NCBI Resources in Bioinformatics
Education

Bio21
10/17/03

NLM / NCBI
• Created as a part of NLM in 1988
  – Establish public databases
    • U.S. National DNA Sequence Database
  – Perform research in computational biology
  – Develop software tools for sequence analysis
  – Disseminate biomedical information

Molecular Databases
• Sequence
• Structure
• Expression
GenBank: NCBI's Primary Sequence Database

Release 137 August 2003
- 27,213,748 Records
- 33,865,022,251 Nucleotides
- 120,000 + Species

- full release every two months
- incremental and cumulative updates daily
- available only through internet


27,213,748 Records
33,865,022,251 Nucleotides
121 Gigabytes

The Growth of GenBank

Release 136: 27.0 million records
34 billion nucleotides
Doubling time = 12 months

RefSeq: NCBI's Derivative Sequence Database

Release 1 June 2003
- Curated transcripts and proteins

<table>
<thead>
<tr>
<th>Release 1</th>
<th>June 2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>64,729</td>
<td>Genomic</td>
</tr>
<tr>
<td>211,803</td>
<td>RNA</td>
</tr>
<tr>
<td>785,143</td>
<td>Protein</td>
</tr>
<tr>
<td>2,005</td>
<td>Species</td>
</tr>
</tbody>
</table>

Third Party Annotation (TPA) Database

- Annotations of existing GenBank sequences
- Allows for community annotation of genomes
- Direct submissions
  - BankIt

NCBI’s SNP Database

- Primary Database and Derivative (RefSNP)
- Single Nucleotide Polymorphism
- Repeat polymorphisms
- Insertion-Deletion Polymorphisms
- 19 Species
- Over 11 million submissions

RefSNP

- Non-redundant
- Computational Analysis
- BLAST hits to
  - genome, mRNA, protein and structure
What is UniGene?

A gene-oriented view of sequence entries
-MegaBlast based automated sequence clustering
-Nonredundant set of gene oriented clusters
-Each cluster a unique gene
-Information on tissue types and map locations
-Includes well-characterized genes and novel ESTs
-Useful for gene discovery and selection of mapping reagents


EST hits A.t. serine protease mRNA

UniGene Collections: Animals

<table>
<thead>
<tr>
<th>Species</th>
<th>Sequences</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens (human)</td>
<td>4,135,222</td>
<td>111,064</td>
</tr>
<tr>
<td>Mus musculus (mouse)</td>
<td>3,382,434</td>
<td>86,905</td>
</tr>
<tr>
<td>Rattus norvegicus (rat)</td>
<td>422,469</td>
<td>52,264</td>
</tr>
<tr>
<td>Bos taurus (cow)</td>
<td>209,914</td>
<td>18,003</td>
</tr>
<tr>
<td>Sus scrofa (pig)</td>
<td>69,048</td>
<td>14,771</td>
</tr>
<tr>
<td>Gallus gallus (chicken)</td>
<td>125,827</td>
<td>5,042</td>
</tr>
<tr>
<td>Danio rerio (zebrafish)</td>
<td>217,019</td>
<td>15,707</td>
</tr>
<tr>
<td>Oryzias latipes (medaka)</td>
<td>59,720</td>
<td>11,366</td>
</tr>
<tr>
<td>Xenopus laevis (frog)</td>
<td>177,814</td>
<td>19,045</td>
</tr>
<tr>
<td>Silurana tropicalis (frog)</td>
<td>110,976</td>
<td>7,674</td>
</tr>
<tr>
<td>Ciona intestinalis (tunicate)</td>
<td>464,060</td>
<td>13,328</td>
</tr>
<tr>
<td>Drosophila melanogaster (fruit fly)</td>
<td>276,922</td>
<td>14,701</td>
</tr>
<tr>
<td>Anopheles gambiae (mosquito)</td>
<td>91,352</td>
<td>3,243</td>
</tr>
<tr>
<td>Caenorhabditis elegans (nematode)</td>
<td>200,436</td>
<td>19,045</td>
</tr>
</tbody>
</table>

