Computational Methods in Genomics

PART ONE

Sami Khuri
Department of Computer Science
San José State University
San José, California, USA
khuri@cs.sjsu.edu
www.cs.sjsu.edu/faculty/khuri

Outline

- Biology Review:
  - Central Dogma of Molecular Biology
- What is Bioinformatics?
  - Human Genome Project (HGP)
  - Importance of Model Organisms
  - Databases and Tools over the Internet
- Pairwise Sequence Alignment
  - Dynamic Programming (2008)
- Multiple Sequence Alignment
- DNA Fragment Assembly Problem
- Gene Prediction (2008)
Understanding Biology I

Nothing in biology makes sense, except in the light of evolution.

*Dobzhansky, Russian geneticist (1900-75)*

Understanding Biology II

- All organisms are (probably) *evolutionarily* related to each other; i.e., descended from a single common ancestor.
- **Living organisms** are “imperfect replication machines”.
- Biology is not an exact science.
“We are our Proteins” Doolittle

Protein Factory

Proteins: basis of how biology gets things done.

A typical protein is 300-500 amino acids long and folds into a 3-dimensional structure which determines its properties.
Central Dogma of Molecular Biology

**Replication**

DNA → DNA (A,C,G,T)

**Transcription**

DNA → RNA (A,C,G,U)

**Translation**

RNA → Protein (A-Y) → {BJOUX}

- Adenine (A)
- Guanine (G)
- Cytosine (C)
- Thymine (T)
- Uracil (U)

20 Amino Acids

©2010 Sami Khuri

---

Central Dogma of Molecular Biology

**DNA replication**

- DNA synthesis (transcription)
- RNA synthesis (transcription)
- Protein synthesis (translation)

**Traits**

- Diseases
- Drug Resistance
- Physiology
- Metabolism

©2010 Sami Khuri
Prokaryotes and Eukaryotes

A **cell** is the fundamental working unit of every living organism.

There are two kinds of cells:

- **prokaryotes**, which are mostly single-celled organisms with **no cell nucleus**: archaea and bacteria.

- **eukaryotes**, which are higher level organisms, and their cells have **nuclei**: animals and plants.
Proteins and Nucleic Acids

All living organisms have a similar molecular chemistry (biochemistry). The main actors in the chemistry of life are molecules called:

– **proteins**: which are responsible for what a living being is and does in a physical sense.
  
  “We are our proteins” R. Doolittle.

– **nucleic acids**: which encode the information necessary to produce proteins and are responsible for passing the “recipe” to subsequent generations.

DNA and RNA

• Living organisms contain two kinds of nucleic acids:
  – **Ribonucleic acid** (RNA)
  – **Deoxyribonucleic acid** (DNA)

• The **central dogma** states that information flows from **DNA** to **RNA** to **protein**.

• The function of a **protein** is determined by its unique three-dimensional structure.
DNA and Chromosomes

- The **human genome**: a complete set of instructions for making an organism, consists of tightly coiled threads of **DNA** and associated protein molecules, organized into structures called **chromosomes**.
- Besides the reproductive cell and red blood cell, every single **cell** in the human body contains the **human genome**.

Autosomal and Sex Chromosomes

- The **human genome** is distributed along 23 pairs of chromosomes
  - 22 autosomal pairs
  - the sex chromosome pair, XX for females and XY for males.
- In each pair, one chromosome is **paternally** inherited, the other **maternally** inherited.
Chromosomes and Genome

The human DNA contains about three billion base pairs (A-T or C-G). Number of chromosomes in a genome is characteristic of a species.

DNA Structure

- A **deoxyribonucleic acid** or DNA molecule is a double-stranded polymer composed of four basic molecular units called nucleotides.
- Each nucleotides comprises
  - a phosphate group
  - a deoxyribose sugar
  - one of four nitrogen bases:
    - purines: **adenine** (A) and **guanine** (G)
    - pyrimidines: **cytosine** (C) and **thymine** (T).
Double Helix

- The binding of two nucleotides forms a base pair.
- The double helix is formed by connecting complementary nucleotides A-T and C-G on two strands with hydrogen bonds.
- Knowledge of the sequence on one strand allows us to infer the sequence of the other strand.
- The bases are arranged along the sugar phosphate backbone in a particular order, known as the DNA sequence, encoding all genetic instructions for an organism.

DNA Phosphodiester Backbone

Figure 1-8

(C) 2010 Sami Khuri
Sugars Present in Nucleic Acids

Pentose sugar present in **RNA** Pentose sugar present in **DNA**

Pairs of Chromosomes in Species

<table>
<thead>
<tr>
<th>Common name</th>
<th>Scientific name</th>
<th>Number of chromosome pairs</th>
<th>Common name</th>
<th>Scientific name</th>
<th>Number of chromosome pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mosquito</td>
<td><em>Culex pipiens</em></td>
<td>3</td>
<td>Wheat</td>
<td><em>Triticum aestivum</em></td>
<td>21</td>
</tr>
<tr>
<td>Housefly</td>
<td><em>Musca domestica</em></td>
<td>6</td>
<td>Human</td>
<td><em>Homo sapiens</em></td>
<td>23</td>
</tr>
<tr>
<td>Garden onion</td>
<td><em>Allium cepa</em></td>
<td>8</td>
<td>Potato</td>
<td><em>Solanum tuberosum</em></td>
<td>24</td>
</tr>
<tr>
<td>Toad</td>
<td><em>Bufo americanus</em></td>
<td>11</td>
<td>Cattle</td>
<td><em>Bos taurus</em></td>
<td>30</td>
</tr>
<tr>
<td>Rice</td>
<td><em>Oryza sativa</em></td>
<td>12</td>
<td>Donkey</td>
<td><em>Equus asinus</em></td>
<td>31</td>
</tr>
<tr>
<td>Frog</td>
<td><em>Rana pipiens</em></td>
<td>13</td>
<td>Horse</td>
<td><em>Equus caballus</em></td>
<td>32</td>
</tr>
<tr>
<td>Alligator</td>
<td><em>Alligator mississippiensis</em></td>
<td>16</td>
<td>Dog</td>
<td><em>Canis familiaris</em></td>
<td>39</td>
</tr>
<tr>
<td>Cat</td>
<td><em>Felis domesticus</em></td>
<td>19</td>
<td>Chicken</td>
<td><em>Gallus domesticus</em></td>
<td>39</td>
</tr>
<tr>
<td>House mouse</td>
<td><em>Mus musculus</em></td>
<td>20</td>
<td>Carp</td>
<td><em>Cyprinus carpio</em></td>
<td>52</td>
</tr>
<tr>
<td>Rhesus monkey</td>
<td><em>Macaca mulatta</em></td>
<td>21</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Banding Pattern of Human Chromosomes 1 to 12

