Molecular Biology 101

Fall 2015
BMI/CS 576
www.biostat.wisc.edu/bmi576/
Colin Dewey
colin.dewey@wisc.edu
Levels of the biological hierarchy

- Ecosystem
  - Lake Monona, Mojave desert, ...

- Population
  - Madison Humans, Monona Muskie

- Organism
  - Pine tree, Lizard, Human...

- Organ
  - Heart, Lungs, Brain...

- Tissue
  - Adipose tissue, Blood, Nerve tissue...

- Cell
  - Fat cell, Blood cell, Muscle cell...

- Organelle
  - Nucleus, Mitochondrion...

- Molecule
  - DNA, RNA, Protein, Lipids...
The Central Dogma
DNA

1 turn = 10 base pairs = 3.4 nanometers

major groove

minor groove
The Double Helix

- DNA molecules usually consist of two strands arranged in the famous double helix
DNA

- DeoxyriboNucleic Acid
- can be thought of as the “blueprint” for an organism
- a linear chain of small molecules called nucleotides
  - four different nucleotides distinguished by the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T)
- is a polymer: large molecule consisting of similar units (nucleotides in this case)
- a single strand of DNA can be thought of as a string composed of the four letters: A, C, G, T

CTGCTGGACC GG GTGCTAGG ACCCTG A CTGCCCGGGGGCGG GGGGT GCGGGGCCCCGCTGAG…
Nucleotides: the subunits of DNA
The four DNA bases

Guanine

Cytosine

Adenine

Thymine
DNA strand: polymer of nucleotides

- in double-stranded DNA
  - A always bonds to T
  - C always bonds to G
The Double Helix

- each strand of DNA has a “direction”
  - at one end, the terminal carbon atom in the backbone is the 5’ carbon atom of the terminal sugar
  - at the other end, the terminal carbon atom is the 3’ carbon atom of the terminal sugar
- therefore we can talk about the 5’ and the 3’ ends of a DNA strand
- in a double helix, the strands are antiparallel (arrows drawn from the 5’ end to the 3’ end go in opposite directions)
DNA dimensions

1 turn = 10 base pairs = 3.4 nanometers

major groove

minor groove
DNA Replication Prior to Cell Division

Parent Strands

A Adenine
T Thymine
G Guanine
C Cytosine

Complementary New Strand

image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
image from the DOE Human Genome Program  
http://www.ornl.gov/hgmis
Chromosomes

- DNA is packaged into individual chromosomes (along with proteins)

- prokaryotes (single-celled organisms lacking nuclei) typically have a single circular chromosome

- eukaryotes (organisms with nuclei) have a species-specific number of linear chromosomes
DNA is tightly packed!
Genomes

- the term genome refers to the complete complement of DNA for a given species
- the human genome consists of 46 chromosomes (23 pairs)
- every cell (except sex cells and mature red blood cells) contains the complete genome of an organism
The Central Dogma

DNA → replication

RNA → transcription

Protein → translation

DNA → replication
# Examples of proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha-keratin</td>
<td>component of hair</td>
</tr>
<tr>
<td>beta-keratin</td>
<td>component of scales</td>
</tr>
<tr>
<td>insulin</td>
<td>regulates blood glucose level</td>
</tr>
<tr>
<td>actin &amp; myosin</td>
<td>muscle contraction</td>
</tr>
<tr>
<td>DNA polymerase</td>
<td>synthesis of DNA</td>
</tr>
<tr>
<td>ATP synthase</td>
<td>makes ATP</td>
</tr>
<tr>
<td>hemoglobin</td>
<td>transport of oxygen</td>
</tr>
<tr>
<td>endonuclease</td>
<td>cuts DNA (restriction enzyme)</td>
</tr>
</tbody>
</table>
Space-Filling Model of Hexokinase
EcoRI – restriction enzyme
Hemoglobin: carrier of oxygen
Mutant $\beta$-globin $\rightarrow$ Sickle blood cells

Fiber of sickle hemoglobin

Sickle and normal blood cells
Normal blood flow
Sickle cell complications
Protein: polymers of amino acids
Proteins

