Protein Sequence Analysis
Domain review

- What is a domain?
  - Part of a sequence that can fold independently, and is present in other sequences as well.
Domain Search

- Motifs (Regular expression & others)
- Profiles
- Profile HMMs

Domains: 1 2 3

Residue interactions: ● with DNA ● with metal
PROSITE (Motif/Profiles)
Also at Sanger
HMMER: profile HMMs for protein sequence analysis

Overview

Profile hidden Markov models (profile HMMs) can be used to do sensitive database searching using statistical descriptions of a sequence family’s consensus. HMMER is a freely distributable implementation of profile HMM software for protein sequence analysis.

The current version is HMMER 2.3.2 (3 Oct 2003), containing minor bugfixes and updates for the May 2003 release of HMMER 2.3.

Documentation

- Text files associated with the HMMER 2.3.2 release:
  [README] [installation] [Release notes] [License summary] [GNU General Public License]
- The HMMER User's Guide. [PDF, 94 pages]
HMMER programs

- **Hmmalign**
  - Align a sequence to an HMM

- **Hmmbuild**
  - Build a model from a multiple alignment

- **Hmmemit**
  - Emits a probabilistic sequence from an HMM

- **Hmmpfam**
  - Search PFAM with a sequence query

- **Hmmsearch**
  - Search a sequence database with an HMM query
Pattern and profile searches

- **InterPro Scan** - Integrated search in PROSITE, Pfam, PRINTS and other family and domain databases
- **ScanProsite** - Scans a sequence against PROSITE or a pattern against Swiss-Prot and TrEMBL
- **MotifScan** - Scans a sequence against protein profile databases (including PROSITE)
- **Frame-ProfileScan** - Scans a short DNA sequence against protein profile databases (including PROSITE)
- **Pfam HMM search**; scans a sequence against the Pfam protein families db [At Washington University or at Sanger Centre]
- **FingerPRINTScan** - Scans a protein sequence against the PRINTS Protein Fingerprint Database
- **FPAT** - Regular expression searches in protein databases
- **ELM** - Eukaryotic Linear Motif resource for functional sites in proteins
- **PRATT** - Interactively generates conserved patterns from a series of unaligned proteins; [at EBI / ExPASy]
- **PPSEARCH** - Scans a sequence against PROSITE (allows a graphical output); at EBI
- **PROSITE scan** - Scans a sequence against PROSITE (allows mismatches); at PBIL
- **PATTINPROT** - Scans a protein sequence or a protein database for one or several pattern(s); at PBIL
- **SMART** - Simple Modular Architecture Research Tool; at EMBL
- **TEIRESIAS** - Generate patterns from a collection of unaligned protein or DNA sequences; at IBM
Post-translational modification

- Residues undergo modification, usually by addition of a chemical group.
- Key mechanism for signal transduction, and many other cellular functions.
- Some modifications might require single residues (Ex: phosphorylation). Others might require a pattern.
<table>
<thead>
<tr>
<th>Post-translational modification prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ChloroP</strong> - Prediction of chloroplast transit peptides</td>
</tr>
<tr>
<td><strong>LipoP</strong> - Prediction of lipoproteins and signal peptides in Gram negative bacteria</td>
</tr>
<tr>
<td><strong>MITOPROT</strong> - Prediction of mitochondrial targeting sequences</td>
</tr>
<tr>
<td><strong>PATS</strong> - Prediction of apicoplast targeted sequences</td>
</tr>
<tr>
<td><strong>PlasMit</strong> - Prediction of mitochondrial transit peptides in <em>Plasmodium falciparum</em></td>
</tr>
<tr>
<td><strong>Predotar</strong> - Prediction of mitochondrial and plastid targeting sequences</td>
</tr>
<tr>
<td><strong>PTS1</strong> - Prediction of peroxisomal targeting signal 1 containing proteins</td>
</tr>
<tr>
<td><strong>SignalP</strong> - Prediction of signal peptide cleavage sites</td>
</tr>
<tr>
<td><strong>NetOGlyc</strong> - Prediction of O-GalNAc (mucin type) glycosylation sites in mammalian proteins</td>
</tr>
<tr>
<td><strong>NetNGlyc</strong> - Prediction of N-glycosylation sites in human proteins</td>
</tr>
<tr>
<td><strong>DictyOGlyc</strong> - Prediction of GlcNAc O-glycosylation sites in <em>Dictyostelium</em></td>
</tr>
<tr>
<td><strong>YinOYang</strong> - O-beta-GlcNAc attachment sites in eukaryotic protein sequences</td>
</tr>
<tr>
<td><strong>big-PI Predictor</strong> - GPI Modification Site Prediction</td>
</tr>
<tr>
<td><strong>DGPI</strong> - Prediction of GPI-anchor and cleavage sites (Mirror site)</td>
</tr>
<tr>
<td><strong>Myristoylator</strong> - Prediction of N-terminal myristoylation by neural networks</td>
</tr>
<tr>
<td><strong>NetPhos</strong> - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins</td>
</tr>
<tr>
<td><strong>NetPicoRNA</strong> - Prediction of protease cleavage sites in picornaviral proteins</td>
</tr>
<tr>
<td><strong>NMT</strong> - Prediction of N-terminal N-myristoylation</td>
</tr>
<tr>
<td><strong>Sulfinator</strong> - Prediction of tyrosine sulfation sites</td>
</tr>
<tr>
<td><strong>SUMOplot</strong> - Prediction of SUMO protein attachment sites</td>
</tr>
</tbody>
</table>
Protein targeting
Protein targeting