Chromosomes 13 to 22
Chromosomes X and Y
Labeling a Chromosome

b) Long arm is labeled q for “queue”
   Short arm is labeled p for “petite”.
Lowest resolution: a few major bands are visible: q1, q2, q3: p1, ..
Higher resolutions show sub-bands: q11, q12 .. and even q11.1 ...

Genes

- A gene is a specific sequence of nucleotide bases along a chromosome carrying information for constructing a protein.
- Genes are part of the chromosomes.
- The distance between genes is often much larger than the genes themselves.
Exons and Introns

In eukaryotes, genes consist of:
- **exons**
  protein-coding regions
- **introns**
  noncoding regions.

Approximately 5-10% of the gene is made up of exons while the rest are **introns**.

Ribonucleic Acid - RNA

- **RNA** is found in the cell and can also carry genetic information.
- While DNA is located primarily in the nucleus, **RNA** can also be found in the cytoplasm: the cellular liquid outside the nucleus.
- **RNA** is built from the nucleotides cytosine, guanine, adenine and uracil (U) (instead of thymine).
- **RNA** has its sugar phosphate backbone containing ribose.
- **RNA** forms a single strand.
Proteins

• 20 different amino acids are used to synthesize proteins.
• The shape and other properties of each protein is dictated by the precise sequence of amino acids in it.
• The function of a protein is determined by its unique three-dimensional structure.

Structure of the Amino Acid

It is the structure of the R group that determines which of the 20 amino acids it is and its special properties.
The Twenty Amino Acids

Orange:
nonpolar and hydrophobic.

The other amino acids are:
polar and hydrophilic - "water loving".

Magenta:
acidic - "carboxy" group in the side chain.

Light blue:
basic - "amine" group in the side chain.

Central Dogma of Molecular Biology

According to the central dogma of molecular biology, there is a single direction of flow of genetic information from the DNA, which acts as the information store, through RNA molecules from which the information is translated into proteins.
Steps of the Central Dogma

Genetic information embodied in DNA is **replicated** into more DNA.

The synthesis of an RNA from a sequence of DNA. The resulting RNA is **mRNA**.

In eukaryotic cells, the mRNA is **spliced** and it migrates from the nucleus to the cytoplasm.

Messenger RNA carries coded information to ribosomes that "read" and use it for **protein synthesis**.
Transcription

Transcription is the process in which one DNA strand: the template strand, is used to synthesize a complementary RNA.

Synthesizing RNA from 5’ to 3’

The Genetic Code

Transfer RNA and Translation

- The translation from nucleotides to amino acid is done by means of **transfer RNA (tRNA)** molecules, each specific for one amino acid and for a particular **triplet** of nucleotides in mRNA called a **codon**.

- The family of tRNA molecules enables the codons in a mRNA molecule to be **translated** into the sequence of amino acids in the protein.
Codons and Anticodons

At least one kind of tRNA is present for each of the 20 amino acids used in protein synthesis.

Each kind of tRNA has a sequence of 3 unpaired nucleotides - the anticodon - which can bind to the complementary triplet of nucleotides - the codon - in an mRNA molecule.

The reading of codons in mRNA requires that the anticodons bind in the opposite direction.

| Anticodon: | 3' AAG 5' |
| Codon: | 5' UUC 3' |

Start and Stop Codons

- The codon AUG serves two related functions
  - It begins most messages; that is, it signals the start of translation placing the amino acid methionine at the amino terminal of the polypeptide to be synthesized.
  - When it occurs within the message, it guides the incorporation of methionine.
- Three **codons**, UAA, UAG, and UGA, act as signals to terminate translation. They are called **STOP codons**.
Translation

The tRNA molecules bind to the ribosome and are the physical link between the mRNA and the growing protein chain.

Steps of Translation: Initiation

- The small subunit of the ribosome binds to a site “upstream” of the start of the message.
- It proceeds downstream until it encounters the **start codon** AUG.
- It is then joined by the large subunit and a special **initiator tRNA**. The initiator tRNA binds to the **P site** on the ribosome.
- In eukaryotes, **initiator tRNA** generally carries methionine (Met).
Steps of Translation: Elongation

An aminoacyl-tRNA able to base pair with the next codon on the mRNA arrives at the A site.

The preceding amino acid is linked to the incoming amino acid with a peptide bond.

Steps of Translation: Termination

- The end of the message is marked by a STOP codon: UAA, UAG, UGG.
- No tRNA molecules have anticodons for STOP codons. A protein release factor recognizes these codons when they arrive at the A site.
- Binding of this protein releases the polypeptide from the ribosome.
- The ribosome splits into its subunits, which can later be reassembled for another round of protein synthesis.
Chain of Amino Acids
Protein Structures

Primary protein structure is sequence of a chain of amino acids

Secondary protein structure occurs when the sequence of amino acids are linked by hydrogen bonds

Tertiary protein structure occurs when certain attractions are present between alpha helices and pleated sheets

Quaternary protein structure is a protein consisting of more than one amino acid chain

3 Reading Frames of mRNA

5’          3’
-------     -------
\underline{CU} \underline{UC} \underline{AC}   \underline{GU} \underline{U} \underline{G}  \underline{A} \underline{C} \underline{C} \underline{U}  \\
|Leu| Ser|Val| Thr|

5’          3’
-------     -------
\underline{UC} \underline{A} \underline{C} \underline{G} \underline{C} \underline{G} \underline{U} \underline{A} \underline{C} \underline{C} \underline{A} \underline{U}  \\
|Ser| Ala|Leu| Pro|

5’          3’
-------     -------
\underline{CU} \underline{CAG} \underline{CGU} \underline{UAC} \underline{CA} \underline{U}  \\
|Gln| Arg|Tyr| His|
Six Reading Frames

G C T A C G G A G C T T C G G A G C
C G A T G C C T C G A A G C C T C G

Sequencing SARS

http://www.bcgsc.ca/bioinfo/SARS
What is Bioinformatics?