- proteins are molecules composed of one or more polypeptides
- a polypeptide is a polymer composed of amino acids
- cells build their proteins from 20 different amino acids
- a polypeptide can be thought of as a string composed from a 20-character alphabet
# Amino Acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Abbreviation</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alanine</td>
<td>Ala</td>
<td>A</td>
</tr>
<tr>
<td>Arginine</td>
<td>Arg</td>
<td>R</td>
</tr>
<tr>
<td>Aspartic Acid</td>
<td>Asp</td>
<td>D</td>
</tr>
<tr>
<td>Asparagine</td>
<td>Asn</td>
<td>N</td>
</tr>
<tr>
<td>Cysteine</td>
<td>Cys</td>
<td>C</td>
</tr>
<tr>
<td>Glutamic Acid</td>
<td>Glu</td>
<td>E</td>
</tr>
<tr>
<td>Glutamine</td>
<td>Gln</td>
<td>Q</td>
</tr>
<tr>
<td>Glycine</td>
<td>Gly</td>
<td>G</td>
</tr>
<tr>
<td>Histidine</td>
<td>His</td>
<td>H</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>Ile</td>
<td>I</td>
</tr>
<tr>
<td>Leucine</td>
<td>Leu</td>
<td>L</td>
</tr>
<tr>
<td>Lysine</td>
<td>Lys</td>
<td>K</td>
</tr>
<tr>
<td>Methionine</td>
<td>Met</td>
<td>M</td>
</tr>
<tr>
<td>Phenylalanine</td>
<td>Phe</td>
<td>F</td>
</tr>
<tr>
<td>Proline</td>
<td>Pro</td>
<td>P</td>
</tr>
<tr>
<td>Serine</td>
<td>Ser</td>
<td>S</td>
</tr>
<tr>
<td>Threonine</td>
<td>Thr</td>
<td>T</td>
</tr>
<tr>
<td>Tryptophan</td>
<td>Trp</td>
<td>W</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>Tyr</td>
<td>Y</td>
</tr>
<tr>
<td>Valine</td>
<td>Val</td>
<td>V</td>
</tr>
</tbody>
</table>
The alphabet of protein: 20 amino acids
Amino Acid Sequence of Hexokinase

1 A A S X D X S L V E V H X X V F I V P P X I L Q A V V S I A
31 T T R X D D X D S A A A S I P M V P G W V L K Q V X G S Q A
61 G S F L A I V M G G G D L E V I L I X L A G Y Q E S S I X A
91 S R S L A A S M X T T A I P S D L W G N X A X S N A A F S S
151 T X Q A X A F S L A X L X K L I S A M X N A X F P A G D X X
181 X X V A D I X D S H G I L X X V N Y T D A X I K M G I I F G
211 S G V N A A Y W C D S T X I A D A A D A G X X G G A G X M X
241 V C C X Q D S F R K A F P S L P Q I X Y X X T L N X X S P X
271 A X K T F E K N S X A K N X G Q S L R D V L M X Y K X X G Q
301 X H X X X A X D F X A A N V E N S S Y P A K I Q K L P H F D
331 L R X X X D L F X G D Q G I A X K T X M K X V V R R X L F L
361 I A A Y A F R L V V C X I X A I C Q K K G Y S S G H I A A X
391 G S X R D Y S G F S X N S A T X N X N I Y G W P Q S A X X S
421 K P I X I T P A I D G E G A A X X V I X S I A S S Q X X X A
451 X X S A X X A
Nucleotides vs. Amino Acids

Nucleotide

Amino Acid

Both made up of “backbone” and “residue” parts
The inaccessible code

DNA is in the nucleus

Proteins are (mostly) made in the cytoplasm

(eukaryotic cell)
The Central Dogma

DNA ➔ replication

RNA ➔ transcription

Protein ➔ translation
• RNA that is transcribed from a gene is called messenger RNA (mRNA)

• RNA polymerase is the enzyme that builds an RNA molecule from a gene
Transcription: DNA→RNA
### RNA vs. DNA structure

<table>
<thead>
<tr>
<th></th>
<th>DNA</th>
<th>RNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>linear polymer</td>
<td>linear polymer</td>
</tr>
<tr>
<td></td>
<td>double-stranded</td>
<td>single-stranded</td>
</tr>
<tr>
<td></td>
<td>deoxyribonucleotide</td>
<td>ribonucleotide</td>
</tr>
<tr>
<td></td>
<td>monomer</td>
<td>monomer</td>
</tr>
</tbody>
</table>

- **DNA**: linear polymer, double-stranded, deoxyribonucleotide monomer, A,C,G,T bases
- **RNA**: linear polymer, single-stranded, ribonucleotide monomer, A,C,G,U bases

