Current Topics in Computer Science: Computational Genomics

CSCI 7000-005
Debra Goldberg
debra.goldberg@cs.colorado.edu
Temporary course website

http://llama.med.harvard.edu/~goldberg/cu
Molecular Biology Primer

Angela Brooks, Raymond Brown, Calvin Chen, Mike Daly, Hoa Dinh, Erinn Hama, Robert Hinman, Julio Ng, Michael Sneddon, Hoa Troung, Jerry Wang, Che Fung Yung
Review of molecular biology for computer scientists
All Life depends on 3 critical molecules

• DNA

• RNA

• Protein
All 3 are specified linearly

- DNA and RNA are constructed from **nucleic acids** (nucleotides)
  - Can be considered to be a string written in a four-letter alphabet (A C G T/U)
- Proteins are constructed from **amino acids**
  - Strings in a twenty-letter alphabet of amino acids
Central Dogma of Biology: DNA, RNA, and the Flow of Information

- **Replication**: DNA can replicate.
- **Transcription**: Information coded in the sequence of base pairs in DNA is passed to molecules of RNA.
- **Translation**: Information in RNA is passed to proteins. It never passes from proteins to nucleic acids.
DNA

- DNA provides a code, consisting of 4 letters.

Letters in DNA code: A C G T

- Each nucleic acid (or base) is always paired with its designated complement on the other strand of the double helix:
  - A and T are complementary
  - C and G are complementary
DNA

• DNA has a double helix structure.

• It is not symmetric. It has a “forward” and “backward” direction. The ends are labeled 5’ and 3’.

• DNA always reads 5’ to 3’ for transcription replication
RNA (ribonucleic acid)

- Similar to DNA chemically
- Usually only a single strand
- Built from nucleotides A, U, G, and C with ribose (ribonucleotides)
  - T(hyamine) is replaced by U(racil)
Types of RNA

- mRNA – carries a gene’s *message* out of the nucleus.
  - The type “RNA” most often refers to.
- tRNA – *transfers* genetic information from mRNA to an amino acid sequence
- rRNA – *ribosomal* RNA. Part of the ribosome.
  - involved in translation.
- siRNA – *small interfering* RNA. Interferes with transcription or translation. Recent discovery.
Transcription

• The process of making RNA from DNA

• Needs a promoter region to begin transcription.
More complex genes

Control regions

Exons

Transcription

Splicing
Terminology

- **Exon**: A portion of the gene that appears in both the primary and the mature mRNA transcripts.
- **Intron**: A portion of the gene that is transcribed but excised prior to translation.
- **Junk DNA**: Any DNA not contained in exons.
  - **NOT** junk
  - Many functions, some known, some unknown
RNA secondary structures

- Some forms of RNA can form secondary structures by “pairing up” with itself. This can change its properties dramatically.
Gene expression

- Human genome is ~ 3 billions base pair long
- Almost every cell in human body contains same set of genes
- But not all genes are used or expressed by those cells
  - Different cell types
  - Different conditions
Proteins: Workhorses of the Cell

- 20 different **amino acids**
- Proteins do essential work for the cell
  - cellular structures
  - enzymes
  - transmit information
- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.
The genetic code: RNA→protein

- Three bases of RNA (called a codon) correspond to one amino acid.
  - Degenerate: several codons for one AA
- Always starts with Methionine and ends with a stop codon

<table>
<thead>
<tr>
<th>FIRST POSITION</th>
<th>SECOND POSITION</th>
<th>THIRD POSITION</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>U</strong>&lt;br&gt;Phenylalanine</td>
<td><strong>C</strong>&lt;br&gt;Serine</td>
<td><strong>A</strong>&lt;br&gt;Tryptophan</td>
</tr>
<tr>
<td>Leucine</td>
<td>Stop</td>
<td>Stop</td>
</tr>
<tr>
<td><strong>C</strong>&lt;br&gt;Leucine</td>
<td><strong>G</strong>&lt;br&gt;Proline</td>
<td><strong>A</strong>&lt;br&gt;Glutamic Acid</td>
</tr>
<tr>
<td><strong>A</strong>&lt;br&gt;Leucine</td>
<td><strong>U</strong>&lt;br&gt;Aspartic Acid</td>
<td><strong>C</strong>&lt;br&gt;Aspartic Acid</td>
</tr>
</tbody>
</table>

* cod start
Terminology

- **Codon**: The sequence of 3 nucleotides in DNA/RNA that encodes for a specific amino acid.

- **mRNA (messenger RNA)**: A ribonucleic acid whose sequence is complementary to that of a protein-coding gene in DNA.
Protein Folding

- Proteins are not linear, they fold into 3D structures

- A protein’s structure determines how the protein can function
Protein Folding

- Proteins fold predominantly into
  - α-helices,
  - β-sheets, and
  - turns

Ubiquitin
Image from wisc.edu
Experimental methods
Analyzing a Genome: 3 steps

- **Copy** DNA many times
  - make it easier to see and detect

- **Cut** it into small fragments

- **Read** small fragments
Polymerase Chain Reaction (PCR)

- **Problem**: Cannot easily detect single molecules of DNA
- **Solution**: PCR massively replicates DNA sequences
  - Doubles the number of DNA fragments at every iteration

1... 2... 4... 8...
Copy DNA: Cloning

- DNA Cloning
  - Insert DNA fragment into the genome of a living organism and watch it multiply.
  - Once you have enough, remove the DNA.
Cutting DNA: Restriction Enzymes

- Restriction Enzymes cut DNA
  - Only cut at special sequences

**Bal I**

---TGGCCA---
---ACCGGT---

---TGG  CCA---
---ACC  GGT---

Blunt ends

**EcoR I**

---GAATTC---
---CTTAAG---

---G      AATTC---
---CTTAA      G---

Staggered ("sticky") ends
Cutting DNA: Restriction Enzymes

- DNA contains thousands of these sites.
- Applying different Restriction Enzymes creates fragments of varying size.

