ICB Fall 2009

G4120: Introduction to Computational Biology

Oliver Jovanovic, Ph.D.
Columbia University
Department of
Microbiology & Immunology

Lecture 6
Introduction to
Protein Analysis
October 29, 2009

Analysis of Protein Sequences

Coding Region Prediction
Start/Stop, Uneven Positional Base Frequency, Hexamer Frequency, Hidden Markov Models (HMM)

Protein-Protein Comparison
Dotplots, Needleman-Wunsch, Smith-Waterman, FastA, BLAST

Functional Region Prediction
Motifs, Profiles, Hidden Markov Models

Secondary Structure Prediction
Kyte-Doolittle, Chou-Fasman, Markov Models, Neural Nets

Tertiary Structure Prediction
Threading, Homology Modeling, Model Verification, Ab-initio Modeling

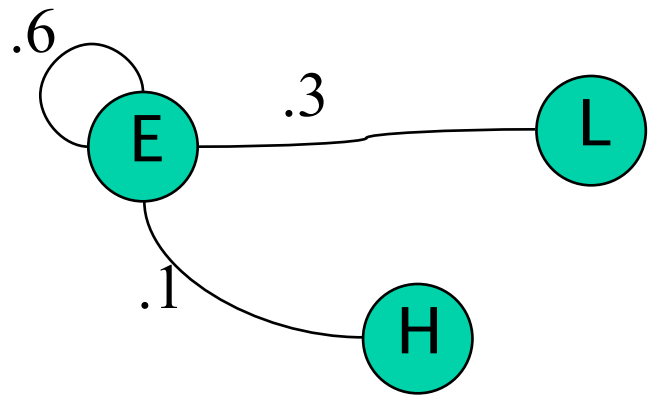
Stochastic Modeling

Stochastic Model

A model involving chance or probability.

Markov State

A Markov state emits a symbol each time you visit it. It connects to other states (and possibly itself), with transition probabilities attached. The sum of the transition probabilities is 1.



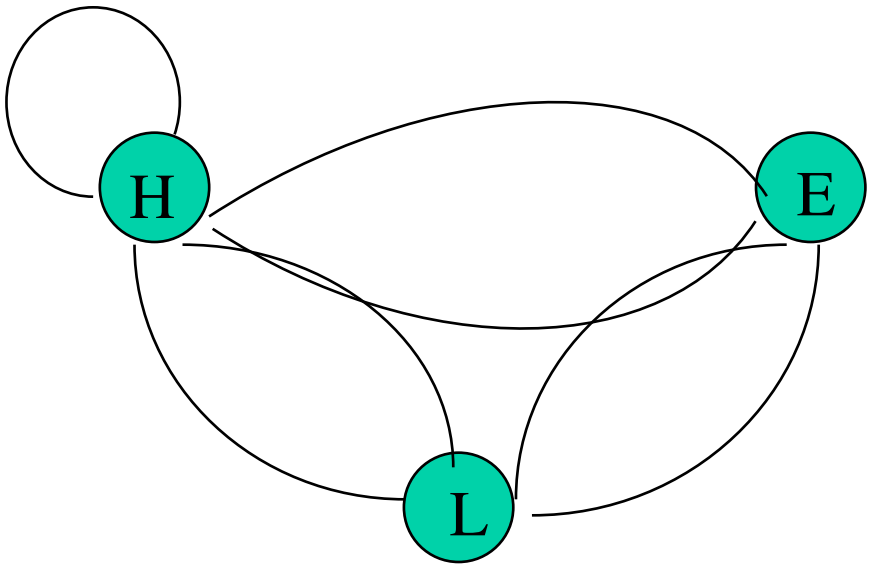
E = Extended
H = Helix
L = Loop

Source
http://www.bioinfo.rpi.edu

Markov Chains

Markov Chain

A Markov chain is an interlinked chain, or network, of states connected by transition probabilities.



