Genome Annotation

Qi Sun
Bioinformatics Facility
Cornell University
Some basic bioinformatics tools

• BLAST

• PSI-BLAST – Position-Specific Scoring Matrix

• HMM - Hidden Markov Model
NCBI BLAST

• How does BLAST work?

• BLAST and Psi-BLAST: Position independent and position specific scoring matrix.
BLAST programs

- 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
How does BLAST work

Step 1. find alignments

The BLAST Search Algorithm

**Step 1**
Query: TGSQSLAALLNKCKTPQGQRLVQWIKQPLMDKNRIERLNVEAFV

<table>
<thead>
<tr>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PQG</td>
<td>18</td>
</tr>
<tr>
<td>PEG</td>
<td>15</td>
</tr>
<tr>
<td>PRC</td>
<td>14</td>
</tr>
<tr>
<td>PKG</td>
<td>14</td>
</tr>
<tr>
<td>PNG</td>
<td>13</td>
</tr>
<tr>
<td>PDG</td>
<td>13</td>
</tr>
<tr>
<td>PHG</td>
<td>13</td>
</tr>
<tr>
<td>PMG</td>
<td>13</td>
</tr>
<tr>
<td>PSG</td>
<td>13</td>
</tr>
</tbody>
</table>

**Step 2**
Neighborhood words

<table>
<thead>
<tr>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PQA</td>
<td>12</td>
</tr>
<tr>
<td>PQN</td>
<td>12</td>
</tr>
</tbody>
</table>

**Step 3**
Query: 325SLLAALLNKCKTPQGQRLLVQWIKQPLMDKNRIERLNVEA365

<table>
<thead>
<tr>
<th>Word</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>+LA++L+</td>
<td>7</td>
</tr>
<tr>
<td>TP</td>
<td>6</td>
</tr>
<tr>
<td>G</td>
<td>6</td>
</tr>
<tr>
<td>R+++W+</td>
<td>5</td>
</tr>
<tr>
<td>P+D+ER+</td>
<td>4</td>
</tr>
<tr>
<td>+A</td>
<td>4</td>
</tr>
</tbody>
</table>

Subject: 290TLASVLDCTVTPMGSRMLKRWLHMVPVRDTRVLLERQQTIGA330

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

Step 2. scoring alignments

Number of Chance Alignments = $2 \times 10^{-73}$
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
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>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>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>-1</td>
<td>1</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>
<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>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>
</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>-3</td>
<td>1</td>
<td>1</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>
</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>-3</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>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>
</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>1</td>
<td>1</td>
<td>-4</td>
<td>-3</td>
<td>-2</td>
<td>11</td>
<td></td>
<td></td>
<td></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>2</td>
<td>-1</td>
<td>-1</td>
<td>-2</td>
<td>3</td>
<td>-3</td>
<td>-2</td>
<td>-2</td>
<td>2</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>0</td>
<td>-3</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>
</tr>
</tbody>
</table>
BLOSUM62 substitution score is position independent

Scores from BLOSUM62, a position independent matrix
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 flavohemoglobins, 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.
Heme Binding Site

Conserved Histidine

blastp

DELTABLAST

- NCBI Discovery Workshops
BLAST is not reliable for alignment of homologous genes between distantly related species.
Search PSSM with DELTA-BLAST

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

HMMs are trained from a multiple sequence alignment.

<table>
<thead>
<tr>
<th>SeqID</th>
<th>Sequence</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q5E94_DGVIN</td>
<td>MREDVATWNYLFKLDSDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_HUMAN</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_MOUSE</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_KEF</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_CHICK</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_RAN</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_BRADB</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_DBM</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_ICPFO</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
<tr>
<td>Q5A0_DIC</td>
<td>MREDVATWNYLFKLSDDDKFKYVAGDNYKMEK</td>
<td>PALAB</td>
</tr>
</tbody>
</table>

**Diagram: 56 globins as a normalanno**

- **Ruler:** 1...10...20...30...40...50...60...70...80...90

- **Data Points:**
  - 5 points between 0 and 50 bits
  - 5 points between 50 and 100 bits
  - 5 points between 100 and 150 bits

