Bioinformatics
Introduction to genomics and proteomics I

ulf.schmitz@informatik.uni-rostock.de

Bioinformatics and Systems Biology Group
www.sbi.informatik.uni-rostock.de
Outline

Genomics/Genetics

1. The tree of life
   • Prokaryotic Genomes
     – Bacteria
     – Archaea
   • Eukaryotic Genomes
     – Homo sapiens

2. Genes
   • Expression Data
Genomics - Definitions

Genetics: is the science of genes, heredity, and the variation of organisms.
- Humans began applying knowledge of genetics in prehistory with the domestication and breeding of plants and animals.
- In modern research, genetics provides tools in the investigation of the function of a particular gene, e.g. analysis of genetic interactions.

Genomics: attempts the study of large-scale genetic patterns across the genome for a given species. It deals with the systematic use of genome information to provide answers in biology, medicine, and industry.
- Genomics has the potential of offering new therapeutic methods for the treatment of some diseases, as well as new diagnostic methods.
- Major tools and methods related to genomics are bioinformatics, genetic analysis, measurement of gene expression, and determination of gene function.
Genes

- A gene coding for a protein corresponds to a sequence of nucleotides along one or more regions of a molecule of DNA.
- In species with double stranded DNA (dsDNA), genes may appear on either strand.
- Bacterial genes are continuous regions of DNA.

**bacterium:**
- A string of $3N$ nucleotides encodes a string of $N$ amino acids.
- Or a string of $N$ nucleotides encodes a structural RNA molecule of $N$ residues.

**eukaryote:**
- A gene may appear split into separated segments in the DNA.
- An exon is a stretch of DNA retained in mRNA that the ribosomes translate into protein.
### Genome size comparison

<table>
<thead>
<tr>
<th>Species</th>
<th>Chrom.</th>
<th>Genes</th>
<th>Base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human (Homo sapiens)</td>
<td>46 (23 pairs)</td>
<td>28-35,000</td>
<td>3.1 billion</td>
</tr>
<tr>
<td>Mouse (Mus musculus)</td>
<td>40</td>
<td>22.5-30,000</td>
<td>2.7 billion</td>
</tr>
<tr>
<td>Puffer fish (Fugu rubripes)</td>
<td>44</td>
<td>31,000</td>
<td>365 million</td>
</tr>
<tr>
<td>Malaria mosquito (Anopheles gambiae)</td>
<td>6</td>
<td>14,000</td>
<td>289 million</td>
</tr>
<tr>
<td>Fruit Fly (Drosophila melanogaster)</td>
<td>8</td>
<td>14,000</td>
<td>137 million</td>
</tr>
<tr>
<td>Roundworm (C. elegans)</td>
<td>12</td>
<td>19,000</td>
<td>97 million</td>
</tr>
<tr>
<td>Bacterium (E. coli)</td>
<td>1</td>
<td>5,000</td>
<td>4.1 million</td>
</tr>
</tbody>
</table>
Genes

**exon:**
A section of DNA which carries the *coding sequence* for a protein or part of it. Exons are separated by intervening, non-coding sequences (called *introns*). In eukaryotes most genes consist of a number of exons.

**intron:**
An intervening section of DNA which occurs almost exclusively within a *eukaryotic* gene, but which is not translated to amino-acid sequences in the gene product. The introns are removed from the pre-mature mRNA through a process called *splicing*, which leaves the *exons* untouched, to form an *active* mRNA.
Examples of the exon:intron mosaic of genes

Globin gene – 1525 bp: 622 in exons, 893 in introns

Ovalbumin gene - ~ 7500 bp: 8 short exons comprising 1859 bp

Conalbumin gene - ~ 10,000 bp: 17 short exons comprising ~ 2,200 bp
Picking out genes in genomes

- Computer programs for genome analysis identify ORFs (open reading frames)

- An ORF begins with an initiation codon $\text{ATG (AUG)}$

- An ORF is a potential protein-coding region

- There are two approaches to identify protein coding regions…
Picking out genes in genomes