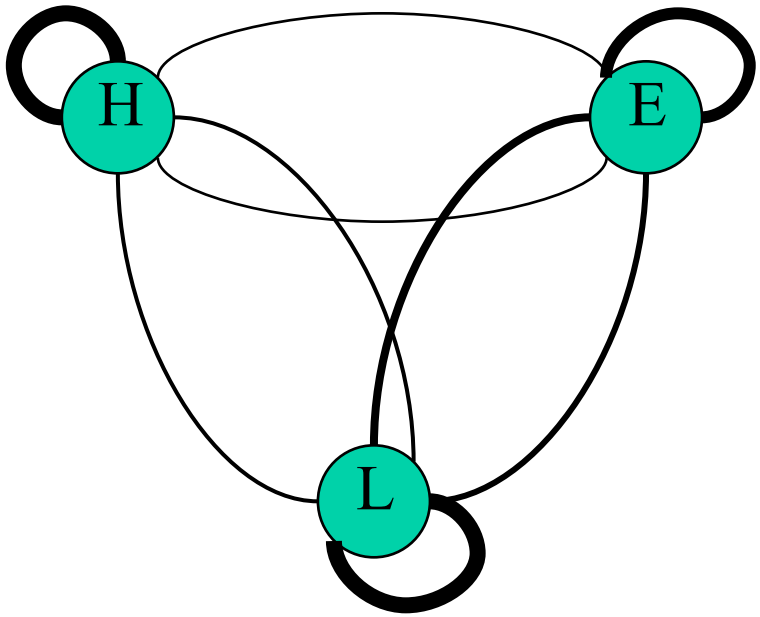
E = Extended
H = Helix
L = Loop

Source
http://www.bioinfo.rpi.edu

Markov Transition Matrices

Transition Matrix

A transition matrix for a first order Markov chain, the simplest kind. The sum of the transition probabilities from each state is 1.



	H	E	L
H	.93	.01	.06
E	.01	.80	.19
L	.04	.06	.90

E = Extended
H = Helix
L = Loop

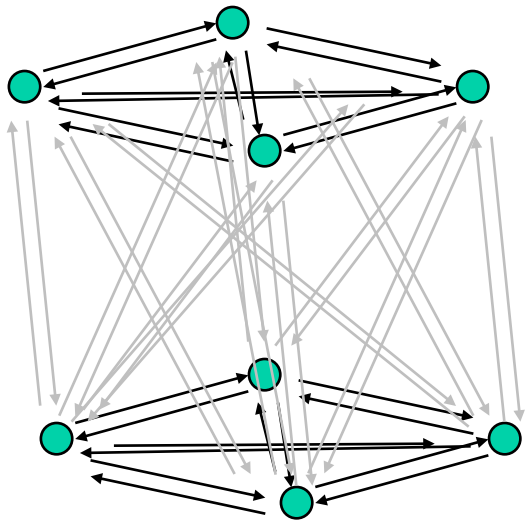
Source
http://www.bioinfo.rpi.edu

Hidden Markov Models

Hidden Markov Model (HMM)

A hidden Markov model consists of two Markov chains connected such that a one to one correspondence between the state and the emitted symbol no longer exists.

Model 1 —



— Transitions between models

Model 2 —

Source
http://www.bioinfo.rpi.edu

Coding Region Prediction

Start/Stop

Searches for start codons followed by a stop codon. Although a coding region must start and stop this way, this cannot predict the likelihood of a region to code for a protein product.
Example: DNA Strider

Uneven Positional Base Frequency

Noncoding regions possess a more random distribution of nucleotides. This method uses the relative abundance of nucleotides in each possible codon position to predict coding regions.
Example: Staden

Hexamer Frequency

The distribution of hexamer frequencies in coding and noncoding regions differs markedly, and can be used to predict coding regions with a high degree of accuracy (70-80%).
Example: SeqStat