Restriction Enzyme “A” Cutting Sites

Restriction Enzyme “B” Cutting Sites

“A” and “B” fragments overlap

Restriction Enzyme “A” & Restriction Enzyme “B” Cutting Sites
Measuring DNA: Electrophoresis

- A gel
- Backbone of DNA is highly negatively charged
  - DNA will migrate in electric field
- Determine DNA fragment sizes
  - Compare their migration in the gel to known size standards
  - Use 2D gel to separate by size and charge
Reading/Sequencing DNA:
Electrophoresis

- Label DNA molecules with radioisotopes or tag with fluorescent dyes
- Group fragments that end in same base (A, C, G, or T)
- Sort in a gel experiment
Gene chips

- Gene chips = DNA chips = microarrays
- Spots of DNA attached to surface
- Each spot has a common 15-30 base long sequence
- Unknown DNA spread across gene chip will hybridize (bind) to complementary sequences
- Amount bound to each spot can be measured
Computational Genomics
What is Bioinformatics?

• Bioinformatics is generally defined as the analysis, prediction, and modeling of biological data with the help of computers.
What is computational biology?

• Different opinions

• Two common definitions:
  • Bioinformatics
  • Subset of bioinformatics that involves developing new computational methods

• Computational genomics:
  • Subset of computational biology dealing with genomes and/or proteomes (genes and/or proteins in the context of the entire organism)
Why computational biology?

- Sequenced DNA doubles every 10-14 months
  - Need computers to efficiently analyze data

- Computing power doubles every 18+ months (Moore’s law)
  - Cannot rely on increased computing power to handle increased genomic data
  - **Need better algorithms!**
Biological Databases

• Vast genomic data is freely available online
  • NCBI GenBank http://ncbi.nih.gov
    Huge collection of databases, including DNA 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
Problems in computational biology

- Permutations
- Graph algorithms
- Pattern matching and discovery
- String similarity
- Clustering
- Optimization
- 3D structure alignment
- Statistical methods, significance
- Randomized algorithms
Data storage

- Use computational algorithms to efficiently store large amounts of biological data
  - Standardize
  - Ontologies
  - Search for 3D protein structures
Assembling genomes

• Assemble the fragments into complete string
  • Not as easy as it sounds.

• SCS Problem (Shortest Common Superstring)
  • Some of the fragments will overlap
  • Fit overlapping sequences together to get the shortest possible sequence that includes all fragment sequences
  • Hamiltonian path problem (traverse all nodes)
  • Eulerian path problem (traverse all edges)
Assembling genomes: Complexities

- DNA fragments contain sequencing errors

- Two complements of DNA
  - Need to take into account both directions of DNA

- Repeat problem
  - 50% of human DNA is repetitive sequences
  - How do you know where it goes?

- Similar problem: peptide (protein) sequencing
  - Mass spectrometry gives weights of fragments
Pattern matching / discovery

• Gene prediction
  • Long open reading frames (ORFs)
    • Long DNA sequences without a “stop” codon
    • E (ORF length) ≈ 21 codons
  • Compare to known genes
  • Hidden Markov models (HMMs)
  • RNA splice sites (intron/exon boundaries)

• Gene Annotation
  • Comparison of similar species
Pattern matching / discovery (cont’d)

- Find known promoter (regulatory) regions
- Find new promoter (regulatory) regions
- Allow for errors
  - Brute force
  - Greedy algorithms
  - Gibbs sampling
- Similarly, find conserved regions in
  - AA sequences [possible active site]
  - DNA/RNA [possible protein binding site]
Sequence similarity searches

• Compare query sequences with all entries in biological databases
  • Measure pairwise similarity
  • Allow mutations/errors, insertions, deletions
  • Longest common (similar) subsequence

• Common tool that does this:

  BLAST
Sequence similarity searches II

- Other considerations
  - Time efficient?
  - Space efficient?
- Find new members of protein family
  - May be distant from other known members
  - Protein family profiles, HMMs
- Make predictions based on sequence
  - Protein/RNA secondary structure folding
  - Protein function
Gene chip analysis

- Image analysis
- Correlated gene expression
  - Clustering
- Determine probe set
  - Small substring of each gene to be tested
  - Unique to only one gene
  - No other similar substrings
Structure to Function

- Protein structure determines possible reactions
- Infer structure from sequence
  - De novo methods: physics based
  - Threading: “fit” known protein structures?
- Infer function from structure
  - Active sites
Comparative genomics

- Learn syntax of DNA (like comparative linguistics)
- Compare interspecies and intraspecies
- Given knowledge of one genome
  - Find similar genes in another (unsequenced) organism
- Sequence of permutations (of restricted types) to convert one genome to another
- Pairwise distances to binary evolutionary tree
  - Find family relationships between species by tracking similarities between species
Network determination

• Determining Regulatory Networks
  • Determine how body reacts to stimuli
  • Which molecules (proteins, others) turn on/off expression of a gene
Predict protein function

- Sequence similarities to known genes
- Similar expression conditions
- Similar interactions
Modeling

- Modeling biological processes tells us if we understand a given process
  - Protein models
  - Regulatory network models
  - Systems biology (whole cell) models
- Because of the large number of variables that exist in biological problems, powerful computers are needed to analyze certain biological questions
The future…

- Computational biology is still in its infancy
- Volume of data means computation in biology is here to stay
- 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?