June, 2003
UniGene Collections: Plants and Others

<table>
<thead>
<tr>
<th>Species</th>
<th>Common Name</th>
<th>Sequences</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>thale cress</td>
<td>221,113</td>
<td>26,813</td>
</tr>
<tr>
<td>Glycine max</td>
<td>soybean</td>
<td>201,710</td>
<td>9,084</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>barrel medic</td>
<td>97,063</td>
<td>5,729</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>wine grape</td>
<td>63,652</td>
<td>7,736</td>
</tr>
<tr>
<td>Lycopersicon esculentum</td>
<td>tomato</td>
<td>70,118</td>
<td>3,749</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>potato</td>
<td>33,859</td>
<td>2,899</td>
</tr>
<tr>
<td>Triticum aestivum</td>
<td>wheat</td>
<td>297,961</td>
<td>20,376</td>
</tr>
<tr>
<td>Hordeum vulgare</td>
<td>barley</td>
<td>263,407</td>
<td>10,272</td>
</tr>
<tr>
<td>Zea mays</td>
<td>maize (corn)</td>
<td>163,541</td>
<td>13,512</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>rice</td>
<td>94,780</td>
<td>19,191</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>sorghum</td>
<td>17,742</td>
<td>802</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>alga</td>
<td>80,771</td>
<td>5,712</td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>slime mold</td>
<td>119,140</td>
<td>4,782</td>
</tr>
</tbody>
</table>

Gene Expression Omnibus

- Expression Data Repository
  - microarray
    - protein
    - nucleotide
  - SAGE
  - Mass Spec

Analysis: GEO DataSets

- ADH2 time course
- Principal Component Analysis

Yeast microarray beer fermentation
MMDB: Molecular Modeling Data Base

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation
  - Inclusion of Taxonomy, Citation, and other information
  - Conversion to ASN.1 data description language
- Structure neighbors determined by Vector Alignment Search Tool (VAST)

Cn3D 4.1: C-SRC

Conserved ATP binding site

Cn3D 4.1: Structural Alignment
Cn3D: Simple Homology Modeling

NCBI’s Conserved Domain Database
- Multiple sequence alignments
- PSI-BLAST–based score matrices
- Sources SMART, PFAM, COGs
- New NCBI curated domains
  - structure informed alignments
- Stats:
  - COGS 4,873
  - Pfam 5,193
  - Smart 653
  - NCBI CDD 316

NCBI CD: Tyrosine Kinase
Using Cn3D to model domains

NCBI Tools

BLAST
VAST

Basic Local Alignment Search Tool

- Calculates similarity for biological sequences
- Finds best local alignments
- Heuristic approach based on Smith-Waterman algorithm
- Searches for matching "words" rather than individual residues
- Uses statistical theory to determine if a match might have occurred by chance
Minimum Requirements for a Hit

**ATGCGCATGCTAATTGGCCTT**

- exact word match
  - one match

**GTQTVEILFYNI**

- neighborhood words
  - two matches

- Nucleotide BLAST requires one exact match
- Protein BLAST requires two neighboring matches within 40 aa

Discontiguous Word Hits: Megablast

<table>
<thead>
<tr>
<th>W</th>
<th>t</th>
<th>Coding</th>
<th>Non-coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>16</td>
<td>1101101101101101</td>
<td>111110110110111</td>
</tr>
<tr>
<td>12</td>
<td>16</td>
<td>1011011011011011</td>
<td>111010110110111</td>
</tr>
<tr>
<td>11</td>
<td>18</td>
<td>1110010110010111</td>
<td>1011011011011011</td>
</tr>
<tr>
<td>12</td>
<td>18</td>
<td>1001010110101011</td>
<td>1110100101010111</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>1001010110101011</td>
<td>1110100101010111</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>1001010110110101</td>
<td>1110100101010111</td>
</tr>
</tbody>
</table>


BLOSUM62: a position independent scoring matrix

| A   | R   | N   | D   | C   | Q   | E   | G   | H   | I   | L   | K   | M   | F   | P   | S   | T   | W   | Y   | V   | X   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

**ow weights**

**weights**

- Negative for less likely substitutions
- Positive for more likely substitutions
### Position Specific Substitution Rates

