Outline:

• 0. History: Major Events in Molecular Biology
• 1. What Is Life Made Of?
• 2. What Is Genetic Material?
• 3. What Do Genes Do?
• 4. What Molecule Code For Genes?
• 5. What Is the Structure Of DNA?
• 6. What Carries Information between DNA and Proteins
• 7. How are Proteins Made?
Outline Cont.

- 8. How Can We Analyze DNA
  - 1. Copying DNA
  - 2. Cutting and Pasting DNA
  - 3. Measuring DNA Length
  - 4. Probing DNA
- 9. How Do Individuals of a Species Differ
- 10. How Do Different Species Differ
  - 1. Molecular Evolution
  - 2. Comparative Genomics
  - 3. Genome Rearrangement
- 11. Why Bioinformatics?
How Molecular Biology came about?

• Microscopic biology began in 1665

• Robert Hooke (1635-1703) discovered organisms are made up of cells

• Matthias Schleiden (1804-1881) and Theodor Schwann (1810-1882) further expanded the study of cells in 1830s

• Robert Hooke

• Matthias Schleiden

• Theodor Schwann
Major events in the history of Molecular Biology 1800 - 1870

1865  Gregor Mendel discover the basic rules of heredity of garden pea.

- An individual organism has two alternative heredity units for a given trait (dominant trait v.s. recessive trait)

1869  Johann Friedrich Miescher discovered DNA and named it nuclein.

Mendel: The Father of Genetics

Johann Miescher
Major events in the history of Molecular Biology  1880 - 1900

• **1881** Edward Zacharias showed chromosomes are composed of nuclein.

• **1899** Richard Altmann renamed nuclein to nucleic acid.

• **By 1900**, chemical structures of all 20 amino acids had been identified
Major events in the history of Molecular Biology 1900-1911

- **1902** - Emil Hermann Fischer wins Nobel prize: showed amino acids are linked and form proteins
  - Postulated: protein properties are defined by amino acid composition and arrangement, which we nowadays know as fact

- **1911** – Thomas Hunt Morgan discovers genes on chromosomes are the discrete units of heredity

- **1911** Pheobus Aaron Theodore Lerene discovers RNA
Major events in the history of Molecular Biology 1940 - 1950

- **1941** – George Beadle and Edward Tatum identify that genes make proteins

- **1950** – Edwin Chargaff find Cytosine complements Guanine and Adenine complements Thymine
Major events in the history of Molecular Biology 1950 - 1952

- **1950s** – Mahlon Bush Hoagland first to isolate tRNA

- **1952** – Alfred Hershey and Martha Chase make genes from DNA
Major events in the history of Molecular Biology 1952 - 1960

- **1952-1953** James D. Watson and Francis H. C. Crick deduced the double helical structure of DNA

- **1956** George Emil Palade showed the site of enzymes manufacturing in the cytoplasm is made on RNA organelles called ribosomes.
Major events in the history of Molecular Biology 1970

- 1970 Howard Temin and David Baltimore independently isolate the first restriction enzyme.

- DNA can be cut into reproducible pieces with site-specific endonuclease called restriction enzymes;
  - the pieces can be linked to bacterial vectors and introduced into bacterial hosts. (gene cloning or recombinant DNA technology)
Major events in the history of Molecular Biology 1970-1977

- **1977** Phillip Sharp and Richard Roberts demonstrated that pre-mRNA is processed by the excision of introns and exons are spliced together.

- Joan Steitz determined that the 5’ end of snRNA is partially complementary to the consensus sequence of 5’ splice junctions.
Major events in the history of Molecular Biology
1986 - 1995

- **1986** Leroy Hood: Developed automated sequencing mechanism

- **1986** Human Genome Initiative announced

- **1990** The 15 year Human Genome project is launched by congress

- **1995** Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published (These maps provide the locations of “markers” on each chromosome to make locating genes easier)
Major events in the history of Molecular Biology 1995-1996

- **1995** John Craig Venter: First bacterial genomes sequenced
- **1995** Automated fluorescent sequencing instruments and robotic operations
- **1996** First eukaryotic genome - yeast - sequenced
Major events in the history of Molecular Biology
1997 - 1999

• 1997 E. Coli sequenced

• 1998 PerkinsElmer, Inc.. Developed 96-capillary sequencer

• 1998 Complete sequence of the Caenorhabditis elegans genome

• 1999 First human chromosome (number 22) sequenced
Major events in the history of Molecular Biology 2000-2001

- **2000** Complete sequence of the euchromatic portion of the *Drosophila melanogaster* genome

- **2001** International Human Genome Sequencing: first draft of the sequence of the human genome published
Major events in the history of Molecular Biology 2003- Present

- **April 2003** Human Genome Project Completed. Mouse genome is sequenced.

- **April 2004** Rat genome sequenced.
Section 1: What is Life made of?
Outline For Section 1:

- All living things are made of Cells
  - Prokaryote, Eukaryote
- Cell Signaling
- What is Inside the cell: From DNA, to RNA, to Proteins
Cells

- Fundamental working units of every living system.
- Every organism is composed of one of two radically different types of cells: prokaryotic cells or eukaryotic cells.
- Prokaryotes and Eukaryotes are descended from the same primitive cell.
  - All extant prokaryotic and eukaryotic cells are the result of a total of 3.5 billion years of evolution.
Cells

- **Chemical composition** by weight
  - 70% water
  - 7% small molecules
    - salts
    - Lipids
    - amino acids
    - nucleotides
  - 23% macromolecules
    - Proteins
    - Polysaccharides
    - lipids
- **biochemical (metabolic) pathways**
- **translation of mRNA into proteins**
Life begins with Cell

- A cell is the smallest structural unit of an organism that is capable of independent functioning.
- All cells have some common features.
All Cells have common Cycles

- Born, eat, replicate, and die
2 types of cells: Prokaryotes v.s. Eukaryotes
According to the most recent evidence, there are three main branches to the tree of life. Prokaryotes include Archaea (“ancient ones”) and bacteria. Eukaryotes are kingdom Eukarya and includes plants, animals, fungi and certain algae.
### Prokaryotes and Eukaryotes, continued

<table>
<thead>
<tr>
<th>Prokaryotes</th>
<th>Eukaryotes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single cell</td>
<td>Single or multi cell</td>
</tr>
<tr>
<td>No nucleus</td>
<td>Nucleus</td>
</tr>
<tr>
<td>No organelles</td>
<td>Organelles</td>
</tr>
<tr>
<td>One piece of circular DNA</td>
<td>Chromosomes</td>
</tr>
<tr>
<td>No mRNA post transcriptional modification</td>
<td>Exons/Introns splicing</td>
</tr>
</tbody>
</table>
Prokaryotes v.s. Eukaryotes
Structural differences

Prokaryotes
- Eubacterial (blue green algae) and archaebacteria
- only one type of membrane--plasma membrane forms
  - the boundary of the cell proper
- The smallest cells known are bacteria
  - Ecoli cell
  - $3 \times 10^6$ protein molecules
  - 1000-2000 polypeptide species.

Eukaryotes
- plants, animals, Protista, and fungi
- complex systems of internal membranes forms
  - organelle and compartments
- The volume of the cell is several hundred times larger
  - Hela cell
  - $5 \times 10^9$ protein molecules
  - 5000-10,000 polypeptide species
# Prokaryotic and Eukaryotic Cells

## Chromosomal differences

### Prokaryotes
- The genome of E.coli contains \(4 \times 10^6\) base pairs
- > 90% of DNA encode protein
- Lacks a membrane-bound nucleus.
  - Circular DNA and supercoiled domain
- Histones are unknown

### Eukaryotes
- The genome of yeast cells contains \(1.35 \times 10^7\) base pairs
- A small fraction of the total DNA encodes protein.
  - Many repeats of non-coding sequences
- All chromosomes are contained in a membrane bound nucleus
  - DNA is divided between two or more chromosomes
- A set of five histones
  - DNA packaging and gene expression regulation
Signaling Pathways: Control Gene Activity

- Instead of having brains, cells make decisions through complex networks of chemical reactions, called pathways
  - Synthesize new materials
  - Break other materials down for spare parts
  - Signal to eat or die
Example of cell signaling

**Signal**
Solute enters the space between the two membranes through large pores in the outer membrane of *E. coli*.

**Receptor**
The EnvZ membrane protein changes shape in response to the high solute concentration, catalyzing the addition of a phosphate from ATP.

**Transduction**
The phosphate from EnvZ is transferred to the OmpR protein...
...and the phosphorylated OmpR changes shape, enabling it to bind to DNA and stimulate transcription of the *ompC* gene.

