CMSC 838T – Lecture 9

- **Bioinformatics databases**
  - Organization & classification of bioinformatic data
  - Identify, format, & retrieval of bioinformatic data

![Diagram of various bioinformatics data sources and retrieval methods]

**What Is a Database?**

- **Computerized storehouse of data (records)**
- **Allows**
  - User-defined queries
  - Extraction of specified records
  - Adding, changing, removing, & merging records
- **Uses standardized formats**
Database Models

- **Defines data organization (schema)**
- **Relational**
  - Entities and relationships stored in tables
  - Predefined schema
  - Examples: Oracle, DB2, MySQL, PostgreSQL
- **Object-oriented**
  - Stores data as objects (i.e., structures with predefined type)
  - Examples: Versant, Jasmine, Objectivity
- **Semi-structured**
  - Schema dynamically defined within data (self-describing)
  - Flexible description of data with complex relationships
  - Example: XML databases

Bioinformatic Databases

- **Outline**
  - Issues
  - Databases
  - Identifiers & formats
  - Searching databases
Bioinformatic Databases

- **Useful information**
  - DNA sequences
  - Conserved DNA domains
  - Genomes
  - Gene expression (ESTs, microarrays)
  - Protein sequences
  - Protein 3D structure
  - Protein families
  - Mutations / polymorphisms / SNPs
  - Metabolic pathways
  - Chemical compounds (ligands)
  - Biomedical literature (journal papers, online books...)

- **Classification schemes**
  - Database design – relational, object-oriented...
  - Data type – DNA, RNA, EST, protein...
  - Organism – bacteria, virus, human...
  - Accessibility – public, academic, commercial
  - Data source – primary, derived
  - Data entry – manually curated, computational derived
  - Focus – sequence-oriented, gene-oriented

- **Resulting in many bioinformatic databases...**
Bioinformatic Database Issues

**Naming**
- Multiple names for same chemical
- Arising from multiple biological disciplines, conventions
- Example

\[
\begin{align*}
\text{Pyruvic acid} & \quad \text{[common or trivial name]} \\
\text{Pyruvate} & \quad \text{[common name for the anionic species]} \\
\text{2-Oxopropanoic acid} & \quad \text{[IUPAC name]} \\
\text{2-Oxopropionic acid} & \quad \text{[systematic name]} \\
\text{2-Oxopropionate} & \quad \text{[systematic name for the anionic species]} \\
\text{alpha-Keto-propanoic acid} & \quad \text{[systematic name]} \\
\text{CH\textsubscript{3}COO} & \quad \text{[line diagram notation]} \\
\text{CH\textsubscript{3}COO} & \quad \text{[line diagram notation for anionic species]}
\end{align*}
\]

is referred in BRENDA as the following:

---

**Redundancy**
- Multiple entries for same DNA / protein sequence
- Arising from multiple experiments & biological disciplines
- Example – redundant GenBank entries for E. coli dUTPase
  - 4 separate publications (X01714, V01578, L10328, AE000441)

**Data annotation & formats**
- Multiple data for single gene
  - Sequence, location, expression, structure, function…
- Resulting in multiple data annotations & formats

**Data integration**
- Combining data from multiple bioinformatic databases
Bioinformatic Databases

**Outline**
- Issues
- Databases
- Identifiers & formats
- Searching databases

**Major Bioinformatic Databases**

- **DNA sequences**
  - GenBank, RefSeq, UniGene

- **Protein sequences**
  - Swiss-Prot, PIR-PSD, GenPept, TrEMBL, NR, RefSeq

- **Protein structure**
  - Protein Data Bank (PDB)

- **Gene expression**
  - Gene Expression Omnibus (GEO)

- **Biomedical publications**
  - PubMed / MedLine
Bioinformatic Data Sources

◆ Primary databases
  - Original submissions by researchers
  - Staff organizes information only
  - Generally sequence oriented
  - Examples
    - GenBank, PDB

Bioinformatic Data Sources

◆ Derived databases
  - Compiled from data in primary databases
  - Manually curated (human selection & correction)
    - Advantages – high quality
    - Disadvantages – high expense, low volume
    - Examples
      - Swiss-Prot, PIR-PSD, RefSeq
  - Computational derivation (automatically generated)
    - Advantages – inexpensive, up-to-date
    - Disadvantages – lower quality
    - Examples
      - GenPept, TrEMBL, UniGene, COGs
Bioinformatic Databases – GenBank