- The Human Genome Project (HGP)
- Mapping
- Model Organisms
- Types of Databases
- Applications of Bioinformatics
- Genome Research

Pathway to Genomic Medicine

- Human Genome Project
  - Sequencing of the human DNA
- ENCODE Project
  - Interpreting the human genome sequence
- HapMap Project
  - Implicating genetic variants with human disease
- Genomic Medicine
  - Personalized medicine
  - Cure for diseases
The Human Genome Project

• The **HGP** is a multinational effort, begun by the USA in 1988, whose aim is to produce a complete physical map of all human chromosomes, as well as the entire human DNA sequence.
  – As part of the project, genomes of other organisms such as bacteria, yeast, flies and mice are also being studied.

• The primary goal of the project is to make a series of descriptive diagrams (called **maps**) of each human chromosome at increasingly finer resolutions.

The HGP Goal

• The ultimate goal of genome research is to find all the **genes** in the **DNA sequence** and to develop tools for using this information in the study of **human biology** and **medicine**.

• **Mapping** involves:
  – dividing the chromosomes into smaller fragments that can be propagated and characterized
  – ordering (mapping) them to correspond to their respective locations on the chromosomes.
Goals of the HGP

- To **identify** all the approximately 20,000-25,000 genes in human DNA,
- To **determine** the sequences of the 3.2 billion chemical base pairs that make up human DNA,
- To **store** this information in databases,
- To **improve** tools for data analysis,
- To **address** the ethical, legal, and social issues (ELSI) that may arise from the project.
HGP Finished Before Deadline

- In 1991, the USA Congress was told that the HGP could be done by 2005 for $3 billion.
- It ended in 2003 for $2.7 billion, because of efficient computational methods.

Other Species

As part of the HGP, genomes of other organisms, such as bacteria, yeast, flies and mice are also being studied.

- Chimps are infected with SIV
- Very rarely progress to AIDS
Model Organisms

- A **model organism** is an organism that is extensively studied to understand particular biological phenomena.
- **Why have model organisms?** The hope is that discoveries made in model organisms will provide insight into the workings of other organisms.
- **Why is this possible?** This works because evolution reuses fundamental biological principles and conserves metabolic, regulatory, and developmental pathways.

<table>
<thead>
<tr>
<th>Name</th>
<th>Genome BP</th>
<th>Genes</th>
<th>Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSV1 (Herpes virus)</td>
<td>$1.5 \times 10^5$</td>
<td>70</td>
<td>1</td>
</tr>
<tr>
<td>Escherichia Coli</td>
<td>$4.6 \times 10^6$</td>
<td>4,300</td>
<td>1</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>$1.2 \times 10^7$</td>
<td>5,900</td>
<td>16</td>
</tr>
<tr>
<td>Caenorhabditis Elegans</td>
<td>$1.0 \times 10^8$</td>
<td>19,100</td>
<td>6</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>$1.8 \times 10^8$</td>
<td>13,600</td>
<td>6</td>
</tr>
<tr>
<td>Arabidopsis Thalania</td>
<td>$1.2 \times 10^8$</td>
<td>25,500</td>
<td>5</td>
</tr>
<tr>
<td>Mus Musculus</td>
<td>$2.5 \times 10^9$</td>
<td>?30,000</td>
<td>20+X/Y</td>
</tr>
<tr>
<td>Homo sapiens</td>
<td>$2.9 \times 10^9$</td>
<td>?30,000</td>
<td>22+X/Y</td>
</tr>
</tbody>
</table>
# Studying Human Diseases

<table>
<thead>
<tr>
<th>Organism</th>
<th>Human Diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>DNA repair; colon cancer and other cancers</td>
</tr>
<tr>
<td>Yeast</td>
<td>Cell cycle; cancer, Werner syndrome</td>
</tr>
<tr>
<td>Drosophila</td>
<td>Cell signaling; cancer</td>
</tr>
<tr>
<td>C. elegans</td>
<td>Cell signaling; diabetes</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>Developmental pathways; cardiovascular disease</td>
</tr>
<tr>
<td>Mouse</td>
<td>Gene expression; Lesch-Nyhan disease, cystic fibrosis, fragile-X syndrome, and many other diseases</td>
</tr>
</tbody>
</table>

Flies have **orthologs** to human diseases-causing genes in categories such as:
- neurological
- renal
- immunological
- endocrine
- cardiovascular
- metabolic
- blood-vessel and
cancerous disorders

Flies can provide insights into human disease at the **systems level**, revealing how different genes interact in vivo.
What is Bioinformatics? Set of Tools

• The use of computers to collect, analyze, and interpret biological information at the molecular level.

• A set of software tools for molecular sequence analysis

What is Bioinformatics? A Discipline

• The field of science, in which biology, computer science, and information technology merge into a single discipline.

Definition of NCBI (National Center for Biotechnology Information)

• The ultimate goal of bioinformatics is to enable the discovery of new biological insights and to create a global perspective from which unifying principles in biology can be discerned.
Bioinformatics and the Internet

• The enormous increase in biological data has made it necessary to use computer information technology to collect, organize, maintain, access, and analyze the data.
• Computer speed, memory, and exchange of information over the Internet has greatly facilitated bioinformatics.
• The bioinformatics tools available over the Internet are accessible, generally well developed, fairly comprehensive, and relatively easy to use.

