Genome Annotation

Following genome assembly and gene calling, genome scientists wish to classify protein coding genes by:
1. Predicted function – InterProScan
2. Predicted function, process and component – Gene Ontology, UniProt annotation transfer
3. Predicted process – pathway analysis

InterPro and InterProScan

- InterPro - a database which integrates predictive information about proteins’ function from partner resources, giving an overview of the families that a protein belongs to and the domains and sites it contains.
- Users who have novel nucleotide or protein sequences to functionally characterise can use the software package InterProScan to run the scanning algorithms from the InterPro database.
- Sequences are submitted in FASTA format. Matches are calculated against all of the required member database’s signatures and the results output in a variety of formats.

Protein Signatures

Protein Signature = an amino acid sequence (not necessarily consecutive) associated with a protein characteristic.

Introduction to InterPro

What value are signatures?

- Better at finding proteins with common function
  - Find more distant homologues than BLAST
- Better at finding proteins with a degree of commonality
  - Associate proteins that share:
    - Function
    - Domains
    - Sequence
    - Structure
- Annotation of protein sequences
  - Define conserved regions of a protein
    - e.g. location and type of domains
    - key structural or functional sites

How are protein signatures made?

Protein family/domain

Build model

Search

Multiple sequence alignment

Significant matches

Refine

Protein signature

E-value of 1e-49
E-value of 5e-42
E-value of 7e-39
E-value of 1e-10
### InterPro Entry

- **Groups similar signature together**: Links related signatures
- **Adds extensive annotation**
- **Linked to other databases**
- **Structural information and viewers**

### The InterPro entry types

- **Family**: Proteins share a common evolutionary origin, as reflected in their related functions, sequences or structure.
- **Domain**: Biological units with defined boundaries.
- **Repeat**: Short sequences typically repeated within a protein.

### Link related signatures - relationships

**Parent - Child**
(subgroup of more closely related proteins)

<table>
<thead>
<tr>
<th>Parent</th>
<th>Child</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFAM</td>
<td>PROSITE</td>
</tr>
<tr>
<td>SMART</td>
<td>PROSITE</td>
</tr>
<tr>
<td>SMART</td>
<td>SMART</td>
</tr>
</tbody>
</table>

- **PTM**: Active site
- **Active Site**: Binding site
- **Conserved Site**: Conserved site

---

### Database

<table>
<thead>
<tr>
<th>Database</th>
<th>Basis</th>
<th>Institution</th>
<th>Focus</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfam</td>
<td>HMM</td>
<td>Sanger Institute</td>
<td>Sequence Alignment</td>
<td><a href="http://pfam.sanger.ac.uk/">http://pfam.sanger.ac.uk/</a></td>
</tr>
<tr>
<td>GenID</td>
<td>HMM</td>
<td>UCL</td>
<td>Structural Alignment</td>
<td><a href="http://pfam.jcvi.org">http://pfam.jcvi.org</a></td>
</tr>
<tr>
<td>Superfamily</td>
<td>HMM</td>
<td>Univ. of Bristol</td>
<td>Evolutionary Domain</td>
<td><a href="http://supfam.jcvi.org">http://supfam.jcvi.org</a></td>
</tr>
<tr>
<td>SMART</td>
<td>HMM</td>
<td>EMBL Heidelberg</td>
<td>Functional Domain</td>
<td><a href="http://smart.embl-heidelberg.de">http://smart.embl-heidelberg.de</a></td>
</tr>
<tr>
<td>TIGRfam</td>
<td>HMM</td>
<td>J. Craig Venter Inst.</td>
<td>Structural Domain</td>
<td><a href="http://gene3d.biochem.ucl.ac.uk/Gene3D/">http://gene3d.biochem.ucl.ac.uk/Gene3D/</a></td>
</tr>
<tr>
<td>Panther</td>
<td>HMM</td>
<td>Univ. S. California</td>
<td>Structural Domain</td>
<td><a href="http://www.pantherdb.org">http://www.pantherdb.org</a></td>
</tr>
<tr>
<td>PfSSF</td>
<td>HMM</td>
<td>PIR, Georgeiah,</td>
<td>Structural Domain</td>
<td><a href="http://pir.georgetown.edu">http://pir.georgetown.edu</a></td>
</tr>
<tr>
<td>PRINTS</td>
<td>HMM</td>
<td>Uni of Manchester</td>
<td>Structural Domain</td>
<td><a href="http://www.bioinf.man.ac.uk/">http://www.bioinf.man.ac.uk/</a></td>
</tr>
<tr>
<td>PROSITE</td>
<td>Pfam &amp; Patterns &amp; Profiles</td>
<td>SIB</td>
<td>Functional Annotation</td>
<td><a href="http://www.ebi.ac.uk/PROSITE/">http://www.ebi.ac.uk/PROSITE/</a></td>
</tr>
<tr>
<td>HAMP</td>
<td>Pfam</td>
<td>SIB</td>
<td>Structural Domain</td>
<td><a href="http://www.ebi.ac.uk/PROSITE/">http://www.ebi.ac.uk/PROSITE/</a></td>
</tr>
</tbody>
</table>

