
- Finding the gene in mice gives clues to where the same gene is located in humans
- Succeeded in identifying location of gene responsible for disorder in mice

Waardenburg’s syndrome

- To locate where corresponding gene is in humans, we have to analyze the relative architecture of genes of humans and mouse
- About 245 genomic rearrangements
- Rearrangement operation in this case: reversals, translocation, fusion, and fission
- Reversal is where a block of genes is flipped within a genomic sequence

Genome Rearrangements.

Turnip and Cabbage

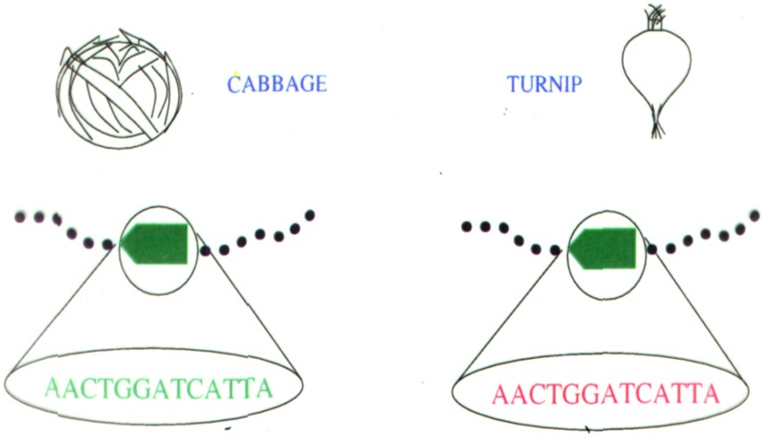
- Cabbages and turnips share a common ancestor

Jeffrey Palmer – 1980s

- discovered evolutionary change in plant organelles by comparing mitochondrial genomes of the cabbage and turnip
- 99% similarity between genes
- These more or less identical gene sequence surprisingly differed in gene order
- This finding helped pave the way to prove that genome rearrangements occur in molecular evolution in mitochondrial DNA

Important discovery

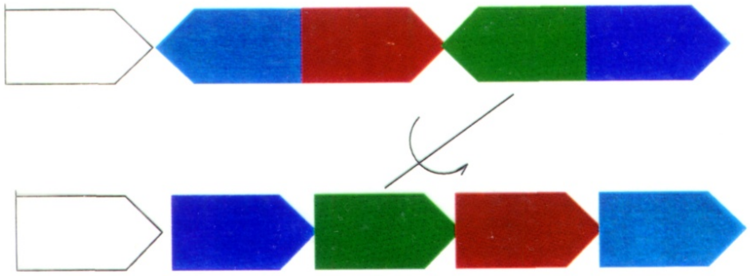
GENE SEQUENCE COMPARISON



AACTGGATCATTA
AACTGGATCATTA

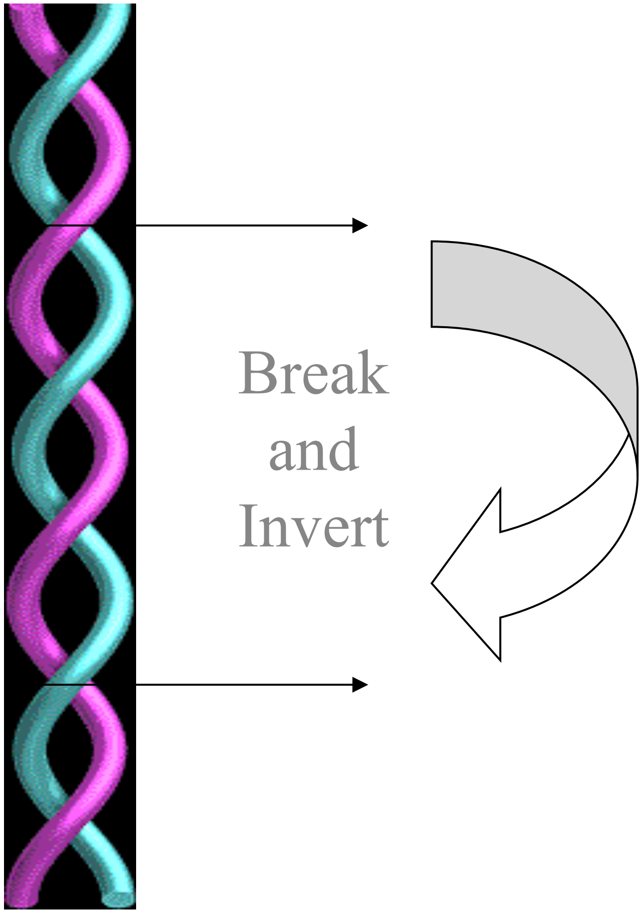
Comparing gene sequences yields no evolutionary information

GENE ORDER COMPARISON



Evolution is manifested as the divergence in Gene Order

DNA Reversal



Break and Invert

5’ATGCCTGTACTA3’

3’TACGGACATGAT5’

↓

5’ATGTACAGGCTA3’

3’TACATGTCCGAT 5’

Bioinformatics
Sequence Driven Problems

- Genomics
  - Fragment assembly of the DNA sequence.
    - Not possible to read entire sequence.
    - Cut up into small fragments using restriction enzymes.
    - Then need to do fragment assembly. Overlapping similarities to matching fragments.
    - N-P complete problem.
  - Finding Genes
    - Identify open reading frames
      - Exons are spliced out.
      - Junk in between genes

Bioinformatics
Sequence Driven Problems

- Proteomics
  - Identification of functional domains in protein’s sequence
    - Determining functional pieces in proteins.
  - Protein Folding
    - 1D Sequence → 3D Structure
    - What drives this process?