**Effects**
OmpC protein inserts into the outer membrane, preventing solute entry and keeping the cell’s exterior osmotically balanced.
Cells Information and Machinery

- 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
- Machinery:
  - Collect and manufacture components
  - Carry out replication
  - Kick-start its new offspring

(A cell is like a car factory)
Overview of organizations of life

• **Nucleus** = library
• **Chromosomes** = bookshelves
• **Genes** = books

• Almost every cell in an organism contains the same libraries and 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

- **Genome**: an organism’s genetic material
- **Gene**: a discrete units of hereditary information located on the chromosomes and consisting of DNA.
- **Genotype**: The genetic makeup of an organism
- **Phenotype**: the physical expressed traits of an organism
- **Nucleic acid**: Biological molecules (RNA and DNA) that allow organisms to reproduce;
More Terminology

- **genome** is an organism’s complete set of DNA.
  - a bacteria contains about 600,000 DNA base pairs
  - human and mouse genomes have some 3 billion.
- human genome has 24 distinct chromosomes.
  - Each chromosome contains many **genes**.
- **Gene**
  - basic physical and functional units of heredity.
  - specific sequences of DNA bases that encode instructions on how to make **proteins**.
- **Proteins**
  - Make up the cellular structure
  - large, complex molecules made up of smaller subunits called **amino acids**.
All Life depends on 3 critical molecules

- **DNAs**
  - Hold information on how cell works

- **RNAs**
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein

- **Proteins**
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body’s major components (e.g. hair, skin, etc.)
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.
DNA, continued

- DNA has a double helix structure which composed of:
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

- DNA always reads from 5’ end to 3’ end for transcription replication:
  5’ ATTTAGGCC 3’
  3’ TAAATCCGG 5’
DNA, RNA, and the Flow of Information

Replication

DNA can replicate.

Transcription

DNA

Information coded in the sequence of base pairs in DNA is passed to molecules of RNA.

Translation

RNA

Information in RNA is passed to proteins. It never passes from proteins to nucleic acids.

Protein
A gene is expressed in two steps
1) Transcription: RNA synthesis
2) Translation: Protein synthesis
DNA the Genetics Makeup

- Genes are inherited and are expressed
  - **genotype** (genetic makeup)
  - **phenotype** (physical expression)

- On the left, is the eye’s phenotypes of green and black eye genes.
Cell Information: Instruction book of Life

- DNA, RNA, and Proteins are examples of strings written in either the four-letter nucleotide of DNA and RNA (A C G T/U)
- or the twenty-letter amino acid of proteins. Each amino acid is coded by 3 nucleotides called codon. (Leu, Arg, Met, etc.)
END of SECTION 1
Section 2: Genetic Material of Life
Outline For Section 2:

• What is Genetic Material?

• *Mendel’s experiments*
  • *Pea plant experiments*

• *Mutations in DNA*
  • Good, Bad, Silent

• *Chromosomes*
  • Linked Genes
  • Gene Order
  • Genetic Maps
  • Chromosomes and sexual reproduction
Mendel and his Genes

• What are genes?
  - physical and functional traits that are passed on from one generation to the next.
• Genes were discovered by Gregor Mendel in the 1860s while he was experimenting with the pea plant. He asked the question:

  Do traits come from a blend of both parent's traits or from only one parent?
The Pea Plant Experiments

- Mendel discovered that genes were passed on to offspring by both parents in two forms: dominant and recessive.

- The dominant form would be the phenotypic characteristic of the offspring.

[Diagram showing Mendel's cross experiment with purple and white flowers, illustrating the dominance of purple flowers with most offspring having that color.]
DNA: the building blocks of genetic material

- DNA was later discovered to be the molecule that makes up the inherited genetic material.
- Experiments performed by Fredrick Griffith in 1928 and experiments with bacteriophages in 1952 led to this discovery. (BILD 1 Lecture, UCSD, Fall 2003)
- DNA provides a code, consisting of 4 letters, for all cellular function.

Letters in DNA code: CAGT
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?
The Good, the Bad, and the Silent

- Mutations can serve the organism in three ways:

  - **The Good**: A mutation can cause a trait that enhances the organism’s function:
    - Mutation in the sickle cell gene provides resistance to malaria.

  - **The Bad**: A mutation can cause a trait that is harmful, sometimes fatal to the organism:
    - 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
Genes are Organized into Chromosomes

- What are chromosomes?
  
  It is a threadlike structure found in the nucleus of the cell which is made from a long strand of DNA. Different organisms have a different number of chromosomes in their cells.

- Thomas Morgan (1920s) - Evidence that genes are located on chromosomes was discovered by genetic experiments performed with flies.
The White-Eyed Male

White-eyed male

X

Red-eyed female
(normal)

white-eyed

Mostly male progeny

Red-eyed

Mostly female progeny

These experiments suggest that the gene for eye color must be linked or co-inherited with the genes that determine the sex of the fly. This means that the genes occur on the same chromosome; more specifically it was the X chromosome.
Linked Genes and Gene Order

• Along with eye color and sex, other genes, such as body color and wing size, had a higher probability of being co-inherited by the offspring ➔ genes are linked.

• Morgan hypothesized that the closer the genes were located on the a chromosome, the more often the genes are co-inherited.
Linked Genes and Gene Order cont…

- By looking at the frequency that two genes are co-inherited, genetic maps can be constructed for the location of each gene on a chromosome.
- One of Morgan’s students Alfred Sturtevant pursued this idea and studied 3 fly genes:

  - **Orange Eyes**
  - **cn-eye**
  - **color**
  - **White Eyes**

Fly pictures from: http://www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html
Linked Genes and Gene Order cont...

- By looking at the frequency that two genes are co-inherited, genetic maps can be constructed for the location of each gene on a chromosome.

- One of Morgan’s students Alfred Sturtevant pursued this idea and studied 3 fly genes:
  - Normal Fly
  - en-eye color
  - b-body color
  - Yellow Body
  - Ebony Body

Fly pictures from: http://www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html
Linked Genes and Gene Order cont...

• By looking at the frequency that two genes are co-inherited, genetic maps can be constructed for the location of each gene on a chromosome.

• One of Morgan’s students Alfred Sturtevant pursued this idea and studied 3 fly genes:

  - eyecolor
  - bodycolor
  - wingsize

Fly pictures from: http://www.exploratorium.edu/exhibits/mutant_flies/mutant_flies.html
What are the genes’ order on the chromosome?

- Mutant $b$, mutant $vg$:
  - Normal fly $\times$
  - 17% progeny have only one mutation

- Mutant $b$, mutant $cn$:
  - Normal fly $\times$
  - 9% progeny have only one mutation

- Mutant $vg$, mutant $cn$:
  - Normal fly $\times$
  - 8% progeny have only one mutation

The genes $vg$ and $b$ are farthest apart from each other.

The gene $cn$ is close to both $vg$ and $b$. 
What are the genes’ order on the chromosome?

<table>
<thead>
<tr>
<th></th>
<th>b</th>
<th>cn</th>
<th>vg</th>
</tr>
</thead>
</table>

This is the order of the genes, on the chromosome, determined by the experiment.
Genetic Information: Chromosomes

- (1) Double helix DNA strand.
- (2) Chromatin strand (DNA with histones)
- (3) Condensed chromatin during interphase with centromere.
- (4) Condensed chromatin during prophase
- (5) Chromosome during metaphase
<table>
<thead>
<tr>
<th>Organism</th>
<th>Number of base pair</th>
<th>number of 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 (insect)</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>
</tbody>
</table>
Sexual Reproduction

- Formation of new individual by a combination of two haploid sex cells (gametes).
- Fertilization - combination of genetic information from two separate cells that have one half the original genetic information.
- Gametes for fertilization usually come from separate parents:
  1. Female - produces an egg
  2. Male produces sperm
- Both gametes are haploid, with a single set of chromosomes
- The new individual is called a zygote, with two sets of chromosomes (diploid).
- Meiosis is a process to convert a diploid cell to a haploid gamete, and cause a change in the genetic information to increase diversity in the offspring.
Meiosis

- Meiosis comprises two successive nuclear divisions with only one round of DNA replication.

- First division of meiosis
  - **Prophase 1**: Each chromosome duplicates and remains closely associated. These are called sister chromatids. Crossing-over can occur during the latter part of this stage.
  - **Metaphase 1**: Homologous chromosomes align at the equatorial plate.
  - **Anaphase 1**: Homologous pairs separate with sister chromatids remaining together.
  - **Telophase 1**: Two daughter cells are formed with each daughter containing only one chromosome of the homologous pair.
Meiosis

• **Second division of meiosis**: Gamete formation
  • **Prophase 2**: DNA does not replicate.
  • **Metaphase 2**: Chromosomes align at the equatorial plate.
  • **Anaphase 2**: Centromeres divide and sister chromatids migrate separately to each pole.
  • **Telophase 2**: Cell division is complete. Four haploid daughter cells are obtained.
• One parent cell produces **four daughter cells**.