- **Graph:**
  - X-axis: Bits
  - Y-axis: Probability

- **Legend:**
  - 50 globins as a normalanno

- **Annotations:**
  - A symbol or number indicating the specific data point or category.
Hidden Markov Model (HMM) is more general than PSSM
Match a sequence to a model
Application: Function Prediction
unknown_protein

MALLYRRMSMLNIIALYLIFLACIVQGSKVQEAEIGKNVSLECASENEAVAKLGNQTKNHTRYKIRTEPLKNSDDGSENDSQDFIKYKNVNLLADVNIKDSGNYCTAQTGQNHSTEFOVRYPYLSKVLQSTPDRIKKIKQOVMYLCLEMYPOQETTRNLKWLKDGSGQEFFLDTSISKLNDTHNFLETFEXITKYKYSGTYKCTVFDDGTEITSKIELTFMVMEPQVQSIDFAMAVAGANKIVLMWTVNDGNDPQKFPFLITQEGAPTFTTYHDGVNSHTSAYLDHFPNNTYFLFIYLRIVGKNSIGNGQTPQPGITLQSYDFIPFQKVEITGATSTI

TICKWNPPOPPIDLITYQVYLTVSEGEVPKFVEEAIYQVSNRLNPYMFDKLFKLTATYEFVRRACSDLKTKCGPWSENVTMGTVKPNLSIQHHNDTVTRGNQNAINwDVPKTPNKGKVSYLHIHLQNPMTSOVRMAGPIKRIDEHPHHKLYVESPVNNTYTVSAILTHKKXGEPATGSLMCPVSTPDAIGRTMWSKVLNDSKYYLKLYPKISERNPICFCYRLYLVIRINNDELPDEPKLNAITAYEQHSDNVTRSSAYJAEEMISSKF

RPEIFLGDEKRSENNIDIRONDIECRMCLETFPLRFKPEIIHPPQGSLNSDSELPISEKDNLKGANLEHAKLIESKLKRDRAVTSDENPSZAVNPNVLHDSRDVFDEIDINSNYTFLEIIVRDRNNA

LMAYSKYFIIDIPATEAEPIQSLNMDYLLSQIGVAKAGAVLLGIVLFIVLWFIHKTKEAEQGTHLTDTRSDSRLALFFRRNHHTEUIGSKGGDGAGCPQHMDMNKVRDKYFROLYHCNPHQKHEYVRFMRRRTTYNDLKEACNYRPP

EQNELIEIMLNLRRQITQOVHYLTKKQSVSYNTVCDLRHEKLLATEIESKQDPHLEIQDFWRTKNCXDDTLKVTQTVWAMCILVQHLRLEC

PFAM

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

http://pfam.sanger.ac.uk/
Genome Annotation

Genome Assembly

Evidence based

Ab initio prediction
Genome Annotation Tools

Procaryotic genomes

Online Services
1. RAST
2. NCBI

Eucaryotic genomes

MAKER
Mark Yandell Lab
University of Utah
To run MAKER, you need the following files:

Genome sequence FASTA file

Transcript sequences:
- Assembled from RNA-seq
- Transcriptome from related species

Protein sequences:
- From related species
- Uniprot/Swissprot
Where to run MAKER:

• Use CyVerse/XSEDE
  https://wiki.cyverse.org/wiki/display/TUT/MAKER+2.31.9+with+CCTOOLS+Jetstream+Tutoria...

• Use BioHPC
  https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=65#c
The MAKER Pipeline:

- Genome Assembly
  - RepeatMasker
    - BLAST
      - Gene Finders
      - Protein Evidence
        - RNA Evidence
          - Exonerate
            - Synthesis
              - Annotate
Use BioHPC

Create and edit control file

```
maker -CTL  #create control file templates
```

Commands

```
maker  #run alignment or prediction software
```

```
SNAP or Augustus  #build model
```

Use BioHPC

Create and edit control file

maker -CTL #create control file templates

Commands

maker #run alignment or prediction software

SNAP or Augustus #build model

GFF3 file
Step 1. Repeat Masking

• Simple repeats: e.g. “AAAAA…AAA”