DNA… Then what?

- DNA → transcription → RNA → translation → Protein
- Ribonucleic Acid (RNA)
  - It is the messenger
    - a temporary copy
  - Why not DNA → Protein.
    - DNA is in nucleus and proteins are manufactured out of the nucleus
    - Adds a proofreading step. (Transcription = DNA→RNA)
- So actually… DNA → pre-mRNA → mRNA → Protein
  - Prokaryotes
    - The gene is continuous. Easy to translate.
  - Eukaryotes
    - Introns and Exons
    - Several Exons in different locations need to be spliced together to make a protein. (Splicing)
    - Pre-mRNA (unspliced RNA)
    - Splicisome cuts the introns out of it making processed mRNA.

Proteins

- Carry out the cell's chemistry
  - 20 amino acids
- A more complex polymer than DNA
  - Sequence of 100 has 20^{100} combinations
  - Sequence analysis is difficult because of complexity issue
  - Only a small number of the possible sequences are actually used in life. (Strong argument for Evolution)
- RNA Translated to Protein, then Folded
  - Sequence to 3D structure (Protein Folding Problem)
  - Translation occurs on Ribosomes
  - 3 letters of DNA → 1 amino acid
    - 64 possible combinations map to 20 amino acids
    - Degeneracy of the genetic code
      - Several codons to same protein

Why Bioinformatics?

- *Sequence Driven Problems*
- *Human and Mouse*
- *Comparative Genomics*
- *Gene Mapping*
- *Cystic Fibrosis*

Why Bioinformatics?

- Bioinformatics is the combination of biology and computing.
- DNA sequencing technologies have created massive amounts of information that can only be efficiently analyzed with computers.
- So far 70 species sequenced
  - Human, rat chimpanzee, chicken, and many others.
- As the information becomes ever so larger and more complex, more computational tools are needed to sort through the data.
  - Bioinformatics to the rescue!!!

What is Bioinformatics?

- Bioinformatics is generally defined as the analysis, prediction, and modeling of biological data with the help of computers

Bio-Information

- Since discovering how DNA acts as the instructional blueprints behind life, biology has become an information science
- Now that many different organisms have been sequenced, we are able to find meaning in DNA through *comparative genomics,* not unlike comparative linguistics.
- Slowly, we are learning the syntax of DNA

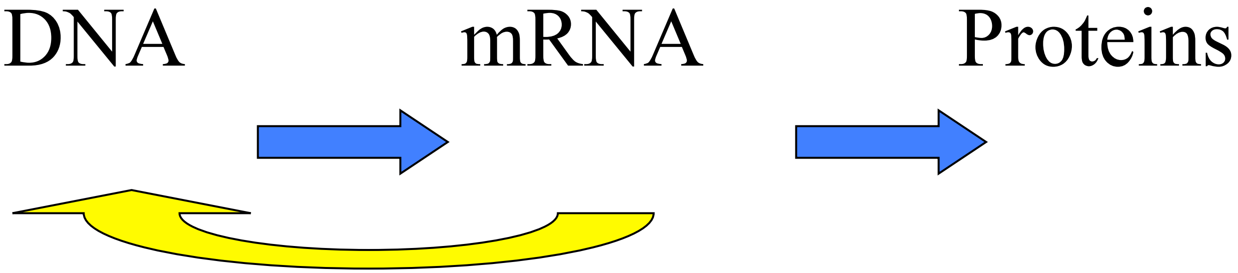
Sequence Information

- Many written languages consist of sequential symbols
- Just like human text, genomic sequences represent a language written in A, T, C, G
- Many DNA decoding techniques are not very different than those for decoding an ancient language

Amino Acid Crack

- An experiment in the early 1900s showed that all proteins are composed of sequences of 20 amino acids
- This led some to speculate that polypeptides held the blueprints of life

Central Dogma



- DNA mRNA Proteins

- DNA in chromosome is transcribed to mRNA, which is exported out of the nucleus to the cytoplasm. There it is translated into protein
- Later discoveries show that we can also go from mRNA to DNA (retroviruses).
- Also mRNA can go through alternative splicing that lead to different protein products.

Structure to Function

- Organic chemistry shows us that the structure of the molecules determines their possible reactions.
- One approach to study proteins is to infer their function based on their structure, especially for active sites.

