Bioinformatics

Proteins II. -
Pattern, Profile, & Structure
Database Searching

Robert Latek, Ph.D.
Bioinformatics, Biocomputing

Proteins I.-III. - Syllabus

- Proteins I.
  - Phylogenetic Trees
  - Multiple Sequence Alignments
- Proteins II.
  - Searching for Homologous Sequences
  - Working with Protein Structure Information
- Proteins III.
  - Comparing Protein Structures
  - Building Structural Models

Last Week

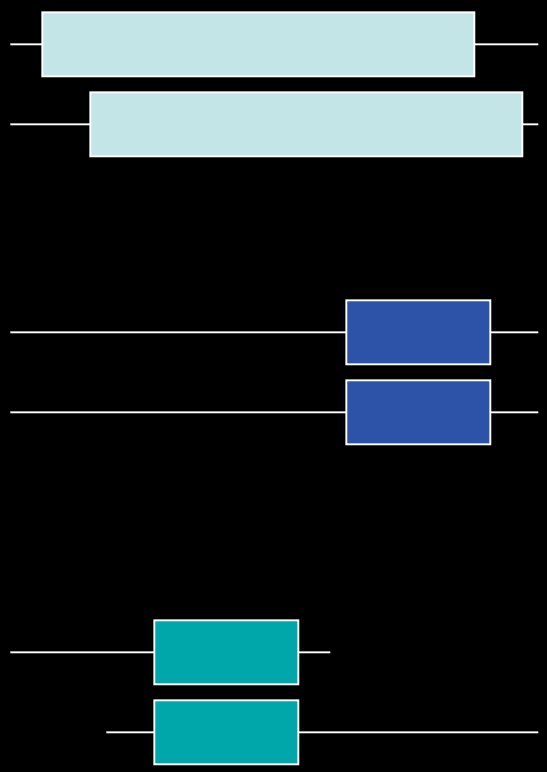
- Phylogenetic Trees
  - Trees show the relationship between sequences
  - Approaches: Maximum Parsimony, Distance, Maximum Likelihood
  - Distance scores should be considered secondary to tree shape
- Multiple Sequence Alignments
  - Approaches
    - Global: Dynamic Programming, Progressive, Iterated
    - Local: Profiles, Block-Based, Motif-Based
  - Manual manipulation of alignments with applications like Jalview is acceptable and recommended (esp. for pattern identification)
  - Scoring is usually the Sum of Pair-wise alignment scores

Proteins II. - Syllabus

- Searching for Homologous Sequences
  - Pattern Searches
    - Patscan
  - Profile Searches
    - PSI-BLAST/HMMER2
- Working with Protein Structure Information
  - Coordinate Files, Databases, Classification
  - Structure Viewers

Protein 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)

Protein Families

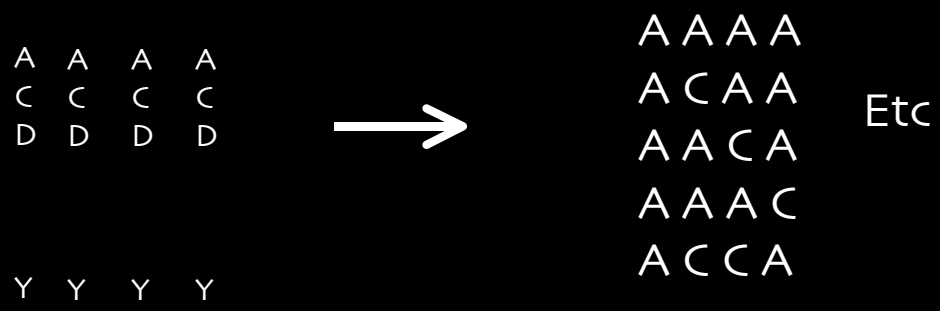
- Protein Family - a group of proteins that share a common function and/or structure, that are potentially derived from a common ancestor (set of homologous proteins)
- Characterizing a Family - Compare the sequence and structure patterns of the family members to reveal shared characteristics that potentially describe common biological properties
- Motif/Domain - sequence and/or structure patterns common to protein family members