![DNA and RNA structures](image-url)
# The Genetic Code

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phenylalanine</td>
<td>UUU</td>
<td>UUC</td>
<td>UAU</td>
<td>UAC</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td>UUA</td>
<td>UUG</td>
<td>UAA</td>
<td>UAG</td>
</tr>
<tr>
<td>C</td>
<td>Leucine</td>
<td>CUU</td>
<td>CUC</td>
<td>CAU</td>
<td>CUG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCC</td>
<td>CCC</td>
<td>CCA</td>
<td>CCG</td>
</tr>
<tr>
<td>A</td>
<td>Isoleucine</td>
<td>AUA</td>
<td>ACC</td>
<td>AAC</td>
<td>ACG</td>
</tr>
<tr>
<td></td>
<td>Methionine;</td>
<td>AUG</td>
<td>ACC</td>
<td>AAA</td>
<td>AAG</td>
</tr>
<tr>
<td></td>
<td>initiation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>codon</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Valine</td>
<td>GUU</td>
<td>GUC</td>
<td>GAU</td>
<td>GUG</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCU</td>
<td>GCC</td>
<td>GCA</td>
<td>GCG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
DNA Genetic Code Dictates Amino Acid Identity and Order

DNA Sequence Is the Genetic Code.

GCA AGA GAT AAT TGT...

Ala Arg Asp Asn Cys...

Growing Protein Chain

image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
Translation

• ribosomes are the machines that synthesize proteins from mRNA

• the grouping of codons is called the reading frame

• translation begins with the start codon

• translation ends with the stop codon
Codons and Reading Frames
Translation

Ribosome

Transfer RNA (tRNA)

Growing polypeptide

This process repeats until reaching a stop codon

Stop codon

Pro

Met

Tyr

Ala

UGCCGUAUGCUCCUU

UGCCGUAUGCUCCUU

GGC

GGC

AUA

AUA

Pro

Met

Tyr

Met
DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code

**Gene A from Person 1**

- Sequence: GCA AGA GAT AAT TGT...
- Protein Products: Ala, Arg, Asp, Asn, Cys...

**Gene A from Person 2**

- Sequence: GCG AGA GAT AAT TGT...
- Codon change made no difference in amino acid sequence
- Protein Products: Ala, Arg, Asp, Asn, Cys...

**Gene A from Person 3**

- Sequence: GCA AAA GAT AAT TGT...
- Codon change resulted in a different amino acid at position 2
- Protein Products: Ala, Lys, Asp, Asn, Cys...

---

image from the DOE Human Genome Program
http://www.ornl.gov/hgmis
Genes

- genes are the basic units of heredity
- a gene is a sequence of bases that carries the information required for constructing a particular protein (polypeptide really)
- such a gene is said to encode a protein
- the human genome comprises ~20,000 protein-coding genes
Gene Density

- not all of the DNA in a genome encodes protein:

  bacteria  ~90% coding gene/kb
  human    ~1.5% coding gene/35kb
RNA Processing in Eukaryotes

- eukaryotes are organisms that have enclosed nuclei in their cells
- in many eukaryotes, genes/mRNAs consist of alternating exon/intron segments
- exons are the coding parts
- introns are spliced out before translation
RNA Splicing

Chromosomal DNA

Transcription (RNA synthesis)

Nuclear RNA

RNA Splicing

Messenger RNA
Protein Synthesis in Eukaryotes vs. Prokaryotes
RNA Genes

- not all genes encode proteins
- for some genes the end product is RNA
  - ribosomal RNA (rRNA), which includes major constituents of ribosomes
  - transfer RNAs (tRNAs), which carry amino acids to ribosomes
  - micro RNAs (miRNAs), which play an important regulatory role in various plants and animals
- etc.
The Dynamics of Cells

• all cells in an organism have the same genomic data, but the genes expressed in each vary according to cell type, time, and environmental factors

• there are networks of interactions among various biochemical entities in a cell (DNA, RNA, protein, small molecules) that carry out processes such as

  • metabolism

  • intra-cellular and inter-cellular signaling

  • regulation of gene expression
Overview of the E. coli Metabolic Pathway Map

image from the KEGG database
The Metabolic Pathway for Synthesizing the Amino Acid Alanine

reactions
metabolites
enzymes (proteins that catalyze reactions)
gen genes encoding the enzymes

image from the Ecocyc database
www.biocyc.org
Gene Regulation Example: the lac Operon

THE LAC OPERON

RNA polymerase

LacI P₁

P O

LacZ LacY LacA

Regions coding for proteins
Regulatory regions
Diffusible regulatory proteins

this protein regulates the transcription of LacZ, LacY, LacA

these proteins metabolize lactose
Gene Regulation Example: the lac Operon

lactose is absent \( \Rightarrow \) the protein encoded by lacI represses transcription of the lac operon
Gene Regulation Example: the lac Operon

lactose is present ⇒ it binds to the protein encoded by lacI changing its shape; in this state, the protein doesn’t bind upstream from the lac operon; therefore the lac operon can be transcribed
Gene Regulation Example: the lac Operon

