Bioinformatics: Network Analysis

Molecular Cell Biology: A Brief Review

COMP 572 (BIOS 572 / BIOE 564) - Fall 2013
Luay Nakhleh, Rice University
The Tree of Life

[Diagram of the Tree of Life showing relationships between different species and kingdoms, including bacteria, archaea, and eukaryotes, with specific examples such as Methanococcus, Thermofilum, Sulfolobus, Haloferax, and human, maize, yeast, Paramecium, Dictyostelium, Euglena, Trypanosoma, Giardia, and Trichomonas.]
Prokaryotic vs. Eukaryotic Cell Structure

Source: Pearson Education, Inc. The Biology Place
# Prokaryotic vs. Eukaryotic Cells

<table>
<thead>
<tr>
<th></th>
<th>Prokaryotes</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>1-10 μm in length</td>
<td>10-100 μm in length</td>
</tr>
<tr>
<td><strong>Nucleus</strong></td>
<td>does not exist</td>
<td>exists, and separated from the cytoplasm</td>
</tr>
<tr>
<td><strong>Intracellular</strong></td>
<td>no compartments</td>
<td>compartments (nucleus, cytosol, mitochondria, etc.)</td>
</tr>
<tr>
<td><strong>organization</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Gene structure</strong></td>
<td>no introns</td>
<td>introns and exons</td>
</tr>
<tr>
<td><strong>Cell division</strong></td>
<td>simple cell division</td>
<td>mitosis or meiosis</td>
</tr>
<tr>
<td><strong>Ribosome</strong></td>
<td>consists of a large 50S subunit and a small 30S subunit</td>
<td>consists of a large 60S subunit and a small 40S subunit</td>
</tr>
<tr>
<td><strong>Reproduction</strong></td>
<td>parasexual recombination</td>
<td>sexual recombination</td>
</tr>
<tr>
<td><strong>Organization</strong></td>
<td>mostly single cellular</td>
<td>mostly multicellular, and with cell differentiation</td>
</tr>
</tbody>
</table>

Source: Systems Biology in Practice, Klipp et al.
The full diversity of life on this planet—from the simplest bacterium to the largest mammal—is captured in a linear code inside all living cells.
DNA

- Deoxyribonucleic Acid

- DNA molecules are linear polymers of just four different nucleotide building blocks.

- Genomic DNA molecules are immensely long, containing millions of bases each, and it is the order of these bases, the nucleotide sequence or base sequence of DNA, which encodes the information for making proteins.
RNA

- Ribonucleic Acid
- RNA molecules are also linear polymers, but are much smaller than genomic DNA.
- Most RNA molecules also contain just four different base types.
- Several classes of RNA molecules are known, some of which have a small proportion of other bases.
The Building Blocks of DNA and RNA

(A) BASE

PHOSPHATE

PHOSPHATE

SUGAR

SUGAR

(B) cytosine uracil adenine thymine

(B) NH₂

NH₂

C

C

N

N

C

C

O

O

O

O

H₂C

H₂C

C

C

O

O

O

O

NH₂

NH₂

C

C

O

O

O

O

H₂C

H₂C

C

C

O

O

O

O


(C) 5' end of chain

phosphodiester linkage

3' OH

3' OH

3' end of chain

5' end of chain

5' end of chain
The Double Helix (DNA)

Watson-Crick base-pairing: A—T, C—G

Each strand of a DNA double helix has a base sequence that is complementary to the base sequence of its partner strand.
DNA Replication

* Hydrogen bonds are noncovalent bonds: the two DNA strands can be easily separated.
* There are a number of processes in which strand separation is required.
* One such process is DNA replication, which is a necessary prelude to cell division.
RNA Structure

- Almost all RNA molecules in living systems are single stranded.
- As a result, RNA has much more structural flexibility than DNA, and some RNAs can even act as enzymes, catalyzing a particular chemical reaction.
Secondary and Tertiary Structures of RNA

The Tetrahymena ribozyme
The Central Dogma

- A single direction of flow of genetic information from the DNA (information store), through RNA, to proteins
- This scheme holds for all known forms of life, with variations in the details of the processes involved in different organisms
- Not all genetic information in the DNA encodes proteins
- RNA can also be the end product, and other regions of the genome have as yet no known function of product
- The genomic DNA encodes all molecules necessary for life, whether they are proteins or RNA or ...
Transcription

(A) One strand of the DNA is involved in the synthesis of an RNA strand complementary to the strand of the DNA.

