Current Topics in Computer Science: Computational Genomics

CSCI 7000-005
Debra Goldberg
debra.goldberg@cs.colorado.edu
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
Life begins with Cell

- Smallest structural unit of an organism that is capable of functioning independently
- All cells have some common features
All Life depends on 3 critical molecules

• **DNA**
  - Holds information on how cell works

• **RNA**
  - Transfers short pieces of information to different parts of cell
  - Provides templates to synthesize into protein

• **Protein**
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)
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
Overview of organizations of life

- Genes = books
- Chromosomes = bookshelves
- Nucleus = library
- Almost every cell in an organism contains the same sets of books.
- Books represent all the information (DNA) that every cell in the body needs so it can grow and carry out its various functions.
Some Terminology

• **Gene**
  • basic physical and functional units of heredity.
  • located on chromosomes
  • specific sequences of DNA bases that encode instructions on how to make **proteins**.

• **Genome**
  • an organism’s genetic material, complete set of DNA.
  • a bacteria contains about 600,000 DNA base pairs
  • human and mouse genomes have some 3 billion.
Nucleic Acid Components

- **Nitrogenous Base:**
  - N is important for hydrogen bonding between bases
  - A – adenine with T – thymine (double H-bond)
  - C – cytosine with G – guanine (triple H-bond)

- **Sugar:**
  - Ribose (5 carbon)
  - Base covalently bonds with 1’ carbon
  - Phosphate covalently bonds with 5’ carbon
  - Normal ribose (OH on 2’ carbon) – RNA
  - deoxyribose (H on 2’ carbon) – DNA
  - dideoxyribose (H on 2’ & 3’ carbon) – used in DNA sequencing

- **Phosphate:**
  - negatively charged
The Purines

- Adenine (A)
  - to 1’ carbon of either pentose

- Guanine (G)
  - to 1’ carbon of either pentose

The Pyrimidines

- Thymine (T)
  - to 1’ carbon of deoxyribose

- Cytosine (C)
  - to 1’ carbon of either pentose
DNA

- Stores all information of life
- 4 “letters” base pairs. AGTC (adenine, guanine, thymine, cytosine) which pair A-T and C-G on complimentary strands.

http://www.lbl.gov/Education/HGP-images/dna-medium.gif
DNA, continued
Basic Structure

Watson-Crick base pair structures

Phosphate

Sugar
DNA, continued

• DNA has a double helix structure. However, it is not symmetric. It has a “forward” and “backward” direction. The ends are labeled 5’ and 3’ after the Carbon atoms in the sugar component.

  5’ AATCGCAAT 3’
  3’ TTAGCGTTA 5’

DNA always reads 5’ to 3’ for transcription replication
DNA: the building blocks of genetic material

- DNA provides a code, consisting of 4 letters, for all cellular function.

Letters in DNA code: ACGT
MUtAsHONS

- The DNA can be thought of as a sequence of the nucleotides: C, A, G, or T.
- What happens to genes when the DNA sequence is mutated?

Normal DNA sequence: ATCTAG

Mutated DNA sequence: ATCGGAG
The Good, the Bad, and the Silent

- **Mutations can serve the organism in three ways:**
  
  - **The Good:** Mutation in the sickle cell gene provides resistance to malaria.
  
  - **The Bad:** Huntington’s disease, a symptom of a gene mutation, is a degenerative disease of the nervous system.
  
  - **The Silent:** A mutation can simply cause no difference in the function of the organism.

Campbell, Biology, 5th edition, p. 255
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.
DNA - replication

- DNA can replicate by splitting, and rebuilding each strand.
- Note that the rebuilding of each strand uses slightly different mechanisms due to the 5’ 3’ asymmetry, but each daughter strand is an exact replica of the original strand.

http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/DNAReplication.html
DNA: The Code of Life

- The structure and the four genomic letters code for all living organisms
- Adenine, Guanine, Thymine, and Cytosine which pair A-T and C-G on complimentary strands.
Human chromosomes
## Chromosomes

<table>
<thead>
<tr>
<th>Organism</th>
<th># base pair</th>
<th># Chromosomes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prokaryotic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Escherichia coli (bacterium)</td>
<td>$4 \times 10^6$</td>
<td>1</td>
</tr>
<tr>
<td><strong>Eukaryotic</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saccharomyces cerevisiae (yeast)</td>
<td>$1.35 \times 10^7$</td>
<td>17</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruitfly)</td>
<td>$1.65 \times 10^8$</td>
<td>4</td>
</tr>
<tr>
<td>Homo sapiens (human)</td>
<td>$2.9 \times 10^9$</td>
<td>23</td>
</tr>
<tr>
<td>Zea mays (corn)</td>
<td>$5.0 \times 10^9$</td>
<td>10</td>
</tr>
<tr>
<td>Potato</td>
<td></td>
<td>24</td>
</tr>
<tr>
<td>Chicken</td>
<td></td>
<td>39</td>
</tr>
<tr>
<td>King crab</td>
<td></td>
<td>104</td>
</tr>
<tr>
<td>Ophioglossum reticulatum (fern)</td>
<td></td>
<td>630</td>
</tr>
</tbody>
</table>
**Definition of a Gene**

