Introduction to Molecular Biology and Genomics

BMI/CS 776
www.biostat.wisc.edu/~craven/776.html
Mark Craven
craven@biostat.wisc.edu
January 2002

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

• can be thought of as the “blueprint” for an organism
• composed of small molecules called nucleotides
• four different nucleotides distinguished by the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T)
• a polymer: large molecule consisting of similar units (nucleotides in this case)

dna

DNA

• a single strand of DNA can be thought of as a string composed of the four letters: A, C, G, T

ctgctggaccgggtgctaggacctgactgcc
cggggccggggggtgcggggcccgctgag…
The Double Helix

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

Watson-Crick Base Pairs

- in double-strand 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)
Chromosomes

- DNA is packaged into individual chromosomes (along with proteins)
- prokaryotes (single-celled organisms lacking nuclei) have a single circular chromosome
- eukaryotes (organisms with nuclei) have a species-specific number of linear chromosomes

Human Chromosomes
Genomes

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

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
Protein Functions

- structural support
- storage of amino acids
- transport of other substances
- coordination of an organism’s activities
- response of cell to chemical stimuli
- movement
- protection against disease
- selective acceleration of chemical reactions

Amino Acids

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Symbol</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>Glu</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>
### Amino Acid Sequence of Hexokinase

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AAS</td>
<td>DXSLVEVX</td>
<td>FPXILQAVV</td>
<td>SIA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>TTR</td>
<td>XDDXDSAAAASIPMVPGWLKVQV</td>
<td>XGSQA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>61</td>
<td>GSF</td>
<td>LAVMGGDLEVIIXLAG</td>
<td>QESIXA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>SRS</td>
<td>LAASMXTTAIPS</td>
<td>DLWGNXAXSNAAFSS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>121</td>
<td>XFS</td>
<td>SXSAGSVP</td>
<td>LGFTXEXAGAKERXVIKQI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>151</td>
<td>TXQ</td>
<td>AAXAFSLAXLKL</td>
<td>ISAMXNAFPAGDX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>181</td>
<td>XXV</td>
<td>ADIXDSHGILXXVNYTDAXIKMGIGF</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>211</td>
<td>SGV</td>
<td>NAAYWCDSTX</td>
<td>IADADAGXGAGXM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>241</td>
<td>VCC</td>
<td>XQDSFRKAFP</td>
<td>SLPQIXYXTNLNXSPX</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>271</td>
<td>AXK</td>
<td>TFEKNSXAKNX</td>
<td>QSRLDVMXYKXXGQ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>301</td>
<td>XXA</td>
<td>XDFXANVENSSYPA</td>
<td>IKQLPHFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>331</td>
<td>LRX</td>
<td>XDLFXGDQ</td>
<td>GIA</td>
<td>AXKTXMXXVVRXLFL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>361</td>
<td>IAA</td>
<td>YAFRVLVVCXI</td>
<td>AICQK</td>
<td>KGYSSEGHHIAAX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>391</td>
<td>GSR</td>
<td>DOYSFGSXS</td>
<td>N</td>
<td>XTNXKNIYGPQPSAXX</td>
<td></td>
<td></td>
</tr>
<tr>
<td>421</td>
<td>KPI</td>
<td>XI</td>
<td>TPAID</td>
<td>GEAAXXVIXSIASSQXXA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>451</td>
<td>XXS</td>
<td>AAXA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hemoglobin

- protein built from 4 polypeptides
- responsible for carrying oxygen in red blood cells

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)
- a gene is said to encode a protein
- the human genome comprises ~ 40,000 genes
  - there is some controversy about this number
Gene Density

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

  microbes  90% coding  gene/kb
  human    3% coding    gene/35kb

The Central Dogma

[Diagram showing the central dogma process: replication, transcription, translation, and protein synthesis.]
RNA

- RNA is like DNA except:
  - backbone is a little different
  - usually single stranded
  - the base uracil (U) is used in place of thymine (T)
- a strand of RNA can be thought of as a string composed of the four letters: A, C, G, U

Transcription

```
DNA  ATGCCGTTAGACCGTTAGCGGAACCTGAC
     TACCGGAAATCTGGCAATCGCCTGGACCTG
  3'   5'

mRNA synthesis
mRNA
AUGCCGUUAGACGCUUAGCGGACCUAGAC
  5'   3'
```

- Top strand coding strand
- Sense strand
- Bottom strand template strand
- Antisense strand
Transcription

- RNA polymerase is the enzyme that builds an RNA strand from a gene
- RNA that is transcribed from a gene is called messenger RNA (mRNA)
  – we’ll talk about other varieties of RNA later in the course

The Genetic Code

[Genetic code table]

- Phenylalanine
- Leucine
- Serine
- Tyrosine
- Stop codon
- Glutamine
- Arginine
- Glycerine
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

Growing poly-peptide

Stop codon

This process repeats until reaching a stop codon
RNA Processing in Eukaryotes

- **eukaryotes** are organisms that have enclosed nuclei in their cells
- in eukaryotes, mRNA consists 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

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

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

DNA → Replication → RNA → Translation → Protein

Transcription