



# A Very Very Very Short Introduction to Protein Bioinformatics

developed by  
Patricia Babbitt, Susan Johns, Leslie King & Sean Mooney  
University of California, San Francisco  
August 22-23, 2002

# Lecture 1

## Overview & Introduction to Database Searching and Pairwise Alignments

Patricia C. Babbitt  
Associate Professor  
Department of Biopharmaceutical Sciences  
& Biological and Medical Information Sciences Graduate Group

August 22, 2002

# What is Bioinformatics?

- Critical underpinning of the movement toward "large-scale biology" (Searls)
- Computer analysis of large assemblages of biological information...
- A subset of the larger field of computational biology (with a closely related field called computational chemistry)
- First appearance as a term in Medline: 1993

# Automation and Discovery Science

- **Advantages**
  - Big picture views
  - Automation saves time and money
  - Automation provides for more systematic, rules-based analyses
- **Disadvantages**
  - Insight is hard to automate
  - Biology is more complex than can be adequately modeled by currently available systems
  - Individual errors are hard to catch and correct

# Emerging Fields in Bioinformatics

- Data storage and retrieval, Database structures, Annotation
- Analysis of genomic/proteomic/other high-throughput information
- Evolutionary model building and phylogenetic analysis
- Architecture and content of genomes
- Complex systems analysis/genetic circuits
- Information content in DNA, RNA, protein sequence and structure
- Metabolic computing
- Data mining using machine learning tools, neural nets, AI
- Nucleic acid and protein sequence analysis

# Tools for Nucleic Acid Informatics

- codon usage, restriction maps
- primer design
- mapping
- identification of coding regions, repeats, translation
- identification of signals associated with gene regulation
- motif identification

# Tools for Protein Informatics

- sequence and structure comparison
- multiple alignments
- phylogenetic tree construction
- composition/pI/mass analysis
- motif/pattern identification
- 2° structure prediction/threading
- TMD prediction/hydrophobicity analysis
- homology modeling
- visualization

# Primary Web Resources

- **European Molecular Biology Laboratory, Germany**  
[http:// www.embl-heidelberg.de](http://www.embl-heidelberg.de)
- **ExPASy Molecular Biology Server, Swiss Institute of Bioinformatics, Switzerland**  
<http://ca.expasy.org/>
- **National Center for Biotechnology Information, USA**  
[http:// www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)  
<http://www3.ncbi.nlm.nih.gov/Entrez/>
- **San Diego Supercomputer Center, USA**  
[http:// www.sdsc.edu](http://www.sdsc.edu)

## Other valuable on-line sites

- **Entrez**

<http://www3.ncbi.nlm.nih.gov/Entrez/>

- **Genome mapping and sequencing**

- **Human genome project:**

<http://www3.ncbi.nlm.nih.gov/genome/guide/>

- **Model organisms:**

<http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>

- **Whole genome analysis:**

<http://www.ncbi.nlm.nih.gov/COG/>

- **Analysis of polymorphisms:**

<http://www.ncbi.nlm.nih.gov/SNP/>

- **Functional genomics:**
  - **Online Mendelian Inheritance in Man (OMIM):**  
<http://www.ncbi.nlm.nih.gov/Omim/>
- **Target identification in drug design, agriculture, biocatalysis:**  
<http://www.labmed.umn.edu/umbbd/index.html>
- **Differential digital display (Cancer genome anatomy project):**  
<http://www.ncbi.nlm.nih.gov/ncicgap/>
- **Array technologies:**  
<http://cmgm.stanford.edu/pbrown/>
- **Metabolic pathways:**  
<http://www.ecocyc.org/>  
<http://www.genome.ad.jp/kegg/>

# Primary databases for 3D structure classification/information

- **Entrez**

<http://www3.ncbi.nlm.nih.gov/Entrez/>

- **Protein Data Bank (PDB)**

<http://www.rcsb.org/pdb/>

- **Structural Classification of Proteins (SCOP)**

<http://scop.mrc-lmb.cam.ac.uk/scop/index.html>

- **CATH: Protein Structure Classification**

[http://www.biochem.ucl.ac.uk/bsm/cath\\_new/index.html](http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html)

# The problem of size, or

Why we have become suddenly and utterly  
dependent on Bioinformatics...

---

## Other issues in large-scale computing

- **Algorithm speed**
  - Example: Removing redundancy from Genbank
  - Example: All-by-all comparison on the PDB
- **Array data**
- **How many SNPs?**
  - A catalog of all sequence differences that may be needed to find the rarest or most complex disease genes would likely require  $15 \times 10^{15}$  entries
- **Proteomics**
  - Example: Large-scale analysis of mass spectrometric data from proteolytic digestion of proteins

## Protein vs. nucleic acid sequence analysis?

- Protein sequence analysis provides greater specificity and less noise than nucleic acid analysis for identification of similarities because of the inherent differences in the message content of nucleic acid and amino acid codes
- Due in part to 4-letter vs. 20-letter code, degeneracy of codon messaging
- But some searches must be done at the nucleotide level...