- **Regulatory regions:** up to 50 kb upstream of +1 site

- **Exons:** protein coding and untranslated regions (UTR)  
  1 to 178 exons per gene (mean 8.8)  
  8 bp to 17 kb per exon (mean 145 bp)

- **Introns:** splice acceptor and donor sites, junk DNA  
  average 1 kb – 50 kb per intron

- **Gene size:** Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.
Central Dogma Revisited

- Transcription: DNA → hnRNA (Nucleus)
- Splicing: hnRNA → mRNA (Spliceosome)
- Translation: mRNA → protein (Ribosome in Cytoplasm)

- Note: Some mRNA stays as RNA (i.e., tRNA, rRNA).
DNA: The Basis of Life

- Humans have about 3 billion base pairs.
  - How do you package it into a cell?
  - How does the cell know where in the highly packed DNA where to start transcription?
    - Special regulatory sequences
  - DNA size does not mean more complex
- Complexity of DNA
  - Eukaryotic genomes consist of variable amounts of DNA
    - Single Copy or Unique DNA
    - Highly Repetitive DNA
Packing it in

Cell

Nucleus

Chromosome

Chromatid

Chromatid

p arm

q arm

Solenoid loops

Nucleosome

Solenoid

Base Pairs

Histones

DNA

Gene

NHGRI
DNA is found in the nucleus of cells in the form of Chromosomes. Each of these is made up of genes, sections of DNA about 1500 nucleotides long.
Packing it in III
Superstructure

Superstructure Implications

- DNA in a living cell is in a highly compacted and structured state

- Transcription factors and RNA polymerase need ACCESS to do their work

- Transcription is dependent on the structural state – SEQUENCE alone does not tell the whole story
Transcriptional Regulation

Diagram showing the transcription process with elements labeled as Promoter region of gene, Coding region of gene, and Transcription "on" signal.
RNA
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)

http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif
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.
Terminology

- **Promoter**: A special sequence of nucleotides indicating the starting point for RNA synthesis.

- **Terminator**: Signal in DNA that halts transcription.

- **RNA Polymerase II**: Multisubunit enzyme that catalyzes the synthesis of an RNA molecule on a DNA template from nucleoside triphosphate precursors.
Transcription

• The process of making RNA from DNA
• Catalyzed by “transcriptase” enzyme
• Needs a promoter region to begin transcription.
• ~50 base pairs/second in bacteria, but multiple transcriptions can occur simultaneously

http://ghs.gresham.k12.or.us/science/ps/sci/ibbio/chem/nucleic/chpt15/transcription.gif
DNA $\rightarrow$ RNA: Transcription

- DNA gets transcribed by a protein known as RNA-polymerase
- This process builds a chain of bases that will become mRNA
Transcription, continued

- Transcription is highly regulated. Most DNA is in a dense form where it cannot be transcribed.
- To begin transcription requires a promoter, a small specific sequence of DNA to which polymerase can bind (~40 base pairs “upstream” of gene)
- Finding these promoter regions is a partially solved computational problem related to motif finding.
- There can also be repressors and inhibitors acting in various ways to stop transcription. This makes regulation of gene transcription complex to understand.
**Transcription:** DNA $\rightarrow$ hnRNA

- Transcription occurs in the nucleus.
- RNA polymerase reads promoter sequence and opens a small portion of the double helix exposing DNA bases.

- RNA polymerase II unwinds helix just ahead of active site
  - During transcription, DNA helix reforms as RNA forms.
  - When the terminator sequence is met, polymerase halts and releases both the DNA template and the RNA.
Definition of a Gene

- Regulatory regions: up to 50 kb upstream of +1 site
- Exons: protein coding and untranslated regions (UTR)
  1 to 178 exons per gene (mean 8.8)
  8 bp to 17 kb per exon (mean 145 bp)
- Introns: splice acceptor and donor sites, junk DNA
  average 1 kb – 50 kb per intron
- Gene size: Largest – 2.4 Mb (Dystrophin). Mean – 27 kb.
**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.
Splicing

![Splicing Diagram](image-url)
Splicing (Eukaryotes)

- Unprocessed RNA is composed of Introns and Exons. Introns are removed before the rest is expressed and converted to protein.
- Sometimes alternate splicings can create different valid proteins.
- A typical Eukaryotic gene has 4-20 introns. Locating them by analytical means is not easy.
RNA secondary structures

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

DNA and RNA can bind with each other.

tRNA linear and 3D view: http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif
Genomic Information

- Cells store all information to replicate itself
  - Human genome is around 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
Proteins
Proteins: Workhorses of the Cell

• 20 different **amino acids**
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.