What do Bioinformaticians do?

• Analyze and interpret data
• Develop and implement algorithms
• Design user interface
• Design database
• Automate genome analysis
• Assist molecular biologists in data analysis and experimental design.
Why Study Bioinformatics?

- Bioinformatics is intrinsically interesting
- Bioinformatics offers the prospect of finding better drug targets earlier in the drug development process.
  - By looking for genes in model organisms that are similar to a given human gene, researchers can learn about protein the human gene encodes and search for drugs to block it.
Databases for Storage and Analysis

- Databases store data that need to be analyzed
- By comparing sequences, we discover:
  - How organisms are related to one another
  - How proteins function
  - How populations vary
  - How diseases occur
- The improvement of sequencing methods generated a lot of data that need to be:
  - stored - organized - curated
  - annotated - managed - networked
  - accessed - assessed

Types of Databases

- **Sequence**
  - Genbank, SwissProt, 3D structure, carbohydrates, organism specific, phylogenetic, sequence patterns
- **Literature**
  - Medline, OMIM, Patents, eJournals
- **Graphical**
  - Swiss2D-Page
- **Expression Analysis Databases**
  - Microarrays
- **Protein Interaction Databases**
  - Pathways
Three Major Databases

- **GenBank** from the NCBI (National Center of Biotechnology Information), National Library of Medicine
- **EBI** (European Bioinformatics Institute) from the European Molecular Biology Library
  http://www.ebi.ac.uk
- **DDBJ** (DNA DataBank of Japan)
  http://www.ddbj.nig.ac.jp

GenBank Taxonomic Sampling

<table>
<thead>
<tr>
<th>Organism</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>62.1%</td>
</tr>
<tr>
<td>Mus musculus</td>
<td>7.7%</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>6.1%</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>3.3%</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>2.9%</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>1.3%</td>
</tr>
<tr>
<td>Rattus norvegicus</td>
<td>0.8%</td>
</tr>
<tr>
<td>Danio rerio</td>
<td>0.6%</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>0.6%</td>
</tr>
</tbody>
</table>
The major DNA databases are updated and synchronized daily.

What does NCBI do?

**NCBI:** established in 1988 as a national resource for molecular biology information.

- it creates public databases,
- it conducts research in computational biology,
- it develops software tools for analyzing genome data, and
- it disseminates biomedical information,

all for the better understanding of molecular processes affecting human health and disease.
GenBank

GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

Interesting Databases

- UCSC Human Genome Browser
  - http://genome.ucsc.edu/
- Organism specific information:
  - Yeast: http://genome-www.stanford.edu/Saccharomyces/
  - Arabidopsis: http://www.tair.org/
  - Mouse: http://www.jax.org/
  - Fruit fly: http://www.fruitfly.org/
  - Nematode: http://www.wormbase.org/
The European Molecular Biology Laboratory (EMBL) was established in 1974.

It is supported by sixteen countries.

EMBL consists of five facilities:

- The main Laboratory in Heidelberg (Germany),
- Outstations in Hamburg (Germany), Grenoble (France) and Hinxton (the U.K.), and an external Research Programme in Monterotondo (Italy).

NCBI – EMBL - DDJB

NCBI, at the NIH campus, USA

EMBL: http://www.embl-heidelberg.de/
European Molecular Biology Laboratory, UK

DDBJ: http://www.ddbj.nig.ac.jp
DNA Databank of Japan

Nucleic acid Databases
Applications of Genome Research

Current and potential applications of Genome Research include:

- Molecular Medicine
- Microbial Genomics
- Risk Assessment
- Bioarcheology, Anthropology, Evolution and Human Migration
- DNA Identification
- Agriculture, Livestock Breeding and Bioprocessing

Molecular Medicine

- Improve the **diagnosis** of disease
- Detect genetic **predispositions** to disease
- Create drugs **based on molecular information**
- Use **gene therapy** and control systems as drugs
- Design **custom drugs** on individual genetic profiles.
Microbial Genomics

- Swift detection and treatment in clinics of disease-causing microbes: pathogens
- Development of new energy sources: biofuels
- Monitoring of the environment to detect chemical warfare
- Protection of citizens from biological and chemical warfare
- Efficient and safe clean up of toxic waste.

DNA Identification I

- Identify potential suspects whose DNA may match evidence left at crime scenes
- Exonerate persons wrongly accused of crimes
- Establish paternity and other family relationships
- Match organ donors with recipients in transplant programs
Louis XVII

Louis XVII: son of Louis XVI and Marie-Antoinette who died from tuberculosis in 1795 at the age of 12

DNA Identification II

- Identify endangered and protected species as an aid to wildlife officials and also to prosecute poachers
- Detect bacteria and other organisms that may pollute air, water, soil, and food
- Determine pedigree for seed or livestock breeds
- Authenticate consumables such as wine and caviar
What have we learned from HGP?

A small portion of the genome codes for proteins, tRNAs and rRNAs.

The small number of genes
Alternative Splicing

Convert all this progress into real riches for science, society, and patients
Objectives of Molecular Biology

- Extract the information in the genomes.
- Understand the structure of the genome.
- Apply this understanding to the diagnosis and treatment of genetic diseases.
- Explain the process of evolution by comparing genomes of related species.

Goals of Modern Molecular Biology

- Read the entire genomes of living things
- Identify every gene
- Match each gene with the protein it encodes
- Determine the structure and function of each protein.
Objectives of Bioinformatics

Development and use of mathematical and computer science techniques to help solving the problems in molecular biology.

Bioinformatics Problems

- Reconstructing long DNA sequences from overlapping string fragments.
- Comparing two or more sequences for similarities.
- Storing, retrieving and comparing DNA sequences and subsequences in databases.
- Exploring frequently occurring patterns of nucleotides.
- Finding informative elements in protein and DNA sequences.
- Finding evolutionary relationships between organisms.
Main Aim of the Problems

- The aim of these problems is to learn about the **functionality** and/or the **structure** of protein without actually having to physically construct the protein itself.
- The research is based on the assumption that similar sequences produce similar proteins.