1. Detection of regions similar to known coding regions from other organisms
   - Regions may encode amino acid sequences similar to known proteins
   - Or may be similar to ESTs (correspond to genes known to be expressed)
   - Few hundred initial bases of cDNA are sequenced to identify a gene

2. Ab initio methods, seek to identify genes from the properties of the DNA sequence itself
   - Bacterial genes are easy to identify, because they are contiguous
   - They have no introns and the space between genes is small
   - Identification of exons in higher organisms is a problem, assembling them another…
Picking out genes in genomes

Ab initio gene identification in eukaryotic genomes

• The initial (5´) exon starts with a transcription start point, preceded by a core promoter site such as the TATA box (~30bp upstream)
  – Free of stop codons
  – End immediately before a GT splice-signal
Picking out genes in genomes

5' splice signal

3' splice signal

5’ UTR

3’ UTR

Promoter

ATG

5’ss

Stop

3’ss

PolyA Signal
Picking out genes in genomes

Ab initio gene identification in eukaryotic genomes

- **Internal exons** are free of stop codons too
  - Begin after an $AG$ splice signal
  - End before a $GT$ splice signal
Picking out genes in genomes

Ab initio gene identification in eukaryotic genomes

- **The final (3´) exon** starts after an AG splice signal
  - Ends with a stop codon (TAA, TAG, TGA)
  - Followed by a polyadenylation signal sequence
Humans have spliced genes...
DNA makes RNA makes Protein

- **DNA**
  - Enhancer
  - Core promoter
  - Start of transcription
  - 5’-non-coding exon
  - ATG-codon
  - Internal exons
  - Stop-codon
  - 3’-non-coding exon
  - Poly-A signal

- **mRNA**
  - pre-mRNA
  - mature mRNA
  - Protein

- **Transcription**
  - 5’ Capping, 3’ Polyadenylation
  - Splicing (removing introns)

- **Translation**
Tree of life

Prokaryotes

Bakteria

Eukarya

Archaeobacteria

Methanogen

Pyrococcus furiosus

Extreme halophiles

Euryarchaeota

Crenoarchaeota

Thermotogales

Animals

Plants

Fungis

Protists

Hyperthermophiles

Mesophiles
Genomics – Prokaryotes

- The genome of a prokaryote comes as a single double-stranded DNA molecule in ring-form:
  - In average 2mm long
  - Whereas the cells diameter is only 0.001mm
  - < 5 Mb
- Prokaryotic cells can have plasmids as well (see next slide).
- Protein coding regions have no introns.
- Little non-coding DNA compared to eukaryotes:
  - In E.coli only 11%
Genomics - Plasmids

- **Plasmids** are circular double stranded DNA molecules that are separate from the *chromosomal* DNA.

- They usually occur in *bacteria*, sometimes in *eukaryotic* organisms.

- Their size varies from 1 to 250 kilo base pairs (kbp). There are from one copy, for large plasmids, to hundreds of copies of the same plasmid present in a single cell.
Prokaryotic model organisms

- **E. coli (Escherichia coli)**
- **Methanococcus jannaschii** (archaeon)
- **Mycoplasma genitalium** (simplest organism known)
Genomics

- DNA of higher organisms is organized into *chromosomes* (human – 23 chromosome pairs)
- not all DNA codes for proteins
- on the other hand some genes exist in multiple copies
- that’s why from the genome size you can’t easily estimate the amount of protein sequence information
Genomes of eukaryotes

- majority of the DNA is in the nucleus, separated into bundles (*chromosomes*)
  - small amounts of DNA appear in organelles (mitochondria and chloroplasts)
- within single chromosomes gene families are common
  - some family members are *paralogues* (related)
    - they have duplicated within the same genome
    - often diverged to provide separate functions in descendants (Nachkommen)
    - e.g. human α and β globin
  - *orthologues* genes
    - are homologues in different species
    - often perform the same function
    - e.g. human and horse *myoglobin*
  - *pseudogenes*
    - lost their function
    - e.g. human *globin* gene cluster
Eukaryotic model organisms

