Introduction to

Bioinformatics and

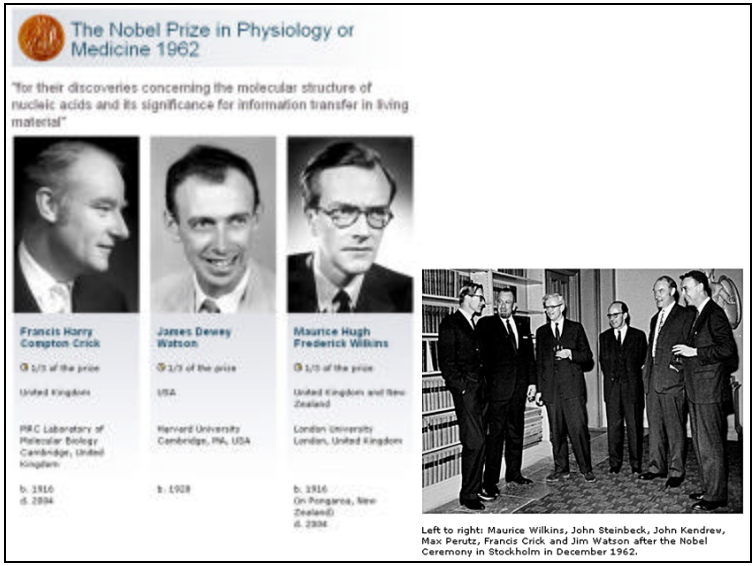
Sequence Alignment

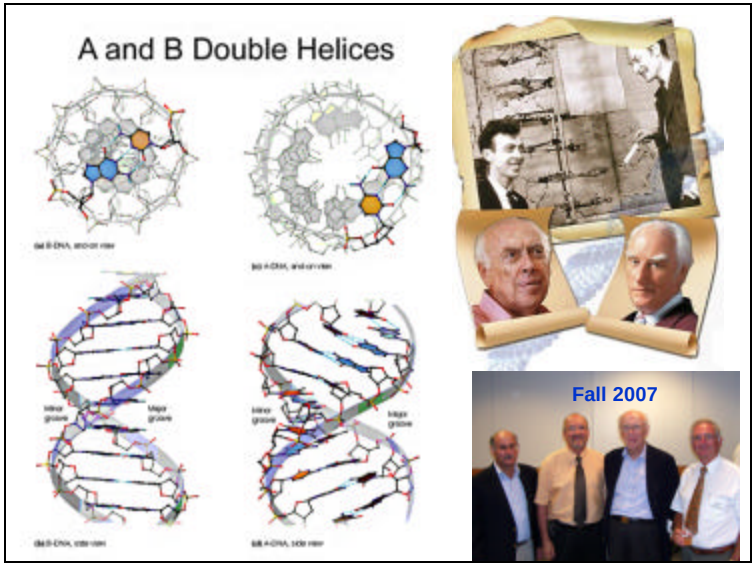
HgbA-human

```
GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
++ ++++:+ ::+ ++ +:++ +      +:   :+ +:+ :
NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

Leghemoglobin, yellow lupin

CH370 / 395G - Biochemistry
Marvin Hackert





Brief Introduction to Bioinformatics

More Terms:

- NCBI / EMBL
- Sequence databases
- FASTA
- Scoring Matrix
- PAM
- BLOSUM
- Smith – Waterman
- BLAST
- PSI – BLAST
- Raw Score
- Probability Value
- E-value
- ClustalW

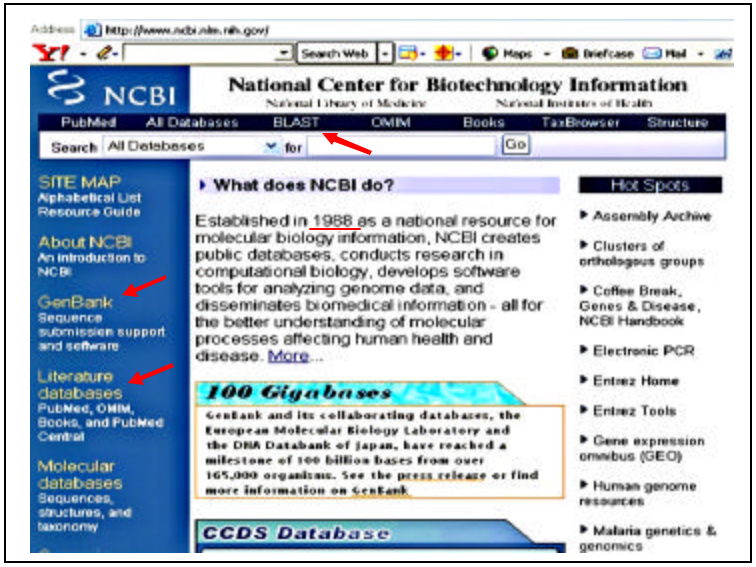
(Sequence Alignments)

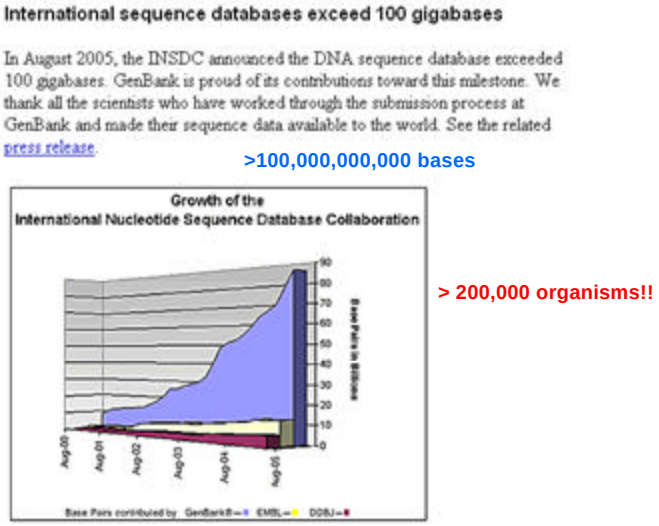
Acknowledgement: This brief introduction on Sequence Alignments is based on information found at web sites such as that at NCBI and EMBL- EBI. I also wish to acknowledge material taken from a handout provided by Dr. Ed Marcotte (Univ. of Texas at Austin) who teaches a course on Bioinformatics (CH391L) and on-line web notes of Michael Yaffe at MIT.

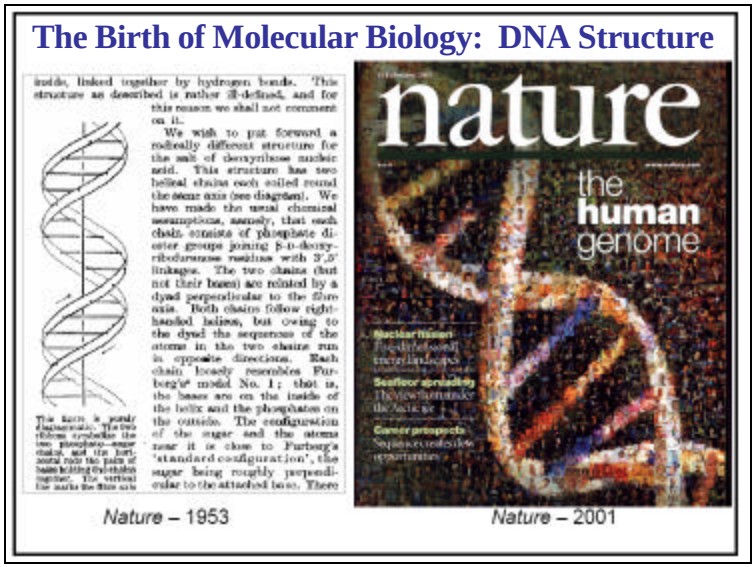
Ref:

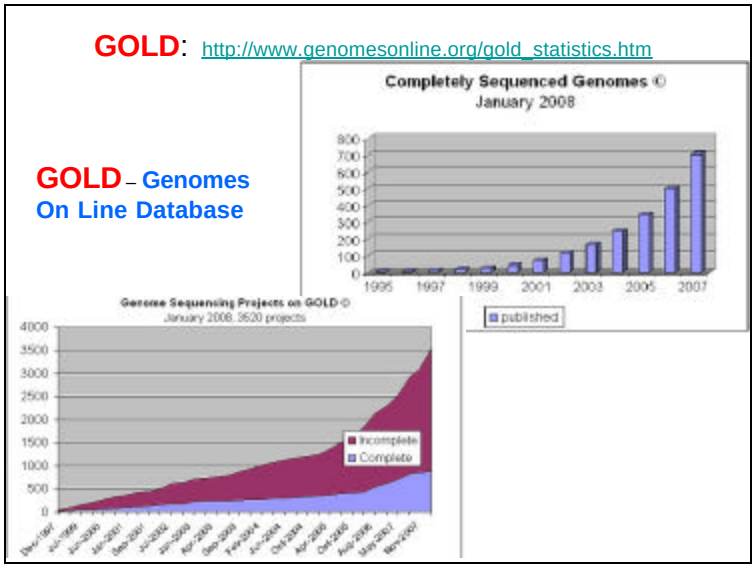
http://www.ncbi.nlm.nih.gov/

http://www.ebi.ac.uk/clustalw/#


NCBI
National Center for Biotechnology Information
PubMed
All Databases
BLAST
OMIM
Books
TaxBrowser
Structure
Search All Databases for Go
SITE MAP
Alphabetical List
Resource Guide
About NCBI
An introduction to NCBI
GenBank
Sequence submission support and software
Literature databases
PubMed, OMIM, Books, and PubMed Central
Molecular databases
Sequences, structures, and taxonomy
What does NCBI do?
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...
100 Gigabases
GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DNA Databank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the press release or find more information on GenBank.
CCDS Database
Hot Spots
Assembly Archive
Clusters of orthologous groups
Coffee Break, Genes & Disease, NCBI Handbook
Electronic PCR
Entrez Home
Entrez Tools
Gene expression omnibus (GEO)
Human genome resources
Malaria genetics & genomics


International sequence databases exceed 100 gigabases
In August 2005, the INSDC announced the DNA sequence database exceeded 100 gigabases. GenBank is proud of its contributions toward this milestone. We thank all the scientists who have worked through the submission process at GenBank and made their sequence data available to the world. See the related press release.
>100,000,000,000 bases
Growth of the International Nucleotide Sequence Database Collaboration
Aug-00
Aug-01
Aug-02
Aug-03
Aug-04
Aug-05
Base Pairs contributed by GenBank EMBL DDBJ
> 200,000 organisms!!


The Birth of Molecular Biology: DNA Structure
nature
the human genome
Nature – 1953
Nature – 2001


GOLD: http://www.genomesonline.org/gold_statistics.htm