## Some information properties of messages for sequence analysis

- A sequence can be described in terms of the # of bits needed to specify its message, where one bit distinguishes between two equally likely things.  
Ex: Where base frequencies are equal, one bit distinguishes a purine from a pyrimidine, two bits are required to uniquely specify a single base among A, T, C, G.
- Information content of a random message can be calculated from the set of relevant symbols' frequencies:

$$I = \sum_{i=1}^n P_i \log_2 P_i$$

where  $P_i$  is the probability of finding the symbol  $i$  at any position

- Using a standard measure for overall amino acid frequencies gives the information content of a random protein sequence as 4.19 bits/residue.
- Thus, for an average size protein domain (150 residues), the message length is ~630 bits and the probability that 2 random sequences would specify the same message is  $2^{-630}$  ( $10^{-190}$ ).  
Database searching for protein similarities is doable, even for fairly short sequences
- BUT, for a transcription binding site of 8-10 bp, the odds of 2 random sequences arriving at the same message is  $10^{-5}$ .  
Database searching for regulatory elements does not work well as databases get larger

# Introduction to Protein Sequence Analysis

- Database searching/pairwise alignments
- Pattern searching and motif analysis
- Multiple alignments and Evaluation using Family/Superfamily Concepts

# Applications

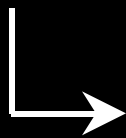
- tracing ancestral connections
- deduction/inference of function
- understanding enzyme mechanisms
- clustering of families, superfamilies
- structural analysis of receptors, molecules involved in cell signaling
- identification of molecular surfaces in protein-protein, protein-DNA interactions
- metabolic computing/comparative genome analysis
- guidance for functional genomics, protein engineering

## References: Database searching

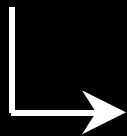
- Altschul et al., "Issues in searching molecular sequence databases"
- Pearson, "Comparison of methods for searching protein sequence databases"
- Altschul, "Amino acid substitution matrices from an information theoretic perspective"
- Pearson & Lipman, (the original FASTA paper) "Improved tools for biological sequence comparison"
- Altschul et al., (the original Blast paper) "Basic local alignment search tool"
- Henikoff & Henikoff, "Amino acid substitution matrices from protein blocks"
- Altschul et al., "Gapped Blast and PSI-Blast: A new generation of protein database programs"

# The underlying assumption used in functional inference...

**Sequence  
Conservation**



**Structure  
Conservation**



**Function  
Conservation**

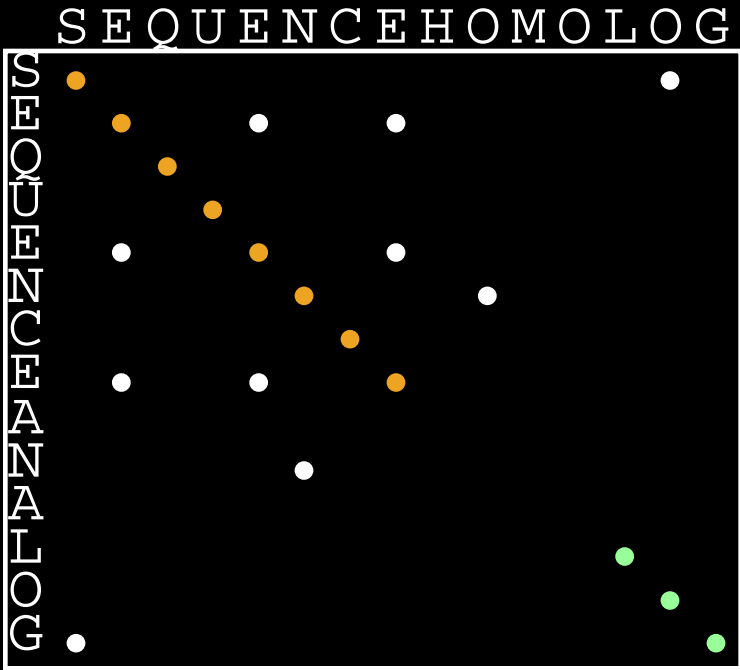
## ...requires comparison of sequences

- The most fundamental operation in protein informatics is finding the best alignment between a query sequence and one or more additional sequences
- Once candidate homologs have been identified, they can be evaluated using statistical methods and structural and biological information
- The correspondence between two aligned sequences can be expressed in a similarity score and/or viewed graphically, e.g., dot plots, alignments, motifs or patterns

## Formalizing the Problem

- Given: two sequences that you want to align
- Goal: find the best alignment that can be obtained by sliding one sequence along the other
- Requirements:
  - a scheme for evaluating matches/mis-matches between any two characters
  - a score for insertions/deletions
  - a method for optimization of the total score
  - a method for evaluating the significance of the alignment

- Dot matrix plots: a simple description of alignment operations illustrating types of relationships between a sequence pair