Daughter cells:
• half the number of chromosomes found in the original parent cell
• crossing over cause genetically difference.
Meiosis

Diagram 1.
END of SECTION 2
Section 3: What Do Genes Do?
Outline For Section 3:

- **Beadle and Tatum Experiment**
- **Design of Life** (gene->protein)
- protein synthesis
  - Central dogma of molecular biology
Beadle and Tatum Experiment

- Experiment done at Stanford University 1941

- The hypothesis: One gene specifies the production of one enzyme

- They chose to work with bread mold (*Neurospora*) biochemistry already known (worked out by Carl C. Lindegren)
  - Easy to grow, maintain
  - Short life cycle
  - Easy to induce mutations
  - Easy to identify and isolate mutants
Beadle and Tatum Experiment Procedure

- 2 different growth media:
  - Complete - consists of agar, inorganic salts, malt & yeast extract, and glucose
  - Minimal - consists of agar, inorganic salts, biotin, disaccharide and fat

- X-ray used to irradiate Neurospora to induce mutation
- Mutated spores placed onto minimal medium
Beadle and Tatum Experiment Procedure

Growth of wild-type *Neurospora* (prototroph) on minimal medium

Exposure of spores to X rays

Growth, crossing with opposite mating type, development of asci in fruiting body

Beadle and Tatum Experiment Procedure

No growth of mutant auxotroph on minimal medium

Growth on enriched medium

Vitamins

Amino acids

Nucleic acid bases

Beadle and Tatum Experiment Procedure

Histidine  Proline  Methionine  Leucine

Tryptophan  Lysine  Arginine  Valine

Minimal medium supplemented with specific amino acids

Beadle and Tatum Experiment Conclusions

• Irradiated Neurospora survived when supplemented with Vitamin B6

• X-rays damaged genes that produces a protein responsible for the synthesis of Vitamin B6

• three mutant strains - substances unable to synthesize (Vitamin B6, Vitamin B1 and Para-aminobenzoic acid) essential growth factors

• crosses between normal and mutant strains showed differed by a single gene

• hypothesized that there was more than one step in the synthesis of Vitamin B6 and that mutation affects only one specific step

• Evidence: One gene specifies the production of one enzyme!
Genes Make Proteins

- genome -> genes -> protein (forms cellular structural & life functional) -> pathways & physiology
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.
END of SECTION 3
Section 4: What Molecule Codes For Genes?
Outline For Section 4:

• *Discovery of the Structure of DNA*
  • *Watson and Crick*

• *DNA Basics*
Discovery of DNA

- DNA Sequences
  - Chargaff and Vischer, 1949
  - DNA consisting of A, T, G, C
    - Adenine, Guanine, Cytosine, Thymine
  - Chargaff Rule
    - Noticing \( #A \approx #T \) and \( #G \approx #C \)
    - A “strange but possibly meaningless” phenomenon.
- Wow!! A Double Helix
    - 1 Biologist
    - 1 Physics Ph.D. Student
    - 900 words
    - Nobel Prize
- Rich, 1973
  - Structural biologist at MIT.
  - DNA’s structure in atomic resolution.
Watson & Crick – “…the secret of life”

- Watson: a zoologist, Crick: a physicist

- “In 1947 Crick knew no biology and practically no organic chemistry or crystallography…” – www.nobel.se

- Applying Chargaff’s rules and the X-ray image from Rosalind Franklin, they constructed a “tinkertoy” model showing the double helix

- Their 1953 Nature paper: “It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”

Watson & Crick with DNA model

Rosalind Franklin with X-ray image of DNA
DNA: The Basis of Life

- **Deoxyribonucleic Acid (DNA)**
  - Double stranded with complementary strands A-T, C-G
- **DNA is a polymer**
  - Sugar-Phosphate-Base
  - Bases held together by H bonding to the opposite strand
Double helix of DNA

- James Watson and Francis Crick proposed a model for the structure of DNA.
  - Utilizing X-ray diffraction data, obtained from crystals of DNA.
- This model predicted that DNA
  - as a helix of two complementary anti-parallel strands,
  - wound around each other in a rightward direction
  - stabilized by H-bonding between bases in adjacent strands.
  - The bases are in the interior of the helix
    - Purine bases form hydrogen bonds with pyrimidine.
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
# Human Genome Composition

## Table 10.1: Major Classes of Eukaryotic DNA and Their Representation in the Human Genome

<table>
<thead>
<tr>
<th>Class</th>
<th>Length</th>
<th>Copy Number in Human Genome</th>
<th>Fraction of Human Genome, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein-coding genes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solitary genes</td>
<td>Variable</td>
<td>1</td>
<td>≈15% (0.8)†</td>
</tr>
<tr>
<td>Duplicated or diverged genes in gene families</td>
<td>Variable</td>
<td>2—1000</td>
<td>≈15% (0.8)†</td>
</tr>
<tr>
<td>Tandemly repeated genes encoding rRNAs, tRNAs, snRNAs, and histones</td>
<td>Variable</td>
<td>20–300</td>
<td>0.3</td>
</tr>
<tr>
<td>Repetitious DNA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple-sequence DNA</td>
<td>1–500 bp</td>
<td>Variable</td>
<td>3</td>
</tr>
<tr>
<td>Interspersed repeats</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA transposons</td>
<td>2–3 kb</td>
<td>300,000</td>
<td>3</td>
</tr>
<tr>
<td>LTR retrotransposons</td>
<td>6–11 kb</td>
<td>440,000</td>
<td>8</td>
</tr>
<tr>
<td>Non-LTR retrotransposons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LINEs</td>
<td>6–8 kb</td>
<td>860,000</td>
<td>21</td>
</tr>
<tr>
<td>SINEs</td>
<td>100–300 bp</td>
<td>1,600,000</td>
<td>13</td>
</tr>
<tr>
<td>Processed pseudogenes</td>
<td>Variable</td>
<td>1—100</td>
<td>≈0.4</td>
</tr>
<tr>
<td>Unclassified spacer DNA</td>
<td>Variable</td>
<td>n.a.†</td>
<td>≈25</td>
</tr>
</tbody>
</table>

*Complete transcription units, including introns.
†Protein-coding exons. The total number of human protein-coding genes is estimated to be 30,000–35,000, but this number is based on current methods for identifying genes in the human genome sequence and may be an underestimate.
‡Not applicable.

END of SECTION 4
Section 5: The Structure of DNA

CSE 181
Raymond Brown
May 12, 2004
Outline For Section 5:

- DNA Components
  - Nitrogenous Base
  - Sugar
  - Phosphate
- Double Helix
- DNA replication
- Superstructure
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

http://www.bio.miami.edu/dana/104/DNA2.jpg
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 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
Basic Structure

Watson-Crick base pair structures

Phosphate
Sugar
Basic Structure Implications

- **DNA is (-) charged due to phosphate:**
  gel electrophoresis, DNA sequencing (Sanger method)

- **H-bonds form between specific bases:**
  hybridization – replication, transcription, translation
  DNA microarrays, hybridization blots, PCR
  C-G bound tighter than A-T due to triple H-bond

- **DNA-protein interactions (via major & minor grooves):**
  transcriptional regulation

- **DNA polymerization:**
  5’ to 3’ – phosphodiester bond formed between 5’ phosphate and 3’ OH
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
Double helix of DNA

- The double helix of DNA has these features:
  - Concentration of adenine (A) is equal to thymine (T)
  - Concentration of cytidine (C) is equal to guanine (G).
  - Watson-Crick base-pairing A will only base-pair with T, and C with G
    - Base-pairs of G and C contain three H-bonds,
    - Base-pairs of A and T contain two H-bonds.
    - G-C base-pairs are more stable than A-T base-pairs
  - Two polynucleotide strands wound around each other.
  - The backbone of each consists of alternating **deoxyribose** and **phosphate groups**
Double helix of DNA
Double helix of DNA

- The DNA strands are assembled in the 5' to 3' direction
  - by convention, we "read" them the same way.
- The phosphate group bonded to the 5' carbon atom of one deoxyribose is covalently bonded to the 3' carbon of the next.
- The purine or pyrimidine attached to each deoxyribose projects in toward the axis of the helix.
- Each base forms hydrogen bonds with the one directly opposite it, forming base pairs (also called nucleotide pairs).
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 Replication

- Phosphates
- Bases
- Deoxyribose
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

The Histone Code

- State of histone tails govern TF access to DNA
- State is governed by amino acid sequence and modification (acetylation, phosphorylation, methylation)

END of SECTION 5
Section 6: What carries information between DNA to Proteins
Outline For Section 6:

- **Central Dogma Of Biology**
- **RNA**
- **Transcription**
- **Splicing hnRNA-> mRNA**
**Central Dogma**
(DNA \( \rightarrow \) RNA \( \rightarrow \) protein)
The paradigm that DNA directs its transcription to RNA, which is then translated into a protein.