(B) The enzyme RNA polymerase reads the DNA and recruits the correct building blocks of RNA to string them together based on the DNA code.
Terminology

- RNA transcribed from a protein-coding gene is called messenger RNA (mRNA)
- When a gene is being transcribed into RNA, the gene is said to be expressed
Overlapping Genes

Although only one segment of the DNA strand is transcribed for any given gene, it is also possible for genes to overlap so that one or both strands at the same location (locus) encode parts of different proteins.

This most commonly occurs in viruses as a means of packing as much information as possible into their very small genomes but it could also occur in mammals (the above figure shows overlapping genes in the human genome).
The genomic DNA sequence contains more information than just the protein sequences. The transcriptional apparatus has to locate the sites where gene transcription should begin, and when to transcribe a given gene. At any one time, a cell is only expressing a few thousand of the genes in its genome. To accomplish this regulated gene expression, the DNA contains control sequences in addition to coding regions (More on this in a few slides).
Translation

- mRNA is translated into protein according to the genetic code, which is the set of rules governing the correspondence of the base sequences in DNA or RNA to the amino acid sequence of a protein.

- Each amino acid is encoded by a set of three consecutive bases (codon)
The Standard Genetic Code

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>Third letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UUU Phe</td>
<td>UGU Cys</td>
</tr>
<tr>
<td></td>
<td>UUC Leu</td>
<td>UGC Stop</td>
</tr>
<tr>
<td></td>
<td>UUA Leu</td>
<td>UGA Stop</td>
</tr>
<tr>
<td></td>
<td>UUG Leu</td>
<td>UGG Trp</td>
</tr>
<tr>
<td>C</td>
<td>CUU Leu</td>
<td>CGU Arg</td>
</tr>
<tr>
<td></td>
<td>CUC Leu</td>
<td>CGC Arg</td>
</tr>
<tr>
<td></td>
<td>CUA Leu</td>
<td>CGA Arg</td>
</tr>
<tr>
<td></td>
<td>CUG Leu</td>
<td>CGG Arg</td>
</tr>
<tr>
<td>A</td>
<td>AUU Ile</td>
<td>AGU Ser</td>
</tr>
<tr>
<td></td>
<td>AUC Ile</td>
<td>AGC Ser</td>
</tr>
<tr>
<td></td>
<td>AUA Ile</td>
<td>AGA Arg</td>
</tr>
<tr>
<td></td>
<td>AUG Ile</td>
<td>AGG Arg</td>
</tr>
<tr>
<td>G</td>
<td>GUU Val</td>
<td>GGU Gly</td>
</tr>
<tr>
<td></td>
<td>GUC Val</td>
<td>GGC Gly</td>
</tr>
<tr>
<td></td>
<td>GUA Val</td>
<td>GGA Gly</td>
</tr>
<tr>
<td></td>
<td>GUG Val</td>
<td>GGG Gly</td>
</tr>
</tbody>
</table>
Reading Frames

- Translation occurs in nonoverlapping sets of three bases.
- There are thus three possible ways to translate any nucleotide sequence, each of which is called a reading frame.
- These three reading frames give three different protein sequences.
- In the actual translation process, the detailed control signals ensure that only the appropriate reading frame is translated into protein.
Reading Frames

5’
C U C A G C G U U A C C A U

—Leu— Ser — Val — Thr —

3’
C U C A G C G U U A C C A U

— Ser — Ala — Leu — Pro —

5’
C U C A G C G U U A C C A U

— Gln — Arg — Tyr — His —

3’
Translation Machinery

- There are three main classes of RNA in the cell: messenger RNA (mRNA), ribosomal RNA (rRNA), and transfer RNA (tRNA).