- The signal-to-noise ratio can be improved using filtering techniques designed to minimize the composition-dependent background
- Example of common filters: over-lapping, fixed-length "windows" for sequence comparison
- To be counted, a comparison must achieve a minimum threshold score summed over the window, derived empirically or from a statistical or evolutionary model of sequence similarity
- The window size and minimum threshold score (often termed "stringency") at which the score is counted can be user-defined

Seq1 = SEQUENCEHOMOLOG  
Seq2 = SEQUENCEANALOG  
Window = 7, Stringency = 42% (3/7 matches)

SEQUENC  
SEQUENCEANALOG (7/7 matches)

SEQUENC  
SEQUENCEANALOG (0/7 matches)

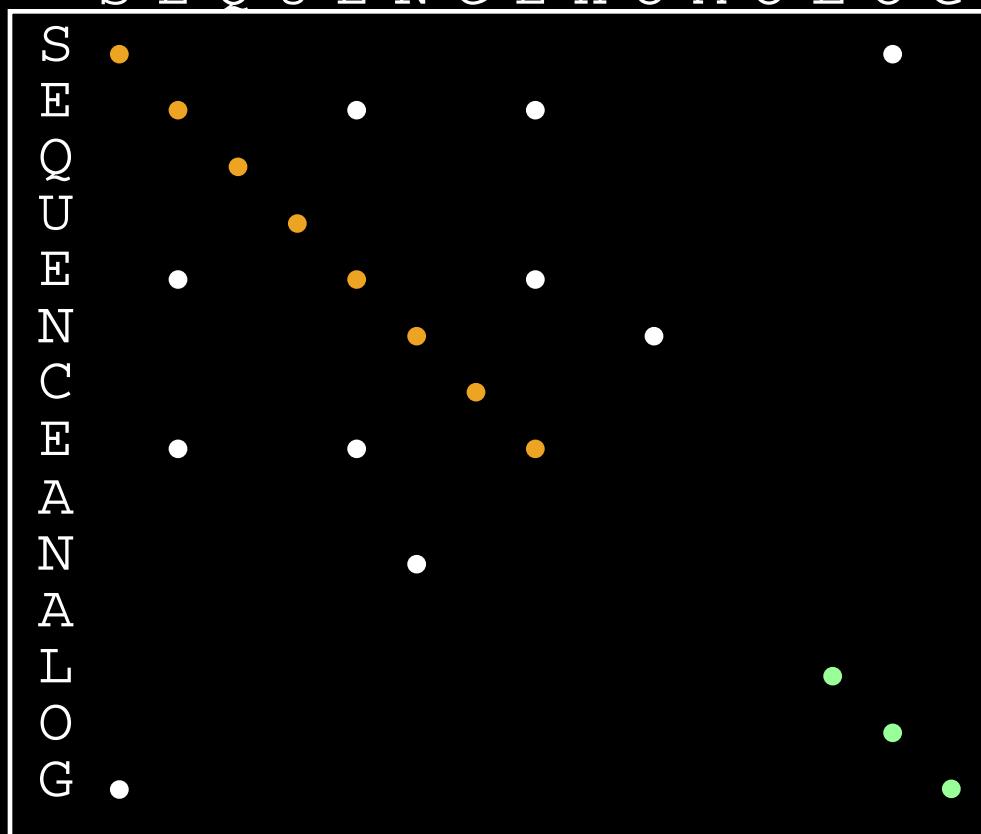
...

EQUENCE  
SEQUENCEANALOG (7/7 matches)

...

HOMOLOG  
SEQUENCEANALOG (3/7 matches)