**Transcription**
(DNA \( \rightarrow \) RNA) The process which transfers genetic information from the DNA to the RNA.

**Translation**
(RNA \( \rightarrow \) protein) The process of transforming RNA to protein as specified by the genetic code.
Central Dogma of Biology

The information for making proteins is stored in DNA. There is a process (transcription and translation) by which DNA is converted to protein. By understanding this process and how it is regulated we can make predictions and models of cells.
RNA

- RNA is similar to DNA chemically. It is usually only a single strand. Thymine (T) is replaced by Uracil (U).
- Some forms of RNA can form secondary structures by “pairing up” with itself. This can have change its properties dramatically.

DNA and RNA can pair with each other.

tRNA linear and 3D view: [http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif](http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif)
RNA, continued

• Several types exist, classified by function
• mRNA – this is what is usually being referred to when a Bioinformatician says “RNA”. This is used to carry a gene’s message out of the nucleus.
• tRNA – transfers genetic information from mRNA to an amino acid sequence
• rRNA – ribosomal RNA. Part of the ribosome which is involved in translation.
Terminology for Transcription

- **hnRNA (heterogeneous nuclear RNA)**: Eukaryotic mRNA primary transcripts whose introns have not yet been excised (pre-mRNA).
- **Phosphodiester Bond**: Esterification linkage between a phosphate group and two alcohol groups.
- **Promoter**: A special sequence of nucleotides indicating the starting point for RNA synthesis.
- **RNA (ribonucleotide)**: Nucleotides A, U, G, and C with ribose.
- **RNA Polymerase II**: Multisubunit enzyme that catalyzes the synthesis of an RNA molecule on a DNA template from nucleoside triphosphate precursors.
- **Terminator**: Signal in DNA that halts transcription.
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.
- RNA and DNA are similar, except that RNA is single stranded and thus less stable than DNA.
  - Also, in RNA, the base uracil (U) is used instead of thymine (T), the DNA counterpart.
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 problem that is 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.
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.
**Transcription:** DNA $\rightarrow$ hnRNA

- Transcription occurs in the nucleus.
- $\sigma$ factor from RNA polymerase reads the promoter sequence and opens a small portion of the double helix exposing the DNA bases.
- RNA polymerase II catalyzes the formation of phosphodiester bond that link nucleotides together to form a linear chain from 5’ to 3’ by unwinding the helix just ahead of the active site for polymerization of complementary base pairs.
  - The hydrolysis of high energy bonds of the substrates (nucleoside triphosphates ATP, CTP, GTP, and UTP) provides energy to drive the reaction.
  - During transcription, the DNA helix reforms as RNA forms.
  - When the terminator sequence is met, polymerase halts and releases both the DNA template and the RNA.
Central Dogma Revisited

- **Base Pairing Rule**: A and T or U is held together by 2 hydrogen bonds and G and C is held together by 3 hydrogen bonds.
- **Note**: Some mRNA stays as RNA (ie tRNA, rRNA).
Terminology for Splicing

- **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.
- **Lariat structure**: The structure that an intron in mRNA takes during excision/splicing.
- **Spliceosome**: A organelle that carries out the splicing reactions whereby the pre-mRNA is converted to a mature mRNA.
Splicing
Splicing: hnRNA $\rightarrow$ mRNA

- Takes place on spliceosome that brings together a hnRNA, snRNPs, and a variety of pre-mRNA binding proteins.

- 2 transesterification reactions:
  1. 2’,5’ phosphodiester bond forms between an intron adenosine residue and the intron’s 5’-terminal phosphate group and a lariat structure is formed.
  2. The free 3’-OH group of the 5’ exon displaces the 3’ end of the intron, forming a phosphodiester bond with the 5’ terminal phosphate of the 3’ exon to yield the spliced product. The lariat formed intron is the degraded.

Figure 2
Spliceosome assembly. U1 binds to the 5’ splice site and U2 to the branch site. A preformed U4-U5-U6 complex then joins this assembly to form a complete spliceosome. [After Stryer “Biochemistry” 1995, figure 33-38, pg. 863]
Splicing and other RNA processing

- In Eukaryotic cells, RNA is processed between transcription and translation.
- This complicates the relationship between a DNA gene and the protein it codes for.
- Sometimes alternate RNA processing can lead to an alternate protein as a result. This is true in the immune system.
Splicing (Eukaryotes)

- Unprocessed RNA is composed of Introns and Extrons. 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.
Posttranscriptional Processing: Capping and Poly(A) Tail

Capping

- Prevents 5’ exonucleolytic degradation.
- 3 reactions to cap:
  1. Phosphatase removes 1 phosphate from 5’ end of hnRNA
  2. Guanyl transferase adds a GMP in reverse linkage 5’
  3. Methyl transferase adds methyl group to guanosine.

Poly(A) Tail

- Due to transcription termination process being imprecise.
- 2 reactions to append:
  1. Transcript cleaved 15-25 past highly conserved AAUAAA sequence and less than 50 nucleotides before less conserved U rich or GU rich sequences.
  2. Poly(A) tail generated from ATP by poly(A) polymerase which is activated by cleavage and polyadenylation specificity factor (CPSF) when CPSF recognizes AAUAAA. Once poly(A) tail has grown approximately 10 residues, CPSF disengages from the recognition site.
END of SECTION 6
Section 7: How Are Proteins Made? (Translation)
Outline For Section 7:

- mRNA
- tRNA
- Translation
- Protein Synthesis
- Protein Folding
**Terminology for Ribosome**

- **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.
mRNA → Ribosome

- mRNA leaves the nucleus via nuclear pores.
- Ribosome has 3 binding sites for tRNAs:
  - **A-site**: position that aminoacyl-tRNA molecule binds to vacant site
  - **P-site**: site where the new peptide bond is formed.
  - **E-site**: the exit site
- Two subunits join together on a mRNA molecule near the 5' end.
- The ribosome will read the codons until AUG is reached and then the initiator tRNA binds to the P-site of the ribosome.
- Stop codons have tRNA that recognize a signal to stop translation. Release factors bind to the ribosome which cause the peptidyl transferase to catalyze the addition of water to free the molecule and releases the polypeptide.
Terminology for tRNA and proteins

- **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 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.
Translation: tRNA

- mRNA is translated in 5’ to 3’ direction and the from N-terminal to C-terminus of the polypeptide.
- Elongation process (assuming polypeptide already began):
  - tRNA with the next amino acid in the chain binds to the A-site by forming base pairs with the codon from mRNA
  - Carboxyl end of the protein is released from the tRNA at the Psite and joined to the free amino group from the amino acid attached to the tRNA at the A-site; new peptide bond formed catalyzed by peptide transferase.
  - Conformational changes occur which shift the two tRNAs into the E-site and the P-site from the P-site and A-site respectively. The mRNA also shifts 3 nucleotides over to reveal the next codon.
  - The tRNA in the E-site is released
- GTP hydrolysis provides the energy to drive this reaction.
Terminology for Protein Folding

- **Endoplasmic Reticulum**: Membraneous organelle in eukaryotic cells where lipid synthesis and some posttranslational modification occurs.
- **Mitochondria**: Eukaryotic organelle where citric acid cycle, fatty acid oxidation, and oxidative phosphorylation occur.
- **Molecular chaperone**: Protein that binds to unfolded or misfolded proteins to refold the proteins in the quaternary structure.
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
Revisiting the Central Dogma

- In going from DNA to proteins, there is an intermediate step where mRNA is made from DNA, which then makes protein
  - This known as **The Central Dogma**
- Why the intermediate step?
  - DNA is kept in the nucleus, while protein synthesis happens in the cytoplasm, with the help of ribosomes
The Central Dogma (cont’d)

The Central Dogma of Molecular Biology
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
- But first, in eukaryotes, a phenomenon called splicing occurs
  - Introns are non-protein coding regions of the mRNA; exons are the coding regions
  - Introns are removed from the mRNA during splicing so that a functional, valid protein can form
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.
Translation, continued