GOLD – Genomes On Line Database
Completely Sequenced Genomes ©
January 2008
published
Genome Sequencing Projects on GOLD ©
January 2008, 3520 projects
Incomplete
Complete

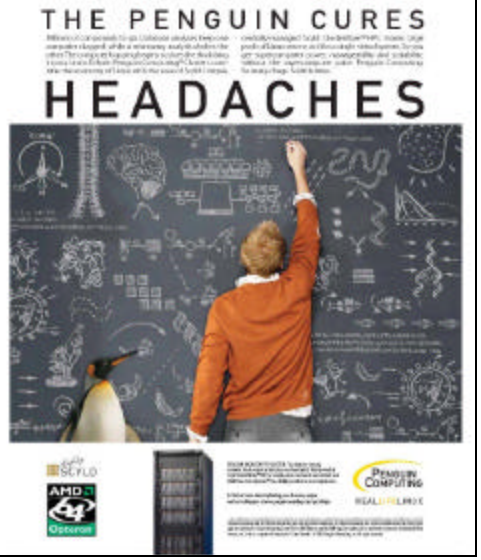
Genomics

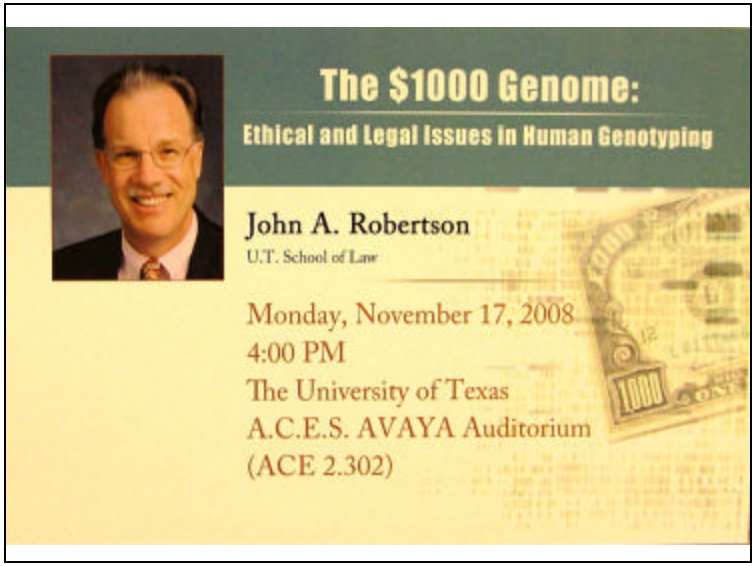
Proteomics

Interactomics

Systems Biology –

None of these fields of research would be possible without **Bioinformatics,** which would not be possible with lots of **computing power!**





Computational biology & Bioinformatics

Computational biology and **bioinformatics** focus on the computational/ theoretical study of biological processes, and much of the disciplines involve constructing models like those above, then testing/validating/proving/applying these models using computers, hence the nickname "*in silico* biology ". The fields are closely related: computational biology is the more inclusive name, and **bioinformatics often refers more specifically to the use of "informatics" tools like databases and data mining.**

Big problems tackled by these fields include:

- **Assembling complete genomes from pieces of sequenced DNA**
- **Finding genes in genomes**
- **Modeling networks & interactions of proteins**
- **Predicting protein/RNA folding, structure, and function**
- **Sequence alignments (BLAST)**

Why Align Sequences?

Identify Protein or Gene from Partial Information

Infer Functional Information

Infer Structural Information

Infer Evolutionary Relationships

Assumes:

conservation of sequence ⟷ conservation of function

BUT: Function carried out at level of proteins, i.e. 3-D structure

Sequence conservation carried out at level of DNA 1-D sequence

Alignments

- **Types:**
  - Local
  - Global
  - Ungapped
  - Gapped (2 types- linear, affine)
    - **Methods:**
      - Dot matrix
      - Dynamic Programming
      - Word, k-tup

(k respective tuples
=1 or 2 proteins / 4-6 DNA)

FASTA & FASTA Format

The FASTA algorithm is a heuristic method for string comparison. It was developed by *Lipman* and *Pearson* in 1985. FASTA compares a query string against a single text string.

- This format contains a one line header followed by lines of sequence data.
- Sequences in fasta formatted files are preceded by a line starting with a " >" symbol.
- The first word on this line is the name of the sequence. The rest of the line is a description of the sequence.

Term	Entry Name	Molecule Type	Gene Name	Sequence Length
e.g.	FOSB_MOUSE	Protein	fosB	338 bp

- The remaining lines contain the sequence itself.
- Blank lines in a FASTA file are ignored, and so are spaces or other gap symbols (dashes, underscores, periods) in a sequence.
- Fasta files containing multiple sequences are just the same, with one sequence listed right after another. This format is accepted for many multiple sequence alignment programs.