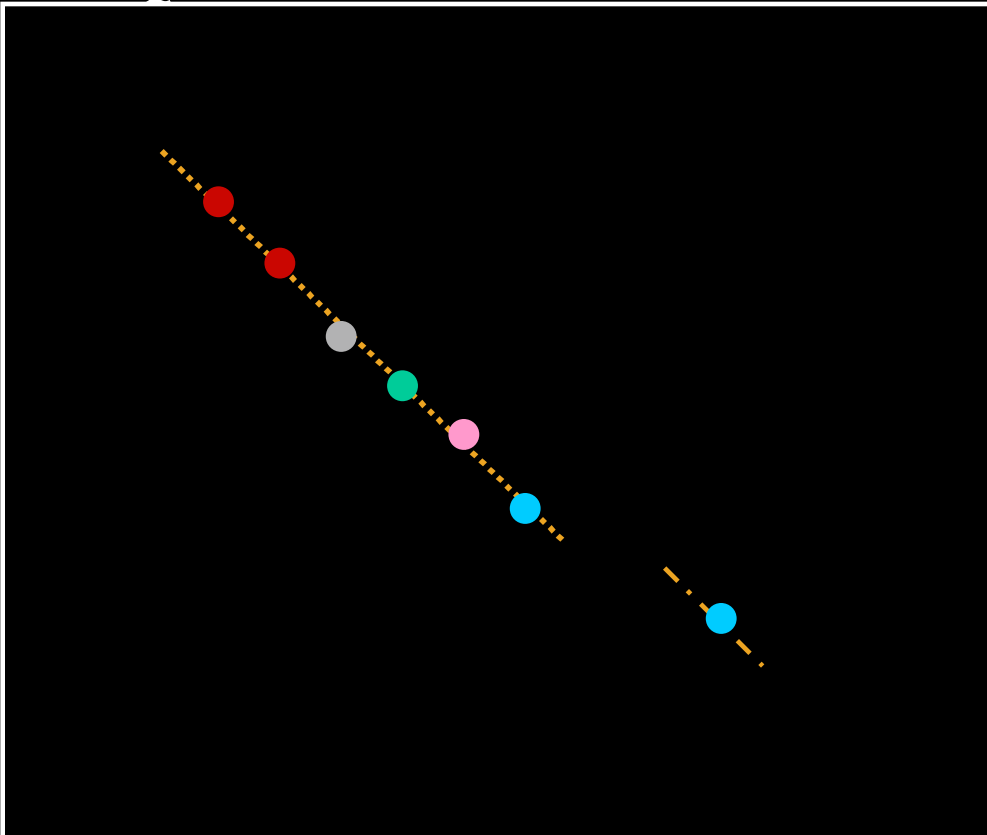
# SEQUENCEHOMOLOG



Stringency = 1/7

# SEQUENCE HOMOLOGY

SEQUENCE ANALOG



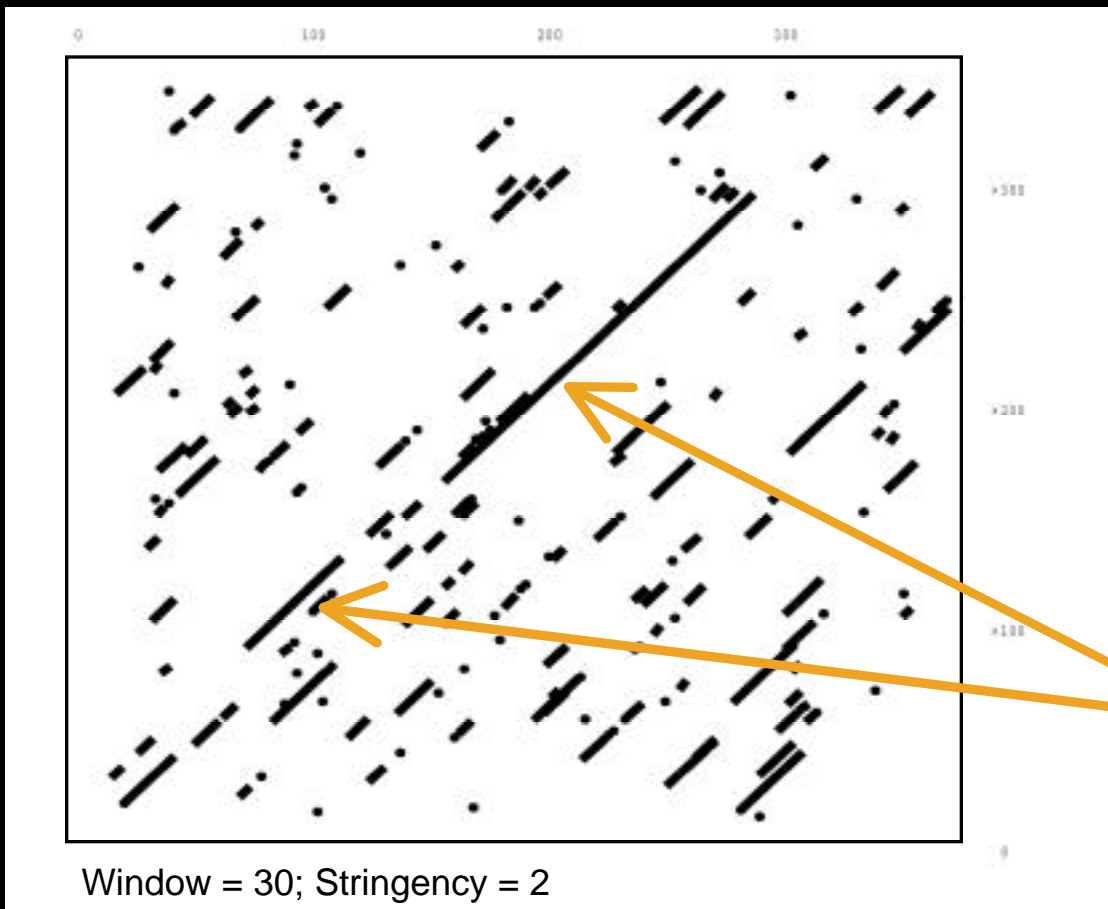
• 7/7

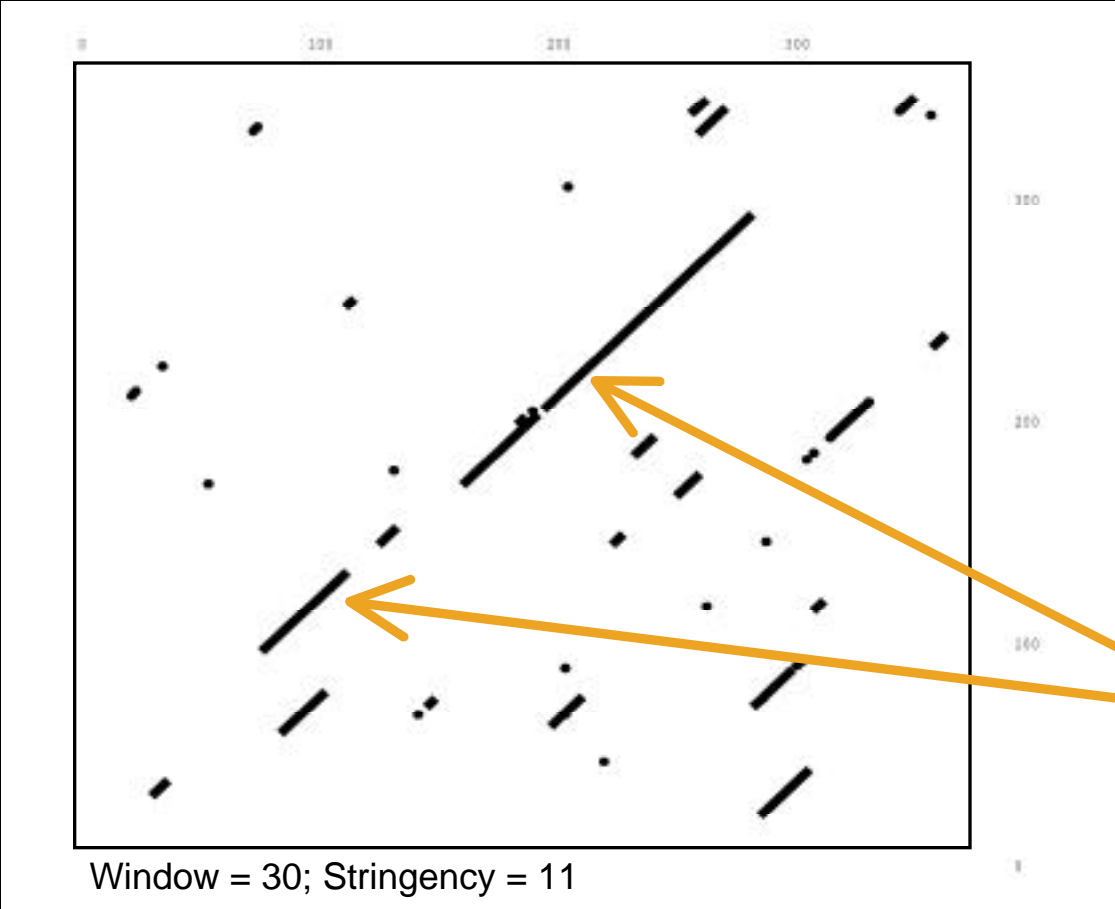
• 6/7

• 5/7

• 4/7

• 3/7





# Scoring Systems

- The degree of match between two letters can be represented in a matrix
- Changing the matrix can change the alignment
  - Simplest: Identity (unitary) matrix
  - Better: Definitions of similarity based on inferences about chemical or biological properties
  - Examples: PAM, Blosum, Gonnet matrices
- The score should have the form:  $p_{ab}/q_a q_b$ , where  $p_{ab}$  is the probability that residue  $a$  is substituted by residue  $b$ , and  $q_a$  and  $q_b$  are the background probabilities for residue  $a$  and  $b$  respectively.
- Handling gaps remains an incompletely solved problem...

## PAM units

- PAM (point accepted mutation) is a unit of evolutionary distance between 2 amino acid sequences\*
- 1 PAM = 1 accepted point-mutation (no insertions or deletions) event per 100 aa
- 200 PAM = 200 point mutations/100 aa (assumes mutations occur multiple times at any given position)
- 2 sequences diverged by 200 PAM 25% identity
- \*PAM is also sometimes defined as "percent accepted mutation"

# PAM matrices

- Substitution matrices used to reflect expected evolutionary change (by point mutations only)
- Given 2 sequences  $i, j$ , for any specific pair of residues  $A_i, A_j$ , the  $(i, j)$  entry in the PAM  $n$  matrix reflects the frequency at which  $A_i$  is expected to replace  $A_j$  in 2 sequences  $n$  PAM units diverged, *i.e.*, use PAM120 matrix to compare 2 protein sequences diverged by 120 PAM units
- Score should be in the form

$$\frac{p_{ij}}{p_i p_j}$$

- Usually presented in *log-odds* form, *i.e.*, probability values are given in logarithmic form