Two Quick Bioinformatics Applications

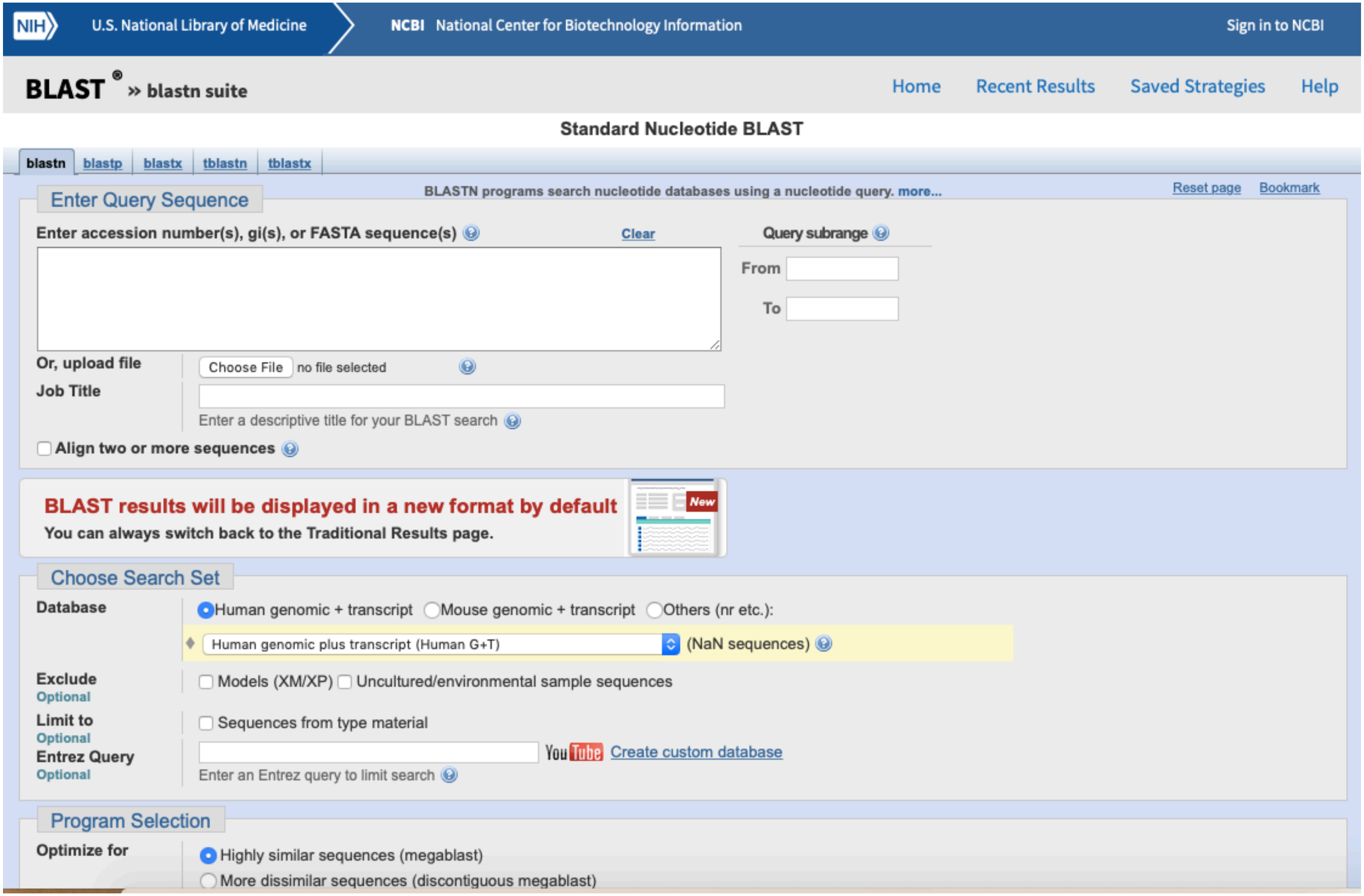
- BLAST (Basic Local Alignment Search Tool)
- PROSITE (Protein Sites and Patterns Database)

BLAST

- A computational tool that allows us to compare query sequences with entries in current biological databases.
- A great tool for predicting functions of a unknown sequence based on alignment similarities to known genes.

BLAST

- https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ® » blastn suite
Home
Recent Results
Saved Strategies
Help

Standard Nucleotide BLAST

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Choose File no file selected

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

BLAST results will be displayed in a new format by default

You can always switch back to the Traditional Results page.

Choose Search Set

Database ● Human genomic + transcript ○ Mouse genomic + transcript ○ Others (nr etc.):

Human genomic plus transcript (Human G+T) (NaN sequences)

Exclude Optional □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional □ Sequences from type material

