Introduction to Data Formats

• Real world, data and formats
• Sequences and alignments
• Trees and Pedigrees
• Motifs and Profiles
• Annotations and 3D structure
• Expression and pathways
• GenBank, SWISS-PROT, GeneCards
Data Formats

• Data: representation of real world
  – Never a perfect copy

• Data format: computer’s representation
  – Computers require formal definitions
  – Almost always represented as text

• Tools are data translators
  – Input one type of data, output another
Nucleotide Sequences

T (for DNA) and U (for RNA) interchangeable

GAATCG
TACTGT
CCATTG
CTCAGA
ATCGTA
CTGTCA

T (for DNA) and U (for RNA) interchangeable
Reverse Complementation

5’ GAATCGTACTGTCCATTGCTCA 3’

3’ ACTCGTTACCTGTCATGCTAAG 5’

5’ TGAGCAATGGGACAGTACGATTC 3’

5’ GAATCGTACTGTCCATTGCTCA 3’
Protein Sequences

KGSQEFWPWPNSHLEIGVKMDVYYS
### Amino Acid Codes

<table>
<thead>
<tr>
<th></th>
<th>Alanine</th>
<th>Glycine</th>
<th>Proline</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Arginine</td>
<td>Histidine</td>
<td>Serine</td>
</tr>
<tr>
<td>R</td>
<td>Asparagine</td>
<td>Isoleucine</td>
<td>Threonine</td>
</tr>
<tr>
<td>N</td>
<td>Aspartic acid</td>
<td>Leucine</td>
<td>Tryptophan</td>
</tr>
<tr>
<td>D</td>
<td>Cysteine</td>
<td>Lysine</td>
<td>Tyrosine</td>
</tr>
<tr>
<td>C</td>
<td>Glutamine</td>
<td>Methionine</td>
<td>Valine</td>
</tr>
<tr>
<td>Q</td>
<td>Glutamic acid</td>
<td>Phenylalanine</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multiple Sequences

• Many ways to encode in a file
  – FASTA format is most common

>HSBGPG Human gene for gla protein (BGP)
GGCAGATTCCCCCCTAGACCCCGCCCGCACCATGGGTCAAGGCAT
GCCCTCCTCCTCATCGCTGGGCACAGCCCAGAGGGTATAAACA
GTGCTGGAGGCTGGGCAGGAGGTTAAAGCA

>HSGLTH1 Human theta 1-globin gene
CCACTGCACTCACCACCGCAACCCCGGCCAACATTTTTTGCTGTTTTTTAG
TAGAGACTAAATACCATATAGTGGAACACCTAAGAGGCAGGGG
CCTTGGATCCAGGGCGA
Alignments

• Comparison between two sequences
  – Add gaps to make equal in length
  – Seek the ‘best’ possible alignment
  – ‘Global’ or ‘local’ alignment
Multiple Alignment

- Comparison between many sequences
- Many encoding formats
  - ALN is popular

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human DSHUCZ</td>
<td>M–ATKAVCVLKGDGPVQGII</td>
</tr>
<tr>
<td>Bovine DSBOCZ</td>
<td>--ATKAVCVLKGDGPVQGTI</td>
</tr>
<tr>
<td>Swordfish SODL</td>
<td>V–L–KAVCVLRGAGETTTGTV</td>
</tr>
<tr>
<td>Drosophila DSF</td>
<td>V–V–KAVCVING–D–AKGTV</td>
</tr>
<tr>
<td>Maize SDMZ</td>
<td>M–V–KAVAVLAGTD–VKGTI</td>
</tr>
<tr>
<td>Yeast DSBYC</td>
<td>V--QAVAVLKGDAGVSGVV</td>
</tr>
</tbody>
</table>
Taxonomy Trees

- **Plant Kingdom**
  - Division
  - Class
  - Monocots
    - Graminales
      - Gramineae (Grasses)
        - Wheat
        - Rye
        - Barley
      - Rice
      - Millet
      - Job's Tears
      - Ragi
      - Teff
      - Sugarcane (Oats)
  - Dicots
    - Caryophyllales
      - Polygonaceae
      - Chenopodiaceae
      - Amaranthaceae
        - Buckwheat
        - Rhubarb
        - Quinoa
        - Spinach
        - Amaranth
        - Pigweed
Phylogenetic Trees

Pedigrees
Motifs

- Common short sequence
- Simple motifs use IUPAC symbols

<table>
<thead>
<tr>
<th>R</th>
<th>A or G</th>
<th>K</th>
<th>G or T</th>
<th>B</th>
<th>C or G or T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>C or T</td>
<td>M</td>
<td>A or C</td>
<td>D</td>
<td>A or G or T</td>
</tr>
<tr>
<td>S</td>
<td>G or C</td>
<td>N</td>
<td>Any base</td>
<td>H</td>
<td>A or C or T</td>
</tr>
<tr>
<td>W</td>
<td>A or T</td>
<td></td>
<td></td>
<td>V</td>
<td>A or C or G</td>
</tr>
</tbody>
</table>