---

Functional: Coding v/s Noncoding

<table>
<thead>
<tr>
<th></th>
<th>Coding Sequence (Genes)</th>
<th>Non-Coding Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifying Computational Tools</td>
<td>Relatively Easy Improving Tools</td>
<td>Very Hard Poor predictive tools</td>
</tr>
<tr>
<td>Signals What to look for</td>
<td>We Have a Good Understanding</td>
<td>Very little is known</td>
</tr>
<tr>
<td>Complementary data we can use</td>
<td>Available – Ex. ESTs and cDNAs</td>
<td>Unavailable</td>
</tr>
</tbody>
</table>
Post Human Genome Project

- Major role for comparative sequence analysis will be the identification of functionally important, non-coding sequences.
- Need to study the relation between Sequence Conservation and Sequence Function.
- Focus on the interpretation of the human genome.
- Learn the functional landscape of the human genome.
- **Challenge**: go from sequence to function
  - i.e., define the role of each gene and understand how the genome functions as a whole.

Pairwise and Multiple Sequence Alignment

- Homology
- Similarity
- Global string alignment
- Local string alignment
- Dynamic programming
- Scoring matrices:
  - PAM and BLOSUM
- BLAST family
Sequence Alignment

- **Sequence alignment** is the procedure of comparing sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences.
  - Comparing two sequences gives us a **pairwise alignment**.
  - Comparing more than two sequences gives us a **multiple sequence alignment**.

Why Do We Align Sequences?

- The basic idea of aligning sequences is that similar DNA sequences generally produce similar proteins.
- To be able to predict the characteristics of a protein using only its sequence data, the structure or function information of known proteins with similar sequences can be used.
- To be able to check and see whether two (or more) genes or proteins are evolutionarily related to each other.
Query Sequence

If a query sequence is found to be significantly similar to an already annotated sequence (DNA or protein), we can use the information from the annotated sequence to possibly infer gene structure or function of the query sequence.

Global and Local Alignments

• **Global Alignment**:
  – Are these two sequences generally the same?

• **Local Alignment**:
  – Do these two sequences contain high scoring subsequences?

• **Local similarities** may occur in sequences with different structure or function that share common substructure or subfunction.
Local Alignments

Thus, the best local alignment achieved from the above Dynamic Programming is:

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>A</td>
<td>G</td>
<td>A</td>
<td>G</td>
</tr>
</tbody>
</table>

Scoring Systems

- Use of the **dynamic programming** method requires a scoring system for
  - the comparison of symbol pairs (**nucleotides** for DNA sequences & **amino acids** for protein sequences),
  - a scheme for insertion/deletion (gap) penalties.
- The most commonly used scoring systems for protein sequence alignments are the log odds form
  - of the **PAM250** matrix and
  - the **BLOSUM62** matrix.
- A number of other choices are available.
Scoring Matrices (I)

• Upon evaluating a sequence alignment, we are really interested in knowing whether the alignment is random or meaningful.
• A scoring matrix (table) or a substitute matrix (table) is a table of values that describe the probability of a residue (amino acid or base) pair occurring in an alignment.

Scoring Matrices (II)

• The alignment algorithm needs to know if it is more likely that a given amino acid pair has occurred randomly or that it has occurred as a result of an evolutionary event.
• Similar amino acids are defined by high-scoring matches between the amino acid pairs in the substitution matrix.
**BLOSUM62 Table**

The unit in this table is the **bit**. Sometime, **half-bits** are used.

**The Roles of the Scoring Matrices**

The quality of the alignment between two sequences is calculated using a **scoring system** that favors the matching of related or identical amino acids and penalizes poorly matched amino acids and gaps.
Comparison: PAM and BLOSUM Matrices

The PAM model is designed to track the evolutionary origins of proteins, whereas the BLOSUM model is designed to find their conserved domains.

<table>
<thead>
<tr>
<th>BLOSUM 80</th>
<th>BLOSUM 62</th>
<th>BLOSUM 45</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAM 1</td>
<td>PAM 120</td>
<td>PAM 250</td>
</tr>
</tbody>
</table>

Less divergent $\rightarrow$ More divergent

BLAST

- Basic Local Alignment Search Tool
- Heuristic method for local alignment
- Designed specifically for database searches
- Idea: Good alignments contain short lengths of exact matches.
The BLAST Family

- **blastp**: compares an amino acid query sequence against a protein sequence database.
- **blastn**: compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx**: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.

Multiple Sequence Alignment

- Progressive Alignment
- Iterative Pairwise
- Guide Tree
- ClustalW
- Co-linearity
- Multiple Sequence Alignment Editors
What is Multiple Alignment

Most simple extension of pairwise alignment

Given:
- Set of sequences
- Match matrix
- Gap penalties

Find:
Alignment of sequences such that an optimal score is achieved.

Uses of Multiple Alignment

A good alignment is critical for further analysis
- Determine the relationships between a group of sequences
- Determine the conserved regions
- **Evolutionary Analysis**
  - Determine the phylogenetic relationships and evolution
- **Structural Analysis**
  - Determine the overall structure of the proteins
Heuristic Algorithms

- Based on a **progressive pairwise** alignment approach
  - ClustalW (*Cluster Alignment*)
  - PileUp (GCG)
  - MACAW
- Builds a global alignment based on **local alignments**
- Builds local multiple alignments
- Based on **Hidden Markov Models**
- Based on **Genetic algorithms**.

Progressive Strategies for MSA

- A common strategy to the MSA problem is to **progressively align** pairs of sequences.
  - A starting pair of sequences is selected and aligned
  - Each subsequent sequence is aligned to the previous alignment.
- **Progressive alignment** is a greedy algorithm.
Iterative Pairwise Alignment

- The greedy algorithm:

  \[
  \text{align some pair} \\
  \text{while not done} \\
  \text{pick an unaligned string “near”} \\
  \text{some aligned one(s)} \\
  \text{align with the previously aligned group}
  \]

- There are many variants to the algorithm.