- Catalyzed by Ribosome
- Using two different sites, the Ribosome continually binds tRNA, joins the amino acids together and moves to the next location along the mRNA
- ~10 codons/second, but multiple translations can occur simultaneously
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 it’s amino acid arrangement and function
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/amino-acids.html](http://www.indstate.edu/thcme/mwking/amino-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.*
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 tend to fold into the lowest free energy conformation.
- Proteins begin to fold while the peptide is still being translated.
- Proteins bury most of its hydrophobic residues in an interior core to form an α helix.
- Most proteins take the form of secondary structures α helices and β sheets.
- 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

- 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 (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
END of SECTION 7
Section 8: How Can We Analyze DNA?
Outline For Section 8:

• 8.1 Copying DNA
  • Polymerase Chain Reaction
  • Cloning

• 8.2 Cutting and Pasting DNA
  • Restriction Enzymes

• 8.3 Measuring DNA Length
  • Electrophoresis
  • DNA sequencing

• 8.4 Probing DNA
  • DNA probes
  • DNA arrays
Analyzing a Genome

• How to analyze a genome in four easy steps.
  • Cut it
    • Use enzymes to cut the DNA in to small fragments.
  • Copy it
    • Copy it many times to make it easier to see and detect.
  • Read it
    • Use special chemical techniques to read the small fragments.
  • Assemble it
    • Take all the fragments and put them back together. This is hard!!!

• Bioinformatics takes over
  • What can we learn from the sequenced DNA.
  • Compare interspecies and intraspecies.
8.1 Copying DNA
Why we need so many copies

- Biologists needed to find a way to read DNA codes.
- How do you read base pairs that are angstroms in size?
  - It is not possible to directly look at it due to DNA’s small size.
  - Need to use chemical techniques to detect what you are looking for.
  - To read something so small, you need a lot of it, so that you can actually detect the chemistry.
- Need a way to make many copies of the base pairs, and a method for reading the pairs.
Polymerase Chain Reaction (PCR)

- Polymerase Chain Reaction (PCR)
  - Used to massively replicate DNA sequences.

- How it works:
  - Separate the two strands with low heat
  - Add some base pairs, primer sequences, and DNA Polymerase
    - Creates double stranded DNA from a single strand.
    - Primer sequences create a seed from which double stranded DNA grows.
  - Now you have two copies.
  - Repeat. Amount of DNA grows exponentially.
    - 1→2→4→8→16→32→64→128→256…
Polymerase Chain Reaction

- **Problem**: Modern instrumentation cannot easily detect single molecules of DNA, making amplification a prerequisite for further analysis.
- **Solution**: PCR doubles the number of DNA fragments at every iteration.
Denaturation

Raise temperature to 94°C to separate the duplex form of DNA into single strands.
Design primers

- To perform PCR, a 10-20bp sequence on either side of the sequence to be amplified must be known because DNA pol requires a primer to synthesize a new strand of DNA.
Annealing

- Anneal primers at 50-65°C
Annealing

• Anneal primers at 50-65°C
Extension

- Extend primers: raise temp to 72°C, allowing Taq pol to attach at each priming site and extend a new DNA strand
Extension

• Extend primers: raise temp to 72° C, allowing Taq pol to attach at each priming site and extend a new DNA strand
Repeat

• Repeat the Denature, Anneal, Extension steps at their respective temperatures…
Polymerase Chain Reaction
Cloning DNA

- DNA Cloning
  - Insert the fragment into the genome of a living organism and watch it multiply.
  - Once you have enough, remove the organism, keep the DNA.
- Use Polymerase Chain Reaction (PCR)
8.2 Cutting and Pasting DNA
Restriction Enzymes

- Discovered in the early 1970’s
  - Used as a defense mechanism by bacteria to break down the DNA of attacking viruses.
  - They cut the DNA into small fragments.
- Can also be used to cut the DNA of organisms.
  - This allows the DNA sequence to be in a more manageable bite-size pieces.
- It is then possible using standard purification techniques to single out certain fragments and duplicate them to macroscopic quantities.
Cutting DNA

- **Restriction Enzymes cut DNA**
  - Only cut at special sequences
- DNA contains thousands of these sites.
- **Applying different Restriction Enzymes creates fragments of varying size.**

![Restriction Enzyme “A” Cutting Sites](image)

![Restriction Enzyme “B” Cutting Sites](image)

“A” and “B” fragments overlap

![Restriction Enzyme “A” & Restriction Enzyme “B” Cutting Sites](image)

**KEY**
- Red = H-bond acceptor
- Blue = H-bond donor
- White = hydrogen atom
- Orange = methyl group

![Eco RI](image)

![Bal I](image)

![Sma I](image)
Pasting DNA

- Two pieces of DNA can be fused together by adding chemical bonds
  - Hybridization – complementary base-pairing
  - Ligation – fixing bonds with single strands
8.3 Measuring DNA Length
Electrophoresis

- A copolymer of mannose and galactose, agaraose, when melted and recooled, forms a gel with pores sizes dependent upon the concentration of agarose.

- The phosphate backbone of DNA is highly negatively charged, therefore DNA will migrate in an electric field.
  - The size of DNA fragments can then be determined by comparing their migration in the gel to known size standards.
Reading DNA

- **Electrophoresis**
  - Reading is done mostly by using this technique. This is based on separation of molecules by their size (and in 2D gel by size and charge).
  - DNA or RNA molecules are charged in aqueous solution and move to a definite direction by the action of an electric field.
  - The DNA molecules are either labeled with radioisotopes or tagged with fluorescent dyes. In the latter, a laser beam can trace the dyes and send information to a computer.
  - Given a DNA molecule it is then possible to obtain all fragments from it that end in either A, or T, or G, or C and these can be sorted in a gel experiment.
  - Another route to sequencing is direct sequencing using gene chips.
Assembling Genomes

- Must take the fragments and put them back together
  - Not as easy as it sounds.
- SCS Problem (Shortest Common Superstring)
  - Some of the fragments will overlap
    - Fit overlapping sequences together to get the shortest possible sequence that includes all fragment sequences
Assembling Genomes

- DNA fragments contain sequencing errors
- Two complements of DNA
  - Need to take into account both directions of DNA
- Repeat problem
  - 50% of human DNA is just repeats
  - If you have repeating DNA, how do you know where it goes?
8.4 Probing DNA

Che Fung Yung
May 12, 2004
DNA probes

- Oligonucleotides: single-stranded DNA 20-30 nucleotides long
- Oligonucleotides used to find complementary DNA segments.
- Made by working backwards---AA sequence----mRNA---cDNA.
- Made with automated DNA synthesizers and tagged with a radioactive isotope.
DNA Hybridization

- Single-stranded DNA will naturally bind to complementary strands.

- Hybridization is used to locate genes, regulate gene expression, and determine the degree of similarity between DNA from different sources.

- Hybridization is also referred to as annealing or renaturation.
Create a Hybridization Reaction

1. Hybridization is binding two genetic sequences. The binding occurs because of the hydrogen bonds [pink] between base pairs.

2. When using hybridization, DNA must first be denatured, usually by using heat or chemical.
3. Once DNA has been denatured, a single-stranded radioactive probe [light blue] can be used to see if the denatured DNA contains a sequence complementary to probe.

4. Sequences of varying homology stick to the DNA even if the fit is poor.
Labeling technique for DNA arrays

RNA samples are labeled using fluorescent nucleotides (left) or radioactive nucleotides (right), and hybridized to arrays. For fluorescent labeling, two or more samples labeled with differently colored fluorescent markers are hybridized to an array. Level of RNA for each gene in the sample is measured as intensity of fluorescence or radioactivity binding to the specific spot. With fluorescence labeling, relative levels of expressed genes in two samples can be directly compared with a single array.
DNA Arrays--Technical Foundations

- An array works by exploiting the ability of a given mRNA molecule to hybridize to the DNA template.

- Using an array containing many DNA samples in an experiment, the expression levels of hundreds or thousands genes within a cell by measuring the amount of mRNA bound to each site on the array.

- With the aid of a computer, the amount of mRNA bound to the spots on the microarray is precisely measured, generating a profile of gene expression in the cell.
An experiment on a microarray

In this schematic:

**GREEN** represents **Control DNA**

**RED** represents **Sample DNA**

**YELLOW** represents a combination of **Control and Sample DNA**

**BLACK** represents areas where **neither the Control nor Sample DNA**

Each color in an array represents either healthy (control) or diseased (sample) tissue. The location and intensity of a color tell us whether the gene, or mutation, is present in the control and/or sample DNA.
DNA Microarray

Millions of DNA strands build up on each location.

