CMPS 6630: Introduction to Computational Biology and Bioinformatics

2017

Ramgopal Mettu
Church-Turing Thesis:
“Universal Model of Computation”

Alan Turing (1912-1954)

Turing Machine (1936)

Avery, Chargaff, Franklin, Pauling, Watson, Crick, Wilkins et al.

X-ray structure of ribosome [Yusupov et al. 2001]

“Post-Genomic Era” [Rual et al. 2005]

Collins/Venter, 2003

+
Human Genome Project

Genome sequencing is only the first step!

From Gene to Structure – and Beyond

- Gene Sequences
- Protein Sequences
- Protein Structures

10,000 unique structures in 10 years

Better understanding of disease-related proteins
Better understanding of protein structure folding, and conservation among different organisms
Predicted protein structures and functions
Targeted drug, design, gene therapy
What is “Computation”?

You are facing a high wall that stretches infinitely in both directions. There is a door in the wall, but you don't know how far away or in which direction. Can you escape? If so, how quickly?

[Ian Parberry, *Problems on Algorithms*]
What is Computation?

- Given a well-defined problem and input, how quickly (in the worst-case) can one produce a solution to the desired accuracy?
- Is there a tradeoff between resource requirements and accuracy?

Tractable vs. Intractable:

set of all Turing-computable problems

GOOD

polynomial-time: $n, n^2, n^3, \ldots$

BAD
What is “Computation”?

You are facing a high wall that stretches infinitely in both directions. There is a door in the wall, but you don't know how far away or in which direction. Can you escape? If so, how quickly?
[Ian Parberry, *Problems on Algorithms*]
“Computational” Biology

• Data Collection/Analysis/Modeling

• Develop problem formulations that are realistic, and are tractable.

• Leverage 50+ years of computational techniques:
  – Combinatorial Optimization
  – Statistics
  – Geometry
  – Software Design
This Course

• DNA/Gene Sequences:
  – Sequence Comparison
  – Sequence Assembly
  – Phylogenetics

• Protein Structure:
  – Secondary/Tertiary Structure Prediction
  – Structural Homology/Alignment/Comparison
  – Drug Discovery/Design

• “Systems” Biology:
  – Microarray Analysis
  – Interaction Networks
  – Metagenomics
Administrative Details

**Time:** TuTh 9:30-10:45

**Office:** 303E Stanley Thomas

**Office Hours:** By appointment

**Webpage:** [www.cs.tulane.edu/~mettu](http://www.cs.tulane.edu/~mettu)

**Course Materials:** Jones/Pevzner and online resources as needed (BioPython etc.).
Class Format

• Homework (40%)
  – 3-5 problem sets
  – short answers and programming
  – 40% of grade

• Midterm (30%)

• Final Project (30%)
  – chosen/assigned after midterm
  – grade based on presentation/writeup
“Tree of Life”
Biotech in 10,000BC

First off – we don’t just look around for our food... we actually grow some of it ourselves, where we live!

Gasp!

Plant and animal domestication is the key. We grow edible plants ourselves, right out of the ground, time after time!

Yum!

Animals, too! We control their reproduction to select desirable characteristics and eliminate bad ones.

Wow! How can we live the Neolithic way?

You can start by joining us in the village! Leave your troubles behind!

Some hunting and gathering may be necessary to maintain dietary variety and avoid famine.

Reshape your environment.

Form complex societies.

Your KEYS to a BETTER LIFE!

Harness Plant Power!

Put Animals To Work For You!

Eat regular meals!

Be sociable.

Settle down.

Build permanent structures.
Gregor Mendel (1822-1884) selectively bred pea plants and studied inheritance of physical characteristics.
Mendel identified a statistical pattern of how “factors” (genes) were inherited.

Mendel’s Laws:
Genes, Inheritance, Dominance, Independence
Mendel’s ideas were rediscovered around 1900 (DeVries, von Tschermak, Correns).