<table>
<thead>
<tr>
<th>Position</th>
<th>Active site serine</th>
<th>Weakly conserved serine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Position Specific Score Matrix (PSSM)

```
<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>N</th>
<th>R</th>
<th>K</th>
<th>H</th>
<th>Q</th>
<th>E</th>
<th>G</th>
<th>D</th>
<th>C</th>
<th>N</th>
<th>G</th>
<th>D</th>
<th>S</th>
<th>S</th>
<th>C</th>
<th>N</th>
<th>G</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>206</td>
<td></td>
<td>207</td>
<td></td>
<td></td>
<td>208</td>
<td></td>
<td>209</td>
<td></td>
<td></td>
<td>210</td>
<td></td>
<td></td>
<td>211</td>
<td></td>
<td>212</td>
<td></td>
<td></td>
<td>213</td>
</tr>
<tr>
<td></td>
<td></td>
<td>204</td>
<td></td>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td></td>
<td></td>
<td>207</td>
<td></td>
<td></td>
<td></td>
<td>208</td>
<td></td>
<td></td>
<td></td>
<td>209</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>207</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>206</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

### Local Alignment Statistics

High scores of local alignments between two random sequences follow the Extreme Value Distribution.

**Expect Value**

\[ E = \text{number of database hits you expect to find by chance} \]

\[ E = K m e^{-\frac{1}{2}} \quad \text{or} \quad E = m n 2^{-S} \]

- \( K \) is scale for search space
- \( \lambda \) is scale for scoring system
- \( S \) is bitscore (\((S - \ln K) / \ln 2\))

The Probability Density Function of the Extreme Value Distribution is:

\[ f(x) = \frac{1}{\lambda} e^{(-x / \lambda)} \]

**Score**

\[ x = \frac{S - \ln K}{\ln 2} \]

**Alignment**

\[ x = \frac{S - \ln K}{\ln 2} \]

\[ E = m n 2^{-S} \]

\[ K = \text{scale for search space} \]

\[ \lambda = \text{scale for scoring system} \]

\[ S = \text{bitscore} = (S - \ln K) / \ln 2 \]

(appplies to ungapped alignments)
The Flavors of BLAST

- **Standard BLAST**
  - traditional contiguous word hit
  - position independent scoring
  - nucleotide, protein and translations

- **Megablast**
  - optimized for large batch searches
  - can use discontiguous words

- **PSI-BLAST**
  - constructs PSSMs automatically
  - searches protein database with PSSMs

- **RPS BLAST**
  - searches a database of PSSMs
  - basis of conserved domain database
For each protein chain, locate SSEs (secondary structure elements), and represent them as individual vectors. Align the vectors.
Submitting a PDB File to VAST

- Pick the correct file format
- Remove all records except ATOM
- This is the best way to convert PDB into MMDB format!

NCBI FieldGuide

WWW Access
Entrez & BLAST

Users per Day

Christmas Day
A Day in the Life of Entrez

<table>
<thead>
<tr>
<th>Database</th>
<th>Searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleotide</td>
<td>125,994</td>
</tr>
<tr>
<td>Protein</td>
<td>29,590</td>
</tr>
<tr>
<td>Journals</td>
<td>25,885</td>
</tr>
<tr>
<td>MeSH</td>
<td>21,946</td>
</tr>
<tr>
<td>UniGene</td>
<td>20,685</td>
</tr>
<tr>
<td>LocusLink</td>
<td>16,442</td>
</tr>
<tr>
<td>SNP</td>
<td>1,053</td>
</tr>
<tr>
<td>UniSTS</td>
<td>506</td>
</tr>
<tr>
<td>CDD</td>
<td>496</td>
</tr>
<tr>
<td>GenBank</td>
<td>486</td>
</tr>
<tr>
<td>域</td>
<td>463</td>
</tr>
<tr>
<td>Prosite</td>
<td>433</td>
</tr>
<tr>
<td>RefSeq</td>
<td>334</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1,809,277</td>
</tr>
</tbody>
</table>

June 23, 2003

Human DNA Mismatch Repair

Links and Neighbors
More Links and Neighbors

OMIM: Human Disease Genes

MLH1 Domain Structure
Mapping Variation Onto Structure

What Data is Available?

- NCBI assembled annotated genomic contigs
  - Genome project data
  - Other primary data
- Reference sequences - mRNA, proteins, transcripts
- Genome Scan gene models
- Mapped variation data
- Integrated maps - RH, genetic, cytogenetic, and sequence
- Clustered and mapped expressed sequences
- Links to outside data sources

Mammalian Genomes
The Mouse Assembly: Build 30

---

NCBI FieldGuide

The Mouse Assembly: Build 30

---

NCBI FieldGuide

Field Guide

---

NCBI FieldGuide

Field Guide Summary

---

NCBI FieldGuide

Field Guide Summary

---

NCBI FieldGuide

Field Guide Summary

---

NCBI FieldGuide

Field Guide Summary

---

NCBI FieldGuide

Field Guide Summary

---