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

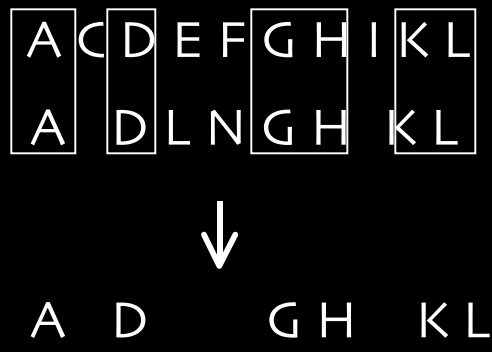
Pattern/Profile Discovery Algorithms

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



Pattern/Profile 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



Sequence Patterns

- 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://web.wi.mit.edu/bio/pub/patscan.html)
  - C 2…4 C 3…3 any(LIVMFYWC) 8…8 H 3…5 H
- Patscan command
  - %scan_for_matches -p pattern_file < /db0/Data/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 an aa will be located at a given position
- Probabilistic pattern constructed from a MSA
- Opportunity to assign penalties for insertions and deletions (not well suited for variable gap lengths)
- PSSM - Position Specific Scoring Matrix
  - Columns of weights for every aa corresponding to each column of a MSA

PSSM Example

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

PSSM Properties

- Score-based sequence representations for searching databases

- Scores Calculation
  - Log odds score representing: Distribution of aa in an MSA column, substitution matrices
- Goal
  - Limit the diversity in each column to improve reliability
- Problems
  - Differing length gaps between conserved positions (unlike patterns)

PSSM Weighting

- Differentially weight sequences to reduce redundancy from non-representative sampling
  - Similar sequences get low weights, diverged sequences get higher weights
  - Maximum discrimination: weights that best discriminate between positives and background
  - Use root to sequence distance to calculate weight
  - Average distance of a sequence to all other sequences

Building Profiles: Hidden Markov Models

- Statistical model that considers 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), then search a database with it
- Limitations
  - Not all potential domains are represented in each family database, therefore, should develop own models/profiles

Searching Family Databases

- 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
- Identify domains and family members
  - Use a sequence to search a database and characterize a pattern and identify its protein family
  - Use a specific pattern 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 a specific patterns
- Derived Databases*
  - Pool other databases into one central resource

*(Higgins)

Curated Family Databases

- Pfam (http://pfam.wustl.edu/hmmsearch.shtml/) **
  - Uses manually constructed seed alignments and HMM PSSMs to automatically extract domains
  - db of protein families and corresponding profile-HMMs
  - 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/)
  - Find fingerprints in your sequence

Clustering Family Databases

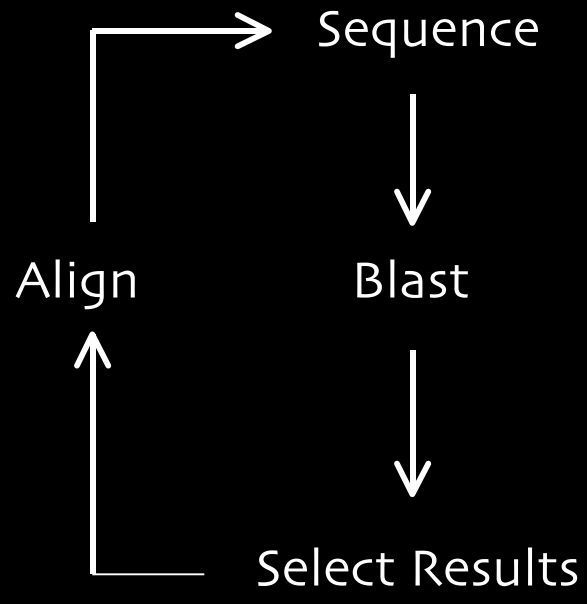
- Search a database against itself and cluster similar sequences into families
- ProDom (http://prodes.toulouse.inra.fr/prodom/doc/prodom.html)
  - Searchable against MSAs and consensus sequences
- Protomap (http://www.protomap.cs.huji.ac.il/)
  - Swiss-Prot based and provides a tree-like view of clustering

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)
  - Not an alignment-based system
- MEME (Locally available through GCG on fladda) (http://meme.sdsc.edu/meme/website/intro.html)

HMM Implementations
Profile Building & Searching

- PSI-BLAST
  http://www.ncbi.nlm.nih.gov/BLAST/
  - Start with a sequence, BLAST it, align select results to query sequence, estimate a profile with the MSA, search DB with the profile
  - Iterate until process stabilizes
  - Focus on domains, not entire sequences
  - Greatly improves sensitivity



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|4502771|ref|NP_001798.1|	(NM_001807) 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|dbj|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|539907|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 Implementations
Profile Building & Searching

- 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>

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

Validating HMMER2 Profiles

- Including false positives into the building process will reduce the accuracy of the profile (increase the noise)
- Validate by
  - Blast results against each other
  - Demonstrate that you can get from one sequence to another