Chromosomes carry genetic information in “homologous” pairs (Sutton, 1902).
After Mendel

Chromosomes have a physically defined size, so the independence rule is not quite true.

How are genes correlated?

Can we “map” where genes lie on chromosomes? How many genes are there?
Prokaryotes are *unicellular* with minimal compartments (e.g. bacteria such as *E. coli*). “Chromosomes” are spread throughout cell.

Eukaryotes have *compartmentalized* cells with *organelles*; cells in eukaryotes *differentiate*. Chromosomes are inside nucleus.
Proteins = Function

- Beadle and Tatum showed correlation between enzymes and genes in the 1940s.

- Using clever analysis of irradiated mold spores, they concluded that genes are connected to enzymes.

- An enzyme is a type of protein; proteins are polypeptides.
Proteins = Function

• So chromosomes control the production of enzymes, but how?

• But what is the mechanism by which a gene is “expressed”?

• Avery-MacLeod-McCarty (1940) showed that DNA ‘controls’ genetic traits.
rough strain (nonvirulent) vs. smooth strain (virulent) vs. heat-killed smooth strain vs. rough strain & heat-killed smooth strain

- rough strain (nonvirulent): mouse lives
- smooth strain (virulent): mouse dies
- heat-killed smooth strain: mouse lives
- rough strain & heat-killed smooth strain: mouse dies
DNA?

Is that a government agency?

Before Avery, scientists had paid little attention to DNA.

They knew it contained the sugar deoxyribose, plenty of phosphate, and four bases.

The four bases are known as A, C, G, and T, which are short for:

A: Adenine

C: Cytosine

G: Guanine

T: Thymine

These were assumed to be present in equal proportions.

After Avery, however, researchers began to look more closely...

Erwin Chargaff found:

1. The composition of DNA varied from one species to another, in particular in the relative amounts of the bases A, C, T, G.

2. In any DNA, the number of A's was the same as the number of T's; similarly, the number of C's was equal to the number of G's.

What did this mean? Chargaff couldn't say.

By studying X-ray pictures of DNA, Rosalind Franklin was able to show that the DNA molecule probably had the corkscrew shape of a helix with two or three chains...

But was it two or three...?
In 1952 James Watson and Francis Crick cracked the puzzle.

By playing with scale-model atoms, they observed that adenine fitted together with thymine, while guanine paired naturally with cytosine.

Each of these two base pairs is nearly flat.

So Watson and Crick proposed to stack them up, one after another, like stairsteps.

Two sugar-phosphate strands wind around the outside.

It's a double helix!!

One complication: the two strands wind in opposite directions; the sugars on one strand are "upside down" compared with those on the other strand—etc!

Each base pair would be held together by hydrogen bonding, a weak attraction that may occur between a hydrogen on one molecule and a non-hydrogen atom of another molecule.

It was also clear A did not fit with C, nor G with T.

You reflect me!!

Sugar phosphate
Chromosomes are composed of DNA!
This model clearly explains Chargaff's observation that the number of T's is equal to the number of A's: T and A are always paired together!

Ditto for G and C!

This is the principle of complementarity: each base can pair with only one other, called its complement.

Watson and Crick got the idea!! They wrote:

"It has not escaped our notice that the pairing... immediately suggests a possible copying mechanism for the genetic material."

In fact, it is the key to the gene's main functions: replication and protein synthesis.
DNA Molecule: Two Views

Sugar

Cytosine and Thymine

Adenine and Guanine

Phosphate group

www.accessexcellence.org
REPLICATION

Gene copying, or DNA replication, as Watson and Crick saw, is simple in principle. Each strand of the double helix contains the information necessary to make its complementary strand.

Schematically, it works like this: When the DNA is ready to multiply, its two strands pull apart:

When a free nucleotide meets its complementary base on the DNA, it sticks, while the 'wrong' nucleotides bounce away.

As the "zipping" enzyme opens the DNA further, more nucleotides are added, and a "clipping" enzyme puts them together, knocking off the extra phosphates.