- **Database type**
  - Nucleotide sequences
  - Primary database

- **Data combined from additional sources**
  - European Molecular Biology Laboratory (EMBL)
  - DNA DataBank of Japan (DDBJ)

- **Current size**
  - Release 134, Feb 2003
  - 23,035,823 sequences
  - 29,358,082,791 nucleotides

---

Bioinformatic Databases – GenBank

- **Types of submissions to database**
  - Genomic DNA
    - High quality complete DNA sequence
  - mRNA / cDNA
    - Partial or complete mRNA (or retranscribed cDNA)
  - Expressed sequence tag (EST)
    - Single-pass partial cDNA sequences from mRNA
  - Sequence tagged sites (STS)
    - Short DNA sequences unique in genome
  - Genomic survey sequence (GSS)
    - Single-pass genomic DNA
  - Third-party annotations of GenBank sequences
Bioinformatic Databases – Proteins

◆ **Protein sequence databases**
  - Once derived from laboratory experiments
  - Now mostly based on predicted ORFs from DNA
    - Manual curation
    - Computational derivation

◆ **Classification**
  - Predicted protein
    - No similarity match to protein of known function
    - Match to EST
  - Hypothetical protein
    - No similarity match to protein of known function
    - No match to EST

Bioinformatic Databases – Swiss-Prot, PIR-PSD

◆ **Database type**
  - Protein sequences
  - Derived database
    - Manually curated (non-redundant, annotated)

◆ **Many annotations**
  - Functions of the protein
  - Post-translational modifications
    - Phosphorylation, acetylation, GPI-anchor, etc...
  - Domains and sites
    - Calcium binding regions, ATP-binding sites, zinc fingers...
  - Secondary & quaternary structure
  - Similarities to other proteins
  - Variants
Bioinformatic Databases – Swiss-Prot, PIR-PSD

◆ **Swiss-Prot statistics**
  - Release 41.2, March 2003
  - 123,721 entries totaling 45,421,741 amino acids
  - Abstracted from 104,046 references
  - Average length 367 amino acids

◆ **PIR-PSD statistics**
  - Release 75.05, March 2003
  - 283,289 entries

Bioinformatic Databases – GenPept, TrEMBL

◆ **Database type**
  - Protein sequences
  - Computationally derived database
    - Predicted coding sequences (CDS) from GenBank, EMBL
    - Candidate sequences for Swiss-Prot, not yet processed

◆ **GenPept**
  - Release 134, February 2003
  - 1,314,007 loci containing 407,394,800 residues

◆ **TrEMBL**
  - Release 23, March 2003
  - 921,952 sequences, 40,914,860 residues
Bioinformatic Databases – Connections

- DNA sequences
  - GenBank
  - EMBL/EBI
  - Genome projects
  - Automatically translated
  - GenPept
  - TrEMBL
  - Manual curation & annotation
  - PIR-PSD
  - SwissProt

Bioinformatic Databases – Protein Data Bank

- **Database type**
  - Protein 3D structures
  - Primary database

- **Statistics**
  - March 2003
  - 20,473 proteins

![Protein Data Bank Statistics Graph](CMSC 838T – Lecture 9)
Bioinformatic Databases – Pfam

◆ **Database type**
  - Protein families
    - Multiple alignments of protein domains, conserved regions
  - Derived database (from Swiss-Prot & TrEMBL)
    - Pfam-A – manually curated (hand-edited MSA)
    - Pfam-B – computationally derived
    - Non-overlapping families from PRODOM database

◆ **Statistics**
  - Release 8.0, February 2003
  - 5193 families in Pfam-A
  - Protein sequence coverage
    - 73% at least one match in Pfam-A
    - 20% at least one match in Pfam-B

Bioinformatic Databases – RefSeq

◆ **Database type**
  - Nucleotide & protein sequences
  - Derived database
    - Human curated (non-redundant, cross-linked)

◆ **Data in RefSeq**
  - Genomic DNA contigs
  - mRNAs & proteins for known genes, gene models
  - Entire chromosomes
  - Multiple organisms

◆ **Statistics**
  - March 2003
  - 17,268 human loci, ~52,000 for all species
Bioinformatic Databases – UniGene

- **Database type**
  - Nucleotide sequences
  - Computationally derived database
    - Partitioned into non-redundant gene-oriented clusters
  - Gene-oriented view

- **Data in UniGene**
  - Clusters of genomic DNA & ESTs
  - Multiple organisms

- **Statistics**
  - March 2003
  - 111,064 human loci, ~500,000 for all species

---

Bioinformatic Databases – Relative Sizes

![Graph showing the relative sizes of different databases](image-url)