- *Saccharomyces cerevisiae* (baker’s yeast)
- *Caenorhabditis elegans* (C. elegans)
- *Drosophila melanogaster* (fruit fly)
- *Arabidopsis thaliana* (flower)
- *Homo sapiens* (human)
The human genome

- ~3.2 x 10⁹ bp (thirty time larger than *C.elegans* or *D.melongaster*)
- coding sequences form only 5% of the human genome
- Repeat sequences over 50%
- Only ~32,000 genes
- Human genome is distributed over 22 chromosome pairs plus X and Y chromosomes
- *Exons* of protein-coding genes are relatively small compared to other known eukaryotic genomes
- *Introns* are relatively long
- Protein-coding genes span long stretches of DNA (dystrophin, coding a 3.685 amino acid protein, is >2.4Mbp long)

- Average gene length: ~8,000 bp
- Average of 5-6 exons/gene
- Average exon length: ~200 bp
- Average intron length: ~2,000 bp
- ~8% genes have a single exon
- Some exons can be as small as 1 or 3 bp.
# The human genome

Top categories in a function classification:

<table>
<thead>
<tr>
<th>Function</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acid binding</td>
<td>2207</td>
<td>14.0</td>
</tr>
<tr>
<td>DNA binding</td>
<td>1656</td>
<td>10.5</td>
</tr>
<tr>
<td>DNA repair protein</td>
<td>45</td>
<td>0.2</td>
</tr>
<tr>
<td>DNA replication factor</td>
<td>7</td>
<td>0.0</td>
</tr>
<tr>
<td>Transcription factor</td>
<td>986</td>
<td>6.2</td>
</tr>
<tr>
<td>RNA binding</td>
<td>380</td>
<td>2.4</td>
</tr>
<tr>
<td>Structural protein of ribosome</td>
<td>137</td>
<td>0.8</td>
</tr>
<tr>
<td>Translation factor</td>
<td>44</td>
<td>0.2</td>
</tr>
<tr>
<td>Transcription factor binding</td>
<td>6</td>
<td>0.0</td>
</tr>
<tr>
<td>Cell Cycle regulator</td>
<td>75</td>
<td>0.4</td>
</tr>
<tr>
<td>Chaperone</td>
<td>154</td>
<td>0.9</td>
</tr>
<tr>
<td>Motor</td>
<td>85</td>
<td>0.5</td>
</tr>
<tr>
<td>Actin binding</td>
<td>129</td>
<td>0.8</td>
</tr>
<tr>
<td>Defense/immunity protein</td>
<td>603</td>
<td>3.8</td>
</tr>
<tr>
<td>Enzyme</td>
<td>3242</td>
<td>20.6</td>
</tr>
<tr>
<td>Peptidase</td>
<td>457</td>
<td>2.9</td>
</tr>
<tr>
<td>Endopeptidase</td>
<td>403</td>
<td>2.5</td>
</tr>
<tr>
<td>Protein kinase</td>
<td>839</td>
<td>5.3</td>
</tr>
<tr>
<td>Protein phosphatase</td>
<td>295</td>
<td>1.8</td>
</tr>
<tr>
<td>Enzyme activator</td>
<td>3</td>
<td>0.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Function</th>
<th>Number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apoptosis inhibitor</td>
<td>132</td>
<td>0.8</td>
</tr>
<tr>
<td>Signal transduction</td>
<td>1790</td>
<td>11.4</td>
</tr>
<tr>
<td>Receptor</td>
<td>1318</td>
<td>8.4</td>
</tr>
<tr>
<td>Transmembrane receptor</td>
<td>1202</td>
<td>7.6</td>
</tr>
<tr>
<td>G-protein link receptor</td>
<td>489</td>
<td>3.1</td>
</tr>
<tr>
<td>Olfactory receptor</td>
<td>71</td>
<td>0.0</td>
</tr>
<tr>
<td>Storage protein</td>
<td>7</td>
<td>0.0</td>
</tr>
<tr>
<td>Cell adhesion</td>
<td>189</td>
<td>1.2</td>
</tr>
<tr>
<td>Structural protein</td>
<td>714</td>
<td>4.5</td>
</tr>
<tr>
<td>Cytoskeletal structural protein</td>
<td>145</td>
<td>0.9</td>
</tr>
<tr>
<td>Transporter</td>
<td>682</td>
<td>4.3</td>
</tr>
<tr>
<td>Ion channel</td>
<td>269</td>
<td>1.7</td>
</tr>
<tr>
<td>Neurotransmitter transporter</td>
<td>19</td>
<td>0.1</td>
</tr>
<tr>
<td>Ligand binding or carrier</td>
<td>1536</td>
<td>9.7</td>
</tr>
<tr>
<td>Electron transfer</td>
<td>33</td>
<td>0.2</td>
</tr>
<tr>
<td>Cytochrome P450</td>
<td>50</td>
<td>0.3</td>
</tr>
<tr>
<td>Tumor suppressor</td>
<td>5</td>
<td>0.0</td>
</tr>
<tr>
<td>Unclassified</td>
<td>4813</td>
<td>30.6</td>
</tr>
</tbody>
</table>