Along each one, a new strand forms in the only possible way:

Along the two strands, we wind up with two copies of the original:

This proceeds along both strands simultaneously—in opposite directions. The 'clipping' enzyme can go only one way, running smoothly down one strand, while backing up the other in a series of steps.
DNA to Proteins

• Genes are encoded by chromosomes, i.e., DNA.

• Genes “control” proteins, which enable function.

• So what is the mechanism that produces proteins from DNA?
DNA is a sequence of *nucleic acids* (4 types).

Proteins are a sequence of *amino acids* (20 types).

What is the mechanism to go from DNA to protein?

DNA must *code* for proteins.
Protein synthesis begins when a region of DNA is teased apart and a molecule of RNA is built along one strand by an enzyme called RNA polymerase. This process is called transcription.

It happens as in DNA replication; each base of the RNA is complementary to the corresponding base on the DNA.

This RNA is called the messenger, or mRNA, because it carries the genetic message from the DNA to the protein factory.

The "words" of the message are triplets of bases -- A-U-G, A-C-A, etc. The technical name for one of these groups is a codon.

Each 3-base codon stands for a single amino acid, and the whole mRNA strand encodes a protein (or several proteins). It's just like a message in code --

**THE GENETIC CODE!**

Cracking this code began in 1961, when Marshall Nirenberg was able to make a special mRNA, whose only base was uracil, repeated over and over -- "poly-U."

From it, he obtained a protein consisting entirely of the amino acid phenylalanine.

So UUU was the codon for phenylalanine...

Next they decoded Poly-A, and Poly-U, Poly-C, Poly-U, Poly-G, Poly-G, etc., etc., etc., until the code was finally broken.

UUU → Phe
AAA → Lys
GGC → Asp
UCC → Pro
UCC → Pro
UUU → Phe
AUA → Ile
CUG → Leu
UUU → Phe
UGG → Val

The complete code table follows!
Pairing

A ↔ T/U
T/U ↔ A
G ↔ C
C ↔ G

DNA

RNA

Ribonucleic acid
<table>
<thead>
<tr>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>UUU = Phe</td>
<td>UCU = Ser</td>
<td>UAU = Tyr</td>
<td>UGU = Cys</td>
</tr>
<tr>
<td>UUC = Phe</td>
<td>UCC = Ser</td>
<td>UAC = Tyr</td>
<td>UGC = Cys</td>
</tr>
<tr>
<td>UUA = Leu</td>
<td>UCA = Ser</td>
<td>UAA = Stop</td>
<td>UGA = Stop</td>
</tr>
<tr>
<td>UUG = Leu</td>
<td>UCG = Ser</td>
<td>UAG = Stop</td>
<td>UGG = Trp</td>
</tr>
<tr>
<td>CUU = Leu</td>
<td>CCU = Pro</td>
<td>CAU = His</td>
<td>CGU = Arg</td>
</tr>
<tr>
<td>CUC = Leu</td>
<td>CCC = Pro</td>
<td>CAC = His</td>
<td>CGC = Arg</td>
</tr>
<tr>
<td>CUA = Leu</td>
<td>CCA = Pro</td>
<td>CAA = Gln</td>
<td>CGA = Arg</td>
</tr>
<tr>
<td>CUG = Leu</td>
<td>CCG = Pro</td>
<td>CAG = Gln</td>
<td>CGG = Arg</td>
</tr>
<tr>
<td>AUU = Ile</td>
<td>ACU = Thr</td>
<td>AAU = Asn</td>
<td>AGU = Ser</td>
</tr>
<tr>
<td>AUC = Ile</td>
<td>ACC = Thr</td>
<td>AAC = Asn</td>
<td>AGC = Ser</td>
</tr>
<tr>
<td>AUA = Ile</td>
<td>ACA = Thr</td>
<td>AAA = Lys</td>
<td>AGA = Arg</td>
</tr>
<tr>
<td>AUG = Met</td>
<td>ACG = Thr</td>
<td>AAG = Lys</td>
<td>AGG = Arg</td>
</tr>
<tr>
<td>GUU = Val</td>
<td>GCU = Ala</td>
<td>GAU = Asp</td>
<td>GGU = Gly</td>
</tr>
<tr>
<td>CUC = Val</td>
<td>GCC = Ala</td>
<td>GAC = Asp</td>
<td>GGC = Gly</td>
</tr>
<tr>
<td>GUA = Val</td>
<td>GCA = Ala</td>
<td>GAA = Glu</td>
<td>GGA = Gly</td>
</tr>
<tr>
<td>GUG = Val</td>
<td>GCG = Ala</td>
<td>GAG = Glu</td>
<td>GGG = Gly</td>
</tr>
</tbody>
</table>

