Getting To Know Your Protein

Comparative Protein Analysis:
Part II. Protein Domain Identification & Classification

Robert Latek, PhD
Sr. Bioinformatics Scientist
Whitehead Institute for Biomedical Research
Comparative Protein Analysis

• Part I. :
  – Phylogenetic Trees and Multiple Sequence Alignments are important tools to understand global relationships between sequences.
  – Tree Building Tools with Different Algorithms
  – Tree Reliability
    • Bootstrapping 1. Randomly re-sample MSA columns to produce a random alignment (equal length as original MSA), 2. Build tree based on random alignment, 3. Predicted branches are significant if they occur in ~ >70% of the trees from multiple, randomized alignments.
    • Use a several tree building algorithms to determine whether they produce similar trees as the original.
Comparative Protein Analysis

• Part II.:
  – How do you identify sequence relationships that are restricted to localized regions?
  – Can you apply phylogenetic trees and MSAs to only sub-regions of sequences?
  – How do you apply what you know about a group of sequences to finding additional, related sequences?
  – What can the relationship between your sequences and previously discovered ones tell you about their function?

• Assigning sequences to Protein Families
Syllabus

• **Protein Families**
  – Identifying Protein Domains
  – Family Databases & Searches

• **Searching for Family Members**
  – Pattern Searches
    • Patscan
  – Profile Searches
    • PSI-BLAST/HMMER2
Proteins As Modules

• Proteins are derived from a limited number of basic building blocks (Domains)

• Evolution has shuffled these modules giving rise to a diverse repertoire of protein sequences

• As a result, proteins can share a global or local relationship

(Higgins, 2000)
Protein Domains

Motifs describe the domain

Janus Kinase 2 Modular Sequence Architecture

SH2 Motif

WIBR Bioinformatics Courses, © Whitehead Institute, 2005
Protein Families

• **Protein Family** - a group of proteins that share a common function and/or structure, that are potentially derived from a common ancestor (set of homologous proteins)

• **Characterizing a Family** - Compare the sequence and structure patterns of the family members to reveal shared characteristics that potentially describe common biological properties

• **Motif/Domain** - sequence and/or structure patterns common to protein family members (trait/feature/characteristic)
Protein Families

Separate Families Can Be Interrelated

Proteins Can Belong To Multiple Families
Creating Protein Families

- Use domains to identify family members
  - Use a sequence to search a database and characterize a pattern/profile
  - Use a specific pattern/profile to identify homologous sequences (family members)
Family Database Resources

- **Curated Databases***
  - Proteins are placed into families with which they share a specific sequence pattern

- **Clustering Databases***
  - Sequence similarity-based without the prior knowledge of specific patterns

- **Derived Databases***
  - Pool other databases into one central resource

- **Search and Browse**
  - **InterPro** [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)

*(Higgins, 2000)*
Curated Family Databases

- **Pfam** (http://pfam.wustl.edu) **
  - Uses manually constructed seed alignments and PSSM to automatically extract domains
  - Db of protein families and corresponding profile-HMMs of prototypic domains
  - Searches report e-value and bits score
- **Prosite** (http://www.expasy.ch/tools/scanprosite/)
  - Hit or Miss -> no stats
- **PRINTS** (http://www.bioinf.man.ac.uk/fingerPRINTScan/)
Clustering Family Databases

- Search a database against itself and cluster similar sequences into families
  - Searchable against MSAs and consensus sequences
- **Protomap** (http://protomap.cornell.edu/)
  - Swiss-Prot based and provides a tree-like view of clustering
Derived Family Databases

- Databases that utilize protein family groupings provided by other resources
- **Blocks** - Search and Make (http://blocks.fhcrc.org/blocks/)
  - Uses Protomap system for finding blocks that are indicative of a protein family (GIBBS/MOTIF)
- **Proclass** (http://pir.georgetown.edu/gfserver/proclass.html)
  - Combines families from several resources using a neural network-based system (relationships)
- **MEME** (http://meme.sdsc.edu/meme/website/intro.html)
Syllabus

• Protein Families
  – Identifying Protein Domains
  – Family Databases & Searches

• Searching for Family Members
  – Pattern Searches
    • Patscan
  – Profile Searches
    • PSI-BLAST/HMMER2
Searching Databases By Family

• BLAST searches provide a great deal of information, but it is difficult to select out the important sequences (listed by score, not family)

• Family searches can give an immediate indication of a protein’s classification/function

