Pairwise Sequence Alignment
Motivation
Comparative genomics
Comparative genomics

- Alignment and Database Search
- Structural Bioinformatics
- Phylogenetics
Comparative genomics

- Alignment and Database Search
- Structural Bioinformatics
- Phylogenetics
WHAT?
WHAT?

- Given any two sequences (DNA or protein)

Seq 1: **VLSPADKTNVKAAWAKVGAHAAGHG**

Seq 2: **VLSEAEWQLVLHVWAKVEADVAGHG**

We are interested to know to what extent are they similar?
The Central Dogma of Molecular Biology

DNA sequence
AATTCATGAAAAATCGTATACTGGTCTGTTACCGG
CAACACTGAGAAAATGGCAGAGCTCATCGCTAAA
GGTATCATCGAATCTGGTAAAGACGTCAACACCA
TCAACGTGTCTGACGTTAACATCGATGAACTGCT
GAACGAAGATATCCTGATCCTGGGTTGCTCTGCC
ATGGGCGATGAAGTTCTCGAGGAAAGCGAATTTG

Protein sequence
MKIVYWSGTGNTEKMAELIAKGIES
GKDVTINIVSDVNIIDELLEDILILGC
SAMGDEVLEEEFEPFIEEISTKISG
KKVALFGSYGWGDGKWMRDFEER
MNGYGCVVETPLIVQNEPDEAEQD
CIEFGKKIANI
DNA and Proteins

Genome: The digital backbone of molecular biology

Proteins: Perform functions encoded in the Genes
WHY DO WE EXPECT THE SEQUENCES TO BE SIMILAR?
Proteins are *homologous* if they are related by divergence from a common ancestor.

**Summary of Darwin’s theory of evolution:**

1. Species are not fixed
2. Common descent
3. Multiplication of species
4. Gradualism
5. Natural selection

(Mayr, 1991)
Sequence Comparison and Homology

When two protein sequences are being compared and the similarity is considered statistically significant, it is highly likely that the two proteins are evolutionary related.

Proteins are \textit{homologous} if they are related by divergence from a common ancestor.

Two kinds of biological homology relationships:

\textit{Orthologs} - Proteins that carry out the same function in different species.

\textit{Paralogs} - Proteins that perform different but related functions within one organism.
Orthologs

Ancestral organism

Speciation

Organism A

Organism B

Orthologs

Organism A

Organism B
Copy of a gene inserted next to the original

Two copies mutate independently
Each can take on separate functions

All or part can be transferred from one part of genome to another
Proteins are *homologous* if they are related by divergence from a common ancestor.

Two sequences can either be homologous or not (there is no point in saying that two sequences are 70% homologous).

On the other hand, two sequence could have various degrees of similarity, e.g. “70% similar”.

Usually, when two protein sequences are being compared and the similarity is considered statistically significant (above a threshold), it is highly likely that the two proteins are evolutionary related.

**similar sequence → same ancestral sequence (Homologues) → share similar structure and function.**
Species X (extinct)

ACAGTACGT

G4C  x steps  y steps  C7T

ACACCTACGT

Species A

ACAGTAATGT

Species B

Sequence Alignment $\rightarrow$ $x + y$ (we don’t know the ancestor Sequence)
WHY COMPARE?
WHY COMPARE?

- Study evolution
- Find crucial features within a sequence
- Identify cause of diseases
- Discover function
If two sequences from different organisms are similar, they may have a common ancestor (Homologues). So sequence alignment (both pairwise and multiple) can help construct the phylogenetic tree.
Find crucial (conserved) features

Conservation of the IGFALS (Insulin-like growth factor) Between human and mouse.

CATATTGCAGTGGTCCCGCGTCAGGCT
TAAATTGCAGTGGTCCCGCGTCAGGCT

VERY SIMILAR
Conserved between many organisms

VERY DIFFERENT
Identify cause of disease

– Comparison of sequences between (highly similar) individuals can detect changes that are related to diseases
Sickle Cell Anemia

Due to 1 swapping an A for a T, causing inserted amino acid to be valine instead of glutamine in hemoglobin

Healthy Individual

>gi|28302128|ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
ACATTTGCTTTCTGACACAATGTTGTTCCACTAGCAACCTCAAACAGACACCAATGGTGCATCTGACTCCTGA
GGAGAAGATCTGCCGTACTGCCCTGTGGGGCAAGGATGAGTTGTTGGTGAGGCGCTGGGGC
AGGCTGCTGGTGGTCTACCTTTGGACCAGAGGTCTTTTGAGTCCTCTGGATCTGCACTCCTGTATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAGTGCTCGGTCGCTTTTAGTATGGAGGCCCTGGGC
TCACCTGGACAACCTCAAGGGCACCCTTTGGGCACACTGAGTGAAGCTGACGTGACACTGTGAACAGCTGCACGGAT
CCTGAGATTTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTATTTGCTAT
GGGGATATTATGAAGGGCCTTGGAGCATCTGGATTCTGCGCTAATAAAAAACATTTATTTGCTAT