Entrez Query Optional

Enter an Entrez query to limit search

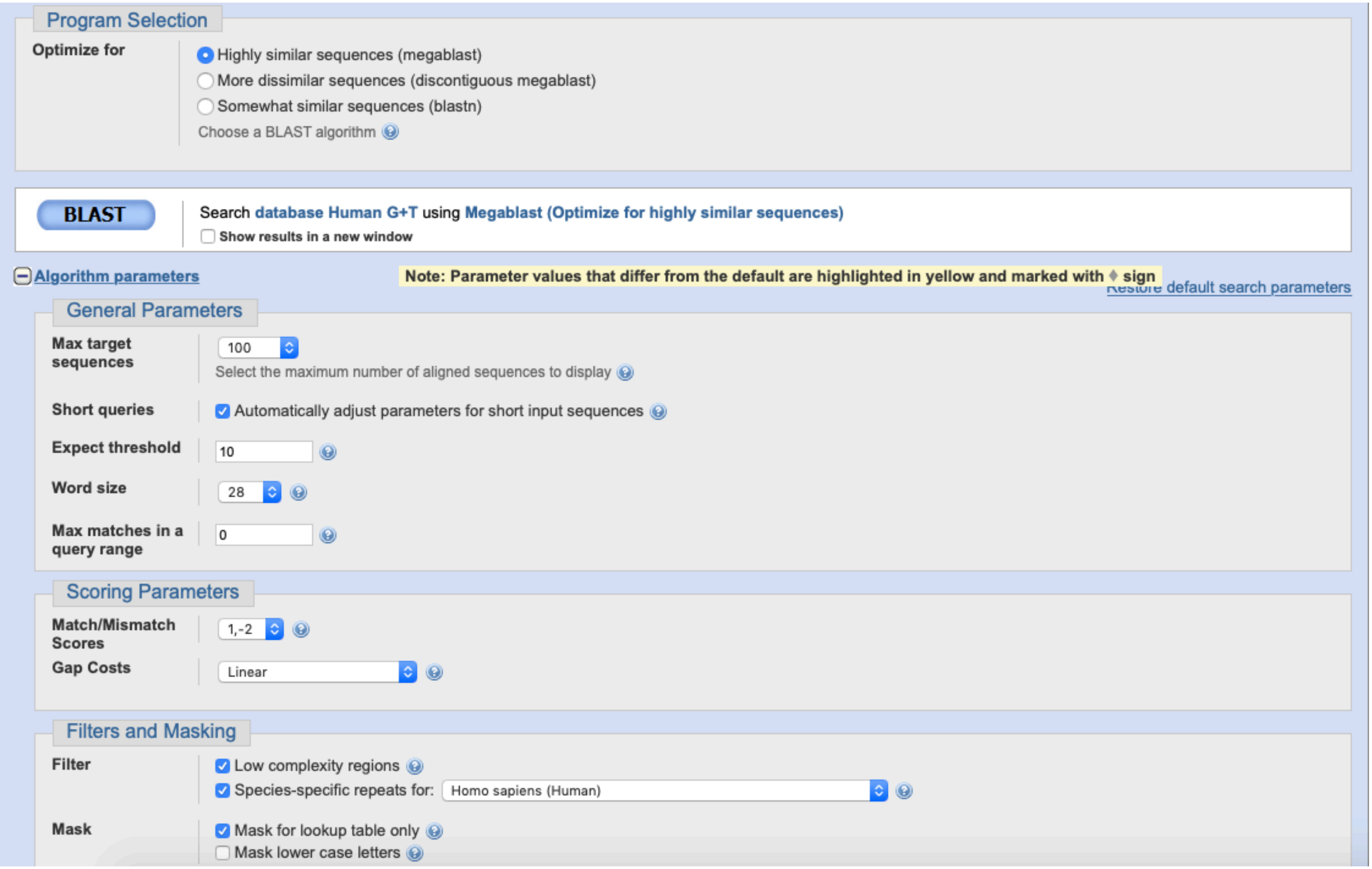
Program Selection

Optimize for ● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

BLAST

- https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn



Program Selection

Optimize for
- ● Highly similar sequences (megablast)
- ○ More dissimilar sequences (discontiguous megablast)
- ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Human G+T using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Algorithm parameters

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Restore default search parameters

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: ☑ Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 28

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2

Gap Costs: Linear

Filters and Masking

Filter: ☑ Low complexity regions

☑ Species-specific repeats for: Homo sapiens (Human)

