Molekularbiologische Datenbanken
Grundlagen der Molekularbiologie

Ulf Leser
Wissensmanagement in der Bioinformatik
50th anniversary

- 25.4.1953
  - James Watson, Francis Crick
  - "Molecular Structure of Nucleic Acids"
  - Nature, 1 page
  - "This structure has two helical chains each coiled around the same axis"
- Based on work by Wilkins & Franklin
Cells
Heritable Information

- Cell nucleus
- Chromosomes
- DNA
- The blueprint of life
**DesoxyriboNucleic Acid**

- Chromosome: string of DNA
- Only 4 different nucleic acids
- Fixed pairs: A-T, G-C
- Same mechanism in all species
Structure of DNA

Quelle: http://www.nhgri.nih.gov
DNA Strands

• DNA single or double stranded
• DNA has fixed orientation
  - read from 5’ to 3’
  - Strands are antiparallel

5’ ... TACTGAA ... 3’
3’ ... ATGACTT ... 5’
DNA Replication

- **Mitosis:** Cell replication
- **Meiosis:** Species replication

![Diagram of DNA replication process with labels for proteins involved in the replication fork.](image)
Species

Tree of Life – the 3 Kingdoms

- Prokaryotes: cell nucleus, differential splicing, gene regulation, complexity
- Model organism
Genome

• Set of all genes of a species is its genome

• Humans
  - App. 3.300.000.000 bp
  - 22 chromosomes + 2 sex chromosomes
  - Length 50–250 MB
Human Genome Project

- DNA: ~ 3.300.000.000 base pairs
- Human Genome Project: Sequencing the complete Human Genome (scheduled 2005)
- World-wide Effort (Germany: since 1996)
- “Finished” in 2000
  - Rapid improvement in technology
  - Commercial companies jumped on it (Celera)
- “Almost” finished in 2001
- “Really” finished in 2003
Sequencing a Genome

- Break it up into very small pieces
- Read the sequence one-by-one
- Assemble the pieces to contigs
What does DNA tell us?

- DNA is just a string: [ACTG]*
- In humans, most DNA is considered junk
- All DNA is replicated, but ...
- ... only DNA that is transformed into „something“ matters
Central Dogma

The Central Dogma of Molecular Genetics
Gene / protein functions

- **Structure**
  - cell wall, membranes, organelles, ...

- **Signal transduction**
  - Signal recognition (outside cell), transduction, and reaction

- **Enzymatic catalysis**
  - Support for chemical reactions: respiration, metabolism, protein construction/deconstruction, ...

- **Transport**
  - mRNA from nucleus to ribosomes, proteins from ribosomes to cell well, ...
Pathways

- Metabolic pathways
- Regulatory pathways
- Signal transduction pathways
DNA and Genes

- Genes are coding regions: DNA -> RNA ( -> Protein)
- Subject to evolutionary principles: mutation, selection
- Mutation might change phenotype - genetic disease
- Complex regulation mechanisms
No. of genes doesn` t mean ...

<table>
<thead>
<tr>
<th>Species</th>
<th>No of genes</th>
<th>Genome size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>30.000</td>
<td>3300 MB</td>
</tr>
<tr>
<td>Fly</td>
<td>13.000</td>
<td>180 MB</td>
</tr>
<tr>
<td>Wheat</td>
<td>16000 MB</td>
<td></td>
</tr>
<tr>
<td>C. elegans</td>
<td>18.000</td>
<td>100 MB</td>
</tr>
<tr>
<td>Mouse</td>
<td>30.000</td>
<td>3000 MB</td>
</tr>
</tbody>
</table>

- Genomic rearrangements & duplications
- Inactivation
- Gene transfer
Genotype - Phenotype

- Genes are different between individuals
  - Alleles
- Genotype: set of genes of an individual
- Phenotype: appearance of an individual
  - Visual features, behavioral features, susceptibility to diseases, ...
- Genotype determines phenotype
  - To what extend?
From Genes to Proteins

- Genes are transcribed into mRNA
- Post-Processing (splicing)
- mRNA is translated into proteins
- Post-processing
- Code not universal