Patterns vs. Profiles

- Patterns
  - Easy to understand (human-readable)
  - Account for different length gaps
- Profiles
  - Sensitivity, better signal to noise ratio
  - Teachable

Protein Structures

- Protein Structure Classification
- Coordinate Files
- Structure Coordinate Databases
- Structure Family Databases
- Structure Visualization

Protein Structure Classification

- Proteins can adopt only a limited number of possible 3D conformations
  - Combinations of α helices, β sheets, loops, and coils
- Completely different sequences can fold into similar shapes
- Protein Structure Classes
  - Class α: bundles of α helices
  - Class β: antiparallel β sheets (sandwiches and barrels)
  - Class α / β: parallel β sheets with intervening α helices
  - Class α + β: segregated α helices and antiparallel β sheets
  - Multidomain
  - Membrane/Cell surface proteins

(Higgins)

Coordinate Files

- Coordinate Data: location of a molecule's atoms in space (XYZ triple)
- XYZ triple is labeled with an atom, residue, chain, and molecule
  - Modified aa are labeled with X, H's not usually listed
- Data Representation
  - Chemistry Rules Approach: connect the dots utilizing a standard rules base to specify bond distances (not consistent among applications)
  - Explicit Bonding Approach: explicit bonding information is specified in the file (very consistent)

Coordinate File Formats

- MMDB http://www.ncbi.nlm.nih.gov/Structure/
  - ASN.1 standard data description language
- PDB
  - Column oriented, “flexible format”
  - Sequence - Explicit SEQRES and Implicit ATOM lines
- mmCIF
  - Chemical Interchange Format - relational db format
  - http://web.wi.mit.edu/proteins/education/1F3J.cif

Structure Coordinate Databases

- RCSB (Research Collaboratory for Structural Bioinformatics) http://www.rcsb.org/
  - Formally know as the Protein Data Bank at Brookhaven National Laboratories
  - Structure Explorer PDB search engine
    - Text and PDB ID (4 letter code) searching
- MMDB (Molecular Modeling Database @NCBI)
  - Compilation of structures represented in multiple formats
  - Provides structure summaries
  - BLAST sequences to search for available structures

Structure Families

- Divide structures into the limited number of possible structure families
  - Homologous proteins can be identified by examining their respective structures for conserved fold patterns
  - Representative members can be used for modeling sequences of unknown structure

Structure Family Databases

- SCOP: Structural Classification Of Proteins
  - based on a definition of structural similarities. Hierarchical levels to reflect evolutionary and structural relationships
- CATH: Classification by Class, Architecture, Topology, and Homology
  - classified first into hierarchical levels like SCOP
- FSSP: Fold classification based on Structure-structure alignment of proteins
  - based on structural alignment of all pair-wise combinations of proteins in PDB by DALI (used to id common folds and place into groups)
- MMDB
  - Aligns 3D structures based on similar arrangements of secondary structural elements (VAST)
- SARF
  - categorized on the basis of structural similarity, categories are similar to other dbs

Resource Links

- Protein data bank (PDB)
  - http://www.rcsb.org/pdb
- Molecular Modeling Database
  - http://www.ncbi.nlm.nih.gov/Structure/
- Structural Classification of Proteins SCOP
  - http://scop.mrc-lmb.cam.ac.uk/scop
- CATH
  - http://www.biochem.ucl.ac.uk/bsm

Next Week

- Proteins I.
  - Phylogenetic Trees
  - Multiple Sequence Alignments
- Proteins II.
  - Searching for Homologous Sequences
  - Working with Protein Structure Information
- Proteins III.
  - Comparing Protein Structures
  - Building Structural Models

Visualizing Structural Information

- Hand edit files
- Different representations of molecule
  - wire, backbone, space-filling, ribbon
- NMR ensembles
  - Models showing dynamic variation of molecules in solution
- VIEWERS
  - RasMol and Chime is the Netscape plug-in
    - http://www.umass.edu/microbio/rasmol/index2.html
  - Cn3D MMDB viewer (See in 3D) with explicit bonding
    - http://www.ncbi.nlm.nih.gov/Structure
  - SwissPDB Viewer
    - http://www.expasy.ch/spdbv/mainpage.html

References

- Bioinformatics: Sequence and genome Analysis. David W. Mount. CSHL Press, 2001.
- Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Andreas D. Baxevanis and B.F. Francis Ouellete. Wiley Interscience, 2001.
- Bioinformatics: Sequence, structure, and databanks. Des Higgins and Willie Taylor. Oxford University Press, 2000.