Steps of ClustalW

<table>
<thead>
<tr>
<th>S_1</th>
<th>S_2</th>
<th>S_3</th>
<th>S_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>9</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Multiple Alignment Step:
1. Aligning S_1 and S_3
2. Aligning S_2 and S_4
3. Aligning (S_1,S_3) with (S_2,S_4).

Cluster Analysis

Dendrogram
ClustalW: An Example

By using the same five sequences and aligning them with CLUSTALW, we get the illustrated results.

DNA Fragment Assembly

- Overlap Graphs
- Shotgun Sequencing
- Repeated Regions
- Sequencing by Hybridization
- Hamiltonian Cycle
- Euler Path
To Sequence

• To sequence a DNA molecule is to obtain the string bases that it contains.
• In large scale DNA sequencing we have to sequence large DNA molecules (thousands of base pairs).

Introduction

• It is impossible to directly sequence contiguous stretches of more than a few hundred bases.
• On the other hand, we know how to cut random pieces of a long DNA molecule and to produce enough copies of the molecule to sequence.
• A typical approach to sequence long DNA molecules is to sample and then sequence fragments from them.
• The problem is that these pieces (fragments) have to be assembled.
Fragment Assembly Problem

- In large scale DNA sequencing, we are given a collection of many fragments of short DNA sequences.
- The fragments are approximate substrings of a very long DNA molecule.
- The Fragment Assembly Problem consists in reconstructing the original sequence from the fragments.

Steps of Fragment Assembly

1. Duplicate DNA
2. Sort and align
3. Sequence
4. Call Bases
   - CCGTACCCACCCCTCCCGCGTTCT
   - CCGGAAAGTCCTCCCGCGGTAGCCG
   - AAGCCTTTTTCCCGAACAGGCTCCCGCGGTAGCCG
   - CCGTACCCCGATCCCCTCCCGG
5. Call Consensus
   - AAGCCTTTTTCCCGAACAGGCTCCCGCGGTAGCCG
   - CCGGAAAGTCCTCCCGCGGTAGCCG
   - CCGTACCCCGATCCCCTCCCGG
Scan of Gel by Sequencer

**Genome**

1. Make random clones.
2. Sequence each clone.
3. Overlap sequence reads.
4. Overlap contigs for complete sequence.
Genome Sequencing Strategies

- Human Genome Project: map-based strategy
  - individual clones subjected to shotgun sequencing
  - shotgun fragments then reassembled
- Celera: whole genome sequence strategy
  - shotgun sequencing
Fig. 1. Sequencing strategies. (Left) The Hierarchical shotgun (HS) strategy involves decomposing the genome into a tiling path of overlapping BAC clones, performing shotgun sequencing on and reassembling each BAC, and then merging the sequences of adjacent clones. The method has the advantage that all sequence contigs and scaffolds derived from a BAC belong to a single compartment with respect to anchoring to the genome. (Right) Whole-genome shotgun (WGS) strategy involves performing shotgun sequencing on the entire genome and attempting to reassemble the entire collection. With the WGS method, each contig and scaffold is an independent component that must be anchored to the genome. In general, many scaffolds may not be anchored without directed efforts. (Contigs are contiguous blocks of sequence; scaffolds are sets of contigs joined by paired reads from both ends of a plasmid insert.)
Complicating Factors

DNA sequencing is very challenging since:

- Real problem instances are very large.
- Many fragments contain errors:
  - Base call errors
  - Chimeras
  - Vector contamination
- The orientation of the fragments is frequently unknown; and both strands must be analyzed.
- There might be a lack of coverage.

Models

- Models of the fragment assembly problem:
  - Shortest Common Superstring
  - Reconstruction
  - Multicontig
- None addresses the biological issues completely.
- Assumption:
  - Fragment collection is free of contamination and chimeras.
Shortest Common Superstring

- The Shortest Common Superstring (SCS):
  One of the first attempts to formalize the Fragment Assembly Problem.
- Look for the shortest superstring from a collection of given strings.
- SCS limitations in representing the fragment assembly problem:
  - Does not account for errors.
  - NP hard problem, hence approximation algorithms are used.

SCS Problem Definition

- **Input**: A collection $F$ of strings
- **Output**: A shortest possible string $S$ such that for every $f$ belonging to $F$, $S$ is a superstring of $f$.
  - $F$ corresponds to the fragments
  - Each fragment is given by its sequence in the correct orientation
  - $S$ is the sequence of the target DNA molecule.
SCS: An Example

Example

- Let $F = \{ACT, CTA, AGT\}$
- **SCS** of $F$, sequence $S = ACTAGT$
- $S$ contains all possible fragments in $F$ as substrings.

FAP Algorithms

- The algorithms we consider:
  - Fragments have no errors
  - Fragments are of known orientation
- Representing overlays:
  - Common superstring correspond to paths in a graph based on the collection of fragments.
  - Properties of these superstrings are translated to properties of paths
- It is easier to relate new problems to graphs due to familiarity and knowledge we have about them.
Overlap Directed Graphs

• Given a set $F$ of fragments, we can construct a directed graph as follows:
  – The vertices of $F$ represent the given DNA fragments.
  – If there is an overlap between the suffix of fragment $F_1$ and the prefix of fragment $F_2$, then an edge is drawn from $F_1$ to $F_2$.
  – Each edge is given a weight corresponding to the length of the overlap.