Mask: ☑ Mask for lookup table only

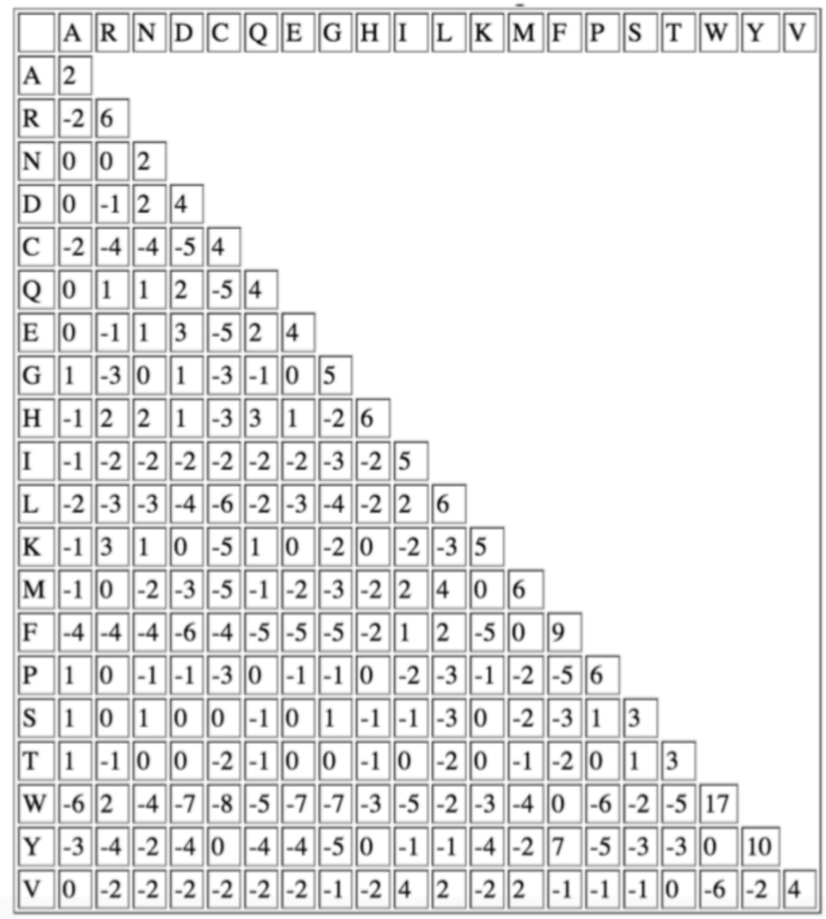
□ Mask lower case letters

Some Early Roles of Bioinformatics

- Sequence comparison
- Searches in sequence databases

Biological Sequence Comparison

- Needleman-Wunsch, 1970
  - Dynamic programming algorithm to align sequences
  - Assign a value A where the aligned pair consists of the same letters (nucleotides, amino acids)
  - If the letters differ, subtract B (edit penalty).
  - If a gap needs to be made, subtract a gap penalty times the number of gaps.



	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	4															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-4	-4	-4	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	-1	-1	-3	0	-1	-1	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	3				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-2	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

	A	G	C	T
A	10	-1	-3	-4
G	-1	7	-5	-3
C	-3	-5	9	0
T	-4	-3	0	8

with gap penalty= -5

```
Read:  AGACTAGTTAC
Ref:   CGA---GACGT
```

-3+7+10+(3)(-5)+7-4+0-1+0 = 1

Early Sequence Matching

- Finding locations of restriction sites of known restriction enzymes within a DNA sequence (very trivial application)
- Alignment of protein sequence with scoring motif
- Generating contiguous sequences from short DNA fragments.
  - This technique was used together with PCR and automated HT sequencing to create the enormous amount of sequence data we have today

Biological Databases

- Vast biological and sequence data is freely available through online databases

- Use computational algorithms to efficiently store large amounts of biological data

- Examples:
- NCBI GeneBank http://ncbi.nih.gov

  Huge collection of databases, the most prominent being the nucleotide sequence database
- Protein Data Bank http://www.pdb.org

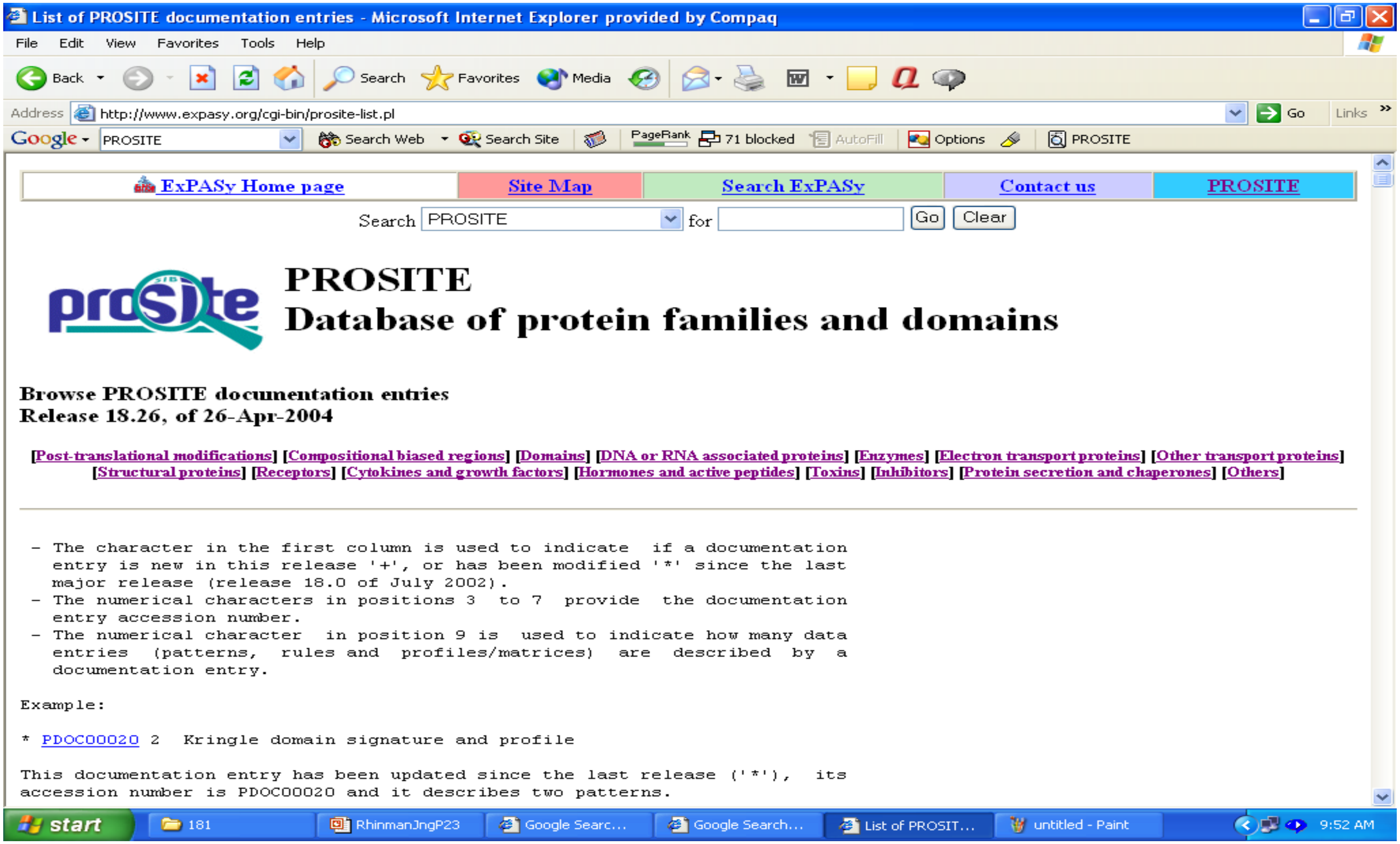
  Database of protein tertiary structures
- SWISSPROT http://www.expasy.org/sprot/
- Database of annotated protein sequences
- PROSITE http://kr.expasy.org/prosite