- rRNA and tRNA are involved in mRNA translation and protein synthesis
The regulation of many processes that interpret the information contained in a DNA sequence relies on the presence of short signal sequences in the DNA.

The general term for these signal sequences is regulatory elements.

For example, the molecules involved in transcription and translation require signals to identify where they should start and stop.

Gene structure and control differ between prokaryotes and eukaryotes.
Transcription Regulation

- The control regions at which RNA polymerase binds to initiate transcription are called promoters.

- RNA polymerase binds more tightly to these regions than to the rest of the DNA and this triggers the start of transcription.
* Bacterial promoters typically occur immediately before the position of the transcription start site (TSS), and contain two characteristic short sequences, or motifs, that are almost the same in the promoters for different genes.
* The termination of transcription is controlled by the terminator signal which in bacteria differs from the promoter in that it is active when transcribed to form the end of the mRNA strand (forms a loop structure that prevents the transcription apparatus from continuing).
* Single type of RNA polymerase transcribes all genes.
Gene Structure in Eukaryotes

* Regulatory elements in eukaryotes are more complex.

* Three types of RNA polymerase transcribe genes: RNA polymerase II transcribes all protein coding genes, where other RNA polymerase types transcribe genes for tRNAs, rRNAs and other types of RNA.
Splicing of an Intron

* The existence of introns necessitates an extra step between transcription and translation, which is known as RNA splicing: (1) the complete gene is initially transcribed into RNA, and (2) the introns are then excised and the exons spliced together to provide a functional mRNA that gives the correct protein sequence when translated. In most protein coding genes, this process is carried out by the spliceosome, which consists of small nuclear RNA (snRNA) and proteins.
In bacteria, functionally related protein-coding sequences are often clustered together into operons. Each operon is transcribed as a single mRNA transcript and the proteins are then separately translated from this one long molecule. This has the advantage that only one control region is required to activate the simultaneous expression of all genes in the operon. Not all bacterial genes are contained in operons; many are transcribed individually and have their own control regions.
Proteins
Levels of Protein Structure

- **Primary Structure:**
  - N terminus—MYCATISEATINGFISHANDMEATANDWATER—C terminus

- **Secondary Structure:**
  - Represents the regular patterns of secondary structure (e.g., alpha helices, beta sheets).

- **Tertiary Structure:**
  - Shows the three-dimensional shape of a single protein molecule.

- **Quaternary Structure:**
  - Refers to the arrangement of two or more polypeptide chains to form a multi-subunit protein.
Side Chains of the Amino Acids

**BASIC SIDE CHAINS**
- Lysine
- Arginine
- Histidine

**NONPOLAR SIDE CHAINS**
- Alanine
- Valine
- Leucine
- Isoleucine
- Methionine
- Tyrosine
- Cysteine

**ACIDIC SIDE CHAINS**
- Aspartic acid
- Glutamic acid

**UNCHARGED POLAR SIDE CHAINS**
- Asparagine
- Glutamine
- Serine
- Threonine
- Glycine

Chemical structures of various amino acids are depicted above.
Organization of the DNA

- The genome is an organism’s complete set of DNA

- Genomes vary widely in size
  - some bacteria have 600,000 base pairs
  - humans have about 3 billion base pairs
  - Except for mature red blood cells, all human cells contain a complete genome

- DNA in the human genome is arranged into 23 pairs of DNA molecules, called chromosomes (physically separate molecules, and vary widely in length)

- Each chromosome contains many genes
Available Data

858 Databases in total
(as classified at the NAR Molecular Biology Database Collection Website, 2006)