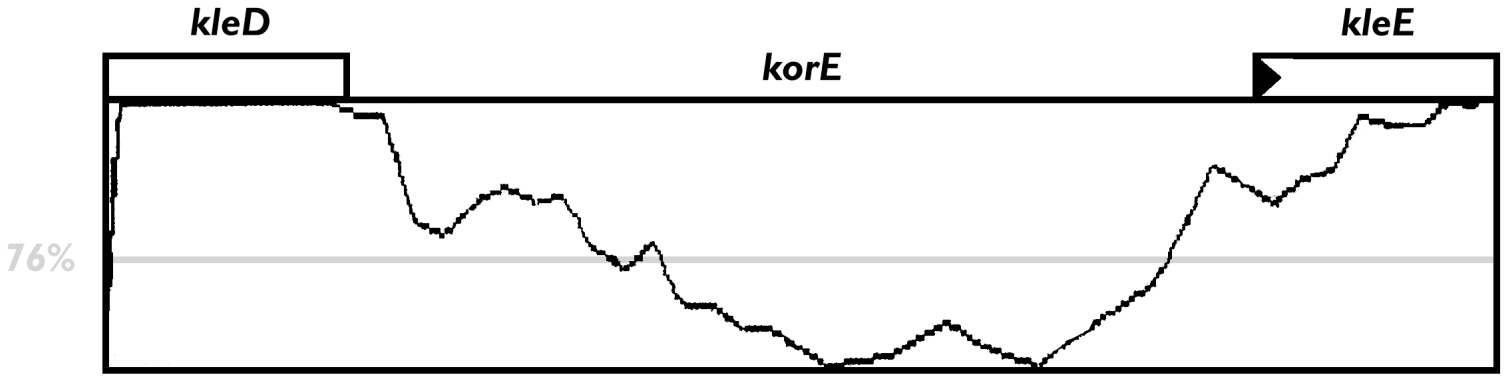
Neural Networks

Neural networks trained on known coding and noncoding regions in a particular species, can be used to predict new coding regions with a very high degree of accuracy (81-96%).
Examples: GrailEXP (**http://compbio.ornl.gov/grailexp/**)
For microbial genomes, Prodigal (**http://compbio.ornl.gov/prodigal/**)

Markov Models

Hidden Markov models, based on known coding and noncoding regions in a particular species, can be used to predict new coding regions with a very high degree of accuracy (92-99%).
Examples: GeneMark (**http://opal.biology.gatech.edu/GeneMark/**)
Glimmer and GlimmerHMM (**http://cbcb.umd.edu/software/**)
GenScan (**http://genes.mit.edu/GENSCAN.html**)

Uneven Positional Base Frequency



GeneMark

GeneMark and GeneMark.hmm

Mark Borodovsky, Georgia Institute of Technology
http://opal.biology.gatech.edu/GeneMark/

GeneMark

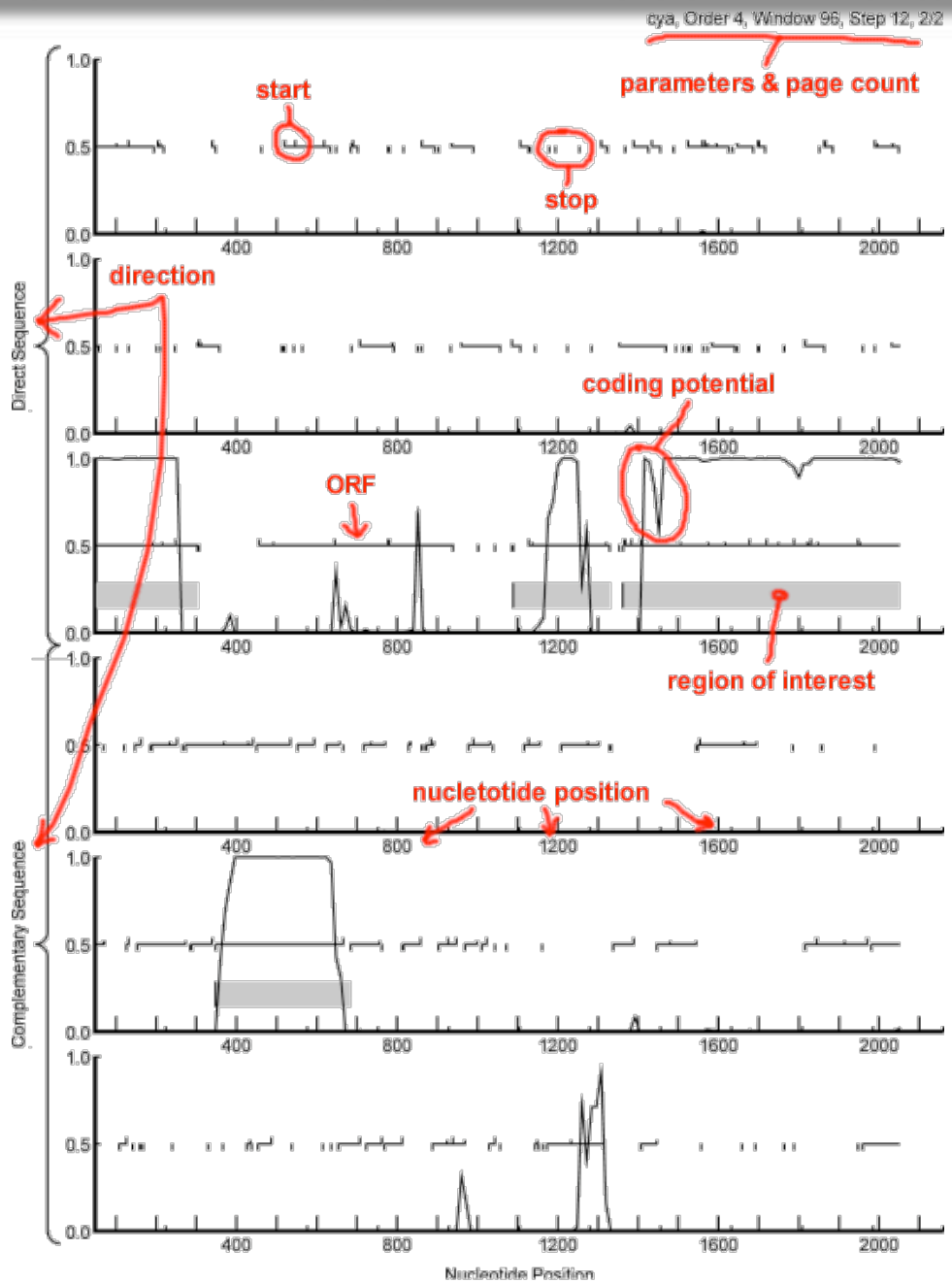
GeneMark evaluates the protein-coding potential of a DNA sequence (within a sliding window) by using Markov models of coding and non-coding regions for various species. This approach is sensitive to local variations of coding potential, and the GeneMark graph shows details of the coding potential distribution along a sequence.