  Database of protein active site motifs

PROSITE Database

- Database of protein active sites.
- A great tool for predicting the existence of active sites in an unknown protein based on primary sequence.

PROSITE



PROSITE
Database of protein families and domains

Browse PROSITE documentation entries
Release 18.26, of 26-Apr-2004

[Post-translational modifications] [Compositional biased regions] [Domains] [DNA or RNA associated proteins] [Enzymes] [Electron transport proteins] [Other transport proteins] [Structural proteins] [Receptors] [Cytokines and growth factors] [Hormones and active peptides] [Toxins] [Inhibitors] [Protein secretion and chaperones] [Others]

```
 - The character in the first column is used to indicate  if a documentation
   entry is new in this release '+', or has been modified '*' since the last
   major release (release 18.0 of July 2002).
 - The numerical characters in positions 3  to 7  provide  the documentation
   entry accession number.
 - The numerical character  in position 9 is  used to indicate how many data
   entries  (patterns,  rules and  profiles/matrices)  are  described  by  a
   documentation entry.

Example:

* PDOC00020 2  Kringle domain signature and profile

This documentation entry has been updated since the last release ('*'),  its
accession number is PDOC00020 and it describes two patterns.
```

Sequence Analysis

- Some algorithms analyze biological sequences for patterns
  - RNA splice sites
  - Open reading frames (ORFs): stretch of codons
  - Amino acid propensities in a protein
  - Conserved regions in
    - AA sequences [possible active site]
    - DNA/RNA [possible protein binding site]
- Others make predictions based on sequence
  - Protein/RNA secondary structure folding

It is Sequenced, What's Next?

- Tracing Phylogeny
  - Finding family relationships between species by tracking similarities between species.
- Gene Annotation (cooperative genomics)
  - Comparison of similar species.
- Determining Regulatory Networks
  - The variables that determine how the body reacts to certain stimuli.
- Proteomics
  - From DNA sequence to a folded protein.

Modeling

- Modeling biological processes tells us if we understand a given process
- Because of the large number of variables that exist in biological problems, powerful computers are needed to analyze certain biological questions

Protein Modeling

- Quantum chemistry imaging algorithms of active sites allow us to view possible bonding and reaction mechanisms
- Homologous protein modeling is a comparative proteomic approach to determining an unknown protein's tertiary structure
- Predictive tertiary folding algorithms are a long way off, but we can predict secondary structure with ~80% accuracy.

The most accurate online prediction tools:

PSIPred

PHD

Regulatory Network Modeling

- Micro array experiments allow us to compare differences in expression for two different states
- Algorithms for clustering groups of gene expression help point out possible regulatory networks
- Other algorithms perform statistical analysis to improve signal to noise contrast

Systems Biology Modeling

- Predictions of whole cell interactions.
  - Organelle processes, expression modeling

- Currently feasible for specific processes (eg. Metabolism in E. coli, simple cells)

  Flux Balance Analysis

The future…

- Bioinformatics is still in it’s infancy
- Much is still to be learned about how proteins can manipulate a sequence of base pairs in such a peculiar way that results in a fully functional organism.
- How can we then use this information to benefit humanity without abusing it?

Sources Cited

- Daniel Sam, "Greedy Algorithm" presentation.
- Glenn Tesler, "Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes" presentation.
- Ernst Mayr, "What evolution is".
- Neil C. Jones, Pavel A. Pevzner, "An Introduction to Bioinformatics Algorithms".
- Alberts, Bruce, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, Peter Walter. Molecular Biology of the Cell. New York: Garland Science. 2002.
- Mount, Ellis, Barbara A. List. Milestones in Science & Technology. Phoenix: The Oryx Press. 1994.
- Voet, Donald, Judith Voet, Charlotte Pratt. Fundamentals of Biochemistry. New Jersey: John Wiley & Sons, Inc. 2002.
- Campbell, Neil. Biology, Third Edition. The Benjamin/Cummings Publishing Company, Inc., 1993.
- Snustad, Peter and Simmons, Michael. Principles of Genetics. John Wiley & Sons, Inc, 2003.