<table>
<thead>
<tr>
<th>First letter</th>
<th>Second letter</th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>Phenylyalanine</td>
<td>UUU</td>
<td>UUC</td>
<td>UAU</td>
<td>UUG</td>
</tr>
<tr>
<td></td>
<td>Leucine</td>
<td>UUA</td>
<td>UAC</td>
<td>UAA</td>
<td>UAG</td>
</tr>
<tr>
<td>C</td>
<td>Leucine</td>
<td>CCU</td>
<td>CCC</td>
<td>CCA</td>
<td>CGG</td>
</tr>
<tr>
<td></td>
<td>Proline</td>
<td>CAU</td>
<td>CAC</td>
<td>CAA</td>
<td>CAG</td>
</tr>
<tr>
<td>A</td>
<td>Isoleucine</td>
<td>AUU</td>
<td>AUC</td>
<td>AUG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methionine; Initiation codon</td>
<td>AUG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Valine</td>
<td>GUU</td>
<td>GUC</td>
<td>GUA</td>
<td>GUG</td>
</tr>
<tr>
<td></td>
<td>Alanine</td>
<td>GCU</td>
<td>GCC</td>
<td>GCA</td>
<td>GCG</td>
</tr>
<tr>
<td></td>
<td>Aspartic acid</td>
<td>GAU</td>
<td>GAC</td>
<td>GAA</td>
<td>GAG</td>
</tr>
<tr>
<td></td>
<td>Glutamic acid</td>
<td>GGU</td>
<td>GGC</td>
<td>GGA</td>
<td>GGG</td>
</tr>
<tr>
<td></td>
<td>Tyrosine</td>
<td>UAU</td>
<td>UAC</td>
<td>UAA</td>
<td>UAG</td>
</tr>
<tr>
<td></td>
<td>Stop codon</td>
<td>UGU</td>
<td>UGC</td>
<td>UGA</td>
<td>UGG</td>
</tr>
<tr>
<td></td>
<td>Histidine</td>
<td>CGU</td>
<td>CGC</td>
<td>CGA</td>
<td>CCG</td>
</tr>
<tr>
<td></td>
<td>Glutamine</td>
<td>CAA</td>
<td>CAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Asparagine</td>
<td>AAA</td>
<td>AAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lysine</td>
<td>AGA</td>
<td>AGG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Serine</td>
<td>AUA</td>
<td>AAC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Arginine</td>
<td>AAG</td>
<td>ACG</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glycine</td>
<td>GGU</td>
<td>GGC</td>
<td>GGA</td>
<td>GGG</td>
</tr>
</tbody>
</table>
Alternative Splicing

One gene – one protein is mostly wrong
Human Genome

- App. 30,000 genes
- Length from 100bp – 2MB (introns+exons)
- Length avg. 1400 bps (only exons)
  - Avg. protein length 447 amino acids
- Avg gene has 9 exons
- Only 3% of human genome is coding
  - Rest „junk“?
  - Many repeats, transposons, low-complexity sequence
  - Regulatory elements
Gene structure

(Protein-coding) Gene Structure in Eukaryotes

Transcription direction

Transcription start site → Exon → Intron → 3' splice site → 5' splice site → Polyadenylation site

Gene

mRNA (spliced) → PolyA tail

5' untranslated region

Coding sequence (ORF) – begins with start codon (ATG), ends with stop codon (TAA, TAG, or TGA)

3' untranslated region
Proteins

• Proteins make (almost) everything
  - Metabolism
  - Signal generation, transportation, recognition, and reaction
  - Cellular structures
  - Regulation of gene expression

• Protein sequence
  - Amino acids
  - 20-letter alphabet
  - 300-500 AA long
Protein Folding

- Proteins fold into 3D-structure
- Complex and dynamic process
- Main problem in bioinformatics
- „Function follows structure“
Protein structure

- Primary: aa sequence
- Secondary: Helixes & sheets
- Tertiary: 3D-Structure
- Quartary: Protein complexes
From Genotype to Phenotyp
The „-omics“ in Life Science

• Genome
  - All DNA sequence in a cell
  - Constant

• Transcriptome
  - All mRNA in a cell at given time
  - Varies greatly with cell type, environmental conditions, developmental stage, sex, ... of cells

• Proteome
  - All proteins in a cell at given time
  - Even more complex than transcriptome: life span, interaction, post-translational modifications, ...

• Metabolome
  - All non-protein things in a cell at given time