GeneMark.hmm

GeneMark.hmm predicts genes and intergenic regions in a sequence as a whole using hidden Markov models with a hidden state network reflecting the "grammar" of gene organization. It identifies the most likely parse of the whole sequence into protein coding genes (with possible introns) and intergenic regions.

The two approaches can be run in parallel, using a combination of GeneMark-P and GeneMark.hmm-P for prokaryotic gene prediction, or GeneMark-E and GeneMark.hmm-E for eukaryotic gene prediction.

Example of GeneMark Results



Source
http://bioweb.pasteur.fr/docs/genemark/images/cyay.gif

Protein-Protein Comparison

Dot Matrix
Can compare proteins to each other using a dot matrix. When doing so, use a small window size (W1-3) and low stringency (S1-2).

Needleman-Wunsch (global)
Performs an optimal global alignment of two protein sequences.

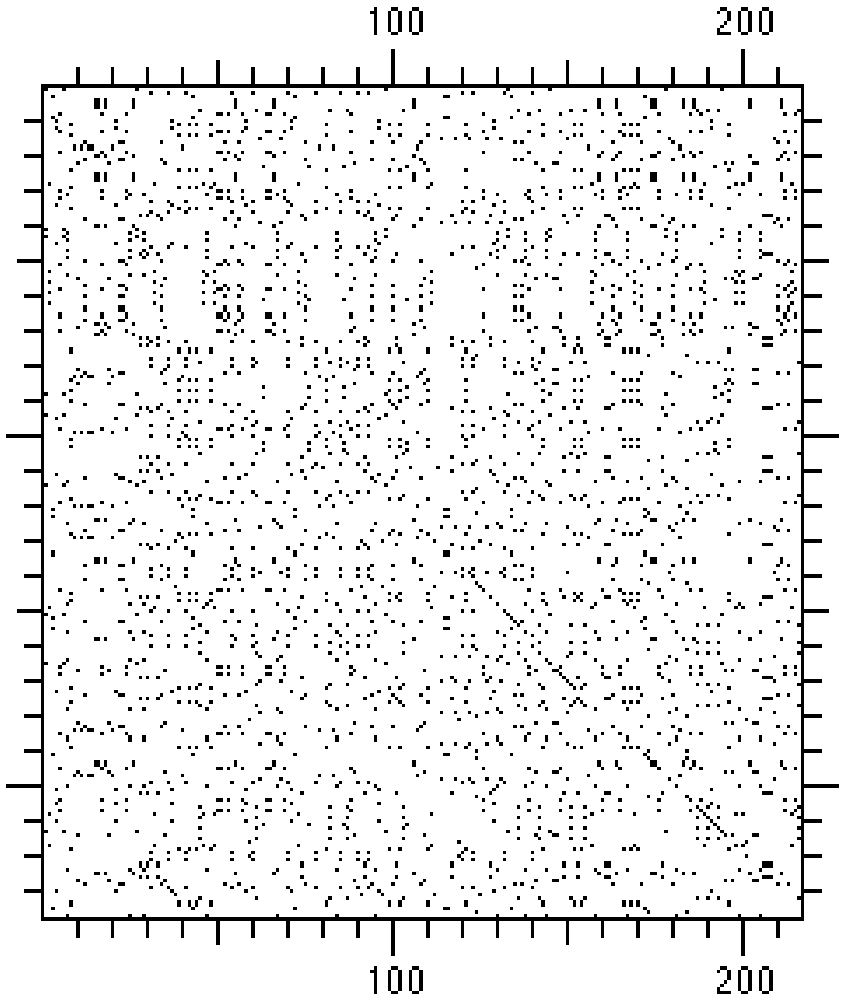
Smith-Waterman (local)
Performs an optimal local alignment of two protein sequences, useful for comparing conserved domains.