• Prebuilt DB: Repbase

• Custom DB: build with RepeatModeler
Control file: `maker_opts.ctl` *

```plaintext
genome=dpp_contig.fasta  # genome sequence

...  

model_org=all  # select a model organism for RepBase masking in RepeatMasker

rmlib=  # provide an organism specific repeat library in fasta format for RepeatMasker

repeat_protein=  # provide a fasta file of transposable element proteins for RepeatRunner

rm_gff=  # pre-identified repeat elements from an external GFF3 file

prok_rm=0  # forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no

softmask=1  # use soft-masking rather than hard-masking in BLAST (i.e. seg and dust filtering)

...  

cpus=1
```

- Keep this file in the directory where you execute the maker command.
Step 2. Train a gene prediction model

Training data set:

- Assembled RNA-seq
- Transcripts from related species
- Proteins from related species

Procedures:

- Alignment with BLAST
- Refine exon-intron junctions with EXONERATE
- Build HMM model with SNAP or AUGUSTUS
MAKER control file for step 2:

```plaintext
genome=dpp_contig.fasta  #genome sequence
...
est=transcriptome.fasta
altest=otherspeciesgene.fasta
protein=protein.fasta
est2genome=1
protein2genome=1
```

Build model with SNAP
Step 3. *Ab initio* Prediction with SNAP or AUGUSTUS

```
genome=dpp_contig.fasta  # genome sequence
...
snaphmm=pyu1.hmm
est2genome=0
protein2genome=0
```

GFF file with predicted genes
Two or more iterations

genome=dpp_contig.fasta # genome sequence
...
  snaphmm=pyu2.hmm
  est2genome=0
  protein2genome=0
Control file in first round of MAKER

```
genome=contig.fasta
est=EST1.fa,EST2.fa
altest=myAltEST.fa
protein=myprotein.fa
model_org=simple
rmlib=myRepeat.fa
repeat_protein=rp.fa
est2genome=1
protein2genome=1
```

Following rounds of MAKER

```
genome=contig.fasta
est=
altest=
protein=
model_org=
rmlib=
repeat_protein=
est2genome=0
protein2genome=0
```
Also included in following rounds of MAKER

maker_gff=pyu_rnd1.all.gff

est_pass=1  #pass on EST alignment in GFF
protein_pass=1  #pass on protein alignment in GFF
rm_pass=1  #pass on repeat alignment in GFF

pred_stats=1  #report AED stats

This way, MAKER would not run BLAST again, instead it will use the alignment from GFF file
A few other notes

1. Use MPI to parallelize the run:

   In control file: `cpus=1`

   `mpiexec -n 40 maker -base output_rnd1`

2. Use both SNAP and AUGUSTUS for prediction.

3. There is a tool to create custom gene, transcript and protein names.
Related to running MAKER on BioHPC

1. Set tmp directory to `/workdir/xxxxx`:
   Create directory:
   ```
   mkdir /workdir/xxxxx/tmp
   ```
   In control file:
   ```
   TMP=/workdir/xxxxx/tmp
   ```
   Use machines with >=40 cores on BioHPC, and use all cores

2. Run mpiexec
   ```
   /usr/local/mpich/bin/mpiexec -n 40 ....
   ```

3. Copy maker and repeatMasker to `/workdir/xxxxx`
   These two directories contain large data files, better to keep them on `/workdir`
Avoid under-fitting and over-fitting: Evaluate results with AED Score (Annotation Edit Distance)

AED=0: Genes models of perfect concordance with the evidence; AED=1: Genes models with no evidence support

Cumulative Distribution of AED

Eilbeck et al. BMC Bioinformatics 2009
Visualization - JBrowse or IGV

JBrowse: https://biohpc.cornell.edu/lab/userguide.aspx?a=software&i=357#c

Can I trust MAKER annotation?
If a gene of interest is missed from annotation:

Run TBLASTN with a closely related protein:
```bash
makeblastdb -in myGenome.fa -parse_seqids -dbtype nucl
tblastn -query myProtein.fa -db myGenome.fa -out output_file
```

Run PFAM on all ORF (slow, and exons only)
```bash
getorf -minsize 100 -sequence myGenome.fa -outseq myorf.fa
pfam_scan.pl -fasta myorf.fa –pfamB mydomain.hmm
```