```
>FOSB_MOUSE Protein fosB  338 bp
MFQAFPGDYDSGSRCSSSPSAESQYLSSVDSFGSPPTAAASQECAGLGEMPGSFVPTVTA
ITTSQDLQWLVQPTLISSMAQSQGQPLASQPPAVDPYDMPGTSYSTPGLSAYSTGGASGS
GGPSTSTTTSGPVSARPARARPRRPREETLTPEEEERRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAHKPGCKIPYEEGPGPGPLAEVRD
IPGSTSAKEDGFGWLLPPPPPPPLPFQSSRDAFPNLTASIFTHSEVQVLGDPFPVVSPSY
TSSFVLTCPEVSAFAGAQRTSGSEQPSDPLNSPSLLAL
```

BLAST – Basic Local Alignment Search Tool

T he ***BLAST*** algorithm was developed for **protein alignments** in comparison to ***FASTA***, which was developed for **DNA sequences**. **BLAST** concentrates on finding regions of high local similarity in alignments without gaps, evaluated by an alphabet-weight scoring matrix.

Many "flavors" of BLAST

Program	Query	Database
BLASTP	aa	aa
BLASTN	nt	nt
BLASTX	nt (⇒ aa)	aa
TBLASTN	aa	nt (⇒ aa)
TBLASTX	nt (⇒ aa)	nt (⇒ aa)
PsiBLAST	aa (aa msa)	aa

(Position-Specific Interative)

Sequence Alignment

The ***Smith-Waterman algorithm*** considers a simple model for protein sequence evolution that allows us to align amino acid sequences of proteins to see if the proteins are related. **BLAST** is designed to **mimic this algorithm**, but BLAST is much faster due to some shortcuts and approximations and clever programming tricks.

This **process of gene evolution** can be modeled as **a stochastic process of gene mutation** followed by a **"selection" process for those sequences still capable of performing their given roles** in the cell. Over enough time, as new species evolve & diverge from related species, this has the result of producing families of related gene sequences, more similar in regions where that particular sequence is critical for the function of the molecule, and less similar in regions less critical for the molecule's function. Frequently, **we observe only the products of millions of years of this process**. Given a set of molecules (DNA, RNA or protein sequences) - **??** How can we **decide if they are similar enough to be considered part of the same family** or if the **observed similarity is just present by random chance**.

Alignments- Good Bad and Ugly

HgbA-human

```
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKL
:+ +::+:::::+ :+++++::+:++ +++++::+::+ ::
GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKL
```

HgbB-human

HgbA-human

```
GSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL
++ ++++:+ ::+ ++ +:++ +          +:   :+ +:+ :
NNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG
```

Leghemoglobin, yellow lupin

SPURIOUS ALIGNMENT

HgbA-human

```
GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD----LHAHKL
::+ ++: +  ++:+:  ++ : +  : +:+ :    ++::+
GSGYLVGDSLTFVDLL--VAQHTADLLAANAALLDEFPQFKAHQE
```

Nematode glutathione S-transferase

To align two sequences, we need:

1. Some way to decide which alignments are better than others. For this, we'll **invent a way to give the alignments** a "**score**" indicating their quality.

→ **"Scoring Matrix"**

2. Some **way to align the proteins** so that they **get the best possible score**.

→ ***Smith-Waterman algorithm***

dynamic programming, ***recursive*** **manner**

3. Then finally, some way to **decide when a score is "good enough"** for us to believe the alignment is biologically significant.

→ **"Scramblings - Expect Values"**

extreme value distribution

What is a scoring matrix?

The aim of a sequence alignment, is to match "the most similar elements" of two sequences. This similarity must be evaluated somehow.

For example, consider the following two alignments:

```
AIWQH        AIWQH
AL-QH        A-LQH
```

They seem quite similar: both contain one "**gap**" and one "**substitution**," just at different positions. However, the first alignment is the better one because isoleucine (I) and leucine (L) are similar sidechains, while tryptophan (W) has a very different structure. This is a **physico-chemical** measure; we might prefer these days to say that leucine simply substitutes for isoleucine more frequently ---without giving an underlying "reason" for this observation.

However we explain it, **it is much more likely that a mutation changed I into L and that W was lost, than W was changed into L and I was lost**. We would expect that a change from I to L would not affect the function as much as a mutation from W to L---but this deserves its own topic.

To **quantify the similarity achieved by an alignment**, ***scoring matrices*** are used: they contain a value for each possible substitution, and the ***alignment score*** is the sum of the matrix's entries for each aligned amino acid pair. For gaps a special *gap score* is necessary ---just add a constant penalty score for each new gap. The ***optimal alignment*** is the one which **maximizes the alignment score**.

Importance of scoring matrices

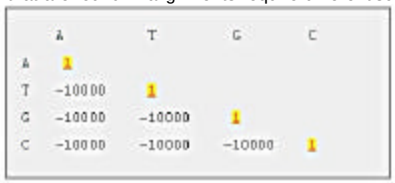
Scoring matrices appear in all analysis involving sequence comparison.
The choice of matrix can strongly influence the outcome of the analysis.
Scoring matrices implicitly represent a particular theory of evdution.
Understanding theories underlying a given scoring matrix can aid in making proper choice.

Types of matrices **Ref:** http://www.ebi.ac.uk/clustalw/#

PAM BLOSSUM GONNET DNA Identity Matrix

Unitary Scoring Matrices

Early sequence alignment programs used **unitary scoring matrix**. A unitary matrix **scores all matches the same and penalizes all mismatches the same**. Although this scoring is sometimes appropriate for DNA and RNA comparisons, **for protein alignments using a unitary matrix amounts to proclaiming ignoran ce about protein evolution and structure** Thirty years of research in aligning protein sequences have shown that different matches and mismatches among the 400 amino acid pairs that are found in alignments require different scores.



Many alternatives to the unitary scoring matrix have been suggested. One of the earliest suggestions was **scoring matrix based on the minimum number of bases that must be changed to convert a codon for one amino acid into a codon for a second amino acid.** This matrix, known as the **minimum mutation distance matrix**, has succeeded in identifying more distant relationships among protein sequences than the unitary matrix approach.

Simplest alignment representation – Dot plot

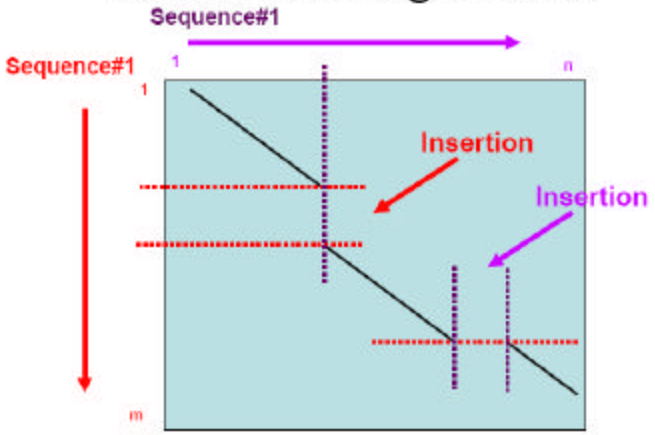
Model: Need a metric of similarity between amino acid pairs

Simplest metric – unitary matrix, identity matrix

Example – self alignment

	G	F	D	S	F	K	R	L	E	F	S	E	V
G	1	0	0	0	0	0	0	0	0	0	0	0	0
F		1	0	0	1	0	0	0	0	1	0	0	0
D			1	0	0	0	0	0	0	0	0	0	0
S				1	0	0	0	0	0	0	1	0	0
F					1	0	0	0	0	1	0	0	0
K						1	0	0	0	0	0	0	0
R							1	0	0	0	0	0	0
L								1	0	0	0	0	0
E									1	0	0	1	0
F										1	0	0	0
S											1	0	0

Dot Matrix Alignments



A Global Alignment

Now align two different sequences

* **Consider other similarity matrices besides identity….**
- ***Chemical similarity*** – binary decision
- ***Amino acid conservation*** in aligned protein families – min. similarity score (+/- window)
- ***Average*** of multiple scoring systems

Evolutionary Distances

The best improvement achieved over the unitary matrix was based on **evolutionary distances**. **Margaret Dayhoff** pioneered this approach in the 1970's. She made an extensive study of the frequencies in which amino acids substituted for each other during evolution. The **studies involved carefully aligning all of the proteins in several families of proteins and then constructing phylogenetic trees for each family**. Each phylogenetic tree was examined for the substitutions found on each branch. This lead to a **table of the relative frequencies with which amino acids replace each other over a short evolutionary period**.

This table and the relative frequency of occurrence of the amino acids in the proteins studied were combined in computing the **PAM (Point Accepted Mutations) family of scoring matrices.**

From a biological point of view **PAM matrices are based on observed mutations**. Thus **they contain information about the processes that generate mutations** as well as the criteria that are important in selection and in fixing a mutation within a population. From a **statistical point of view PAM matrices, and other log-odds matrices, are the most accurate description of the changes in amino acid composition** that are expected after a given number of mutations that can be derived from the data used in creating the matrices. Thus the highest scoring alignment is statistically the most likely to have been generated by evolution rather than by chance.

Log-odds scoring

Log-odds matrices: Each score in the matrix is the **logarithm of an odds ratio**. The odds ratio used is the **ratio of the number of times residue "A" is observed to replace residue "B" divided by the number of times residue "A" would be expected to replace residue "B" if replacements occurred at random**.

Deriving realistic substitution matrices:

First need to know frequency of one amino acid substituting for another in related proteins [=P(ab)] c/w the chance that substituting one for the other occurred by chance, based on the relative frequencies of each amino acid in proteins, q(a) and q(b). Call this the "odds ratio": P(ab)/q(a)q(b)

If we do this for all positions in an alignment, then the total probability will be the product of the odds ratios at each position....but multiplication is computationally expensive....so....take the log (odds ratio) and add them instead.

The **BLOSUM family of matrices** developed by Steven and Jorja Henikoff are one of these newly developed log-odds scoring matrices. The **improved performance of the BLOSUM** matrices can be attributed to **many more protein sequences known now,** thus they incorporate many more observed amino acid substitutions, and because the **substitutions used in constructing the BLOSUM matrices are restricted to those substitutions found within well conserved blocks in a multiple sequence alignment**.

PAM (Percent Accepted Mutation)

A **unit** introduced by M.O. **Dayhoff** et al. to quantify the amount of **evolutionary change** in a protein sequence. **1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence.** A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

PAM matrices are based on global alignments of closely related proteins.

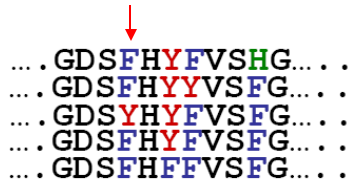
71 groups of protein sequences, 85% similar
1572 amino acid changes.
Functional proteins →"Accepted" mutations by natural selection

PAM1 matrix means 1% divergence between proteins - i.e. 1 amino acid change per 100 residues. Some texts re-state this as the probability of each amino acid changing into another is ~ 1% and probability of not changing is ~99%

The optimal alignment of two very similar sequences with PAM 500 may be less useful than that with PAM 50.

Construction of a Dayhoff Matrix: PAM1

Step 1: Measure pairwise substitution frequencies for each amino acid within families of related proteins



900 Phe (F)....+ another 100 probable Phe but...
100 Phe (F) → 80 Tyr (Y), 3 Trp (W), 2 His (H)....

Gives f_{ab}, i.e. $f_{FY}=80$
$f_{FW}=3$

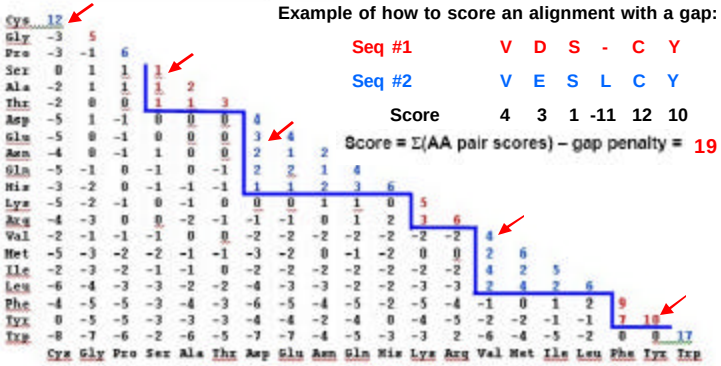
....by evolution!

Amino Acid Change	PAM 1 Score	PAM 250 Score
F→A	0.0002	0.04
F→R	0.0001	0.01
F→N	0.0001	0.02
F→D	0.0000	0.01
F→C	0.0000	0.01
F→Q	0.0000	0.01
F→E	0.0000	0.01
F→G	0.0001	0.03
F→H	0.0002	0.02
F→I	0.0007	0.05
F→L	0.0013	0.13
F→K	0.0000	0.02
F→M	0.0001	0.02