Tagged probes become hybridized to the DNA chip’s microarray.

http://www.affymetrix.com/corporate/media/image_library/image_library_1.affx
DNA Microarray

Microarray is a tool for analyzing gene expression that consists of a glass slide.

Each blue spot indicates the location of a PCR product. On a real microarray, each spot is about 100um in diameter.
Photolithography

- Light directed oligonucleotide synthesis.
- A solid support is derivatized with a covalent linker molecule terminated with a photolabile protecting group.
- Light is directed through a mask to deprotect and activate selected sites, and protected nucleotides couple to the activated sites.
- The process is repeated, activating different set of sites and coupling different based allowing arbitrary DNA probes to be constructed at each site.
**Affymetrix GeneChip® Arrays**

A combination of photolithography and combinatorial chemistry to manufacture GeneChip® Arrays. With a minimum number of steps, Affymetrix produces arrays with thousands of different probes packed at extremely high density.Enable to obtain high quality, genome-wide data using small sample volumes.

May 11, 2004

http://www.affymetrix.com/technology/manufacturing/index.affx
Affymetrix GeneChip® Arrays

Data from an experiment showing the expression of thousands of genes on a single GeneChip® probe array.
END of SECTION 8
Section 9: How Do Individuals of a Species Differ?
Outline For Section 9:

- *Physical Variation and Diversity*
- *Genetic Variation*
How Do Individuals of Species Differ?

- Genetic makeup of an individual is manifested in traits, which are caused by variations in genes.

- While 0.1% of the 3 billion nucleotides in the human genome are the same, small variations can have a large range of phenotypic expressions.

- These traits make some more or less susceptible to disease, and the demystification of these mutations will hopefully reveal the truth behind several genetic diseases.
The Diversity of Life

- Not only do different species have different genomes, but also different individuals of the same species have different genomes.
- No two individuals of a species are quite the same – this is clear in humans but is also true in every other sexually reproducing species.
- Imagine the difficulty of biologists – sequencing and studying only one genome is not enough because every individual is genetically different!
Physical Traits and Variances

- Individual variation among a species occurs in populations of all sexually reproducing organisms.
- Individual variations range from hair and eye color to less subtle traits such as susceptibility to malaria.
- Physical variation is the reason we can pick out our friends in a crowd, however most physical traits and variation can only be seen at a cellular and molecular level.
Sources of Physical Variation

- Physical Variation and the manifestation of traits are caused by variations in the genes and differences in environmental influences.
- An example is height, which is dependent on genes as well as the nutrition of the individual.
- Not all variation is inheritable – only genetic variation can be passed to offspring.
- Biologists usually focus on genetic variation instead of physical variation because it is a better representation of the species.
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.
- Although there is a finite number of possible variations, the number is so high ($4^{3,000,000}$) that we can assume no two individual people have the same genome.
- What is the cause of this genetic variation?
Sources of Genetic Variation

- **Mutations** are rare errors in the DNA replication process that occur at random.
- When mutations occur, they affect the genetic sequence and create genetic variation between individuals.
- Most mutations do not create beneficial changes and actually kill the individual.
- Although mutations are the source of all new genes in a population, they are so rare that there must be another process at work to account for the large amount of diversity.
Sources of Genetic Variation

- **Recombination** is the shuffling of genes that occurs through sexual mating and is the main source of genetic variation.

- Recombination occurs via a process called **crossing over** in which genes switch positions with other genes during meiosis.

- Recombination means that new generations inherit random combinations of genes from both parents.

- The recombination of genes creates a seemingly endless supply of genetic variation within a species.
How Genetic Variation is Preserved

- **Diploid** organisms (which are most complex organisms) have two genes that code for one physical trait – which means that sometimes genes can be passed down to the next generation even if a parent does not physically express the gene.

- **Balanced Polymorphism** is the ability of natural selection to preserve genetic variation. For example, natural selection in one species of finch keeps beak sizes either large or small because a finch with a hybrid medium sized beak cannot survive.
Variation as a Source of Evolution

- Evolution is based on the idea that variation between individuals causes certain traits to be reproduced in future generations more than others through the process of Natural Selection.
- **Genetic Drift** is the idea that the prevalence of certain genes changes over time.
- If enough genes are changed through mutations or otherwise so that the new population cannot successfully mate with the original population, then a new species has been created.
- Do all variations affect the evolution of a species?
Neutral Variations

• Some variations are clearly beneficial to a species while others seem to make no visible difference.

• **Neutral Variations** are those variations that do not appear to affect reproduction, such as human fingerprints. Many such neutral variations appear to be molecular and cellular.

• However, it is unclear whether neutral variations have an effect on evolution because their effects are difficult, if not impossible to measure. There is no consensus among scientists as to how much variation is neutral or if variations can be considered neutral at all.
The Genome of a Species

- It is important to distinguish between the genome of a species and the genome of an individual.
- The genome of a species is a representation of all possible genomes that an individual might have since the basic sequence in all individuals is more or less the same.
- The genome of an individual is simply a specific instance of the genome of a species.
- Both types of genomes are important – we need the genome of a species to study a species as a whole, but we also need individual genomes to study genetic variation.
Human Diversity Project

- The Human Diversity Project samples the genomes of different human populations and ethnicities to try and understand how the human genome varies.
- It is highly controversial both politically and scientifically because it involves genetic sampling of different human races.
- The goal is to figure out differences between individuals so that genetic diseases can be better understood and hopefully cured.
END of SECTION 9
Section 10: How Do Different Species Differ?
Outline For Section 10:

• **Section 10.1 – Molecular Evolution**
  • What is Evolution
  • Molecular Clock
  • New Genes

• **Section 10.2 – Comparative Genomics**
  • Human and Mouse
  • Comparative Genomics
  • Gene Mapping
  • Cystic Fibrosis

• **Section 10.3 – Genome Rearrangements**
  • Gene Order
  • DNA Reversal
Section 10.1 The Biological Aspects of Molecular Evolution
What is evolution?

- A process of change in a certain direction (*Merriam – Webster Online*).
- **In Biology**: The process of biological and organic change in organisms by which descendants come to differ from their ancestor (*Mc GRAW – HILL Dictionary of Biological Science*).
- **Charles Darwin** first developed the Evolution idea in detail in his well-known book *On the Origin of Species* published in 1859.
Some Conventional Tools For Evolutionary Studies

- **Fossil Record**: some of the biota found in a given stratum are the descendants of those in the previous stratum.
- **Morphological Similarity**: similar species are found to have some similar anatomical structure; For example: horses, donkeys and zebras.
- **Embryology**: embryos of related kinds of animals are astoundingly similar.
Molecular Clock

- Introduced by Linus Pauling and his collaborator Emile Zuckerkandl in 1965.
- They proposed that the rate of evolution in a given protein (or later, DNA) molecule is approximately constant overtime and among evolutionary lineages.
• Observing hemoglobin patterns of some primates, They found:
  - The gorilla, chimpanzee and human patterns are almost identical.
  - The further one gets away from the group of Primates, the primary structure that is shared with human hemoglobin decreases.
  - α and β chains of human hemoglobin are homologous, having a common ancestor.

Human Hemoglobin, A 2-α and 2-β tetramer.
• Linus and Pauling found that $\alpha$-chains of human and gorilla differ by 2 residues, and $\beta$-chains by 1 residues.
• They then calculated the time of divergence between human and gorilla using evolutionary molecular clock.
• Gorilla and human $\beta$ chain were found to diverge about 7.3 years ago.
Molecular Evolution

- Pauling and Zuckerkandl research was one of the pioneering works in the emerging field of *Molecular Evolution*.
- *Molecular Evolution* is the study of evolution at molecular level, genes, proteins or the whole genomes.
- Researchers have discovered that as somatic structures evolves (*Morphological Evolution*), so does the genes. But the *Molecular Evolution* has its special characteristics.
Molecular Evolution Cont.

- Genes and their proteins products evolve at different rates.
  
  For example, histones changes very slowly while fibrinopeptides very rapidly, revealing function conservation.

- Unlike physical traits which can evolved drastically, genes functions set severe limits on the amount of changes.