FastA (heuristic)
An older fast heuristic algorithm for comparing proteins.

BLAST (heuristic)
The most popular fast heuristic algorithm for protein comparison.

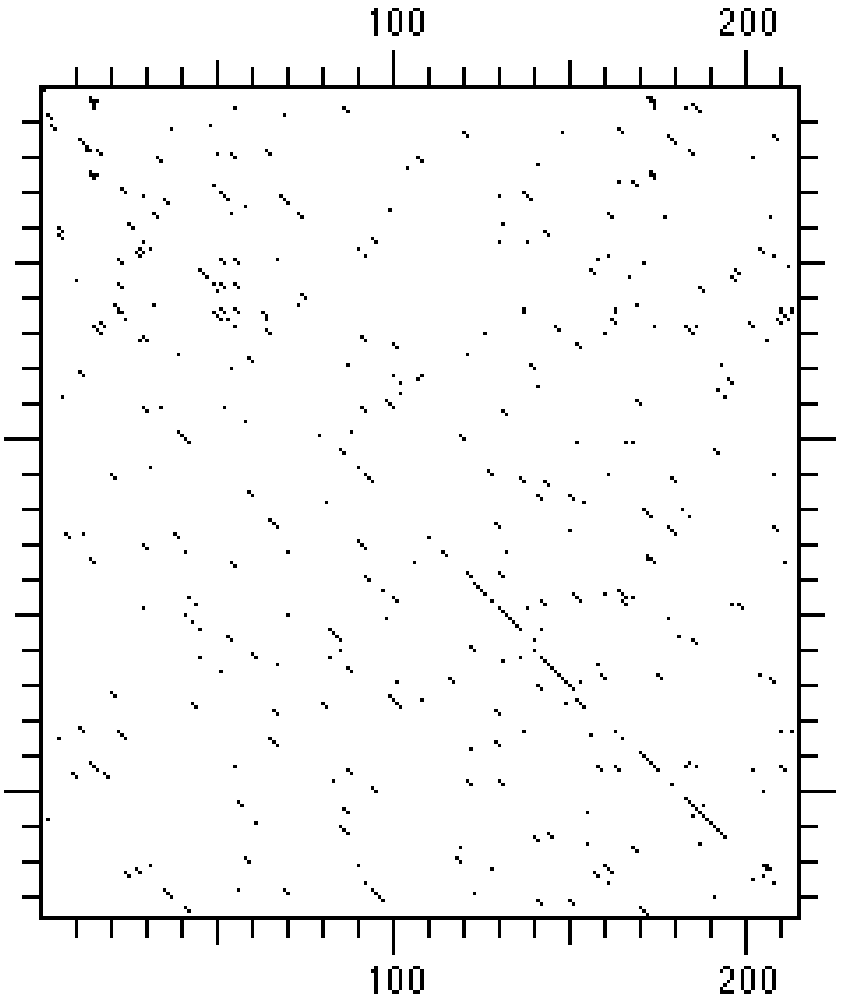
BLAT (heuristic)
A very fast heuristic algorithm related to BLAST and easy to run locally.