## Derivation of ideal PAM matrices\*

- Using many sets of 2 aligned sequences, for each amino acid pair  $A_i, A_j$ , count the # of times  $A_i$  aligns with  $A_j$  and divide that number by the total # of amino acid pairs in all of the alignments, resulting in the frequency,  $f(i,j)$
- Let  $f_i$  and  $f_j$ , respectively, denote the frequencies at which  $A_i$  and  $A_j$  appear in the sets of sequences
- Then the  $(i,j)$  entry for the ideal PAM matrix is

$$\log \frac{f(i, j)}{f(i) f(j)}$$

\*adapted from [Algorithms on Strings, Trees, and Sequences](#), Dan Gusfield, 1997

## Actual Derivation of PAM matrices

- Originally compiled from a group of sequences >85% identical that could be unambiguously aligned  
(M.O.Dayhoff, R.M. Schwartz, B.C. Orcutt, in Atlas of Protein Sequence and Structure, 5:345-352 (1978))
- These sequences were close in length and the few insertions/deletions could be placed correctly
- A PAM-1 matrix was calculated from these data
- Assumes more distantly related proteins can be described by a series of uncorrelated mutations consistent with the PAM-1 matrix such that a PAM-N matrix is derived by multiplying PAM-1 by itself N times

# Guidelines for using PAM matrices

from Altschul, "Amino acid substitution matrices from an information theoretic perspective"



A
2
0
-2
0
0
-4
1
-1
-1
-1
-2
-1
0
1
0
-2
1
1
0
-6
-3
0

K
-1
1
-5
0
0
-5
-2
0
-2
5
-3
0
1
-1
1
3
0
0
-2
-3
-4
0

W
-6
-5
-8
-7
-7
0
-7
-3
-5
-3
-2
-4
-4
-6
-5
2
-2
-5
-6
17
0
-6

\*

## Issues with PAM matrices

- Actually work quite well, with PAM-250 still used routinely for finding distant homologs
- BUT there are some clear problems with the model...
  - PAM model assumes all residues are equally mutable
  - Model devised using the most mutable positions rather than the most conserved positions, *i.e.*, those that reflect chemical and structural properties of importance
  - Derived from a biased set of sequences: small globular proteins available in the database in 1978

# BLOSUM (Blocks Substitution) Matrices

- Derived from the BLOCKS database, which, in turn is derived from the PROSITE library  
<http://blocks.fhcrc.org/blocks/>, <http://www.expasy.ch/prosite/>
- BLOCKS generated from multiply aligned sequence segments without gaps clustered at various similarity thresholds and corrected to avoid sampling bias
- Derived from data representing highly conserved sequence segments from divergent proteins rather than data based on very similar sequences (as with PAM matrices)

## Derivation of BLOSUM matrices

- Many sequences from aligned families are used to generate the matrices
- Sequences identical at  $>X\%$  are eliminated to avoid bias from proteins over-represented in the database
- Specific matrices refer to these clustering cut-offs, *i.e.*, BLOSUM62 reflects observed substitutions between segments  $<62\%$  identical
- In analogy to PAM matrices, a log-odds matrix is calculated from the frequencies  $A_{ij}$  of observing residue  $i$  in one cluster aligned against residue  $j$  in another cluster

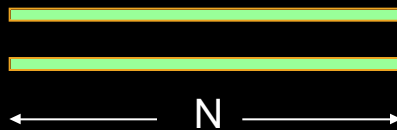
## BLOSUM vs. PAM Matrices

- BLOSUM matrices have replaced PAM matrices as the default matrices at many database searching sites (Blast, FASTA servers)
- Both PAM-120 and BLOSUM62 work best for moderately diverged proteins and may miss similarities outside their optimum performance windows
- PAM provides the only easily accessible alternative for short sequences (no appropriate version of Blosum available)
- Best solution is to provide a range of scoring systems, which is currently the practice for most primary servers
- Setting appropriate gap penalties can have a large effect on matrix performance

# Optimizing the Score: Brute-force Approach

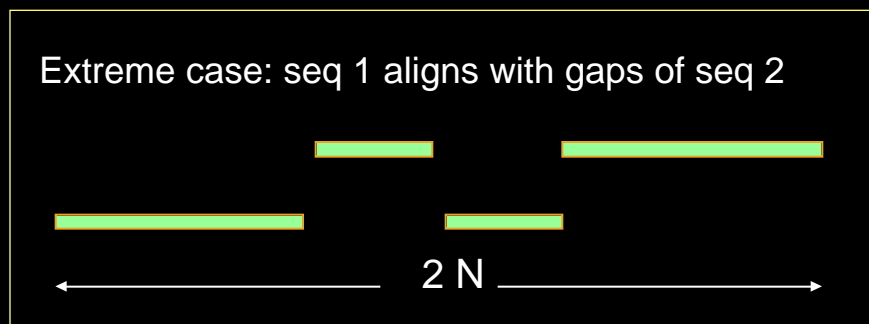
- Considering two sequences, both of length  $N$ :
  - If gaps or local alignments are not considered, there is only one optimal solution

