Searching for similar proteins in a Database

BLAST  HMM  Threading
Profile hidden Markov models for biological sequence analysis

http://hmmer.wustl.edu
40pt sH
41pt sH
42r haS
43r ot

Alignment

Model

Search for matches
The programs in **HMMER**:

- **hmmpfam**
  
  Search an HMM database for matches to a query sequence.

- **hmmbuild**
  
  build a model from a multiple sequence alignment

- **hmmsearch**
  
  Search a sequence database for matches to an HMM
HMM libraries:

**PFAM**: [http://pfam.wustl.edu/](http://pfam.wustl.edu/)
An HMM library based on the Swissprot 40 and SP-TrEMBL 18 protein sequence databases. 3882 protein families in current version.

**SMART**: [http://smart.embl-heidelberg.de/](http://smart.embl-heidelberg.de/)
More than 500 extensively annotated domain families
### The input and output:

<table>
<thead>
<tr>
<th>Model</th>
<th>Seq-from HMM</th>
<th>Seq-to HMM</th>
<th>from HMM-to Score</th>
<th>E-value</th>
<th>Alignment</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>!! fn3</td>
<td>139</td>
<td>221</td>
<td>1</td>
<td>84</td>
<td>58.1</td>
<td>1.2e-14</td>
</tr>
<tr>
<td>!! fn3</td>
<td>233</td>
<td>317</td>
<td>1</td>
<td>84</td>
<td>59.4</td>
<td>5.1e-15</td>
</tr>
<tr>
<td>!! fn3</td>
<td>328</td>
<td>410</td>
<td>1</td>
<td>84</td>
<td>36.3</td>
<td>4.4e-08</td>
</tr>
<tr>
<td>!! fn3</td>
<td>421</td>
<td>501</td>
<td>1</td>
<td>84</td>
<td>58.4</td>
<td>9.8e-15</td>
</tr>
<tr>
<td>!! fn3</td>
<td>512</td>
<td>591</td>
<td>1</td>
<td>84</td>
<td>27.0</td>
<td>3e-05</td>
</tr>
<tr>
<td>!! fn3</td>
<td>599</td>
<td>677</td>
<td>1</td>
<td>84</td>
<td>78.9</td>
<td>6.9e-21</td>
</tr>
<tr>
<td>!! fn3</td>
<td>689</td>
<td>778</td>
<td>1</td>
<td>84</td>
<td>40.8</td>
<td>2e-09</td>
</tr>
<tr>
<td>!! fn3</td>
<td>789</td>
<td>869</td>
<td>1</td>
<td>84</td>
<td>14.8</td>
<td>0.0063</td>
</tr>
<tr>
<td>!! fn3</td>
<td>880</td>
<td>955</td>
<td>1</td>
<td>84</td>
<td>67.6</td>
<td>1.7e-17</td>
</tr>
<tr>
<td>!! fn3</td>
<td>974</td>
<td>1060</td>
<td>1</td>
<td>84</td>
<td>58.4</td>
<td>1e-14</td>
</tr>
<tr>
<td>!! Y_phosphatase</td>
<td>1312</td>
<td>1542</td>
<td>1</td>
<td>274</td>
<td>393.6</td>
<td>1.3e-115</td>
</tr>
</tbody>
</table>
Program 1: hmmpfam

Search an HMM database for matches to a query sequence.

Query sequence:
C:\cbsu\module1\hmmer_projects\exe1\unknown_proteins

Database:
pfam
Exercise 1: Identifying domains in an unknown protein

1. Check the files in the directory
   C:\cbsu\module1\hmmer_projects\exe1
   • sequence file: unknown_proteins
   • database file: pfam_test
   • program file: hmmpfam.exe, parse_pfam.pl

2. Run the program: (for help: hmmpfam –h)
   
   hmmpfam -E 1e-10 -A 10 pfam_test unknown_proteins > 
   pfamresult.txt

3. Parse the result into a spreadsheet
   
   parse_pfam.pl pfamresult.txt pfamresult.xls
Evaluating the significance of a hit:

1. E-value: $\leq 0.1$
   (10% chance that you would've seen a hit this good in a search of random sequences)

2. Raw score $\geq$ GA (the scores used as cutoffs in constructing Pfam)

3. Raw score $> \log_2($number of seqs in the database$)$ (20 for the nr)
Parallel HmmPfam at CBSU
contact: cbsu@tc.cornell.edu
Program 2, 3. `hmmbuild` & `hmmsearch`

Build a model and search the sequence database for motifs that fit the model.

Sequence alignment

Model

More sequence motifs that fit this model
Exercise 2: Identifying all putative genes that are regulated by crp (Cyclic AMP receptor)

Available resources:

1. E coli genome sequence

2. a list of crp binding sites determined by DNA footprinting.
Automated Process: Finding **Known Regulators** in Genome Sequences using **Hidden Markov Models**

**Training Sets:**
- RegulonDB
- Church Guesses
- Bench Data

**Initial Alignment**

**Forward Model**

**Reverse Complement Model**

**Genomes/Assemblies**

**Forward Hits**

**Reverse Complement Hits**

**HMMER**

- Slide provided by Dr. Angela Baldo
Exercise 2: Building models

1. Check the files in the directory.
   
   C:\cbsu\module1\hmmer_projects\exe2
   
   • sequence alignment: crp-Church.aln
   
   • database file: ecoli_k12
   
   • program file: hmmbuild.exe, hmmcalibrate.exe, hmmsearch.exe

2. Build the model: (for help: hmmbuild –h)
   
   hmmbuild crpmodel crp-Church.aln

3. Calibrate the model:
   
   hmmcalibrate crpmodel

4. Search the genome: (for help: hmmsearch –h)
   
   hmmsearch crpmodel ecoli_k12 > searchresult.txt