Pairwise Sequence Alignment

Outline

- Introduction to sequence alignment
- Pair wise sequence alignment
  - The Dot Matrix
  - Scoring Matrices
  - Gap Penalties
  - Dynamic Programming

Introduction to sequence alignment

Sequence Alignment is the identification of residue-residue correspondences.

- It is the basic tool of bioinformatics.

Sequence Alignment

- Question: Are two sequences related?
- Compare the two sequences, see if they are similar

- Example: ***pear*** and ***tear***
- Similar words, different meanings

Protein Evolution

“For many protein sequences, evolutionary history can be traced back 1-2 billion years”

-William Pearson

- When we align sequences, we assume that they share a common ancestor
  - They are then homologous
- Protein fold is much more conserved than protein sequence
- DNA sequences tend to be less informative than protein sequences

Use Protein Sequences for Similarity Searches

1) 4 DNA bases vs. 20 amino acids - less chance similarity
2) Similarity of AAs can be scored
   - # of mutations, chemical similarity, PAM matrix
3) Protein databanks are much smaller than DNA databanks
   -less random matches.
4) Similarity is determined by pairwise alignment of different sequences

Pairwise Alignment

- The alignment of two sequences (DNA or protein) is a relatively straightforward computational problem.
  - There are lots of possible alignments.
- Two sequences can always be aligned.
- Sequence alignments have to be scored.
- Often there is more than one solution with the same score.

Biological Sequences

- Similar biological sequences tend to be related
- Information:
  - Functional
  - Structural
  - Evolutionary

- Common mistake:
  - **sequence similarity is not homology!**
- Homologous sequences: derived from a common ancestor

Relation of sequences

- **Homologs**: similar sequences in 2 different organisms derived from a common ancestor sequence.
- **Orthologs**: Similar sequences in 2 different organisms that have arisen due to a speciation event. Functionality Retained.
- **Paralogs**: Similar sequences within a single organism that have arisen due to a gene duplication event.
- **Xenologs**: similar sequences that have arisen out of horizontal transfer events (symbiosis, viruses, etc)

Relation of sequences

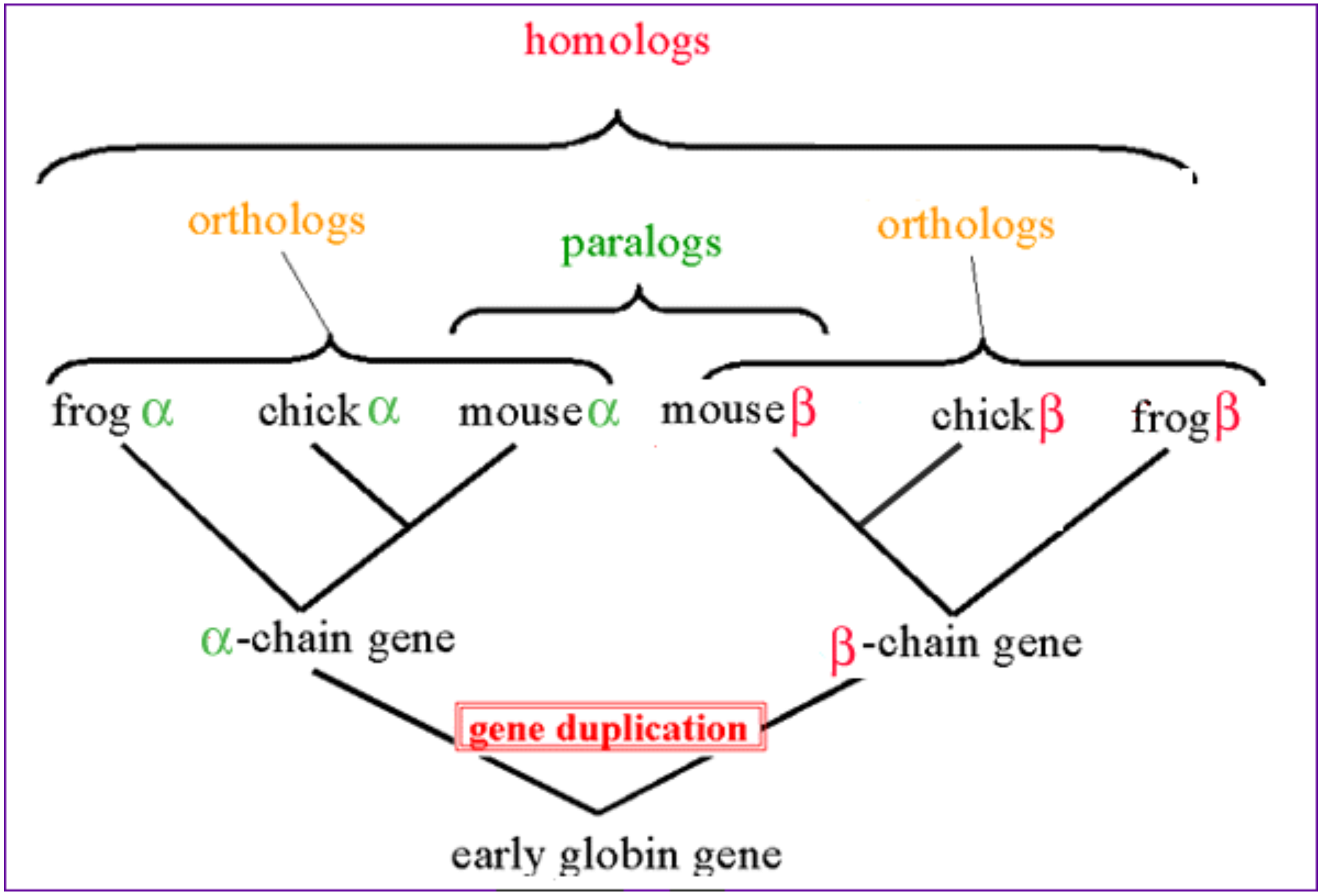


Image Source: http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Orthology.html