Example: **CHGW** matches **CCGA, CTGT, CAGC**
Motif Profiles

- Statistical model for a motif

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td>0.3</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td>0</td>
<td>0.4</td>
<td>0.9</td>
<td>0.2</td>
<td>0</td>
</tr>
<tr>
<td><strong>G</strong></td>
<td>0.5</td>
<td>0.2</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td><strong>T</strong></td>
<td>0.2</td>
<td>0.3</td>
<td>0</td>
<td>0.7</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**GTCTA**: $0.5 \times 0.3 \times 0.9 \times 0.7 \times 0.5 = 0.0473$

**TTGCT**: $0.2 \times 0.3 \times 0.1 \times 0.2 \times 0.5 = 0.0006$
Genomic Annotations

• Physical structure
  – Centromere, telomeres

• Genes
  – Introns, exons, alternative splice sites

• Binding sites
  – Transcription factors

• Variable sites
  – SNPs, repeats
Protein Annotations

- Domains
- Hydrogen bonds between peptides
- $\alpha$-helices, $\beta$-sheets

**Pred:** CCEEEEEEEEEEEHHHHHHHHHHHHEEEEEEEEEECCC

**AA:** PPPILFGLSLSLEVTTFDNLVLARFSVRSVSLDVDT
Protein 3D structure

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.982</td>
<td>-9.774</td>
<td>1.377</td>
</tr>
<tr>
<td>9.623</td>
<td>-9.833</td>
<td>1.984</td>
</tr>
<tr>
<td>8.913</td>
<td>-11.104</td>
<td>1.521</td>
</tr>
<tr>
<td>9.187</td>
<td>-11.630</td>
<td>0.461</td>
</tr>
<tr>
<td>8.814</td>
<td>-8.614</td>
<td>1.546</td>
</tr>
<tr>
<td>7.372</td>
<td>-8.754</td>
<td>2.039</td>
</tr>
<tr>
<td>7.339</td>
<td>-8.625</td>
<td>3.562</td>
</tr>
<tr>
<td>8.370</td>
<td>-8.307</td>
<td>4.131</td>
</tr>
<tr>
<td>6.284</td>
<td>-8.846</td>
<td>4.132</td>
</tr>
<tr>
<td>7.998</td>
<td>-11.599</td>
<td>2.304</td>
</tr>
<tr>
<td>7.266</td>
<td>-12.832</td>
<td>1.907</td>
</tr>
<tr>
<td>6.096</td>
<td>-12.456</td>
<td>1.005</td>
</tr>
</tbody>
</table>
## Expression Levels

<table>
<thead>
<tr>
<th>Gene</th>
<th>normal</th>
<th>hot</th>
<th>cold</th>
</tr>
</thead>
<tbody>
<tr>
<td>uch1</td>
<td>-2.0</td>
<td>0.0</td>
<td>0.924</td>
</tr>
<tr>
<td>gut2</td>
<td>0.398</td>
<td>0.402</td>
<td>-1.329</td>
</tr>
<tr>
<td>fip1</td>
<td>0.225</td>
<td>0.225</td>
<td>-2.151</td>
</tr>
<tr>
<td>msh1</td>
<td>0.676</td>
<td>0.685</td>
<td>-0.564</td>
</tr>
<tr>
<td>vma2</td>
<td>0.41</td>
<td>0.414</td>
<td>-1.285</td>
</tr>
<tr>
<td>meu26</td>
<td>0.353</td>
<td>0.286</td>
<td>-1.503</td>
</tr>
<tr>
<td>git8</td>
<td>0.47</td>
<td>0.47</td>
<td>-1.088</td>
</tr>
<tr>
<td>sec7b</td>
<td>0.39</td>
<td>0.395</td>
<td>-1.358</td>
</tr>
<tr>
<td>apn1</td>
<td>0.681</td>
<td>0.636</td>
<td>-0.555</td>
</tr>
<tr>
<td>wos2</td>
<td>0.902</td>
<td>0.904</td>
<td>-0.149</td>
</tr>
<tr>
<td>sec1</td>
<td>0.5</td>
<td>0.737</td>
<td>-1.0</td>
</tr>
<tr>
<td>spf31</td>
<td>1.171</td>
<td>0.946</td>
<td>0.228</td>
</tr>
<tr>
<td>slp1</td>
<td>0.378</td>
<td>0.364</td>
<td>-1.404</td>
</tr>
<tr>
<td>shm2</td>
<td>0.502</td>
<td>0.512</td>
<td>-0.994</td>
</tr>
</tbody>
</table>
Pathways
Data Formats: Summary