Equal length w/o gap



- The computational time required to compute the optimal alignment =  $N^2$

- But when gaps or local alignments are considered, things get complicated because we have to repeat the calculation  $2N$  times to allow for the possibility of gaps at each position of each sequence
- Requires time proportional to  $N^{4N}$
- Even when nonsensical alignments are removed (aligning gaps with gaps), for  $N = 300$  residues,  $\sim 10^{88}$  comparisons are required



# Optimizing the Score: Dynamic Programming

- Requires computational time proportional to  $N^2$
- Original version often termed the “Needleman-Wunsch” algorithm  
(Needleman, S.B. and Wunsch, C.D. J. Mol. Biol. 48 (1970) 443-453)
- Addresses the problem for GLOBAL alignments; still has to deal with gaps

## Next step forward: local alignments

- **Implemented by Smith & Waterman**  
(Smith & Waterman. *J. Mol. Biol.* 147 (1981) 195-197)
- Finds the two “most similar” segments to generate an alignment from parts of the two sequences
- Modifications of dynamic programming algorithm:
  - The scoring system must include negative scores for mismatches
  - 0 = the minimum score recorded in the score matrix
  - The end of the optimal path can be anywhere in the matrix, not just in the last row or column

# Statistical Significance

- A good way to determine if an alignment score has statistical meaning is to compare it with the score generated from the alignment of two random sequences
- A model of 'random' sequences is needed. The simplest model chooses the amino acid residues in a sequence independently, with background probabilities

(Karlin & Altschul (1990) Proc. Natl. Acad. Sci. USA, 87 (1990) 2264-2268)

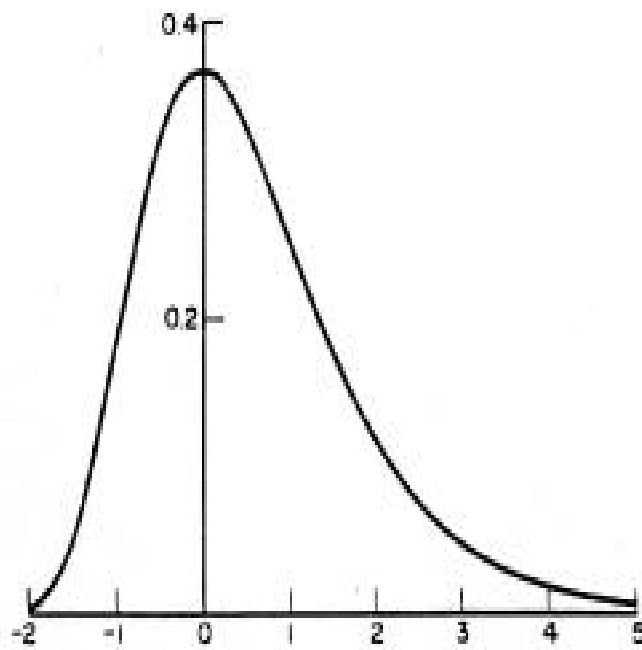


Figure 5. The probability density function of the extreme value distribution (74), with characteristic value  $\nu = 0$  and decay constant  $\lambda = 1$ .

## A most important caveat...

- For database searches, the **ONLY** criteria available to judge the likelihood of a structural or evolutionary relationship between 2 sequences is an estimate of statistical significance
- For a medium-sized protein using default parameters (Blosum62,  $E = 10$ ), the cut-off for statistical significance is  $P = 10^{-7} - 10^{-5}$   
(for the relationship between E and P, see <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>)
- Statistical significance and biological significance are **NOT** necessarily the same

