Sequence Based Function Annotation

Qi Sun
Bioinformatics Facility
Biotechnology Resource Center
Cornell University
Usage scenarios for sequence based function annotation

• Function prediction of newly cloned genes.

• Identify homologs of genes in a different species
Usage scenarios for sequence based function annotation

Genomic scale function annotation for non-model organisms

- RNA-seq data
  - Assembly (Trinity)
    - ORF prediction (Trinity)
      - Function prediction
  - Genomic sequencing
    - Assembly (SOAP de novo)
      - Gene prediction (Maker)
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
>unknown_protein_1
MVHLTDAEKAAVSClwGKVNSDEVGGEALGRLLVYYPWTQR
YFDSFGDLSSASAIMGNAKVKAHGKKVITAFNDGLNHLDSL
KGTFASLSHELCDKHLVDPENFLLGNMIVIVLVGHHLGKDF
TPAAQAADFQKVAVGVALAHKYH
```

**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;
- ...

•
**NCBI BLAST**

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

Step 1. Create alignments between HSPs (High-scoring Segment Pair)

The BLAST Search Algorithm

query word $w = 3$

<table>
<thead>
<tr>
<th>Step1</th>
<th>Query: TGSQSLAALLNKCKTPQGQRQQLVNVQPKQPQLMDKNR1EELNLVLVEAFV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PQG 18</td>
</tr>
<tr>
<td></td>
<td>PQA 12</td>
</tr>
</tbody>
</table>

Step 2 neighborhood words

Step 3

Query: 325SLAALLNKCKTPQGQRQQLVNVQPKQPQLMDKNR1EELNLVLVEA365
       +LA++L++TPGR++W++P+D+ER+A

Subject: 290TLASVLDCTVTPMGSRMLKRWLPVDRDTRVLLERQQTIGA330

High-scoring Segment Pair (HSP)
How does BLAST work

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

Number of Chance Alignments = 2 \times 10^{-73}

Gap \(-5 + 4(2)\) = -13

Match = +2

Mismatch = -3

NCBI Discovery Workshops
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>-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>
</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>-4</td>
<td>-2</td>
<td>-2</td>
<td>-3</td>
<td>-2</td>
<td>-3</td>
<td>-2</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>-3</td>
<td>-1</td>
<td>-2</td>
<td>-1</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<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>
</table>
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) microorganismal flavohemoglobin, which are linked to C-terminal FAD-dependent reductase domains, (3) homodimeric bacterial hemoglobins, such as from Vitreoscilla, (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
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
Heme Binding Site

Conserved Histidine

blastp

DELTA-BLAST

- NCBI Discovery Workshops
Heme Binding Site

Conserved Histidine

BLAST is not reliable for alignment of homologous genes between distantly related species.

- NCBI Discovery Workshops
BLAST does Local Alignment
(Basic Local Alignment Search Tool)

Local Alignment vs Global Alignment

HSP-1

HSP-2

(Bowtie, BWA, ClustalW, et al)
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

- NCBI Discovery Workshops
Nucleotide Databases: List

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)
<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Category</th>
<th>GO:ID</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G035341</td>
<td>molecular_function</td>
<td>GO:0008270</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td></td>
<td>molecular_function</td>
<td>GO:0046872</td>
<td>metal ion binding</td>
</tr>
<tr>
<td></td>
<td>cellular_component</td>
<td>GO:0005622</td>
<td>intracellular</td>
</tr>
<tr>
<td></td>
<td>cellular_component</td>
<td>GO:0019005</td>
<td>SCF ubiquitin ligase complex</td>
</tr>
<tr>
<td></td>
<td>biological_process</td>
<td>GO:0009733</td>
<td>response to auxin</td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>molecular_function</td>
<td>GO:0003677</td>
<td>DNA binding</td>
</tr>
<tr>
<td></td>
<td>cellular_component</td>
<td>GO:0005634</td>
<td>nucleus</td>
</tr>
<tr>
<td></td>
<td>cellular_component</td>
<td>GO:0005694</td>
<td>chromosome</td>
</tr>
<tr>
<td></td>
<td>biological_process</td>
<td>GO:0006259</td>
<td>DNA metabolic process</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0034641</td>
<td>cellular nitrogen compound metabolic process</td>
</tr>
<tr>
<td>Gene ID</td>
<td>Function/Component</td>
<td>GO ID</td>
<td>Description</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------------------------------</td>
<td>-----------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>GRMZM2G035341</td>
<td>molecular function</td>
<td>GO:0008270</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td></td>
<td>molecular function</td>
<td>GO:0046872</td>
<td>metal ion binding</td>
</tr>
<tr>
<td></td>
<td>cellular component</td>
<td>GO:0005622</td>
<td>intracellular</td>
</tr>
<tr>
<td></td>
<td>SCF ubiquitin ligase complex</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>biological process</td>
<td>GO:0006259</td>
<td>DNA metabolic process</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DNA metabolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cell-cell signaling</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>signal transduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>transport</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>stress response</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cell cycle and proliferation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>death</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cell adhesion</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>protein metabolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RNA metabolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>DNA metabolism</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>developmental processes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>cell organization and biogenesis</td>
</tr>
</tbody>
</table>

Gene Ontology
Gene Ontology

[Diagram of Gene Ontology categories: Molecular Function, Cellular Component, Biological Process]
The Necessity for GO Slim
The Necessity for GO
The Necessity for GO Slim

To download premade GO Slim:

Create your own GO Slim:
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

BLAST2GO step

http://cbsuapps.tc.cornell.edu

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

   Software page: http://cbsu.tc.cornell.edu/lab/labsoftware.aspx
   BLAST2GO page: http://cbsu.tc.cornell.edu/lab/doc/instruction_blast2go.htm