Total                                           | 15683  | 100.0|
The human genome

- Repeated sequences comprise over 50% of the genome:
  - *Transposable* elements, or *interspersed* repeats include *LINEs* and *SINEs* (almost 50%)
  - Retroposed *pseudogenes*
  - Simple ‘*stutters*’ - repeats of short oligomers (*minisatellites* and *microsatellites*)
  - *Segment duplication*, of blocks of ~10 - 300kb
  - Blocks of *tandem repeats*, including gene families

<table>
<thead>
<tr>
<th>Element</th>
<th>Size (bp)</th>
<th>Copy number</th>
<th>Fraction of genome %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short Interspersed Nuclear Elements (SINEs)</td>
<td>100-300</td>
<td>1.500.000</td>
<td>13</td>
</tr>
<tr>
<td>Long Interspersed Nuclear Elements (LINEs)</td>
<td>6000-8000</td>
<td>850.000</td>
<td>21</td>
</tr>
<tr>
<td>Long Terminal Repeats</td>
<td>15.000 -110.000</td>
<td>450.000</td>
<td>8</td>
</tr>
<tr>
<td>DNA Transposon fossils</td>
<td>80-3000</td>
<td>300.000</td>
<td>3</td>
</tr>
</tbody>
</table>
The human genome

- All people are different, but the DNA of different people only varies for 0.2% or less.
- So, only up to 2 letters in 1000 are expected to be different.
- Evidence in current genomics studies (Single Nucleotide Polymorphisms or SNPs) imply that on average only 1 letter out of 1400 is different between individuals.
- means that 2 to 3 million letters would differ between individuals.
Functional Genomics

From gene to function

- Genome
- Expressome
- Proteome
- Metabolome
DNA makes RNA makes Protein:
Expression data

- More copies of mRNA for a gene leads to more protein
- mRNA can now be measured for all the genes in a cell at once through **microarray technology**
- Can have 60,000 spots (genes) on a single gene chip
- Color change gives intensity of gene expression (over- or under-expression)
Genes and regulatory regions

regulatory mechanisms organize the expression of genes

– genes may be turned on or off in response to concentrations of nutrients or to stress
– control regions often lie near the segments coding for proteins
– they can serve as binding sites for molecules that transcribe the DNA
– or they bind regulatory molecules that can block transcription
Can diseases be characterized by patterns of gene activity?

- clustering
- supervised machine learning
Outlook – coming lecture

Proteomics

– Proteins
  – post-translational modification
  – Key technologies

• Maps of hereditary information
• SNPs (Single nucleotide polymorphisms)
• Genetic diseases
Thanks for your attention!