Overlap Graphs

• Note that the Overlap Graph:
  – Is a multigraph since we can have more than one edge between any 2 vertices in the graph
  – There is an edge between any 2 vertices with weight zero
• To find the target DNA sequence, we look for a Hamiltonian path: A path that visits each vertex exactly once.
• We choose the Hamiltonian path with the largest sum of edges.
Example 2: Overlap Multigraph

\[ F_1 = AGACCGCGTATAG \]
\[ F_2 = CGTATAGCATCAG \]
\[ F_3 = TAGCATCAGCATGACGCGT \]
\[ F_4 = GCGTAG \]
\[ F_5 = TAGCACTA \]

Reconstruct the target DNA sequence from the given fragments

AGACCGCGTATAG
CGTATAGCATCAG
TAGCATCAGCATGACGCGT
GCGTAG
TAGCACTA
CGTATAGCATCAGGCCCTA
The Greedy Algorithm

- Edges are processed in non increasing order by weight.
- Continuously add the heaviest available edge as long as it does not upset the construction of the Hamiltonian path given the previously chosen edges.
- The procedure ends when there are exactly n-1 edges, or when the accepted edges induce a connected subgraph.

Example: Greedy Algorithm Fails

- F={ATGC, GCC, TGCAT}

Order the edges by weight
(ATGC,TGCAT) = 3
(ATGC, GCC) = 2
(TGCAT, ATGC) = 2

The greedy algorithm will choose first
(ATGC, TGCAT) = 3 and then
is forced to select an edge
with weight 0 to complete the path:
(ATGC, TGCAT) (TGCAT, GCC)

Instead the solution should be
(TGCAT, ATGC) = 2
(ATGC, GCC) = 2
Sequencing by Hybridization

Universal DNA Array detects all the k-mers in given DNA sample (red dots)

probes - all possible k-mers

SBH: An Example

DNA array (DNA chip) with $4^3$ probes

Target DNA: AAATGCG
Sequencing by Hybridization

- **Spectrum** \((T, l)\): The set of all possible \((n - l + 1)\) \(l\)-mers in a string \(T\) of length \(n\)
- The order of individual elements in **Spectrum** \((T, l)\) does not matter
- **Example**: \(T = \text{ATGC}\text{G}\text{T}\text{GGCA}\)

\[
\text{Spectrum}(T, 3) = \{\text{ATG, TGC, GCG, CGT, GTG, TGG, GGC, GCA}\}
\]

The SBH Problem

- **Goal**: Reconstruct a string \(T\) from its \(l\)-mer composition

- **Input**: A set \(S\), representing all \(l\)-mers from an (unknown) string \(T\)

- **Output**: String \(T\) such that **Spectrum**\((T, l) = S\)
**SBH: An Example**

\[ S = \{\text{ACG, CGC, GCA, CAT, ATC}\} \]

**DNA Sample**

\[
\begin{array}{cccc}
A & C & G & C \\
G & C & A & T \\
A & T & C & T \\
\end{array}
\]

**Spectrum for k=3**

\[ T \text{ is such that} \]

Spectrum \((T, 3) = \{\text{ACG, CGC, GCA, CAT, ATC}\} \]

In other words, Spectrum\((T,3) = S \)

---

**SBH and Eulerian Path**

- Given a spectrum \( S \), draw a directed graph where:
  - Each vertex represents a \((k-1)\)-prefix or \((k-1)\)-suffix of \(k\)-mers in \( S \)
  - Each edge is a \(k\)-mer from \( S \) connecting a vertex representing a \((k-1)\)-prefix and a \((k-1)\)-suffix.
- Find a Eulerian path of \( G \), and reconstruct the sequence from the path

**Example:**

- Spectrum= \{ACG, ATC, CAT, CGC, GCA\}
- Edges: ACG, ATC, CAT, CGC and GCA
- Vertices: AC, CG, AT, TC, CA, and GC.
SBH and Eulerian Path (I)

- **Example:**
  - Spectrum= \{ACG, ATC, CAT, CGC, GCA\}
- **Draw the vertices:**
  AC, AT, CA, CG, GC, TC (alphabetical order)
  Draw edge from vertex AC to vertex CG \(\rightarrow\) edge ACG
  Draw edge from vertex AT to vertex TC \(\rightarrow\) edge ATC

![Eulerian Path Diagram (I)](image)

SBH and Eulerian Path (II)

- **Spectrum= \{ACG, ATC, CAT, CGC, GCA\}
- **Draw the vertices:**
  AC, AT, CA, CG, GC, TC (alphabetical order)
  Draw edge from vertex AC to vertex CG \(\rightarrow\) edge ACG
  Draw edge from vertex AT to vertex TC \(\rightarrow\) edge ATC
  Draw edge from vertex CA to vertex AT \(\rightarrow\) edge CAT
  Draw edge from vertex CG to vertex GC \(\rightarrow\) edge CGC
  Draw edge from vertex GC to vertex CA \(\rightarrow\) edge GCA

![Eulerian Path Diagram (II)](image)
SBH and Eulerian Path

• An Eulerian Path is a path which visits each edge of the graph once
  – Eulerian path: AC → CG → GC → CA → AT → TC
  – Sequence: ACGCATC
  – Multiple paths are possible

Uniqueness

Spectrum={ATG, TGC, GCG, CGT, GTG, TGG, GGC, GCA }

ATGCGTGGCA
ATGGCGTGCA

Adapted from Shuai Cheng Li: CS482/682
Challenges of SBH

- The solution may not be unique
  - For example: Obtain an Eulerian cycle instead of a path $\rightarrow$ multiple solutions
- The input data, the Spectrum $S$, may contain errors
  - For example: false positives, false negatives, uncertain frequency of k-mers
- Multiple parallel edges $\rightarrow$ ambiguous solutions

Some Solutions

- Several solutions were proposed to solve the problems
  - Positional Eulerian Path (PEP) by Hanninenhalli et al. 1996
  - Positional Sequencing by Hybridization (PSBH)
    - add extra information to probes
  - Interactive Protocols by Skiena et al. 1995
  - Gapped probes by Preparata et al. 2000 and Frieze et al. 1999
  - Analog-Spectrum by Preparata 2004
- Note that we consider the simple case were the spectrum yields an Euler path.
Gene Prediction

- Exons
- Introns
- Splicing
- Promoters
- Enhancers
- Silencers
- Hidden Markov Models
- VEIL
- GenScan