Protein-Protein Dot Matrix Analysis (W1, S1)



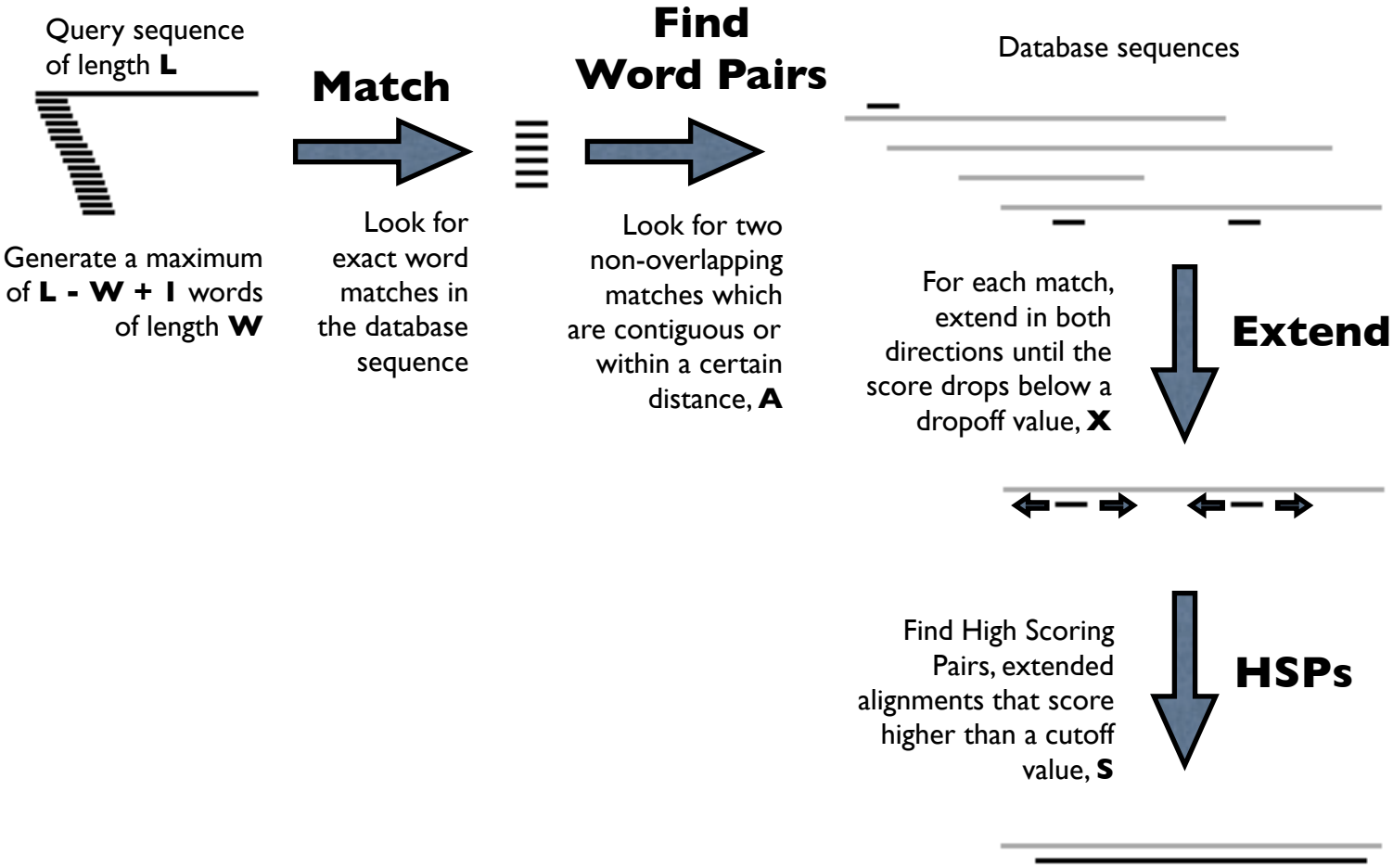
DNA Strider Settings
Vertical scale: phage lambda cI
Horizontal scale: phage P22 c2
Window size: 1
Stringency: 1