• Use Family Database search tools to identify domains and family members
Patterns & Profiles

- Techniques for searching sequence databases to uncover common domains/motifs of biological significance that categorize a protein into a family
- **Pattern** - a deterministic syntax that describes multiple combinations of possible residues within a protein string
- **Profile** - probabilistic generalizations that assign to every segment position, a probability that each of the 20 aa will occur
Pattern Discovery Algorithms

• Pattern Driven Methods
  – Enumerate all possible patterns in solution space and try matching them to a set of sequences

```
1  2  3  4
A     A     A     A
C     C     C     C
D     D     D     D
... ...
Y     Y     Y     Y

1  2  3  4
Poss.1 A     A     A     A
Poss.2 A     C     A     A
Poss.3 A     A     C     A
Poss.4 A     A     A     C
Poss.5 A     C     C     A
Etc.
```
Pattern Discovery Algorithms

• Sequence Driven Methods
  – Build up a pattern by pair-wise comparisons of input sequences, storing positions in common, removing positions that are different
Pattern Building

- Find patterns like “pos1 xx pos2 xxxx pos3”
  - Definition of a non-contiguous motif

Pattern Properties

• Specification
  – a single residue K, set of residues (KPR), exclusion {KPR}, wildcards X, varying lengths x(3,6) -> variable gap lengths

• General Syntax
  – C-x(2,4)-C-x(3)-[LIVMFYWC]-x(8)-H-x(3,5)-H

• Patscan Syntax
  http://jura.wi.mit.edu/bio/education/bioinfo/homework/hw8/patscan.txt
  – C 2…4 C 3…3 any(LIVMFYWC) 8…8 H 3…5 H

• Pattern Database Searching
  – %scan_for_matches -p pattern_file < nr > output_file
Sequence Pattern Concerns

• Pattern descriptors must allow for approximate matching by defining an acceptable distance between a pattern and a potential hit
  – Weigh the sensitivity and specificity of a pattern

• What is the likelihood that a pattern would randomly occur?
Sequence Profiles

- **Consensus** - mathematical probability that a particular **aa** will be located at a given position
- **Probabilistic** pattern constructed from a MSA
- Opportunity to assign penalties for insertions and deletions, but not well suited for variable gap lengths
- **PSSM** - (Position Specific Scoring Matrix)
  - Represents the sequence profile in tabular form
  - Columns of weights for every **aa** corresponding to each column of a MSA
Profile Discovery/Analysis

- Perform global MSA on group of sequences
- Move highly conserved regions to smaller MSAs
- Generate scoring table with log odds scores
  - Each column is independent
  - Average Method: profile matrix values are weighted by the proportion of each amino acid in each column of MSA
  - Evolutionary Method: calculate the evolutionary distance (Dayhoff model) required to generate the observed amino acid distribution
**PSSM Example**

(i.e. Distribution of aa in an MSA column)

**Resulting Consensus:** \(\text{ITLS}\)

---

### Target sequences

1. ITISS
2. TDLS
3. VTMG
4. ITIG
5. VGFS
6. IELT
7. TTTS
8. ITLS

---

| PSSM | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
| 1    | 8 | -2| 5 | 4 | 5 | 5 | -4| 24| 0 | 15| 13| 1 | 1 | 1 | -7| 2 | 22| 21| -18| -6|
| 2    | 13| -5| 24| 18| -18| 19| 7 | 1 | 7 | -7| -4| 14| 11| 10| -1| 9 | 29| 3 | -28| -14|
| 3    | 5 | -5| 3 | 4 | 13| 4 | 2 | 8 | -4| 14| 12| 8 | -5| 0 | -10| 0 | 10| 10 | -1 | 5 |
| 4    | 17| 17| 13| 10| -12| 29| -5| -5| -5| -14| -9| 12| 10| 0 | -2| 34| 19| 1 | -8 | -15|
PSSM Properties

• Score-based sequence representations for searching databases

• Goal
  – Limit the diversity in each column to improve reliability

• Problems
  – Differing length gaps between conserved positions (unlike patterns)
PSI-BLAST Implementation

- **PSI-BLAST**
  - Start with a sequence, BLAST it, align select results to query sequence, estimate a profile with the MSA, search DB with the profile - constructs PSSM
  - Iterate until process stabilizes
  - Focus on domains, not entire sequences
  - Greatly improves sensitivity (but may affect specificity)

[Diagram showing the PSI-BLAST process flow]