F→F	0.9946	0.32
F→P	0.0001	0.02
F→S	0.0003	0.03
F→T	0.0001	0.03
F→W	0.0001	0.01
F→Y	0.0021	0.15
F→V	0.0001	0.05

These are the M_{ab} values! i.e. the chance that one amino acid will replace another at 250 PAMs in two proteins that are evolutionarily related to each other!

SUM = 1.0

PAM 250 Amino Acid Similarity Matrix



Example of how to score an alignment with a gap:

Seq #1	V	D	S	-	C	Y
Seq #2	V	E	S	L	C	Y
Score	4	3	1	-11	12	10

Score = Σ(AA pair scores) – gap penalty = **19**

The **colored regions** in the figure above mark one possible grouping of such positive scores. **These regions provide an objective basis for defining conservative substitutions**, namely as amino acids that replace each other more frequently than would be expected from random replacements.

But we have to use the right matrix!!!

PAM 250 matrix – 250% expected change

Sequences still ~ 15-30 % similar, i.e. Phe will match Phe ~ 32% of the time
Ala will match Ala ~ 13% of the time

Expected % similarity

Other PAM matrices: PAM 120 – 40%
PAM 80 – 50%
PAM 60 – 60%
} Use for similar sequences

PAM250 – 15-30% similarity.

Use the correct PAM matrix for alignments based on how similar the sequences to be aligned are! But wait…..how do we know that in the first place? Usually don't!!!!.
So…… try PAM200, PAM120, PAM60, PAM80, and PAM30 matrix and use the one that gives the highest ungapped aligment score

Alternative amino acid matrices

Problems with Dayhoff:
- Based on amino acids, not nucleotides.
- Assumes evolutionary model with explicit phylogenetic relationships, and circular arguments: alignment → matrices; matrices → new alignments.
- Based on a small set of closely related molecules.

- Gonnett, Cohen & Benner
  -All against All database matching using DARWIN
  1,700,000 matches
  Compile mutation matrices at different PAMs DIRECTLY

- BLOSUM = Blocks Amino Acid Substitution Matrices-Henikoff&Henikoff 1992
  -based on a much larger dataset from ~500 Prosite families identified by Bairoch using conserved amino acid patterns "blocks" that define each family.

  Typically used for multiple sequence alignment.
  AA substitutions noted, log odds ratios derived.

  for example…Block patterns 60% identical give rise to Blosum60 matrix, etc….i.e. conservation of functional blocks based on un-gapped alignments.
  Blosum62 - best match between information content and amount of data
  Not based on explicit evolutionary model

BLOSUM matrices are based on local alignments.

BLOSUM (BLOcks SUbstitution Matrix): BLOSUM 62 is a matrix calculated **from comparisons of sequences with no less than 62% divergence.**

BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

Differences between PAM and BLOSUM

PAM matrices are based on an **explicit evolutionary model** (that is, replacements are counted on the branches of a phylogenetic tree), whereas the **BLOSUM** matrices are based on an **implicit** rather than explicit **model of evolution**.

The sequence variability in the alignments used to count replacements. The **PAM** matrices are based on mutations observed throughout a **global alignment**, this includes both highly conserved and highly mutable regions. The **BLOSUM** matrices are **based only on highly conserved regions** in series of alignments forbidden to contain gaps.

BLOSUM62 Substitution Scoring Matrix. The BLOSUM 62 matrix is a 20 x 20 matrix in which every possible identity and substitution is **assigned a score based on the observed frequencies of such occurences in alignments of related proteins. Identities are assigned the most positive scores. Frequently observed substitutions also receive positive scores** and **seldom observed substitutions are given negative scores.**

Blosum 45 Amino Acid Similarity Matrix

	Gly	Pro	Asp	Glu	Asn	His	Gln	Lys	Arg	Ser	Thr	Ala	Met	Val	Ile	Leu	Phe	Tyr	Trp	Cys
Gly	7																			
Pro	-2	9																		
Asp	-1	-1	7																	
Glu	-2	0	2	6																
Asn	0	-2	2	0	6															
His	-2	-2	0	0	1	10														
Gln	-2	-1	0	2	0	1	6													
Lys	-2	-1	0	1	0	-1	1	5												
Arg	-2	-2	-1	0	0	0	1	3	7											
Ser	0	-1	0	0	1	-1	0	-1	-1	4										
Thr	-2	-1	-1	-1	0	-2	-1	-1	-1	2	5									
Ala	0	-1	-2	-1	-1	-2	-1	-1	-2	1	0	5								
Met	-2	-2	-3	-2	-2	0	0	-1	-1	-2	-1	-1	6							
Val	-3	-3	-3	-3	-3	-3	-3	-2	-2	-1	0	0	1	5						
Ile	-4	-2	-4	-3	-2	-3	-2	-3	-3	-2	-1	-1	2	3	5					
Leu	-3	-3	-3	-2	-3	-2	-2	-3	-2	-3	-1	-1	2	1	2	5				
Phe	-3	-3	-4	-3	-2	-2	-4	-3	-2	-2	-1	-2	0	0	0	1	8			
Tyr	-3	-3	-2	-2	-2	2	-1	-1	-1	-2	-1	-2	0	-1	0	0	3	8		
Trp	-2	-3	-4	-3	-4	-3	-2	-2	-2	-4	-3	-2	-2	-3	-2	-2	1	3	15	
Cys	-3	-4	-3	-3	-2	-3	-3	-3	-3	-1	-1	-1	-2	-1	-3	-2	-2	-3	-5	12

The PAM family

- PAM matrices are based on global alignments of closely related proteins.
- The PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence.
- Other PAM matrices are extrapolated from PAM1.

The BLOSUM family

- BLOSUM matrices are based on local alignments.
- BLOSUM 62 is a matrix calculated from comparisons of sequences with no less than 62% divergence.
- All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
- BLOSUM 62 is the default matrix in BLAST 2.0. Though it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

BLOSUM 80	BLOSUM 62	BLOSUM 45
PAM 1	PAM 120	PAM 250
Less divergent ←	→	*More divergent*

The relationship between BLOSUM and PAM substitution matrices. BLOSUM matrices with higher numbers and PAM matrices with low numbers are both designed for comparisons of closely related sequences. BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related proteins. If distant relatives of the query sequence are specifically being sought, the matrix can be tailored to that type of search.

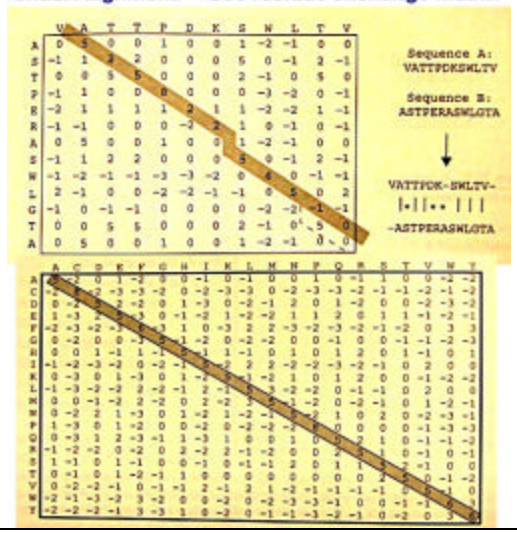
Sequnce Analysis: Which scoring method should I use?

Comparable Blosum and PAM Tables

Blosum Tables	(Entropy)	PAM Tables	(Entropy)	Percent Sequence Identity PAM Tables
Blosum 90	(1.18)	PAM 100	(1.18)	43
Blosum 80	(0.99)	PAM 120	(0.98)	38
Blosum 60	(0.66)	PAM 160	(0.70)	30
Blosum 52	(0.52)	PAM 200	(0.51)	25
Blosum 45	(0.38)	PAM 250	(0.36)	20

The **entropy** as defined by **information theory** is the **average amount of information per position in a sequence alignment that is available to determine whether or not the sequences are homologous**. This amount of entropy is available only if the similarity scores used in the database search or alignment are matched for the appropriate degree of sequence divergence.

Initial Alignment - Use residue exchange matrix



An Alignment Algorithm

If we had all the time in the world, we could just make all possible alignments, score them all, & choose the best. But realistically, that won't work, since even for **two 100 amino acid sequences**, there are **10^{59} possible alignments**. So, the following approach was developed.

The particular class of algorithm we'll use is called ***dynamic programming***, which refers to a set of algorithms that allow the optimal solutions to be found for problems that can be defined in a ***recursive* manner**. That is, the **problems are broken into subproblems**, which are **in turn broken into subproblems**, etc, until the simplest subproblems can be solved. For sequence alignments, this sequential dependency takes a form where the choice of optimal alignment of a sequence of length *n* is found from the solution to the optimal alignment of a sequence of length *n*-1 plus the alignment of the *n*th symbol, and the optimal alignment of the *n*-1 case is a function of the *n*-2 case, and so on. **Dynamic programming was developed by Richard Bellman 40-50 years ago, but then "rediscovered" by biologists aligning sequences in the 1970's.**

There are 2 types of alignments that we could make: ***global*** and *local*

Global alignments will require a forced match between every symbol of one string with some symbol (or gap) of the second string, e.g.