  Thought Humans and Chimpanzees lineages separated at least 6 million years ago, many genes of the two species highly resemble one another.
Beta globins:

- Beta globin chains of closely related species are highly similar:
- Observe simple alignments below:

  Human β chain: MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLL
  Mouse β chain: MVHLTDAEKAAVNGLWGKVNPDDVGGEALGRLL

  Human β chain: VVYPWTQRFESFGDSLSTPDAVMGNPVKAHGGKVVLG
  Mouse β chain: VVYPWTQRYFDSEXGDSASAIMGNPKVKAHGKVGIN

  Human β chain: AFSDGLAHLDNLKGTFAHTLSLRELHCDKLHVDPENFRLLG
  Mouse β chain: AFNDGLKHLDNLKGTFAHLSLRELHCDKLHVDPENFRLLG

  Human β chain: VLVCLAHFFGKKEFTPVPQAYQKVAGVANALAHKYH
  Mouse β chain: MIVLVGLHHLGKKEFTPACAQAFQKVAGVASALAHKYH

There are a total of 27 mismatches, or \((147 - 27) / 147 = 81.7 \% \) identical
Beta globins: Cont.

Human β chain: **MVH L TPEEKSAVT ALWGKVNV DEVGGEALGRLL**
Chicken β chain: **MVHWT AEEKQL I TGLWGKVNV AECAEALARLL**

Human β chain: **VVYPWTQRFF ESFGDLSTPDAVMGNPKVKAHGKKVLG**
Chicken β chain: **IVYPWTQRFF ASFGNLSSPTA I LGNPVMRAHGKKVLT**

Human β chain: **AFSDGLAH LDNLKGTFATLSELHCDKLHVDPENFRLLGN**
Chicken β chain: **SFDAVK NLNDI NTSQ LSELHCDKLHVDPENFRLLGD**

Human β chain: **VLVC VLAH HFGKEFTPPV QAA Y QKVAGVANALAHKYH**
Mouse β chain: **IL LI VLAHFSK DFTEPCQA W QKLVRVVAH A RKYH**

- There are a total of **44** mismatches, or \( (147 - 44) / 147 = 70.1\% \) identical
- As expected, mouse β chain is ‘closer’ to that of human than chicken’s.
Molecular evolution can be visualized with phylogenetic tree.
Origins of New Genes.

- All animals lineages traced back to a common ancestor, a protist about 700 million years ago.
Section 10.2: Comparative Genomics
How Do Different Species Differ?

- As many as 99% of human genes are conserved across all mammals
- The functionality of many genes is virtually the same among many organisms
- It is highly unlikely that the same gene with the same function would spontaneously develop among all currently living species
- The theory of evolution suggests all living things evolved from incremental change over millions of years
Mouse and Human overview

- Mouse has $2.1 \times 10^9$ base pairs versus $2.9 \times 10^9$ in human.
- About 95% of genetic material is shared.
- 99% of genes shared of about 30,000 total.
- The 300 genes that have no homologue in either species deal largely with immunity, detoxification, smell and sex*

*Scientific American Dec. 5, 2002
Human and Mouse

Significant chromosomal rearranging occurred between the diverging point of humans and mice.

Here is a mapping of human chromosome 3.

It contains homologous sequences to at least 5 mouse chromosomes.
Comparative Genomics

• What can be done with the full Human and Mouse Genome? One possibility is to create “knockout” mice – mice lacking one or more genes. Studying the phenotypes of these mice gives predictions about the function of that gene in both mice and humans.
Comparative Genomics

- By looking at the expression profiles of human and mouse (a recent technique using Gene Chips to detect mRNA as genes are being transcribed), the phenotypic differences can be attributed to genes and their expression.

A gene chip made by Affymetrix. The well can contain probes for thousands of genes.

Imaging of a chip. The amount of fluorescence corresponds to the amount of a gene expressed.
Comparative Genome Sizes

- The genome of a protist Plasmodium falciparum, which causes malaria, is 23 Mb long.
- Human genome is approximately 150 times larger, mouse > 100 times, and fruit fly > 5 times larger.
- Question: How genomes of old ancestors get bigger during evolution?
Mechanisms:

- Gene duplications or insertions
Comparative Genomics

• Knowing the full sequence of human and mouse genomes also gives information about gene regulation. Because the promoter regions tend to remain conserved through evolution, looking for similar DNA upstream of a known gene can help identify regulatory sites. This technique gets more powerful the more genomes can be compared.
Gene Mapping

- Mapping human genes is critically important
  - Insight into the evolutionary relationship of human to other vertebrate species
  - Mapping disease gene create an opportunity for researchers to isolate the gene and understand how it causes a disease.

Genomics: the sub discipline of genetics devoted to the mapping, sequencing, and functional analysis of genomes
Gene Mapping

- The procedure for mapping chromosomes was invented by Alfred H. Sturterant.
  - Analysis of experiment data from Drosophila
- Experimental data demonstrated that genes on the same chromosome could be separated as they went through meiosis and new **combination** of genes is formed.
- Genes that are tightly linked seldom recombine, whereas genes that are loosely linked recombine
Gene Mapping

- Genetic maps of chromosomes are based on recombination frequencies between markers.
- Cytogenetic maps are based on the location of markers within cytological features such as chromosome banding patterns observed by microscope.
- Physical maps of chromosomes are determined by the molecular distances in base pairs, kilobase pairs, or mega base pairs separating markers.
- High-density maps that integrate the genetic, cytological and physical maps of chromosomes have been constructed for all of human chromosomes and for many other organisms.
Gene Mapping

• Recombinant DNA techniques have revolutionized the search for defective genes that cause human disease.
• Numerous major “disease genes” have already been identified by positional cloning.
  • Huntington’s disease (HD gene)
  • Cystic fibrosis (CF gene)
  • Cancer
Cystic fibrosis

- Symptoms:
  - excessively salty sweat
  - The lungs, pancreas, and liver become clogged with thick mucus, which results in chronic infections and eventual malfunction
  - Mucus often builds up in the digestive tract, causing malnourishment
  - Patients often die from infections of the respiratory system.
Cystic Fibrosis

- In 1989, Francis Collins and Lap-Chee Tsui
  - identified the CF gene
  - characterized some of the mutation that cause this disease.
- A cDNA (complimentary DNA) library was prepared from mRNA isolated from sweat gland cells growing in culture and screened by colony hybridization
- CF gene product is similar to several ion channels protein,
  - which form pores between cells through which ions pass.
- Mutant CFTR protein does not function properly
  - salt accumulates in epithelial cells and mucus builds up on the surfaces of the cells.
Cystic Fibrosis

- Chromosome walking and jumping and complementary DNA hybridization were used to isolate DNA sequences, encompassing more than 500,000 base pairs, from the cystic fibrosis region on the long arm of human chromosome 7.
- neither gene therapy nor any other kind of treatment exists
- doctors can only ease the symptoms of CF
  1. antibiotic therapy combined with treatments to clear the thick mucus from the lungs.
  2. For patients whose disease is very advanced, lung transplantation may be an option.
Waardenburg’s syndrome

- Genetic disorder
- Characterized by loss of hearing and pigmentary dysphasia
- Found on human chromosome 2
Waardenburg’s syndrome

- A certain breed of mice (with splotch gene) that had similar symptoms caused by the same type of gene in humans
- Mice and Human genomes very similar → but easier to study mice
- Finding the gene in mice gives clues to where the same gene is located in humans
- Succeeded in identifying location of gene responsible for disorder in mice
Waardenburg’s syndrome

• To locate where corresponding gene is in humans, we have to analyze the relative architecture of genes of humans and mouse

• About 245 genomic rearrangements

• Rearrangement operation in this case: reversals, translocation, fusion, and fission

• Reversal is where a block of genes is flipped within a genomic sequence
Section 10.3 Genome Rearrangements.
Turnip and Cabbage

- Cabbages and turnips share a common ancestor
Jeffrey Palmer – 1980s

- discovered evolutionary change in plant organelles by comparing mitochondrial genomes of the cabbage and turnip
- 99% similarity between genes
- These more or less identical gene sequence surprisingly differed in gene order
- This finding helped pave the way to prove that genome rearrangements occur in molecular evolution in mitochondrial DNA
Important discovery

Comparing gene sequences yields no evolutionary information

Evolution is manifested as the divergence in Gene Order
DNA Reversal

Break and Invert

5’ ATGCCTGTATACTA 3’
3’ TACGGGACATGAT 5’

5’ ATGTCAGGCCCTA 3’
3’ TACATGTCCTCGAT 5’
Bioinformatics

Sequence Driven Problems

• Genomics
  • Fragment assembly of the DNA sequence.
    • Not possible to read entire sequence.
    • Cut up into small fragments using restriction enzymes.
    • Then need to do fragment assembly. Overlapping similarities to matching fragments.
    • N-P complete problem.

• Finding Genes
  • Identify open reading frames
    • Exons are spliced out.
    • Junk in between genes
Bioinformatics

Sequence Driven Problems

• Proteomics
  • Identification of functional domains in protein’s sequence
    • Determining functional pieces in proteins.
  • Protein Folding
    • 1D Sequence $\rightarrow$ 3D Structure
    • What drives this process?
DNA... Then what?