WIBR Bioinformatics Courses, © Whitehead Institute, 2005
PSI-BLAST Sample Output

Sequences with E-value WORSE than threshold

- gi|9629055|ref|NP_044074.1| (NC_001731) MC123R [Molluscum contag... 37 0.16
- gi|8176554|gb|AAB35488.2| (S79774) bile salt-dependent lipase; B... 36 0.25
- gi|450277|ref|NP_001798.1| (NM_01807) carboxyl ester lipase (b... 35 0.86
- gi|231629|sp|P19835|BAL_HUMAN Bile-salt-activated lipase precurs... 35 0.89
- gi|15242929|ref|NP_200612.1| (NM_125189) putative protein [Arabi... 34 1.1
- gi|9759529|gb|BAB10995.1| (AB024029) gene_id:K21L19.3~unknown p... 34 1.3
- gi|180482|gb|AAA52014.1| (M85201) cholesterol esterase [Homo sap... 33 1.8
- gi|118706|sp|P21173|DNAA_MICLU Chromosomal replication initiator... 32 4.6
- gi|126679|sp|P16110|LEG3_MOUSE GALECTIN-3 (GALACTOSE-SPECIFIC LE... 32 4.9
- gi|52851|emb|CAA34206.1| (X16074) L-34 protein (AA 1-264) [Mus sp.] 32 5.0
- gi|539207|pir|A45983 lactose-binding lectin Mac-2 - mouse 32 5.0
- gi|387111|gb|AAA37311.1| (J03723) carbohydrate binding protein 3... 32 5.4
- gi|9506427|ref|NP_062019.1| (NM_019146) bassoon [Rattus norvegic... 32 5.5
HMM Building

- **Hidden Markov Models** are Statistical methods that consider all the possible combinations of matches, mismatches, and gaps to generate a consensus (Higgins, 2000)
- Sequence ordering and alignments are not necessary at the onset (but in many cases alignments are recommended)
- Ideally use at least 20 sequences in the training set to build a model
- Calibration prevents over-fitting training set (i.e. Ala scan)
- Generate a model (profile/PSSM), then search a database with it
HMM Implementation

- **HMMER2** (http://hmmer.wustl.edu/)
  - Determine which sequences to include/exclude
  - Perform alignment, select domain, excise ends, manually refine MSA (pre-aligned sequences better)
  - Build profile
    - `%hmmbuild [-options] <hmmfile output> <alignment file>`
  - Calibrate profile (re-calc. Parameters by making a random db)
    - `%hmmcalibrate [-options] <hmmfile>`
  - Search database
    - `%hmmsearch [-options] <hmmfile> <database file> > out`
HMMER2 Output

- Hmmsearch returns e-values and bits scores
- Repeat process with selected results
  - Unfortunately need to extract sequences from the results and manually perform MSA before beginning next round of iteration

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
<th>Score</th>
<th>E-value</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi16131263</td>
<td>phosphoglycolat</td>
<td>168.4</td>
<td>2.9e-45</td>
<td>1</td>
</tr>
<tr>
<td>gi24114648</td>
<td>phosphoglycolat</td>
<td>167.8</td>
<td>4.2e-45</td>
<td>1</td>
</tr>
<tr>
<td>gi15803888</td>
<td>phosphoglycolat</td>
<td>167.8</td>
<td>4.2e-45</td>
<td>1</td>
</tr>
<tr>
<td>gi26249979</td>
<td>Phosphoglycolat</td>
<td>166.4</td>
<td>1.1e-44</td>
<td>1</td>
</tr>
</tbody>
</table>
Patterns vs. Profiles

• Patterns
  – Easy to understand (human-readable)
  – Account for different length gaps

• Profiles
  – Sensitivity, better signal to noise ratio
  – Teachable
Domain ID & Searching

- Family/Domain Search
  - http://pfam.wustl.edu

- Pattern Search
  - scan_for_matches (Patscan)
    - scan_for_matches -p pattern_file < /cluster/db0/Data/yeast.aa > output_file

- Profile Search
  - HMMER2
    - hmmbuild [-options] <hmmfile output> <alignment file>

WIBR Bioinformatics Courses, © Whitehead Institute, 2005
Exercises

• Use PFAM to identify domains within your sequence
• Scan your sequences with ProSite to find a pattern to represent the domain
• Use the ProSite pattern to search the non-redundant db
• Use PSI-BLAST to build a sequence profile and search the non-redundant db
References