```
ACGTTATGCATGACGTA
-C---ATGCAT----T-
```

Local alignments will correspond to the best matching subsequences (including gaps). For the above example, this corresponds to:

```
ATGCAT
ATGCAT
```

We'll look at **local alignments**, since these are what are used in almost any sequence alignment algorithm you might choose. This approach (in biology) is named the ***Smith-Waterman algorithm*** after Temple Smith & Mike Waterman, Journal of Molecular Biology vol. 147, 195-197 (1981).

Recursion and Dynamic Programming

Aligning two protein sequences without gaps – roughly an O(mn) problem. With gaps – becomes computationally astronomical, and cannot be done by direct comparison methods. ($= 2^{2L}/\sqrt{(2\pi L)}$; L=sequence length)

Alternative is to compare all possible pairs of characters (matches and mismatches, and also take gaps into account as well, while keeping the number of comparisons manageable. The approach is called dynamic programming. Mathematically proven to produce optimal alignment

Need a substitution or similarity matrix and some way to account for gaps.

GAPS / Gap peanaties

In most alignment and search programs, the **gap penalty** consists of **two terms,** the **cost to open the gap** and the **cost to extend the gap .**

Utility	Details
FASTA3, BLAST2, CLUSTALW, ScanPS and MPsrch.	**GAPOPEN or OPENGAP or OPEN GAP PENALTY:** Penalty for the first residue in a gap (e.g.fasta defaults: -12 by with proteins, -16 for DNA). **GAPEXT or EXTENDGAP or EXTEND GAP PENALTY:** Penalty for additional residues in a gap (e.g. fasta defaults: -2 with proteins, -4 for DNA).

Ref: http://www.ebi.ac.uk/clustalw/#

To do Dynamic Programming:

First write one sequence across the top, and one down along the side

j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8					
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

S_{ij}

So scoring Sij requires that we know S(i-1, j-1) and S(i, j-1) and S(i-1, j)… Therefore *recursive.* We use the solutions Of smaller problems to solve larger ones. AND we store how we got to the Sij score, i.e. the intermediate solutions in a tabular matrix. Computer scientists call this dynamic programming, where "programming" means the matrix, not some kind of computer code.

To do Dynamic Programming:

First write one sequence across the top, and one down along the side

j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4				
2	E	-16					
3	S	-24					
4	L	-32					
5	C	-40					
6	Y	-48					

Global alignments: Needleman-Wunsch-Sellers $O(n^2)$ using linear gap penalty

S_{ij} = max of:

- $S_{i-1, j-1} + \sigma(x_i, y_j)$ (diagonal)
- $S_{i-1, j} - A$ (from left to right)
- $S_{i, j-1} - A$ (from top to bottom)

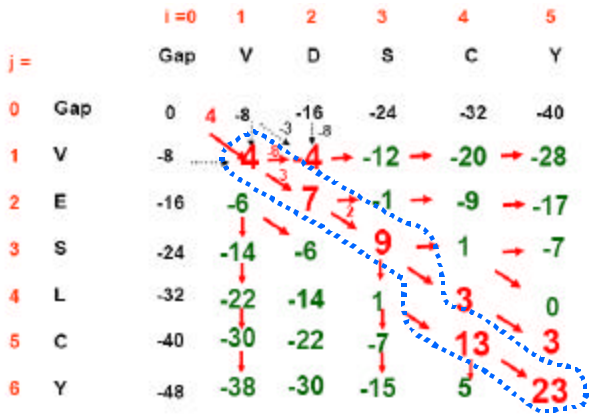
To do Dynamic Programming:

First write one sequence across the top, and one down along the side

j =		i =0	1	2	3	4	5
		Gap	V	D	S	C	Y
0	Gap	0	-8	-16	-24	-32	-40
1	V	-8	4	-4	-12	-20	-28
2	E	-16	-6	7	-1	-9	-17
3	S	-24	-14	-6	9	1	-7
4	L	-32	-22	-14	1	3	0
5	C	-40	-30	-22	-7	13	3
6	Y	-48	-38	-30	-15	5	23

The Traceback:

After the alignment square is finished, start at the lower right and work backwards following the arrows to see how you got there…



Examples of aligned protein sequences:

Shown are 3 pairs of sequences, showing aligned sequences of proteins named FlgA1, FlgA2, FlgA3, and HvcPP. Between each pair the perfect matches and close matches (shown by + symbols, indicating chemically similar amino acids) are written.

Two biologically related proteins with similar sequences:

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
        ++K+K+GRLDTLPP  +L+ N     A+SLR ++  QP+     R+ W +KAGQ V V+A G+
FlgA2 TLQDIKMKQGRLDTLPPGALLEPNFAQGAVSLRQINAGQPLTRNMLRRLWIIKAGQDVQVLALGE   (186)
```

Also biologically related (& fold up into the same 3D protein structure):

```
FlgA1 EAGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQAWRVKAGQRVNVIASGD
       A   +         P        +L    I+ R L P + I     R+AW V+ G  V V
FlgA3 LAALKQVTLIAGKHKPDAMATHAEELQGKIAKRTLLPGRYIPTAAIREAWLVEQGAAVQVFFIAG    (50)
```

But these are biologically unrelated (& fold up into unrelated structures):

```
FlgA1 AGNVKLKRGRLDTLPPRTVLDINQLVDAISLRDLSPDQPIQLTQFRQA-WRVKAGQRVNVIASGD
      AG+V  K G +  + PRT  ++       I+     P  PI     +++A WRV A + V V+  GD
HvcPP AGHV--KNGTMRIVGPRTCSNVWNGTFPINATTTGPSIPIPAPNYKKALWRVSATEYVEVVRVGD   (128)
```

The problem we face is how to distinguish the biologically meaningless match (FlgA1-HvcPP) from the biologically meaningful ones (FlgA1-FlgA2 and FlgA1-FlgA3)?