sp P31467 YIEH_ECOLI	Begin: 1 End: 180	HYPOTHETICAL 24.7 KD PROTEIN IN TNAB-BGLB I...	36	0.10
sp O14165 YDX1_SCHPO	Begin: 34 End: 201	HYPOTHETICAL 27.1 KD PROTEIN C4C5.01 IN CHR...	31	2.6
sp Q39565 DYHB_CHLRE	Begin: 3911 End: 4032	DYNEIN BETA CHAIN, FLAGELLAR OUTER ARM	29	7.6
sp P77625 YFBT_ECOLI	Begin: 143 End: 187	HYPOTHETICAL 23.7 KD PROTEIN IN LRHA-ACKA I...	29	10.0
sp Q40297 FCPA_MACPY	Begin: 146 End: 176	FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...	29	13
sp Q40296 FCPB_MACPY	Begin: 146 End: 176	FUCOXANTHIN-CHLOROPHYLL A-C BINDING PROTEIN...	29	13
sp P52183 ANNU_SCHAM	Begin: 119 End: 168	ANNULIN (PROTEIN-GLUTAMINE GAMMA-GLUTAMYLTR...	29	13
sp P37934 MAY3_SCHCO	Begin: 435 End: 552	MATING-TYPE PROTEIN A-ALPHA Y3	27	29
sp O06219 MURE_MYCTU	Begin: 255 End: 371	UDP-N-ACETYLMURAMOYLALANYL-D-GLUTAMATE--2,6...	27	29
sp P08419 EL2_PIG	Begin: 182 End: 245	ELASTASE 2 PRECURSO	27	38
sp Q11034 Y07S_MYCTU	Begin: 163 End: 218	HYPOTHETICAL 69.5 KD PROTEIN CY02B10.28C	27	38
sp P00577 RPOC_ECOLI	Begin: 1290 End: 1401	DNA-DIRECTED RNA POLYMERASE BETA' CHAIN (T	27	38
sp P32282 RIR1_BPT4	Begin: 239 End: 266	RIBONUCLEOSIDE-DIPHOSPHATE REDUCTASE ALPHA C...	27	50
sp P17346 LEC2_MEGRO	Begin: 36 End: 121	LECTIN BRA-2	27	50
sp P54947 YXEH_BACSU	Begin: 24 End: 51	HYPOTHETICAL 30.2 KD PROTEIN IN IDH-DEOR IN...	27	50
sp P30139 THIG_ECOLI	Begin: 43 End: 79	THIG PROTEIN	27	50
sp P95649 CBBY_RHOSH	Begin: 96 End: 189	CBBY PROTEIN	27	50
sp Q43154 GSHC_SPIOL	Begin: 228 End: 327	GLUTATHIONE REDUCTASE, CHLOROPLAST PRECURSO...	26	66
sp P34132 NT6A_HUMAN	Begin: 191 End: 215	NEUROTROPHIN-6 ALPHA (NT-6 ALPHA)	26	66
sp P34134 NT6G_HUMAN	Begin: 115 End: 144	NEUROTROPHIN-6 GAMMA (NT-6 GAMMA)	26	66

# Database searching

- The first and most common operation in protein informatics...and the only way to access the information in large databases
- Primary tool for inference of homologous structure and function
- Improved algorithms to handle large databases quickly
- Provides an estimate of statistical significance
- Generates alignments
- Definitions of similarity can be tuned using different scoring matrices and algorithm-specific parameters

# BLAST and FASTA

- The rigorous Needleman-Wunsch and Smith-waterman algorithms are too slow for large database searches
- There are two major heuristic algorithms (BLAST and FASTA) to speed up the searching
- However, these compromise speed and sensitivity and neither of them guarantees to find the best alignment
- BUT, these are the primary search engines used by the majority of scientists today and their excellent performance justifies such use
- Pairwise comparisons limit information content

# FASTA suite

- "Fast" search algorithm generates global alignments, allows gaps
- Good documentation (Pearson)  
<http://vega.crbm.cnrs-mop.fr/help/fasta-help.html>
- Extensively updated since first release
  - more rigorous statistical analysis has been added
  - multiple variants available
  - FASTA3 is the current implementation

# BLAST suite

- Original "fast" search algorithm generates local alignments without gaps (Blast 1.4)
- Newer versions (Blast 2.0x) accommodates gaps
- Documentation
  - **Manual:** [http://www.ncbi.nlm.nih.gov/BLAST/blast\\_help.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_help.html)
  - **FACS:** [http://www.ncbi.nlm.nih.gov/BLAST/blast\\_FAQs.html](http://www.ncbi.nlm.nih.gov/BLAST/blast_FAQs.html)
  - **Tutorial:** <http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html>
- Other subtypes recently available for aligning 2 sequences, motif searching, domain matching

# BLAST flavors

- **blastp** compares an amino acid query sequence against a protein sequence database
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands)
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database

## Psi-Blast: Extending our reach...

- Generalizes BLAST algorithm to use a position-specific score matrix in place of a query sequence and associated substitution matrix for searching the databases
- Position-specific score matrix is generated from the output of an initial Gapped Blast search, *i.e.*, uses a profile or motif defined in the initial Blast search in place of a single query sequence and matrix for subsequent searches of the database
- Results in a database search tuned to the specific sequence characteristics representative of the sequence set of interest

## Steps in a Psi-Blast search\*

- Constructs a multiple alignment from a Gapped Blast search and generates a profile from any significant local alignments found
- The profile is compared to the protein database and PSI-BLAST estimates the statistical significance of the local alignments found, using "significant" hits to extend the profile for the next round
- PSI-BLAST iterates step 2 an arbitrary number of times or until convergence

\*Adapted from the PSI-BLAST tutorial at NCBI

# PSI-BLAST information at NCBI

- **Access**

<http://www.ncbi.nlm.nih.gov/BLAST/>

- **Tutorial**

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-2.html>

- **A short explanation of PSI-BLAST statistics**

<http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-3.html>

- **See also:**

- Park J; Karplus K; Barrett C; Hughey R; Haussler D; Hubbard T; Chothia C. "Sequence comparisons using multiple sequences detect three times as many remote homologs as pairwise methods." (1998) *J. Mol. Biol.*, 284:1201-10