- **GenBank**: Computationally Derived
- **GenPept**, **TriEMBL**, **UniGene**, **PR-PSD**, **Swiss-Prot**, **RefSeq**, **PDB**, **Pfam**: Manually Curated
Bioinformatic Databases – PubMed

- **Database type**
  - Biomedical papers
  - Manually curated database

- **Data searched by PubMed**
  - MedLine papers
  - Additional physics, chemistry, life science journals
  - Abstracts, citations, some full articles
  - Over 11 million journal articles dating back to 1960’s

Bioinformatic Databases – Others

- **Gene expression**
  - ArrayExpress, Gene Expression Omnibus (GEO)

- **Multi-organism genomes**
  - Entrez Genome, HomoloGene, COGs, TIGR

- **Genetic variation & genetic diseases**
  - dbSNP, OMIM, CGAP

- **Metabolic pathways**
  - WIT, KEGG

- **Many more…**
  - Listed in journal “Nucleic Acids Research” each January
Bioinformatic Search Tools – Entrez

- **Search / retrieval tool for multiple linked databases**
  - Papers: biomedical literature (PubMed)
  - Nucleotide: sequence database (GenBank)
  - Protein: sequence database
  - Structure: 3D macromolecular structures
  - Genome: complete genome assemblies
  - OMIM: Online Mendelian Inheritance in Man
  - Taxonomy: organisms in GenBank
  - ProbeSet: gene expression and microarray datasets

![Diagram of Entrez search and retrieval processes]

Bioinformatic Search Tools – Entrez

- **Mapviewer (component of Entrez Genome)**
  - View & search complete genome by chromosome position
  - Display & zoom into chromosome maps

![Master Map: Genes On Sequence](chart)

- **Master Map: Genes On Sequence**
  - Total Genes On Chromosome: 736
  - Region Displayed: 0-623M bp
  - Download View Sequence/Findgene
  - Genes Labeled: 38 Total Genes in Region: 736

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Gene Name</th>
<th>Chromosome Location</th>
<th>Cytogenetic Location</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACTA2</td>
<td>Actin, gamma pseudogene 3</td>
<td>20p13.3-p13</td>
<td>C</td>
<td>actin, gamma pseudogene 3</td>
</tr>
<tr>
<td>MFSD6</td>
<td>Mitochondrial ribosomal protein</td>
<td>20p13</td>
<td>C</td>
<td>mitochondrial ribosomal protein</td>
</tr>
<tr>
<td>ITGB2</td>
<td>Rho GTPase-activating protein 2</td>
<td>20p13.3-p13</td>
<td>C</td>
<td>Rho GTPase-activating protein 2</td>
</tr>
<tr>
<td>ZAG1</td>
<td>ZAG1</td>
<td>20p12.1-p12.23</td>
<td>C</td>
<td>jagged 1 (Alagille syndrome)</td>
</tr>
<tr>
<td>PTFAP1</td>
<td>PTFAP1</td>
<td>20q12</td>
<td>C</td>
<td>prefibrosis, alpha pseudogene</td>
</tr>
<tr>
<td>LEPRE2</td>
<td>LEPRE2</td>
<td>20q13.1</td>
<td>C</td>
<td>leptin-related protein 2</td>
</tr>
<tr>
<td>FTO</td>
<td>FTO</td>
<td>20q11.2</td>
<td>C</td>
<td>fat mass phenotype, obesity</td>
</tr>
<tr>
<td>ASIP</td>
<td>ASIP</td>
<td>20q11.2-q12</td>
<td>C</td>
<td>agouti signaling protein, melanocortin</td>
</tr>
</tbody>
</table>
Bioinformatic Databases

✿ Outline
  - Issues
  - Databases
  - Identifiers & formats
  - Searching databases

Bioinformatic Database Identifiers

✿ Common identifiers for bioinformatic data
  - Locus name
  - Accession numbers
  - GenInfo ID
  - PubMed ID
Database Identifiers – Locus Names

- **Original identifiers of GenBank records**
  - LOCUS line in GenBank entries

- **Originally**
  - First 3 letters of organism followed by code for gene

- **Example**
  - HUMBB for human ß-globin region

- **Problems**
  - Unmaintainable due to growth of data
  - Homologous genes not named the same