1) H E A G A W G H E E

2) A W - H E

How do we know when a score is "good enough"?

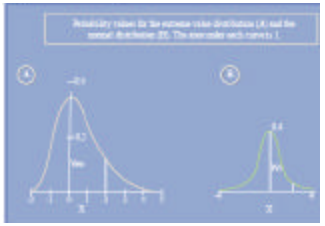
Two elements of aligning sequences:

scoring the alignments (by generating substitution matrices)

constructing the optimal scoring alignments by dynamic programming.

After we get an alignment, we have to decide if score is "good enough" to be significant. One way to this is to ask how hard it is to get that score from random alignments. Suppose we **"scrambled" one of the sequences, and found the best alignment with the other sequence.** The algorithm will always give us an alignment, even though the score is not very good. Still, let's **do the scrambling and alignment process 1000 times**. If we look at those scores, and **never see a score as good as the real one, we can say that the real one has a 1 in a 1000 chance of happening just by luck**. If we did this 1,000,000 times and still didn't see a score that good, we would begin to feel pretty confident in our alignment being significant.

Could do a **million random tests** after an alignment, and that should give a correct feeling for how good the alignment was. However, in practice, we can get away with just doing a **few random trials**, then mathematically modeling the scores we get out to save having to do a million such trials. The histogram of scores turns out to have a particular, predictable shape known as the ***extreme value distribution*** (also called the Gumbel distribution). Visually, the extreme value distribution looks this:



This distribution can be described by an equation of the form:

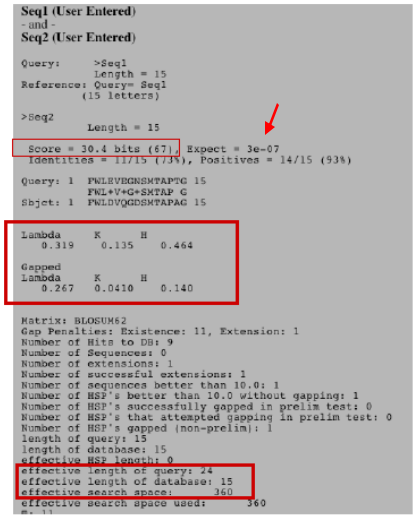
$$p(\max \text{ score} \le X) \approx e^{-kNe^{\lambda(X-m)}}$$

where N is the number of scrambled y's tested, m is the mean value of the high scores from the scrambling experiment, and k and l are numbers that characterize the shape of the particular **extreme value distribution** that comes from aligning x to y. In practice, k and l can be fit from the scores from a few random alignments,

- **Two ways to get the *K* and λ parameters:**

  1- For many amino acid substitution matrices, Altschul and Gish have tabulated their score distribution for 10,000 random amino acid sequences using various gap penalties

  2- Even better! Calculate the distribution for the two sequences you are aligning by keeping one of them fixed and scrambling the other one – this preserves BOTH sequence length and amino acid composition!



Raw score = 67

Bit score: $S=\frac{\lambda R - \ln K}{\ln(2)}$

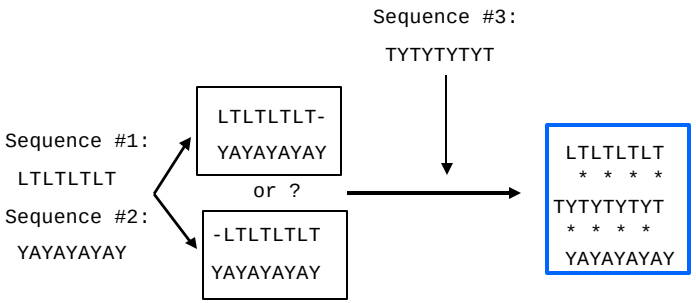
$S=\frac{(.267)(67)-\ln(.0410)}{0.693} = 30.4$

E~ *K*mne$^{-\lambda S}$ which is Equivalent to:

$E=mn2^{-s}$

$E=(24)(15)(2^{-30.4})=2.54e\text{-}07$

Advantages of Multiple Sequence Alignments



Multiple Sequence Alignments

For 3 sequences....

```
ARDFSHGLLENKLLGCDSMRWE
GRDYKMALLEQWILGCD-MRWD
SRDW--ALIEDCMV-CNFFRWD
```

An O(mnj) problem !

Consider sequences each 300 amino acids

2 sequences – $(300)^2$
3 sequences – $(300)^3$ Just became exponential!
but for v sequences – $(300)^v$

Still takes too long for more than three sequences…need a better way!

- **Progressive Methods of Multiple Sequence Alignment**

ClustalW

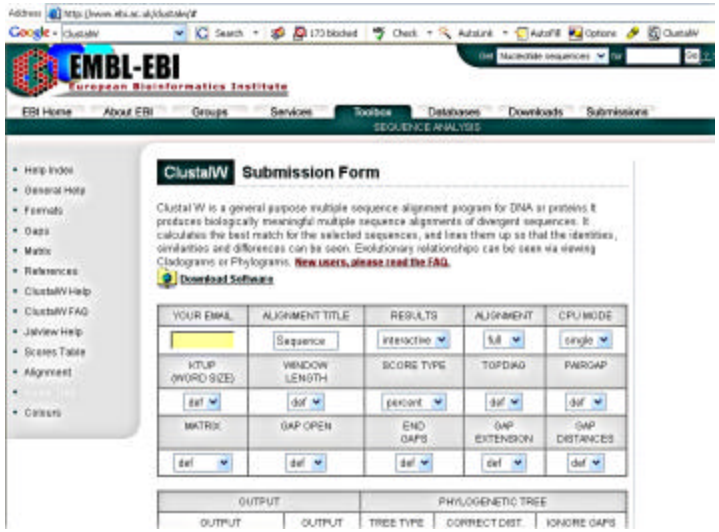
Higgins and Sharp 1988

- 1- Do pairwise analysis of all the sequences (you choose similarity matrix).
- 2- Use the alignment scores to make a phylogenetic tree.
- 3- Align the sequences to each other guided by the phylogenetic relationships in the tree.