Sequence Alignment

The concept

- An alignment is a mutual arrangement of two sequences.
- It exhibits where the two sequences are similar, and where they differ.
- An 'optimal' alignment is one that exhibits the most correspondences, and the least differences.
- sequences that are similar probably have the same function

Sequence Alignment

Terms of sequence comparison

- **Sequence identity**
  - exactly the same Amino Acid or Nucleotide in the same position
- **Sequence similarity**
  - substitutions with similar chemical properties
- **Sequence homology**
  - general term that indicates evolutionary relatedness among sequences
  - sequences are homologous if they are derived from a common ancestral sequence
  - one speaks of percentage of sequence homology

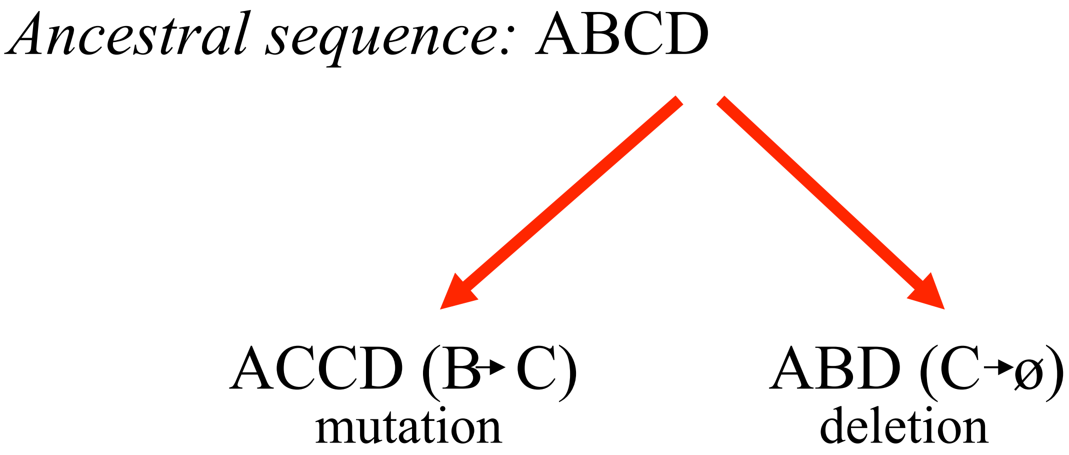
Sequence Alignment

Things to consider:

- to find the best alignment one needs to examine all possible alignments
- to reflect the quality of the possible alignments one needs to score them
- there can be different alignments with the same highest score
- variations in the scoring scheme may change the ranking of alignments

sequence alignment

Evolution:



Ancestral sequence: ABCD

ACCD (B→C)
mutation

ABD (C→ø)
deletion

ACCD
AB—D

or

ACCD
A—BD

Pairwise Alignment

true alignment

sequence alignment

A protein sequence alignment

```
MSTGAVLIY--TSILIKECHAMPAGNE-----
---GGILLFHRTHELIKESHAMANDEGGSNNS
   *  *    *  **** ***
```

A DNA sequence alignment

```
attcgttggcaaatcgcccctatccggccttaa
att---tggcggatcg-cctctacgggcc----
***   ****  **** **    ******
```

Edit Distance

- Sequence similarity: function of edit distance between two sequences

```
P E A R
  | | |
T E A R
```

Hamming Distance

- Minimum number of letters by which two words differ
- Calculated by summing number of mismatches
- Hamming Distance between PEAR and TEAR is 1

Gapped Alignments

- Biological sequences
  - Different lengths
  - Regions of insertions and deletions

- Notion of gaps (denoted by ‘-’)

```
A L I G N M E N T
  | | |   | | | |
- L I G A M E N T
```

Possible Residue Alignments

- Match
- Mismatch (substitution or mutation)
- Insertion/Deletion (INDELS – gaps)

Alignments

- Which alignment is best?

```
A - C - G G - A C T
|   |   |       | |
A T C G G A T _ C T

  A T C G G A T C T
  |   | | |     | |
  A - C G G - A C T
```

Alignment Scoring Scheme

- Possible scoring scheme:

  match: +2

  mismatch: -1

  indel –2

- Alignment 1: 5 * 2 – 0(1) – 4(2) = 10 – 0 – 8 = 2
- Alignment 2: 6 * 2 – 1(1) – 2 (2) = 12 – 1 – 4 = 7

Alignment Methods

- Visual
- Brute Force
- Dynamic Programming
- Word-Based (k tuple)