Sequence Based Function Annotation

Qi Sun
Bioinformatics Facility
Biotechnology Resource Center
Cornell University
Workflow of genomic projects for non-model organisms

- RNA-seq data
- Assembly (Trinity)
- ORF prediction (Trinity)
- Genomic sequencing
- Assembly (SOAP de novo)
- Gene prediction (Maker)
- Function prediction
Sequence Based Function Annotation

1. Given a sequence, how to predict its biological function?

2. How to describe the function of a gene?

3. How to work with 50,000 genes?
1. Given a protein sequence, how to predict its function?

```plaintext
>unknow_protein_1
MVHLTDAEKAAVSCLWGKVNSDEVGGGEALGRLLVVYPWTQR
YFDSFGDLSSASAIMGNAKVKAHGKKVITAFNDGLNHLDSL
KGTFASLSELHCDKLHVDPFNFRLLGNMIVIVLGHHLGKDFTPAAQAADFQKVVAGVATALAHKYH
```

**Common approaches**

- Identify the homologous gene in a different species with good function annotation; (BLAST, et al.)
- Identify conserved motif; (PFAM, InterProScan, et al.)

**Alternative approaches**

- Protein 3D structure prediction (threading methods)
- Co-expression network modules; (Genevestigator)
- Linkage or association mapping;
- ...
BLAST does Local Alignment
(Basic Local Alignment Search Tool)

Local Alignment vs Global Alignment

HSP-1 vs HSP-2

(Bowtie, BWA, ClustalW, et al)

Distribution of 41 Blast Hits on the Query Sequence

Mouse-over to show deline and scores. Click to show alignments

Color Key for Alignment Scores

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;40</td>
<td>Black</td>
</tr>
<tr>
<td>40-50</td>
<td>Blue</td>
</tr>
<tr>
<td>50-60</td>
<td>Green</td>
</tr>
<tr>
<td>60-200</td>
<td>Red</td>
</tr>
<tr>
<td>&gt;=200</td>
<td>Orange</td>
</tr>
</tbody>
</table>

0 100 200 300 400 500 600
NCBI BLAST

- How does BLAST work?

- BLAST and Psi-BLAST: Position independent and position specific scoring matrix.
How does BLAST work

Step 2. Score each alignment, and report the top alignments

Number of Chance Alignments = 2 \times 10^{-73}
BLOSUM62, a position independent matrix

<table>
<thead>
<tr>
<th></th>
<th>Ala</th>
<th>Arg</th>
<th>Asn</th>
<th>Asp</th>
<th>Cys</th>
<th>Gln</th>
<th>Glu</th>
<th>Gly</th>
<th>His</th>
<th>Ile</th>
<th>Leu</th>
<th>Lys</th>
<th>Met</th>
<th>Phe</th>
<th>Pro</th>
<th>Ser</th>
<th>Thr</th>
<th>Trp</th>
<th>Tyr</th>
<th>Val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>4</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>Arg</td>
<td>-1</td>
<td>5</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>Asn</td>
<td>-2</td>
<td>0</td>
<td>6</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>Asp</td>
<td>-2</td>
<td>2</td>
<td>1</td>
<td>6</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>Cys</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>9</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>Gln</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>5</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>Glu</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-4</td>
<td>2</td>
<td>5</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>Gly</td>
<td>0</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>6</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>His</td>
<td>-2</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>-3</td>
<td>0</td>
<td>0</td>
<td>-2</td>
<td>8</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>Ile</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-4</td>
<td>-3</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>-1</td>
<td>2</td>
<td>0</td>
<td>-1</td>
<td>-3</td>
<td>1</td>
<td>1</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-2</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>2</td>
<td>-1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-3</td>
<td>0</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-3</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>0</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>-3</td>
<td>-3</td>
<td>-4</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>1</td>
<td>11</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>-2</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>0</td>
<td>-3</td>
<td>-3</td>
<td>-1</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-3</td>
<td>3</td>
<td>1</td>
<td>-2</td>
<td>1</td>
<td>-1</td>
<td>-2</td>
<td>2</td>
<td>0</td>
<td>-3</td>
<td>-1</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Ala  Arg  Asn  Asp  Cys  Gln  Glu  Gly  His  Ile  Leu  Lys  Met  Phe  Pro  Ser  Thr  Trp  Tyr  Val
How does BLAST work