Steps in doing a Multiple Sequence Alignment:

1) Get desired sequence in FASTA format.
2) NCBI web site – BLAST run (PSI BLAST)
3) Select best sequences to use in alignment (diversity, not always best scores)
4) EMBL web site – ClustalW run

```
>CgX SEQUENCE
MPTYTCWSQRIRISREAKQRIAEAITDAHHELAHAPKYLVQVIFNEVEPDSYFIAAQS
ASENHIWVQATIRSGRTEKQKEELLLRLTQEIALILGIPNEEVWVYITEIPGSNMTEY
GRLLMEPGEEEKWFNSLPEGLRERLTELEGSSE
```



4-OT– (Tautomerase/MIF Superfamily)

- with Professor Chris Whitman (Pharmacy)

Christian P. Whitman

```
          EEEEEEETT   HHHHHHHHHHHHHHHHHHT  GGG EEEEEEE GGG EETTEETTTT
4OT     1 PIAQIHILEG_RSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASKVRR    62
CHMI    1 PHFIVECSDNIREEADLPGLFAKVNPTLAATGIFPLAGIRSRVHWVDTWQMADGQHDYAFVHM..-125
MIF     1 PMFIVNTNVP_RASVPEGFLSELTQQLAQATGK_PAQYIAVHVVPDQLMTFSGTNDPCALCSL..-114
            *         *       *             *
```

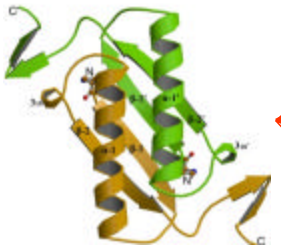
A) B) C)

Ref: Taylor, A.B., Czerwinski, R.M., Johnson, W.H., Whitman, C.P., an d Hackert, M.L., "Crystal Structure of 4 -Oxalocrotonate Tautomerase Inactivated by 2 -Oxo-3-pentynotate at 2.4 Å resolution: Analysis and Implications for the Mechanism of Inactivation and Catalysis" *Biochemistry*, **37**, 14692-14700 (1998).

β–α–β Motif

New Activities

New Structures



- **4-OT - Tautomerase**
- 4OT Homologues
- CHMI - Isomerase
- MIF - Cytokine / Hormone
- Dehalogenase
- Decarboxylase

$a_6\ a_3\ a_2\ (ab)_3$

Sample Psi-BLAST Output

Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database searchprograms", Nucleic Acids Res. 25:3389-3402.
RID: 1012187428-16844-19639
Query= Pseudomonas putida - 4-OT (62 letters)
1 piaqihileg rsdeqketli revseaisrs ldapltsvrv iitemakghf giggelaskv rr
Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF
857,413 sequences; 270,034,499 total letters
****
Sequences with E-value BETTER than threshold

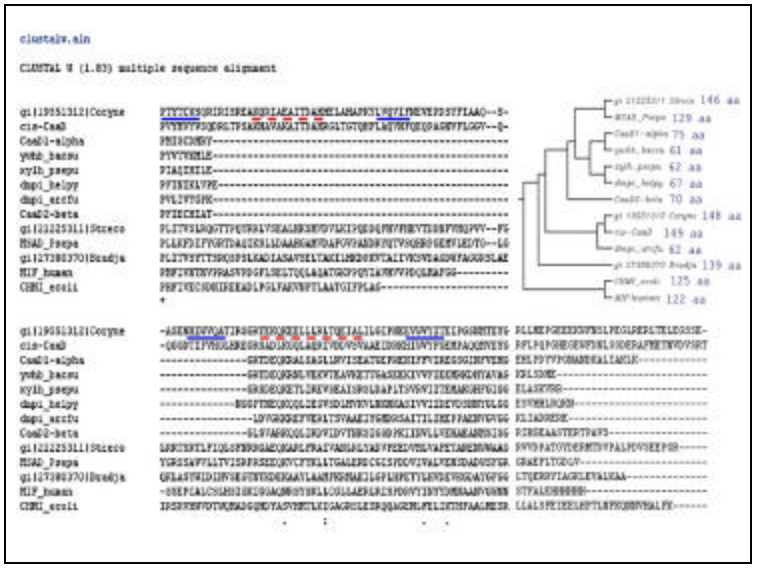
Round 1 - 30 Hits / Round 2 57 hits / Round 3 - 66 Hits

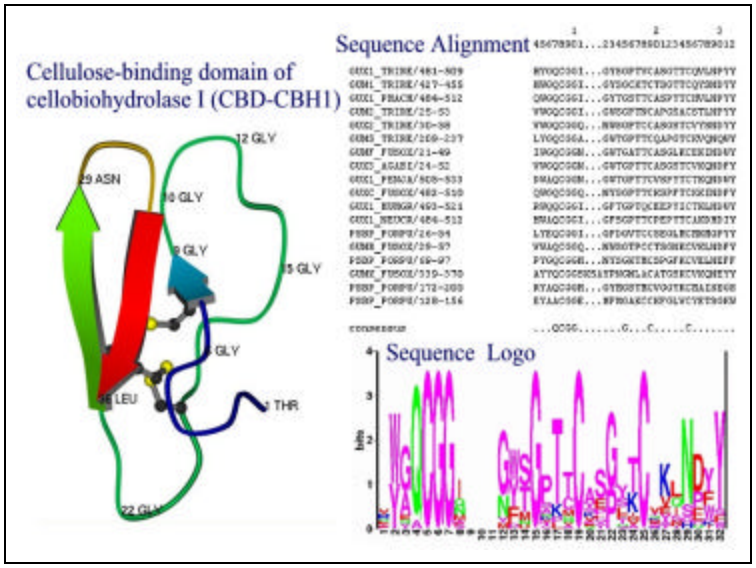
Sequences with E-value BETTER than threshold

Sequences producing significant alignments:		Score (bits)	E Value
gi|6624277|dbj|BAA88507.1|	(AB029044) 4-oxalocrotonate isomerase...	81	2e-15
gi|16124116|ref|NP_407429.1|	(NC_003143) putative tautomerase [Y...	78	2e-14
gi|14715457|dbj|BAB62059.1|	(D85415) 4-oxalocrotonate tautomeras...	78	2e-14
gi|15642664|ref|NP_232297.1|	(NC_002505) 5-carboxymethyl-2-hydro...	44	3e-04
gi|15801678|ref|NP_287696.1|	(NC_002655) ydcE gene product [Esch...	44	3e-04
gi|16079011|ref|NP_389834.1|	(NC_000964) similar to hypothetical...	43	8e-04

****
Sequences with E-value WORSE than threshold

gi|15894207|ref|NP_347556.1|	(NC_003030) Protein related to MIFH...	38	0.014
gi|14600626|ref|NP_147143.1|	(NC_000854) MRSA protein [Aeropyrum...	37	0.047
gi|17562710|ref|NP_506003.1|	(NM_073602) macrophage migration in...	35	0.16
gi|5051891|gb|AAD38354.1|	(AF119571) macrophage migration inhibi...	30	4.4
gi|14600626|ref|NP_147143.1|	(NC_000854) MRSA protein [Aeropyrum...	30	4.6
gi|5327268|emb|CAB46354.1|	(AJ012740) macrophage migration inhib...	30	8.1





Homology Modeling - Adjusting an alignment based on structure.

		1	2	3	4	5	6	7	8	9	10	11	12	13
Template		PHE	ASP	ILE	CYS	ARG	LEU	PRO	GLY	SER	ALA	GLU	ALA	VAL
Model (bad)	1	PHE	ASN	VAL	CYS	ARG	ALA	PRO	---	---	---	GLU	ALA	ILE
Model (good)	2	PHE	ASN	VAL	CYS	ARG	---	---	---	ALA	PRO	GLU	ALA	ILE