- In 1970, Gunter Blobel showed that proteins have an N-terminal signal sequence which directs proteins to the membrane.
- Proteins have to be transported to other organelles: nucleus, mitochondria,…
- Can we computationally identify the ‘signal’ which distinguishes the cellular compartment?
For transmembrane proteins, can we predict the transmembrane, outer, and inner regions?
Post-translational modification prediction

- **ChloroP** - Prediction of chloroplast transit peptides
- **LipoP** - Prediction of lipoproteins and signal peptides in Gram negative bacteria
- **MITOPROT** - Prediction of mitochondrial targeting sequences
- **PATS** - Prediction of apicoplast targeted sequences
- **PlasMit** - Prediction of mitochondrial transit peptides in Plasmodium falciparum
- **Predotar** - Prediction of mitochondrial and plastid targeting sequences
- **PTS1** - Prediction of peroxisomal targeting signal 1 containing proteins
- **SignalP** - Prediction of signal peptide cleavage sites

- **NetOGlyc** - Prediction of O-GalNAc (mucin type) glycosylation sites in mammalian proteins
- **NetNGlyc** - Prediction of N-glycosylation sites in human proteins
- **DictyOGlyc** - Prediction of GlcNAc O-glycosylation sites in Dictyostelium
- **YinOYang** - O-beta-GlcNAc attachment sites in eukaryotic protein sequences

- **big-PI Predictor** - GPI Modification Site Prediction
- **DGPI** - Prediction of GPI-anchor and cleavage sites (Mirror site)
- **Myristoylation** - Prediction of N-terminal myristoylation by neural networks
- **NetPhos** - Prediction of Ser, Thr and Tyr phosphorylation sites in eukaryotic proteins
- **NetPicoRNA** - Prediction of protease cleavage sites in picornaviral proteins
- **NMT** - Prediction of N-terminal N-myristoylation
- **Sulfinator** - Prediction of tyrosine sulfation sites
- **SUMOplot** - Prediction of SUMO protein attachment sites

Topology prediction

- **PSORT** - Prediction of protein subcellular localization
- **TargetP** - Prediction of subcellular location

- **DAS** - Prediction of transmembrane regions in prokaryotes using the Dense Alignment Surface method
- **HMMTOP** - Prediction of transmembrane helices and topology of proteins (Hungarian Academy of Sciences)
- **PredictProtein** - Prediction of transmembrane helix location and topology (Columbia University)
- **SOSUI** - Prediction of transmembrane regions (TUAT; Tokyo Univ. of Agriculture & Technology)
- **TMAP** - Transmembrane detection based on multiple sequence alignment (Karolinska Institut; Sweden)
- **TMHMM** - Prediction of transmembrane helices in proteins (CBS; Denmark)
- **TMpred** - Prediction of transmembrane regions and protein orientation (EMBnet-CH)
- **TopPred.2** - Topology prediction of membrane proteins (Stockholm University)
Multiple alignment tools

- CLUSTALW [At EBI, PBIL or at EMBnet-CH]
- T-Coffee [At EMBnet Switzerland or at CMBI]
- ALIGN - at Genestream (IGH)
- DIALIGN - Multiple sequence alignment based on segment-to-segment comparison, at University of Bielefeld, Germany
- Match-Box - at University of Namur, Belgium
- MSA - at Washington University

- Multalin [At INRA or at PBIL]
- MUSCA - Multiple sequence alignment using pattern discovery, at IBM

- AMAS - Analyse Multiply Aligned Sequences
- Bork's alignment tools - Various tools to enhance the results of multiple alignments (including consensus building).
- CINEMA - Color Interactive Editor for Multiple Alignments
- ESPript - Tool to print a multiple alignment

- WebLogo - Sequence logos at Berkeley/USA
- plogo - Sequence logos at CBS/Denmark
- GENIO/logo - Sequence logos at Stuttgart/Germany
- WebLogo - Sequence logos at Cambridge/UK
Tools for secondary structure prediction

- Each residue must be given a state: Helix, Loop, Strand
- HMMs/Neural networks are used to predict

<table>
<thead>
<tr>
<th>Secondary structure prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AGADIR</strong> - An algorithm to predict the helical content of peptides</td>
</tr>
<tr>
<td><strong>APSSP</strong> - Advanced Protein Secondary Structure Prediction Server</td>
</tr>
<tr>
<td><strong>GOR</strong> - Garnier et al., 1996</td>
</tr>
<tr>
<td><strong>HNN</strong> - Hierarchical Neural Network method (Guermeur, 1997)</td>
</tr>
<tr>
<td><strong>Ipred</strong> - A consensus method for protein secondary structure prediction at University of Dundee</td>
</tr>
<tr>
<td><strong>JUFO</strong> - Protein secondary structure prediction from sequence (neural network)</td>
</tr>
<tr>
<td><strong>nnPredict</strong> - University of California at San Francisco (UCSF)</td>
</tr>
<tr>
<td><strong>PredictProtein</strong> - PHDsec, PHDacc, PHDhtm, PHDtopology, PHDthreader, MaxHom, EvalSec from Columbia University</td>
</tr>
<tr>
<td><strong>Prof</strong> - Cascaded Multiple Classifiers for Secondary Structure Prediction</td>
</tr>
<tr>
<td><strong>PSA</strong> - BioMolecular Engineering Research Center (BMERC) / Boston</td>
</tr>
<tr>
<td><strong>PSIpred</strong> - Various protein structure prediction methods at Brunel University</td>
</tr>
<tr>
<td><strong>SOPMA</strong> - Geourjon and Deléage, 1995</td>
</tr>
<tr>
<td><strong>SSpro</strong> - Secondary structure prediction using bidirectional recurrent neural networks at University of California</td>
</tr>
</tbody>
</table>
Next topic: Gene finding