- Real world, data and data formats
- Nucleotide, protein sequences
- Sequence comparisons
  - Alignments, motifs, profiles
- Relationships between entities
  - Phylogenetic trees, pedigrees, pathways
- Annotations
- Tools are data translators
GenBank

- DNA/RNA sequence database
  - USA National Institute of Health
  - Synchronized daily with Europe, Japan
- Founded in 1982
  - Contains > 20 million sequences
  - Total size > 20 billion base pairs
  - >200 full species genomes
- Public submission
A GenBank Entry (1)

• Unique IDs
  – ‘Accession’ is permanent for an entry
  – ‘GenBank Identifier’ specifies one version

• Verbal description
  – Organism, gene/source
  – GenBank division, keywords
  – Date, version
  – Submitter, references
## GenBank Divisions

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRI</td>
<td>Primate</td>
<td>PHG</td>
<td>Bacteriophage</td>
</tr>
<tr>
<td>ROD</td>
<td>Rodent</td>
<td>SYN</td>
<td>Synthetic</td>
</tr>
<tr>
<td>MAM</td>
<td>Other mammal</td>
<td>UNA</td>
<td>Unannotated</td>
</tr>
<tr>
<td>VRT</td>
<td>Vertebrate</td>
<td>EST</td>
<td>Expressed seq. tag</td>
</tr>
<tr>
<td>INV</td>
<td>Invertebrate</td>
<td>PAT</td>
<td>Patented</td>
</tr>
<tr>
<td>PLN</td>
<td>Plant, fungus, algae</td>
<td>STS</td>
<td>Sequence tag site</td>
</tr>
<tr>
<td>BCT</td>
<td>Bacterial</td>
<td>GSS</td>
<td>Genome survey</td>
</tr>
<tr>
<td>VRL</td>
<td>Viral</td>
<td>HTG</td>
<td>High-throughput</td>
</tr>
</tbody>
</table>
A GenBank Entry (2)

• The sequence itself
  – Summary of base pair frequencies

• Features (at point within sequence)
  – Genes, exons, introns, translation
  – Promoters, binding sites
  – Repeats, stem and loop
  – Variation: RFLPs and SNPs
  – Known sequence tag locations
Searching GenBank

• Structureless and structured
  – All information fields are searchable

• Limits
  – Only recent sequences
  – Specific database only
  – Exclude drafts, patented, etc…

• Searching by sequence?
  – Using BLAST tools
SWISS-PROT

• Protein sequence database
  – Geneva University and Europe’s EBI
  – Some curation to minimize redundancy
• Founded in 1986
  – Contains > 120,000 entries
  – Total size > 46 million amino acids
• TrEMBL translated nucleotide database
  – For all coding sequences in GenBank
A SWISS-PROT Entry

• Core data
  – References
  – Biological source

• Annotations
  – Function and associated diseases
  – Post-translational modifications
  – Domains and binding sites
  – Secondary, quaternary structure

<table>
<thead>
<tr>
<th>General information about the entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entry name</td>
</tr>
<tr>
<td>Primary accession number</td>
</tr>
<tr>
<td>Secondary accession numbers</td>
</tr>
<tr>
<td>Entered in Swiss-Prot in</td>
</tr>
<tr>
<td>Sequence was last modified in</td>
</tr>
<tr>
<td>Annotations were last modified in</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name and origin of the protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein name</td>
</tr>
<tr>
<td>Synonyms</td>
</tr>
<tr>
<td>Gene name</td>
</tr>
</tbody>
</table>
Searching SWISS-PROT

• Using SRS (sequence retrieval system)
  – European equivalent to Entrez

• Other search types
  – Retrieve accession or ID
  – Unstructured (‘full text’)
  – Structured (‘advanced’)
  – Taxonomy browser

• BLAST to search by sequence
GeneCards

• Human gene database
  – Weizmann Institute in Israel
• Founded in 1997
  – Describes > 14,000 known genes
  – > 21,000 predicted genes
  – Also: pseudogenes, gene clusters, etc…
• Data mining approach
  – Data sourced from 36 other databases
A GeneCards Entry (1)

- Aliases and descriptions
  - HUGO standard nomenclature
- Chromosomal location
  - Links to genome browsers
- Expression levels
  - Different tissues
A GeneCards Entry (2)

- Available sequences
  - Homologues in other organisms
- Proteins
  - Protein families
- Mutations and other variation
  - Related diseases
- Publications
  - News articles