- this example provides a simple illustration of how a cell can regulate (turn on/off) certain genes in response to the state of its environment

- an operon is a sequence of genes transcribed as a unit

- the lac operon is involved in metabolizing lactose

  - it is “turned on” when lactose is present in the cell

  - the lac operon is regulated at the transcription level

- the depiction here is incomplete; for example, the level of glucose in the cell also influences transcription of the lac operon
## Selected milestones in genome sequencing

<table>
<thead>
<tr>
<th>Year</th>
<th>Common Name</th>
<th>Species</th>
<th># of Chromosomes</th>
<th>Size (base pairs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1995</td>
<td>Bacterium</td>
<td><em>Haemophilus influenzae</em></td>
<td>1</td>
<td>$1.8 \times 10^6$</td>
</tr>
<tr>
<td>1996</td>
<td>Yeast</td>
<td><em>Saccharomyces cerevisiae</em></td>
<td>16</td>
<td>$1.2 \times 10^7$</td>
</tr>
<tr>
<td>1998</td>
<td>Worm</td>
<td><em>Caenorhabditis elegans</em></td>
<td>6</td>
<td>$1.0 \times 10^8$</td>
</tr>
<tr>
<td>1999</td>
<td>Fruit Fly</td>
<td><em>Drosophila melanogaster</em></td>
<td>4</td>
<td>$1.3 \times 10^8$</td>
</tr>
<tr>
<td>2000</td>
<td>Human</td>
<td><em>Homo sapiens</em></td>
<td>23</td>
<td>$3.1 \times 10^9$</td>
</tr>
<tr>
<td>2002</td>
<td>Mouse</td>
<td><em>Mus musculus</em></td>
<td>20</td>
<td>$2.6 \times 10^9$</td>
</tr>
<tr>
<td>2004</td>
<td>Rat</td>
<td><em>Rattus norvegicus</em></td>
<td>21</td>
<td>$2.8 \times 10^9$</td>
</tr>
<tr>
<td>2005</td>
<td>Chimpanzee</td>
<td><em>Pan troglodytes</em></td>
<td>24</td>
<td>$3.1 \times 10^9$</td>
</tr>
</tbody>
</table>

Sequence is freely available at:

- **UCSC** - [http://genome.ucsc.edu](http://genome.ucsc.edu)
But Wait, There’s More…

- > 1000 other publicly available databases pertaining to molecular biology (see pointer to Nucleic Acids Research directory on course home page)
- GenBank
  > 187 million sequence entries
  > 200 billion bases
- UniProtKB / Swis-Prot
  > 50 million protein sequence entries
  > 16 billion amino acids
- Protein Data Bank
  111,749 protein (and related) structures

* all numbers current as of 9/15
More Data: High-Throughput Experiments

- RNA abundances
- protein abundances
- small molecule abundances
- protein-protein interactions
- protein-DNA interactions
- protein-small molecule interactions
- genetic variants of an individual (e.g. which DNA base does the individual have at a few thousand selected positions)
- something (e.g. viral replication) measured across thousands of genetic variants
- etc.
Example HT Experiment

- this figure depicts one yeast gene-expression data set
- each row represents a gene
- each column represents a measurement of gene expression (mRNA abundance) at some time point
  - red indicates that a gene is being expressed more than some baseline; green means less

Figure from Spellman et al., Molecular Biology of the Cell, 9:3273-3297, 1998
More Data: Interactions

- each node represents a gene product (protein)
- blue edges show direct protein-protein interactions
- yellow edges show interactions in which one protein binds to DNA and affects the expression of another

Figure from Ideker et al., Science 292(5518):929-934, 2001
Significance of the Genomics Revolution

- data driven biology
  - functional genomics
  - comparative genomics
  - systems biology
- molecular medicine
  - identification of genetic components of various maladies
  - diagnosis/prognosis from sequence/expression
- gene therapy
- pharmacogenomics
  - developing highly targeted drugs
  - toxicogenomics
  - elucidating which genes are affected by various chemicals
Bioinformatics Revisited

Representation/storage/retrieval/ analysis of biological data concerning

- sequences (DNA, protein, RNA)
- structures (protein, RNA)
- functions (protein, sequence signals)
- activity levels (mRNA, protein, metabolites)
- networks of interactions (metabolic pathways, regulatory pathways, signaling pathways)

of/among biomolecules