So, after mRNA has been transcribed, how are codons translated into, for example, an enzyme?
The actual translators of the genetic code are a group of RNA molecules called transfer RNA, or tRNA. Owing to pairing among its bases, tRNA's twist themselves into this key shape.

The loop end of tRNA has three unpaired bases. This "anticodon" may bind with the complementary codon of mRNA. At the "tail" end of tRNA is a site for attaching a single amino acid.

For each anticodon, there is an enzyme which recognizes it and attaches the appropriate amino acid to its tRNA.

Once they are linked, the anticodon binds to the complementary codon of message.

Schematically, this is the way a string of bases is translated into a sequence of amino acids. However, the cell needs one more piece of equipment to make it work: the ribosome.
HOW PROTEINS ARE MADE

THE FINAL INGREDIENT IN THE PROTEIN-MAKING APPARATUS IS AN OBJECT THAT HOLDS EVERYTHING IN PLACE.

THIS IS THE RIBOSOME, A DOUBLE BALL OF ABOUT 50 PROTEINS WRAPPED UP WITH RNA. THIS RNA IS CALLED RIBOSOMAL RNA, tRNA FOR SHORT.

THE RIBOSOME HAS TWO SLOTS IN WHICH MOLECULES OF tRNA CAN FIT SNUGLY.

X-ray structure of ribosome [Noller et al. 1999]
NOW TO MAKE A PROTEIN: WHEN THE mRNA READS OUT THE DNA SEQUENCE, IT ENTERS A SEA OF RIBOSOMES.

One half at a time, a ribosome binds onto the mRNA.

The binding site is located at or near the codon A-U-G.

Thus, A-U-G is always the first “word” of every message.

A-U-G and the next codon each bond with complementary tRNA’s, which fit into the slots on the ribosome.

The amino acids are linked; the “empty” tRNA is discarded; and so the ribosome moves along the message, piling up amino acids, which fold themselves into a protein.

Each tRNA carries an amino acid (AA), the first one always being methionine, which goes with A-U-G.

An enzyme in the ribosome links the two amino acids, and the first tRNA floats away.

Another tRNA and amino acid bind on.

The ribosome then moves down three more bases.
Amino Acids

Sidechain

Acid

Alpha Carbon
Amino Acids

Sidechain

Backbone
Backbone dihedral angles essentially define the geometry of the protein backbone.

Side-chains have a variable number of dihedrals angles, depending on composition.
Protein Structure

Primary Sequence: Linear String of Amino Acids

Secondary structure: regular $\alpha$-helices and $\beta$-strands

Global Fold
Structure = Function
Structure = Function
Structure = Function
A GLU to VAL mutation at 6th amino acid in the β-subchains causes hemoglobin to aggregate, resulting in sickle-cell anemia.
Recap: Central “Dogma”
“Endogenous retroviruses” are thought to make up 8% of the human genome!
Mitochondria are aerobic “energy generators.”

Cell-Mitochondrial “endosymbiosis” is hypothesized.

Mitochondrial DNA is used for accurate “genomic geography”.

Figure 1