Comparative Protein Analysis

Part II: Protein Domain Identification & Classification

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

- 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

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 > 50% of the trees from multiple, randomized alignments.
  - Use a several tree building algorithms to determine whether they produce similar trees as the original.

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

Syllabus

- Protein Families
  - Identifying Protein Domains
  - Family Databases & Searches
- Searching for Family Members
  - Pattern Searches
    - Patscan
  - Profile Searches
    - PSI-BLAST/HMMER2
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)

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)

Curated Family Databases

- **Pfam** (http://pfam.wustl.edu)**
  - Uses manually constructed seed alignments and PSSM to automatically extract domains
  - dh 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/)

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/ (Hughey, 2000)

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.thec.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)

<table>
<thead>
<tr>
<th>Name</th>
<th>Combined profile</th>
<th>Motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td>mem.exp.157</td>
<td>2.5±e-1</td>
<td></td>
</tr>
</tbody>
</table>

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

- 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
  
  1. CYD -- C A F T L R Q S A V M R K H A R E H
  6. C L H N T C T A F W R Q K K D D T V H S L H

  C xxxx C xxxx [LIVMFW] xxxxxxx H xxxx H


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?

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
  - 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

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

1. I T E E
2. T L E L
3. V T M G
4. I E L I
5. G F T G
6. L E L T
7. T E L E
8. B I L L

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

Target sequences

Resulting Consensus: I T L S

PSSM

1  2  3  4  5  6  7  8  9 10
A 1 2 3 4 5 6 7 8 9 10
C 11 12 13 14 15 16 17 18 19 20
D 1 2 3 4 5 6 7 8 9 10
E 11 12 13 14 15 16 17 18 19 20
F 11 12 13 14 15 16 17 18 19 20
G 11 12 13 14 15 16 17 18 19 20
H 11 12 13 14 15 16 17 18 19 20
I 1 2 3 4 5 6 7 8 9 10
K 11 12 13 14 15 16 17 18 19 20
L 11 12 13 14 15 16 17 18 19 20
M 11 12 13 14 15 16 17 18 19 20
N 11 12 13 14 15 16 17 18 19 20
P 1 2 3 4 5 6 7 8 9 10
Q 11 12 13 14 15 16 17 18 19 20
R 11 12 13 14 15 16 17 18 19 20
S 11 12 13 14 15 16 17 18 19 20
T 11 12 13 14 15 16 17 18 19 20
V 1 2 3 4 5 6 7 8 9 10
W 11 12 13 14 15 16 17 18 19 20
Y 11 12 13 14 15 16 17 18 19 20

C R A T E

Position
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)

HMM Implementation

- HMMER2
  - 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
    - `[%hmssearch [options] <hmmfile] <database file> > out`

HMMER2 Output

- Hmmssearch 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

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

PSI-BLAST Sample Output

- 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)

PSI-BLAST Implementation

- Sequence
  - Align
  - Blast
  - Select Results
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>`

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