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTPFEEKSAVTALWKGKNDEVGGGEALGRLLVYVPWTQRFFESFGDLSPTAVMGNPKVKAHKKKGVLG
AFSDGLAHLDNLKGTATLSELHDKLHVDPENFRLLGNVLVCVLAHFKEFPPVQAAYQKVAVAGVAN
ALAHKYH
Diseased Individual

>gi|28302128|ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
ACATTTGCTTCTGACACAACTGTGTTTCACTAGCAACCTCAAACAGACACCA__GTGTCATCTGACTGACTCCTGA
GTTGAAGTCTGCCGTTACTGCCCTGTTGGGGCAAGGTGAAGTTGGTGGTGGAGGCGCTGGGCC
AGGCTGCTGGTGTCTACCTTTTGACCACAGAGTTCTTTTGGATCTGCTCCTGATG
CTGTATGGGCAACCAAGCCTAAGGTGAAGGGCTCATGGCAAGAAAGTGCTCGTGCCTTTTAGTGATGGCCTTGGC
TCACCTGGACAACCTCAAGGGCACCTTGGCCACACTGAAGTGCTGCTGACTGTGACAAGCTGACGATC
CCTGAGAAGTTCAGGCTCTCTGGCAACCGTCTGTCTGTGCTGTGCTGGGCACTACATTAATGGCAAAAGATTCA
CCGGGATATTATGAGGGGCTTGCATCTGCTTCTGGCCTATTAAATTACGGTATTTCTGGTCCCTAAGGTCACTACTAAACT

>gi|4504349|ref|NP_000509.1| beta globin [Homo sapiens]
MVHLTP
VEKSAVTALWGBKVINDEVGGEGALGRLLLVYVPWTQRFFESFGDLSTPDAMGNPKVKAHGGKVLG
AFSDGLAHLDNLKGTATLSELHCDKLHVPENFRLLGNVLCVLAHHFGKEFTPVPQAAYVKVAVAN
ALAHKYH
Discover function

Sequences that are similar probably have the same function
Finding sequence similarities with genes of known function is a common approach to infer a newly sequenced gene’s function.

1. In 1984 Russell Doolittle and colleagues found similarities between a cancer-causing gene and normal growth factor (PDGF) gene.

2. Another success story of sequence alignment is in the identification of the Cystic Fibrosis Gene.
Cystic Fibrosis

Cystic fibrosis (CF) is a chronic and frequently fatal genetic disease of the body's mucus glands (abnormally high level of mucus in glands). CF primarily affects the respiratory systems in children.

Mucus is a slimy material that coats many epithelial surfaces and is secreted into fluids such as saliva.
Cystic Fibrosis: Inheritance

In early 1980s biologists hypothesized that CF is an autosomal recessive disorder caused by mutations in a gene that remained unknown till 1989.

Heterozygous carriers are asymptomatic.

Must be homozygously recessive in this gene in order to be diagnosed with CF.
Cystic Fibrosis: Finding the Gene

1. Human chromosome 7 long arm
2. Cystic fibrosis gene
3. CFTR protein product
   - Model of protein in cell membrane
     - Out
     - In
Cystic Fibrosis and the CFTR Protein

• CFTR (Cystic Fibrosis Transmembrane conductance Regulator) protein is acting in the cell membrane of epithelial cells that secrete mucus.

• These cells line the airways of the nose, lungs, the stomach wall, etc.
Mechanism of Cystic Fibrosis

• The **CFTR protein** (1480 amino acids) regulates a chloride ion channel
• Adjusts the “wateriness” of fluids secreted by the cell
• Those with cystic fibrosis are missing one single amino acid in their CFTR
• **Mucus ends up being too thick, affecting many organs**
Cystic Fibrosis and the CFTR Protein

• CFTR (Cystic Fibrosis Transmembrane conductance Regulator) protein is acting in the cell membrane of epithelial cells that secrete *mucus*.

• These cells line the airways of the nose, lungs, the stomach wall, etc.
Finding Similarities between the Cystic Fibrosis Gene and ATP binding proteins

ATP binding proteins are present on cell membrane and act as transport channel.

In 1989 biologists found similarity between the cystic fibrosis gene and ATP binding proteins.

A plausible function for cystic fibrosis gene, given the fact that CF involves sweat secretion with abnormally high sodium level.
Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA

John R. Riordan; Johanna M. Rommens; Bat-sheva Kerem; Noa Alon; Richard Rozmahel; Zbyszko Grzelczak; Julian Zieleni; Si Lok; Natasa Plavsic; Jia-Ling Chou; Mitchell L. Drumm; Michael C. Iannuzzi; Francis S. Collins; Lap-Chee Tsui

Cystic Fibrosis: Mutation Analysis

If a high % of cystic fibrosis (CF) patients have a certain mutation in the gene and the normal patients don’t, then that could be an indicator of a mutation that is related to CF

A certain mutation was found in 70% of CF patients, convincing evidence that it is a predominant genetic diagnostics marker for CF

This genetic marker was discovered in 1997…