• **Proteins do all essential work** for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - Mediate information flow within a cell and among cellular communities.

• **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.
Proteins

• Complex organic molecules made up of amino acid subunits
• 20* different kinds of amino acids. Each has a 1 and 3 letter abbreviation.
• [http://www.indstate.edu/thcme/mwking/amine-acids.html](http://www.indstate.edu/thcme/mwking/amine-acids.html) for complete list of chemical structures and abbreviations.
• Proteins are often enzymes that catalyze reactions.
• Also called “poly-peptides”

*Some other amino acids exist but not in humans.*
Uncovering the code

- Scientists conjectured that proteins came from DNA; but how did DNA code for proteins?
- If one nucleotide codes for one amino acid, then there’d be $4^1$ amino acids.
- However, there are 20 amino acids, so at least 3 bases codes for one amino acid, since $4^2 = 16$ and $4^3 = 64$.
  - This triplet of bases is called a “codon”
  - 64 different codons and only 20 amino acids means that the coding is degenerate: more than one codon sequence code for the same amino acid.
Translation

- The process of going from RNA to polypeptide.
- Three base pairs of RNA (called a codon) correspond to one amino acid based on a fixed table.
- Always starts with Methionine and ends with a stop codon.
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.
- **Ribosome**: The organelle that synthesizes polypeptides under the direction of mRNA.
- **rRNA (ribosomal RNA)**: The RNA molecules that constitute the bulk of the ribosome and provides structural scaffolding for the ribosome and catalyzes peptide bond formation.
- **tRNA (transfer RNA)**: The small L-shaped RNAs that deliver specific amino acids to ribosomes according to the sequence of a bound mRNA.
RNA → Protein: Translation

• Ribosomes and *transfer-RNAs* (tRNA) run along the length of the newly synthesized mRNA, decoding one codon at a time to build a growing chain of amino acids ("peptide")
• The tRNAs have anti-codons, which complimentarily match the codons of mRNA to know what protein gets added next
Terminology

• **Anticodon**: The sequence of 3 nucleotides in tRNA that recognizes an mRNA codon through complementary base pairing.

• **C-terminal**: The end of the protein with the free COOH.

• **N-terminal**: The end of the protein with the free NH3.
Purpose of tRNA

- The ribosome continually binds tRNA, joins the amino acids together and moves to the next location along the mRNA.
- The proper tRNA is chosen by having the corresponding anticodon for the mRNA’s codon.
- The tRNA then transfers its aminoacyl group to the growing peptide chain.
- For example, the tRNA with the anticodon UAC corresponds with the codon AUG and attaches methionine amino acid onto the peptide chain.
Purpose of tRNA

(a) Amino acid attachment site
(b) Anticodon
(c) Amino acid attachment site

©1999 Addison Wesley Longman, Inc.
Terminology

- **Molecular chaperone**: Protein that binds to unfolded or misfolded proteins to refold the proteins in the quaternary structure.
Translation, continued

http://wong.scripps.edu/PIX/ribosome.jpg
The Central Dogma (revisited)
Protein Synthesis: Summary

- There are twenty amino acids, each coded by three-base-sequences in DNA, called “codons”
  - This code is degenerate
- The **central dogma** describes how proteins derive from DNA
  - DNA \(\rightarrow\) mRNA \(\rightarrow\) (splicing?) \(\rightarrow\) protein
- The protein adopts a 3D structure specific to its amino acid arrangement and function
Polypeptide v. Protein

- A protein is a polypeptide, however to understand the function of a protein given only the polypeptide sequence is a very difficult problem.
- Protein folding an open problem. The 3D structure depends on many variables.
- Current approaches often work by looking at the structure of homologous (similar) proteins.
- Improper folding of a protein is believed to be the cause of mad cow disease.

http://www.sanger.ac.uk/Users/sgj/thesis/node2.html for more information on folding
Protein Folding

- Proteins are not linear structures, though they are built that way
- The amino acids have very different chemical properties; they interact with each other after the protein is built
  - This causes the protein to start fold and adopting it’s functional structure
  - Proteins may fold in reaction to some ions, and several separate chains of peptides may join together through their hydrophobic and hydrophilic amino acids to form a polymer
Protein Folding

- Proteins tend to fold into the lowest free energy conformation.
- Proteins begin to fold while the peptide is still being translated.
- Molecular chaperones, hsp60 and hsp 70, work with other proteins to help fold newly synthesized proteins.
- Much of the protein modifications and folding occurs in the endoplasmic reticulum and mitochondria.
Protein folding – secondary structure

- Proteins bury most of its hydrophobic residues in an interior core to form an α helix.
Protein folding – secondary structure

• Most proteins take the form of secondary structures: \(\alpha\) helices and \(\beta\) sheets.
Protein Folding (cont’d)

- The structure that a protein adopts is vital to its chemistry.
- Its structure determines which of its amino acids are exposed to carry out the protein’s function.
- Its structure also determines what substrates it can react with.