Database Identifiers – Accession Numbers

- **No biological meaning**

- **Originally**
  - Uppercase letter followed by 5 digits: **U00002**

- **Currently**
  - Two uppercase letters followed by six digits: **BC037153**
  - May include version number for entry: **BC037153.1**

- **Stable way of identifying GenBank entries**

- **Now being used for both DNA and proteins**
Database Identifiers – GenInfo (gi) IDs

- Identifier for a particular sequence only
  - Each entry gets a unique gi number
- Example
  - GI: 22477487
- Not subject to versioning
  - Entry always remains the same
- Different / new versions of the same sequence
  - Manage using accession numbers

Database Identifiers – PubMed IDs (PMID)

- Identifies articles managed by NCBI
- Reliable, stable link to citation
- Example
  - PMID: 12205585
Bioinformatic Database Formats

- Data is stored / presented in a variety of formats
  - FASTA
  - GenBank
  - SwissProt
  - ASN.1
  - XML

Database Format – FASTA

- Used by FASTA tools
- Comment line followed by sequence data
  - No annotation, just sequence
- Example

```
>gi|1040960|gb|U35641.1|MMU35641 Mus musculus Brca1 mRNA.
GGCACGAGGATCAGCACCTCTCTCTGGGCTTTCCTGCTGCTGCTGGGCTTGAGTAG
GGATCTTTTTTCTGGAGAAAATGGCAAGAAATAGGGTCTAGTCTACGAAGAAAATCTTCTAGGTU
CGTCTGTTTGGAACTGATCAAAGAACCTGTTTCCACAAAGTGTGACCACATTT
TGCAAATTTGATGCTGAAACTTCTTAACCAGAAGAAAGGGCCTTCACAATGTCCT
TTTGTAAGAGAATGAGATAACCAAAAGGAGCCTACAGGGAAAGCAAGTTCAGTCCAG
```
Database Format – GenBank

- Flat file format used by GenBank
  - Annotation, author, version, etc...

- Example (just the top)

  LOCUS     MMU35641
  DEFINITION Mus musculus Brca1 mRNA, complete cds.
  ACCESSION U35641
  VERSION   U35641.1  GI:1040960
  KEYWORDS  .
  ORGANISM  house mouse strain=C57Bl/6.
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
  REFERENCE 1 (bases 1 to 5538)
  AUTHORS Sharan, S.K., Wims, M. and Bradley, A.
  TITLE     Murine Brca1: sequence and significance for human missense
            mutations
  MEDLINE   96177660
  PUBMED    8634698

Database Format – SWISS-PROT

- Defined by SWISS-PROT database
  - Includes annotation, other info

- Example

  ID       BRC1_MOUSE  STANDARD;  PRT;  1812 AA.
  AC       P48754; G60957; G60983;
  DT       01-FEB-1996 (Rel. 33, Created)
  DT       01-NOV-1997 (Rel. 35, Last sequence update)
  DT       16-OCT-2001 (Rel. 40, Last annotation update)
  DE       Breast cancer type 1 susceptibility protein homolog.
  GN       BRCA1.
  OS       Mus musculus (Mouse).
  OC       Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
          Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Mus.
  OX       NCBI_TaxID=10090;
  RN       [1]
  RP       SEQUENCE FROM N.A.
  RC       SEQUENCE FROM N.A.
  STRAIN=C57BL/6; TISSUE=Embryo;
  RX       MEDLINE=96177659; PubMed=8634697;
  RT       "Mouse Brca1: localization sequence analysis and identification of
          evolutionarily conserved domains."
Database Format – ASN.1

- **International standard**
- Semi-structured format
- Base format for NCBI data

**Example**

```
Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  descr {
    title "Mus musculus Brca1 mRNA, and translated products" ,
    source {
      org {
        taxname "Mus musculus" ,
        db {
          db "taxon" ,
          tag
          id 10090 } } ,
      orgname {
        name
        binomial {
          genus "Mus" ,
          species "musculus" } , ...
```

Database Format – XML

- **eXtensible Markup Language**
- Open standard for semi-structured data, uses tags like HTML
- Document split into content (XML), style (XSL), linking (XLL)