Protein-Protein Dot Matrix Analysis (W3, S2)



DNA Strider Settings
Vertical scale: phage lambda cI
Horizontal scale: phage P22 c2
Window size: 3
Stringency: 2

Protein BLAST Algorithm Illustrated



Protein Sequence Comparison

BLAST Protein Sequence Comparison

- A variety of BLAST programs are featured by NCBI for protein-protein comparison, including blastp (protein vs. protein), PSI-BLAST (position specific iterated) and PHI-BLAST (pattern hit initiated) and tblastn (protein vs. translated database).
- The default word size for protein BLAST searches is 3, this can be changed to 2 for more stringent, but slower search.
- The choice of Dayhoff substitution matrix can be important. The default matrix for NCBI BLAST protein comparison is BLOSUM62, which is optimized for long query sequences (over 85 aa) and known close homologies. When searching with short query sequences or distant homologies, be sure to try other matrices.

Rules of Thumb

- Proteins that are more than 30% identical throughout their entire lengths are likely homologous.
- Proteins that are 20-30% identical throughout their entire lengths may or may not be homologous (the "gray zone").
- Proteins that are less than 20% identical throughout their entire lengths are not likely homologous.
- Matches that are more than 50% identical in a 20-40 amino acid region occur frequently by chance.

Protein Functional Region Prediction

Motif

Uses a pattern derived from a number of known examples of a functional protein region. As this yields only a single consensus sequence, it is less accurate than the Profile or HMM methods.
Example: PROSITE (**http://us.expasy.org/prosite/**)

Profiles

Profiles are **statistical matrices** based on a family of known functional protein regions. They are more accurate than searching with a single consensus sequence.
Examples: Pfam (**http://pfam.sanger.ac.uk/**)
CDD (**http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml**)

Hidden Markov Models (HMM)

Hidden Markov models can be used to create statistical descriptions of a functional protein sequence family's consensus, which can then be used to accurately search for related functional domains.
Examples: SMART (**http://smart.embl-heidelberg.de/**)
HMMER (**http://hmmer.janelia.org/**)

Other Resources

The NCBI (**http://www.ncbi.nlm.gov**) and the Center for Biological Sequence Analysis (**http://www.cbs.dtu.dk/services/**) maintain resources for identifying protein sequence features. In addition, ExPASy maintains an extensive list of protein resources curated by Amos Bairoch (**http://www.expasy.org/links.html**).

Protein Domain Databases

CDD (Conserved Domain Database)

CDD is an NCBI database that contains conserved domains based on recurring sequence patterns or motifs derived from two popular collections, Smart and Pfam, as well as contributions from NCBI, such as COG. The source databases also provide descriptions and links to citations. Since conserved domains correspond to compact structural units, CDs contain links to 3D-structure via Cn3D whenever possible. Conserved Domains are indexed for retrieval by keywords; links between Conserved Domains and Proteins, PubMed, and Taxonomy have been added. Conserved Domains are also linked to other Conserved Domains by two different neighboring mechanisms. "Similar" domains are defined as those giving overlapping annotations on sets of protein sequences, "Co-occuring" domains are defined as those giving non-overlapping annotations on sets of protein sequences.

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

CD-Search (Conserved Domain Search)

CD-Search identifies conserved domains in a protein sequence by employing the reverse position-specific BLAST algorithm. The query sequence is compared to a position-specific score matrix prepared from the underlying conserved domain alignment. Hits may be displayed as a pairwise alignment of the query sequence with a representative domain sequence, or as a multiple alignment. CD-Search now is run by default in parallel with protein BLAST searches.

http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi

CDART (Conserved Domain Architecture Retrieval Tool)

CDART allows one to search for proteins with similar domain architectures. It uses precomputed CD-Search results to quickly identify proteins with a set of domains similar to that of the query.

http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi

Source

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Protein Secondary Structure Prediction

Kyte-Doolittle

A hydropathy plot that can indicate potential transmembrane or surface regions in proteins.
Accuracy: Scores -4.5 hydrophilic to 4.5 hydrophobic, poor for beta sheets (**DNA Strider**)

Chou-Fasman

A statistical approach to secondary structure prediction based on observed frequencies.
Accuracy: ~60% (**http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=misc1**)