---

### Foundations of InterPro

- **Integration of signatures**
- Manual curation

---

### InterPro Consortium

Consortium of 11 major signature databases

---

### Introduction to InterPro

- EMBL-EBI
InterPro Search

wwwdev.ebi.ac.uk/interpro

InterProScan – Searching New Sequence

wwwdev.ebi.ac.uk/interpro

InterProScan

- The InterProScan web service allows the user to only submit one sequence at a time can put the submission commands into a script, so multiple sequences will be chained.
  - http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_rest (REST)
  - http://www.ebi.ac.uk/Tools/webservices/services/pfa/iprscan_soap (SOAP)
- Or download the standalone version to your local servers

InterProScan New Search Results

wwwdev.ebi.ac.uk/interpro

InterProScan

Version 5.2-45.0 Nov 2013)

- Supported Platforms
  - 64 bit Linux
  - 32 bit Linux
- There are no versions planned for Windows or Apple (MAC OS X) operating systems.
- This is due to constraints in the various third-party binaries that InterProScan runs.
InterProScan 5.0

- How is InterProScan 5 different to InterProScan 4?
  - New analysis type: Phobius for transmembrane and signal peptide prediction
  - New feature: ability to map InterPro results back to the original nucleotide sequences that were submitted
  - New feature: option to look up biological pathways that the protein is potentially involved in
  - New output formats: “IMPACT” XML format and GFF3.0
  - Improved graphical (HTML and SVG) representations of the protein matches
  - InterProScan 4.8 is no longer supported or updated.

- InterP Roc Scan 5 allows components of the analysis to be farmed out on an LSF or SGE cluster.

- Large quantity of manually-curated entries
  - 133,026 signatures, 32,535 integrated into 25,326 entries
  - Total number of GO terms mapped to InterPro entries - 27,511

- Regularly updated
  - 8 week release schedule
  - New signatures added
  - Scanned against latest version of UniProtKB

Why use InterPro?

- InterPro as a tool for Automatic Annotation
  - Automated clean-up of annotation from original nucleotide sequence entry
  - Additional value added by using automatic annotation
  - Recognises common annotation belonging to a closely related family within UniProtKB/Swiss-Prot
  - Identifies all members of this family using pattern/motif/HMMs in InterPro
  - Transfers common annotation to related family members in TrEMBL

- Mapping genes to pathways
  - Reactome a database of manually curated human pathways
  - Pathways steps (reactions) used to infer orthologous events in 20 non-human species (including plants, animals, protists, fungi, metazoa and bacteria)
  - An open-source suite of tools – help given to create species-specific Reactomes

Data Expansion – Projecting to Other Species

Human

\[ A + ATP \rightarrow A-P + ADP \]

Mouse

\[ A + ATP \rightarrow A-P + ADP \]

Drosophila

\[ A + ATP \rightarrow \text{Reaction not inferred} \]

No orthologue - Protein not inferred
Front Page  http://www.reactome.org

Pathway Hierarchy Panel

The Pathway Browser

Boxes are proteins, protein sets, mixed sets or complexes.
Ovals are small molecules (or sets of)
Green boxes are proteins or sets, blue are complexes.

Expression data overlay

http://www.reactome.org
Summary

- Multiple resources available to support genome annotation
- Help is at hand if you need it. First step – visit us at stand 303 to discuss