Step 2. Score each alignment – protein alignment

Number of Chance Alignments = \(4 \times 10^{-50}\)

Scores from BLOSUM62, a position independent matrix

- NCBI Discovery Workshops
Scores from BLOSUM62, a position independent matrix

- NCBI Discovery Workshops
PSSM Alignment: Globins

Globins are heme proteins, which bind and transport oxygen. This family summarizes a diverse set of homologous protein domains, including: (1) tetrameric vertebrate hemoglobins, which are the major protein component of erythrocytes and transport oxygen in the bloodstream, (2) microorganism flavohemoglobins, which are linked to C-terminal FAD-dependent reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vibrio cholerae, (4) plant leghemoglobins (symbiotic hemoglobins, involved in nitrogen metabolism in plant rhizomes), (5) plant non-symbiotic hexacoordinate globins and hexacoordinate globins from bacteria and animals, such as neuroglobin, (6) invertebrate hemoglobins, which may occur in tandem-repeat arrangements, and (7) monomeric myoglobins found in animal muscle tissue.

Conserved Histidine

- NCBI Discovery Workshops
PSSM Viewer

Histidine scored differently at two positions

- NCBI Discovery Workshops
Build PSSM with PSI-BLAST

• PSI-BLAST

  1. Iteration 1: Regular BLASTP (BLOSSOM62) to identify a list of closely related proteins. Build PSSM from these proteins.
  2. Iteration 2: Use the PSSM built from Iteration 1 to score alignment in this Iteration.
  3. Repeat multiple iterations.
**Build PSSM with DELTA-BLAST**

DELTA-BLAST employs a subset of NCBI's Conserved Domain Database (CDD) to construct PSSM

---

<table>
<thead>
<tr>
<th>Enter Query Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter accession number(s), g(i), or FASTA sequence(s)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Or, upload file</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choose File No file chosen</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Job Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enter a descriptive title for your BLAST search</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Align two or more sequences</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Choose Search Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database</td>
</tr>
<tr>
<td>‘Reference proteins (refseq_protein)’</td>
</tr>
<tr>
<td>Organism Optional</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
</tr>
<tr>
<td>Exclude</td>
</tr>
<tr>
<td>Exclude Optional</td>
</tr>
<tr>
<td>Models (PDB) Uncultured/environmental sample sequences</td>
</tr>
<tr>
<td>Entrez Query Optional</td>
</tr>
<tr>
<td>Enter an Entrez query to limit search</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Program Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
</tr>
<tr>
<td>□ NCST (protein-protein BLAST)</td>
</tr>
<tr>
<td>□ PSI-BLAST (Position-Specific Iterated BLAST)</td>
</tr>
<tr>
<td>□ PHI-BLAST (Pattern Hit Initiated BLAST)</td>
</tr>
<tr>
<td>□ DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BLAST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search database ‘Reference proteins (refseq_protein)’ using DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)</td>
</tr>
<tr>
<td>□ Show results in a new window</td>
</tr>
</tbody>
</table>

**Note:** Parameter values that differ from the default are highlighted in yellow and marked with + sign.
Heme Binding Site

Conserved Histidine

blastp

DELTA-BLAST

- NCBI Discovery Workshops
BLAST and BLAST–like programs

- Traditional BLAST (formerly blastall) nucleotide, protein, translations
  - blastn nucleotide query vs. nucleotide database
  - blastp protein query vs. protein database
  - blastx nucleotide query vs. protein database
  - tblastn protein query vs. translated nucleotide database
  - tblastx translated query vs. translated database
- Megablast nucleotide only
  - Contiguous megablast
    - Nearly identical sequences
  - Discontiguous megablast
    - Cross-species comparison
Nucleotide Databases: List