- **DNA** → transcription → **RNA** → translation → **Protein**
- Ribonucleic Acid (RNA)
  - It is the messenger
  - a temporary copy
  - Why not DNA → Protein.
    - DNA is in nucleus and proteins are manufactured out of the nucleus
    - Adds a proofreading step. (Transcription = DNA→RNA)
- So actually… DNA → pre-mRNA → mRNA → Protein
  - Prokaryotes
    - The gene is continuous. Easy to translate.
  - Eukaryotes
    - Introns and Exons
    - Several Exons in different locations need to be spliced together to make a protein. (Splicing)
    - Pre-mRNA (unspliced RNA)
    - Splicisome cuts the introns out of it making processed mRNA.
Proteins

- Carry out the cell's chemistry
  - 20 amino acids

- A more complex polymer than DNA
  - Sequence of 100 has $20^{100}$ combinations
  - Sequence analysis is difficult because of complexity issue
  - Only a small number of the possible sequences are actually used in life. (Strong argument for Evolution)

- RNA Translated to Protein, then Folded
  - Sequence to 3D structure (Protein Folding Problem)
  - Translation occurs on Ribosomes
  - 3 letters of DNA → 1 amino acid
    - 64 possible combinations map to 20 amino acids
    - Degeneracy of the genetic code
      - Several codons to same protein
Radiodurans

- Survives Larger Radiation Doses
  - Survives by orders of magnitudes more than other organisms
- DNA is cut by radiation
  - Radiodurans can reconstruct its DNA after being cut.
  - Basically, fragment assembly in vivo.
- We cut it too, but is hard for us
  - How did they do that???
END of SECTION 10
Section 11: Why Bioinformatics?

Julio Ng, Robert Hinman
CSE 181 Projects 2,3
April 20, 2004
Outline For Section 11:

- **Sequence Driven Problems**
- **Human and Mouse**
- **Comparative Genomics**
- **Gene Mapping**
- **Cystic Fibrosis**
Why Bioinformatics?

• Bioinformatics is the combination of biology and computing.

• DNA sequencing technologies have created massive amounts of information that can only be efficiently analyzed with computers.

• So far 70 species sequenced
  • Human, rat chimpanzee, chicken, and many others.

• As the information becomes ever so larger and more complex, more computational tools are needed to sort through the data.
  • Bioinformatics to the rescue!!!
What is Bioinformatics?

- Bioinformatics is generally defined as the analysis, prediction, and modeling of biological data with the help of computers.
Bio-Information

• Since discovering how DNA acts as the instructional blueprints behind life, biology has become an information science.

• Now that many different organisms have been sequenced, we are able to find meaning in DNA through *comparative genomics*, not unlike comparative linguistics.

• Slowly, we are learning the syntax of DNA.
Sequence Information

• Many written languages consist of sequential symbols
• Just like human text, genomic sequences represent a language written in A, T, C, G
• Many DNA decoding techniques are not very different than those for decoding an ancient language
The *Rosetta Stone*

- The Rosetta Stone allowed linguists to solve the code of Egyptian Hieroglyphics.
- The Greek language inscribed gave clues to what the Hieroglyphs meant.
- This is an example of *comparative linguistics*.
Linear B

- At the beginning of the twentieth century, archeologists discovered clay tablets on the island of Crete
- This unknown language was named “Linear B”
- It was thought to write in an ancient Minoan Language, and was a mystery for 50 years
Linear B

- The same time the structure of DNA is deciphered, Michael Ventris solves Linear B using mathematical code breaking skills.
- He notes that some words in Linear B are specific for the island, and theorizes those are names of cities.
- With this bit of knowledge, he is able to decode the script, which turns out to be Greek with a different alphabet.
Amino Acid Crack

- Even earlier, an experiment in the early 1900s showed that all proteins are composed of sequences of 20 amino acids.
- This led some to speculate that polypeptides held the blueprints of life.
Central Dogma

- DNA → mRNA → Proteins

- DNA in chromosome is transcribed to mRNA, which is exported out of the nucleus to the cytoplasm. There it is translated into protein.

- Later discoveries show that we can also go from mRNA to DNA (retroviruses).

- Also mRNA can go through alternative splicing that lead to different protein products.
Structure to Function

• Organic chemistry shows us that the structure of the molecules determines their possible reactions.

• One approach to study proteins is to infer their function based on their structure, especially for active sites.
Two Quick Bioinformatics Applications

• BLAST (Basic Local Alignment Search Tool)
• PROSITE (Protein Sites and Patterns Database)
BLAST

• A computational tool that allows us to compare query sequences with entries in current biological databases.

• A great tool for predicting functions of a unknown sequence based on alignment similarities to known genes.
Your request has been successfully submitted and put into the Blast Queue.

Query = (183 letters)

The request ID is 1082998002-15402-91580503850.BLASTQ3

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.
Some Early Roles of Bioinformatics

- Sequence comparison
- Searches in sequence databases
Biological Sequence Comparison

- **Needleman-Wunsch, 1970**
  - Dynamic programming algorithm to align sequences
Early Sequence Matching

- Finding locations of restriction sites of known restriction enzymes within a DNA sequence (very trivial application)
- Alignment of protein sequence with scoring motif
- Generating contiguous sequences from short DNA fragments.
  - This technique was used together with PCR and automated HT sequencing to create the enormous amount of sequence data we have today
Biological Databases

- Vast biological and sequence data is freely available through online databases
- Use computational algorithms to efficiently store large amounts of biological data

Examples

  Huge collection of databases, the most prominent being the nucleotide sequence database
- **Protein Data Bank** [http://www.pdb.org](http://www.pdb.org)
  Database of protein tertiary structures
  Database of annotated protein sequences
  Database of protein active site motifs
PROSITE Database

- Database of protein active sites.
- A great tool for predicting the existence of active sites in an unknown protein based on primary sequence.
PROSITE

PROSITE
Database of protein families and domains

Browse PROSITE documentation entries
Release 18.26, of 26-Apr-2004

- The character in the first column is used to indicate if a documentation entry is new in this release ('+'), or has been modified ('*') since the last major release (release 18.0 of July 2002).
- The numerical characters in positions 3 to 7 provide the documentation entry accession number.
- The numerical character in position 9 is used to indicate how many data entries (patterns, rules and profiles/matrices) are described by a documentation entry.

Example:
* PDOC00020 2 Kringle domain signature and profile

This documentation entry has been updated since the last release ('*'), its accession number is PDOC00020 and it describes the patterns.
Sequence Analysis

• Some algorithms analyze biological sequences for patterns
  • RNA splice sites
  • ORFs
  • Amino acid propensities in a protein
  • Conserved regions in
    • AA sequences [possible active site]
    • DNA/RNA [possible protein binding site]

• Others make predictions based on sequence
  • Protein/RNA secondary structure folding
It is Sequenced, What’s Next?

- Tracing Phylogeny
  - Finding family relationships between species by tracking similarities between species.
- Gene Annotation (cooperative genomics)
  - Comparison of similar species.
- Determining Regulatory Networks
  - The variables that determine how the body reacts to certain stimuli.
- Proteomics
  - From DNA sequence to a folded protein.
Modeling

• Modeling biological processes tells us if we understand a given process
• Because of the large number of variables that exist in biological problems, powerful computers are needed to analyze certain biological questions
Protein Modeling

• Quantum chemistry imaging algorithms of active sites allow us to view possible bonding and reaction mechanisms
• Homologous protein modeling is a comparative proteomic approach to determining an unknown protein’s tertiary structure
• Predictive tertiary folding algorithms are a long way off, but we can predict secondary structure with ~80% accuracy.
  The most accurate online prediction tools:
    PSIPred
    PHD
Regulatory Network Modeling

• Micro array experiments allow us to compare differences in expression for two different states
• Algorithms for clustering groups of gene expression help point out possible regulatory networks
• Other algorithms perform statistical analysis to improve signal to noise contrast
Systems Biology Modeling

• Predictions of whole cell interactions.
  • Organelle processes, expression modeling

• Currently feasible for specific processes (eg. Metabolism in E. coli, simple cells)
  Flux Balance Analysis
The future…

• Bioinformatics is still in it’s infancy
• Much is still to be learned about how proteins can manipulate a sequence of base pairs in such a peculiar way that results in a fully functional organism.
• How can we then use this information to benefit humanity without abusing it?
Sources Cited

- Daniel Sam, “Greedy Algorithm” presentation.
- Glenn Tesler, “Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes” presentation.
- Ernst Mayr, “What evolution is”.