**Example**

```
<?xml version="1.0"?>
<!DOCTYPE GBSeq PUBLIC "-//NCBI//NCBI GBSeq/EN"
<GBSet>
<GBSeq>
  <GBSeq_locus>MMU35641</GBSeq_locus>
  <GBSeq_length>5538</GBSeq_length>
  <GBSeq_strandedness value="not-set">0</GBSeq_strandedness>
  <GBSeq_moltype value="mrna">5</GBSeq_moltype>
  <GBSeq_topology value="linear">1</GBSeq_topology>
  <GBSeq_division>ROD</GBSeq_division>
  <GBSeq_update-date>18-OCT-1996</GBSeq_update-date>
  <GBSeq_create-date>25-OCT-1995</GBSeq_create-date>
  <GBSeq_definition>Mus musculus Brca1 mRNA, complete cds</GBSeq_definition>
  <GBSeq_primary-accession>U35641</GBSeq_primary-accession>
  <GBSeq_accession-version>U35641.1</GBSeq_accession-version>
```
Processing Data in Bioinformatic Databases

◆ Format conversion
  - Frequently tools handle only one of the data formats
  - Use software to transform between formats
    • ReadSeq, SeqIO
◆ Perl (Practical Extraction and Report Language)
  - Portable C-like interpreted scripting language
  - Powerful pattern matching, string processing operations
  - Frequently used to extract / process bioinformatic data
◆ BioPerl
  - Collection of Perl classes designed for bioinformatic tools
  - Sequence analysis, alignment, format conversion, I/O, automate bioinformatic analyses, parse results, create GUIs, manage persistent storage in RDMBS...

Bioinformatic Databases

◆ Outline
  - Issues
  - Databases
  - Identifiers & formats
  - Searching databases
Bioinformatic Databases – Usage

- NCBI Protein information usage survey

Using Bioinformatic Databases

- Primary use of bioinformatics
  - Finding similar sequences
  - BLAST!

1) insert sequence

2) click button!
Using Bioinformatic Databases

Versions of BLAST

- BLASTN
  - Nucleic acids against nucleic acids
- BLASTP
  - Protein query against protein database
- BLASTX
  - Translated nucleic acids against protein database
- TBLAST
  - Protein query against translated nucleic acid database
- TBLASTX
  - Translated nucleic acids against translated nucleic acids

Databases – Searching w/ BLAST

 QUERY SEQUENCE                  DATABASE

 Nucleic Acid  blastn
               Blastn
               blastn
               blastx
               Blastx
               Tblastx
               Tblastx

 Peptide/Protein  blastip

 Conceptual protein translations

 Nucleic Acids

 Proteins/Pepitides

 Conceptual protein translations

 ...
Databases – Searching w/ BLAST

- **BLAST result**
  - Graphic display

![Color Key for Alignment Scores](image)

- **Example**
  - Matching sequences w/ bit-score & E-value
  - Hyperlinks to database entry for sequence

<table>
<thead>
<tr>
<th>gi</th>
<th>gb</th>
<th>Bit Score</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>17330420</td>
<td>BH384278.1</td>
<td>153</td>
<td>3e-36</td>
</tr>
<tr>
<td>17320126</td>
<td>BH373984.1</td>
<td>140</td>
<td>9e-34</td>
</tr>
<tr>
<td>17338337</td>
<td>BH392196.1</td>
<td>112</td>
<td>8e-25</td>
</tr>
<tr>
<td>20373967</td>
<td>BH771010</td>
<td>105</td>
<td>1e-21</td>
</tr>
<tr>
<td>17314411</td>
<td>BH368367.1</td>
<td>104</td>
<td>2e-21</td>
</tr>
<tr>
<td>17332712</td>
<td>BH386570.1</td>
<td>64</td>
<td>3e-21</td>
</tr>
</tbody>
</table>

Hyperlinks to sequences

 CMSC 838T – Lecture 9
Searching w/ BLAST – Interpreting Results

◆ **High quality hits**
  - Matching sequences with high E-values, % identity
  - >25% identity may imply similar function, 3D structure
  - Caveat – similarity does not guarantee homology

◆ **Low quality hits**
  - No matching sequences
  - Few matching sequences, with low E-values, % identity
  - Absence of match does not always mean no homology
    - Check sequence format
    - Change search parameters (scoring matrix, gap penalties)
    - Check (low complexity) sequence filtering
    - Try PSI-BLAST for distant homologs

Bioinformatic Databases – Summary

◆ **Observations**
  - Lots of useful information
  - Complex relationships, naming schemes
  - High volume of nucleotide sequence data (GenBank)
  - Protein sequence data mostly derived from sequence data
    - Computational derivation (GenPept, TrEMBL)
    - Manual curation (PIR-PSD, Swiss-Prot)
  - Attempts to organize sequence data
    - Eliminate redundancy, add annotation (RefSeq, UniGene)
  - Many tools attempt to link useful bioinformatic data