**Problem**: Given a genomic DNA sequence, identify where the genes are.

**Input**: A genomic DNA sequence.

**Output**: Location of gene elements in the raw, genomic DNA sequence, including (for eukaryotes):
- exons
- introns
Gene Prediction

<table>
<thead>
<tr>
<th>Organism</th>
<th>Coding (protein)</th>
<th>RNA</th>
<th>Non-coding</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>E. coli</em></td>
<td>85%</td>
<td>13%</td>
<td>2%</td>
</tr>
<tr>
<td><em>Drosophila</em></td>
<td>82%</td>
<td>17%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Yeast <em>S. cerevisiae</em></td>
<td>70%</td>
<td>28%</td>
<td>2%</td>
</tr>
<tr>
<td>0.5%</td>
<td>1.5%</td>
<td>98%</td>
<td></td>
</tr>
<tr>
<td>0.5%</td>
<td>0.01%</td>
<td>99.5%</td>
<td></td>
</tr>
<tr>
<td>Nematode <em>C. elegans</em></td>
<td>71%</td>
<td>28%</td>
<td>2%</td>
</tr>
<tr>
<td>Human</td>
<td>98%</td>
<td>2%</td>
<td>0.5%</td>
</tr>
<tr>
<td>Lunfish (dipnoi)</td>
<td>71%</td>
<td>28%</td>
<td>0.5%</td>
</tr>
</tbody>
</table>

Coding (protein), RNA, Non-coding

Gene: From DNA to Protein

Promoter

Transcription factor binding sites

TATA box

CCAAT box

5' cap

Transcription

Exon 1

Intron 1

Exon 2

Intron 2

Exon 3

3' DNA

Transcriptional element

Primary transcript

5'UTR

PolyA tail

AA-A

Mature mRNA

CDS

PolyA site

Splicing

Start codon

Stop codon

Nonsense (TAA, TGA, TAG)

Translation

Protein

Downstream element
Alternative Splicing

- Alternative pathways of splicing can produce different mRNAs and, subsequently, different proteins from the same primary transcript.
- The altered forms of the same protein that are generated by alternative splicing are usually used in different cell types or at different stages of development.
The Alpha-Tropomyosin Gene

E. Coli Promoter Sequences
Gene as a Docking Site

A depiction of a gene within DNA as a series of docking sites for proteins and RNAs

Gene Prediction Methods

- Use metaphor of parsing from Linguistics and Computer Science.
- Use neural networks
- Use Markov models of sequence elements
- Use mixed probabilistic models of sequence elements (best performance).
Markov Model Assumptions (I)

- A set \( Q \) of \( N \) states, denoted by \( 1, 2, \ldots, N \)
- An observable sequence, \( O: \)
  \[ o_1, o_2, \ldots, o_t, \ldots, o_T \]
- An unobservable sequence, \( q: \)
  \[ q_1, q_2, \ldots, q_t, \ldots, q_T \]
- First order Markov model:
  \[
P(q_t = j \mid q_{t-1} = i, q_{t-2} = k, \ldots) = P(q_t = j \mid q_{t-1} = i)
  \]

Markov Model Assumptions (II)

- An initial probability distribution:
  \[
  \pi_i = P(q_1 = i) \quad 1 \leq i \leq N
  \]
  where \( \sum_{i=1}^{N} \pi_i = 1 \)
- Stationary condition:
  \[
P(q_t = j \mid q_{t-1} = i) = P(q_{t+1} = j \mid q_{t+1-1} = i)
  \]
State Transition Probabilities

State transition probability matrix:

\[
A = \begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1N} \\
a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2N} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{iN} \\
\vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\
a_{N1} & a_{N2} & \cdots & a_{Nj} & \cdots & a_{NN}
\end{bmatrix}
\]

where:

\[
a_{ij} = P(q_t = j \mid q_{t-1} = i) \quad 1 \leq i, j \leq N
\]

\[
a_{ij} \geq 0, \quad \forall i, j
\]

\[
\sum_{j=1}^{N} a_{ij} = 1, \quad \forall i
\]

Hidden Markov Model

- **N**: the number of hidden states
  - A set of states \( Q = \{1, 2, \ldots, N\} \)
- **M**: the number of symbols
  - A set of symbols \( V = \{1, 2, \ldots, M\} \)
- **A**: the state-transition probability matrix
  - \( a_{i,j} = P(q_{t+1} = j \mid q_t = i) \quad 1 \leq i, j \leq N \)
- **B**: Emission probability distribution; \( k \) is a symbol:
  - \( B_j(k) = P(o_t = k \mid q_t = j) \quad 1 \leq i, j \leq M \)
- The initial state distribution \( \pi \):
  - \( \pi_i = P(q_1 = i) \quad 1 \leq i \leq N \)
  - The entire model \( \lambda \):
    - \( \lambda = (A, B, \pi) \)
Three Basic Questions

1. **EVALUATION** – given observation \( O=(o_1, o_2, \ldots, o_T) \) and model \( \lambda = (A, B, \pi) \), efficiently compute \( P(O \mid \lambda) \).
   - Given two models \( \lambda \) and \( \lambda' \), this can be used to choose the better one.
     - Forward Algorithm or Backward Algorithm

2. **DECODING** - given observation \( O=(o_1, o_2, \ldots, o_T) \) and model \( \lambda \) find the optimal state sequence \( q=(q_1, q_2, \ldots, q_T) \).
   - Optimality criterion has to be decided (e.g., maximum likelihood)
     - Viterbi Algorithm

3. **LEARNING** – given \( O=(o_1, o_2, \ldots, o_T) \), estimate model parameters \( \lambda = (A, B, \pi) \) that maximize \( P(O \mid \lambda) \).
   - EM and Baum-Welch Algorithms

Important Considerations

- For the user:
  - Know the algorithm
  - Know well the weaknesses and strengths of the program
  - Know how to interpret a particular score given by the program

- For the developer:
  - Know the current state of the art to be able to compare the program and recognize the weaknesses that need to be addressed.