- Human genomic + transcript
- Mouse genomic + transcript
- Others (nr etc.):
- Nucleotide collection (nr/nt)
  - Genomic plus Transcript
    - Human genomic plus transcript (Human G+T)
    - Mouse genomic plus transcript (Mouse G+T)
  - Other Databases
    - Nucleotide collection (nr/nt)
      - Reference RNA sequences (refseq_rna)
      - Reference genomic sequences (refseq_genomic)
      - NCBI Genomes (chromosome)
      - Expressed sequence tags (est)
      - Genomic survey sequences (gss)
      - High throughput genomic sequences (HTGS)
      - Patent sequences (pat)
      - Protein Data Bank (pdb)
      - Human ALU repeat elements (alu_repeats)
      - Sequence tagged sites (dbsts)
      - Whole-genome shotgun contigs (wgs)
      - Transcriptome Shotgun Assembly (TSA)
      - 16S ribosomal RNA sequences (Bacteria and Archaea)

Services
megablast
blastn
tblastn
tblastx
Non-redundant protein

**nr** (non-redundant protein sequences)
- GenBank CDS translations
- NP_, XP_ **refseq_protein**
- Outside Protein
  - PIR, **Swiss-Prot**, PRF
  - **PDB** (sequences from structures)

**pat** protein patents

**env_nr** metagenomes
  (environmental samples)
Hidden Markov Model (HMM) is more general than PSSM
HMMs are trained from a multiple sequence alignment.
Match a sequence to a model
Application: Function Prediction
**PFAM**

a pre-constructed HMM model database for protein function domain prediction

http://pfam.sanger.ac.uk/
How to describe the function of a gene?

- Description line in free text.
- Controlled vocabulary (Gene Ontology)
- Pathway (KEGG)
Gene Ontology
To download premade GO Slim:

Create your own GO Slim:
http://oboedit.org/docs/html/Creating_Your_Own.GO_Slim_in_OBO_Edit.htm
Maintained GO slim sets

- Generic GO slim
- Plant slim
- Yeast slim
- Protein Information Resource
- Metagenomics slim
High throughput gene function prediction

• **BLAST2GO**
  Function prediction based on BLAST match to known proteins.
  [http://www.blast2go.com](http://www.blast2go.com)

• **Interproscan**
  Function prediction mostly based on PFAM and other motif scanning tools.
  [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)
BLAST2GO Annotation Steps

- **BLAST:** BLAST against “NCBI nr” or Swissprot database;

- **Mapping:** Retrieve GO from annotated homologous genes;

- **Annotation:** Assign GO terms to query sequences.

- **InterProScan (optional):** Integrate with InterProScan results.
BLAST2GO

- Do each steps separately on different computers.

BLAST step

[Image of BLAST step]

http://cbsuapps.tc.cornell.edu

BLAST2GO step

[Image of BLAST2GO step]

BioHPC Lab computer through VNC
Computing Resource at Cornell

BioHPC Web : Web based job submission
  • Web base GUI interface
  • Limited applications

BioHPC Lab : A cloud-like computing Service
  • Linux based system;
  • command line operation;
Using BRC Bioinformatics Facility Resource

1. **Office hour**
   1pm to 3pm every Monday, 618 Rhodes Hall
   Signup at: http://cbsu.tc.cornell.edu/lab/office1.aspx

2. **Step-by-step instruction using software on BioHPC computers.**
   Software page: http://cbsu.tc.cornell.edu/lab/labsoftware.aspx
   BLAST2GO page: http://cbsu.tc.cornell.edu/lab/doc/instruction_blast2go.htm