PSA

A Markov model based approach to secondary structure prediction with detailed output.
Accuracy: ~70% (**http://bmerc-www.bu.edu/psa/**)

Predict Protein

A secondary structure prediction method based on a consensus of several complementary prediction methods, including PHD, which uses jury decision between a number of neural networks, enhanced by multiple sequence alignment information.
Accuracy: ~75% (**http://www.predictprotein.org/**)

Meta Predict Protein

Potentially combines a large number of different methods, including JPred and PHD, to detect functional motifs, transmembrane helices and other regions of interest as well as predict protein secondary structure and inter-residue contacts.
http://www.predictprotein.org/meta.php

Protein Tertiary Structure Prediction

Homology Modeling

Builds a model of a protein based on homologies to proteins of known structure. Can produce good results when proteins with significant homology and known structure exist.

Examples: Modeller (**http://salilab.org/modeller/modeller.html**)
SWISS-MODEL (**http://swissmodel.expasy.org/**)

Threading

Compares the fitness of protein sequence to assume various known tertiary structures. It assumes a particular fold, then evaluates the quality of the resulting structure. Can identify distantly related structural homologs and verify homology models.

Examples: 3D-PSSM, 123D, PHD (**http://www.predictprotein.org/**)

Model Verification

Checks the fitness of a protein sequence to assume a modeled fold.

Examples: VERIFY-3D, PROCHECK (**http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html**)

Ab-initio Structure Modeling

Predicts a model of a protein directly from the sequence. To date, limited accuracy, but improving.

Examples: RAMP, ROBETTA (**http://robetta.bakerlab.org/**)

Homology Modeling, Step-by-Step

1. **Identify related protein sequences** (BLAST, FastA)

2. **Align related protein sequences** (CLUSTALW, CLUSTALX)

3. **Construct model of core** (use conserved regions of existing structure or sequence to form core)

4. **Construct model of loops** (use known loop conformations or predictions)

5. **Construct model of side-chains** (use known rotamers or predictions)

6. **Evaluate model** (PROCHECK, Verify-3D)

3D Protein Structure Databases

MMDB (Molecular Modeling DataBase)

MMDB is the NCBI protein stucture database. It consists a subset of experimentally determined three-dimensional structures obtained from the Protein Data Bank (PDB) which have had errors and ambiguities removed, and then were converted to ASN.1 (Abstract Syntax Notation 1) format. The data is available thru Entrez or the free Cn3D 3D structure viewer. MMDB currently contains over 10,000 structure records, with approximately 80% of the structures determined by X-ray diffraction studies, the rest by NMR or other experimental methods. Links are provided to Medline records and the NCBI taxonomy databases through PDBeast. Related sequences are provided by BLAST, related structures are provided by VAST.

http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

VAST (Vector Alignment Search Tool)

The structural data of proteins in MMDB are compared against each other using the VAST algorithm for detecting significantly similar substructures. Entrez or Cn3D can be used to retrieve structures which seem highly similar to the query protein structure, in much the same way as sequence neighbors computed by BLAST. This will retrieve almost all structures with an identical 3D "fold", even in distantly related proteins, though it may occasionally miss a few or report chance similarities.

VAST functions by reducing x, y, z coordinate data for all alpha helices and beta sheets in a protein into vectors, then creating pairs of vectors called secondary structure elements (SSEs), which it attempts to superimpose. It is a heuristic approach, not an optimal one, and loses some information by converting substructures to vectors, but is extremely fast.

http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html

3D Visualization of Proteins

Cn3D

Cn3D "See in 3D" is a simple free 3D molecular structure viewer from the NCBI. It views MMDB ASN.1 formatted files. Cn3D may provide a useful first image of a structure, and is supported by NCBI.

Cn3D 4.1 for OS X, Windows or Unix

http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml

http://www.ncbi.nlm.nih.gov/Structure/CN3D/test_launch.prt

1) Copy Cn3D.app to /Applications
2) Firefox ➞ **http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dinstall.shtml** ➞ **Click here to test Cn3D 4.1!** ➞ Open with ➞ /Applications/Cn3D.app ➞ Do this automatically for files like this from now on ➞ OK
3) To use in Safari, in the Finder control click on test_launch.prt ➞ Open With ➞ Cn3D.app

Cn3D Tutorial

http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3dtut.shtml