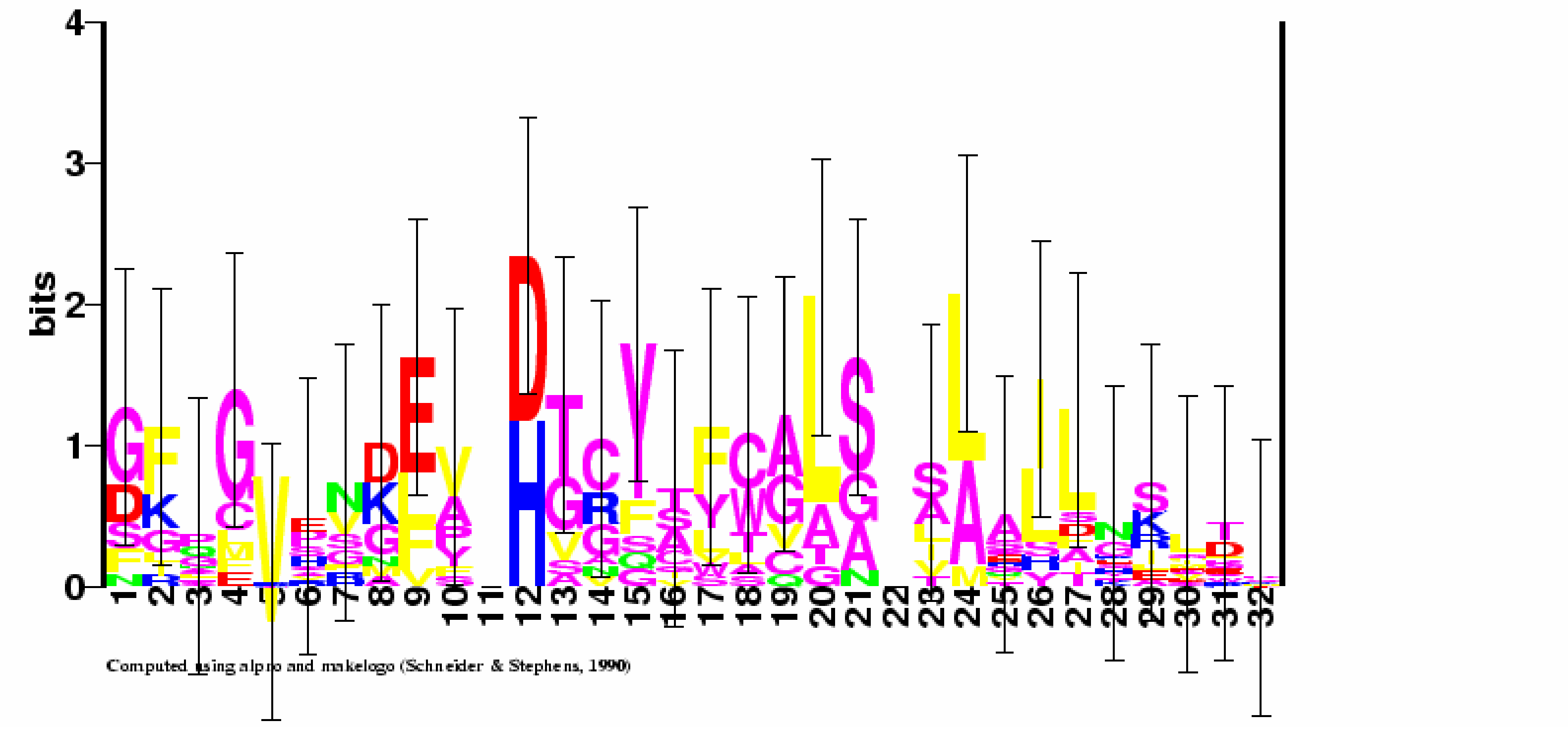
7.91 – Lecture #3 Michael Yaffe

More Multiple Sequence Alignment

-and-

Motif Scanning, Database Searching



Computed using alpro and makelogo (Schneider & Stephens, 1990)

Outline

- Multiple Sequence Alignment - Carillo & Lipman, Clustal(W)
- Position-Specific Scoring Matrices (PSSM)
- Information content, Shannon entropy
- Sequence logos
- Hidden Markov Models
- …Other approaches: Genetic algorithms, expectation maximization, MEME,Gibbs sampler
- FASTA, Blast searching, Smith-Waterman
- Psi-Blast

Reading - Mount p. 139-150, 152-157, 161-171, 185-198

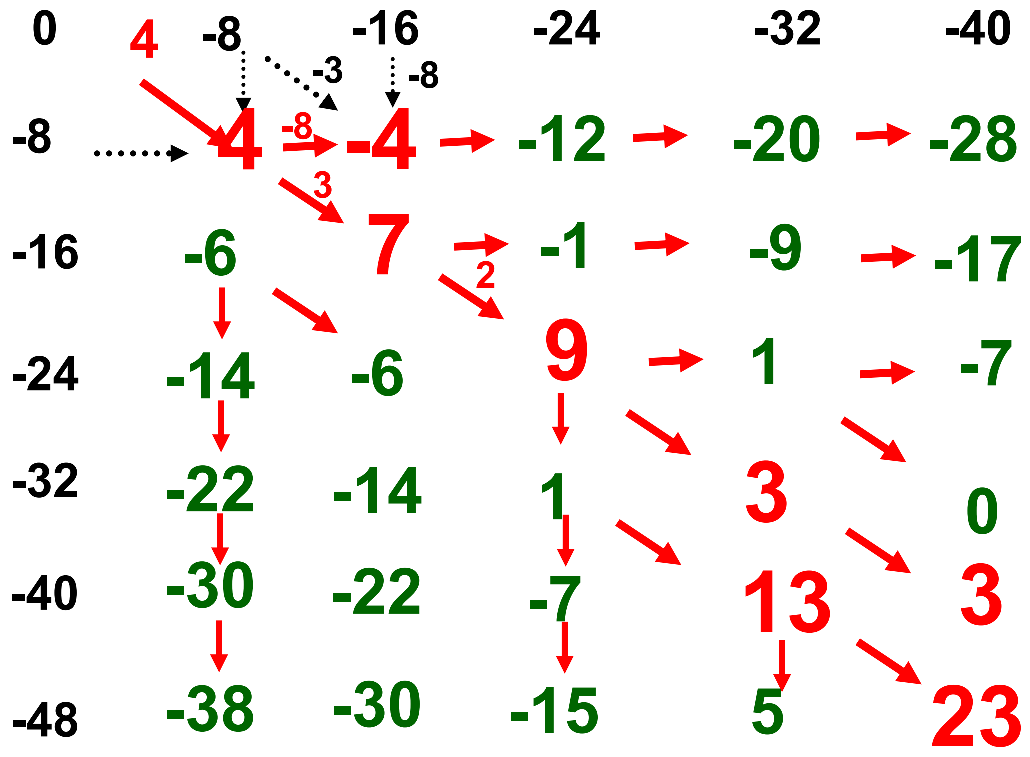
Multiple Sequence Alignments

- Sequences are aligned so as to bring the greatest number of single characters into register.
- If we include gaps, mismatches, then even dynamic programming becomes limited to ~ 3 sequences unless they are very short….need an alternative approach…

Why?

Consider the 2 sequence comparison

…..an O(mn) problem – order n^2



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

For 3 sequences….

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

An O(mnj) problem !

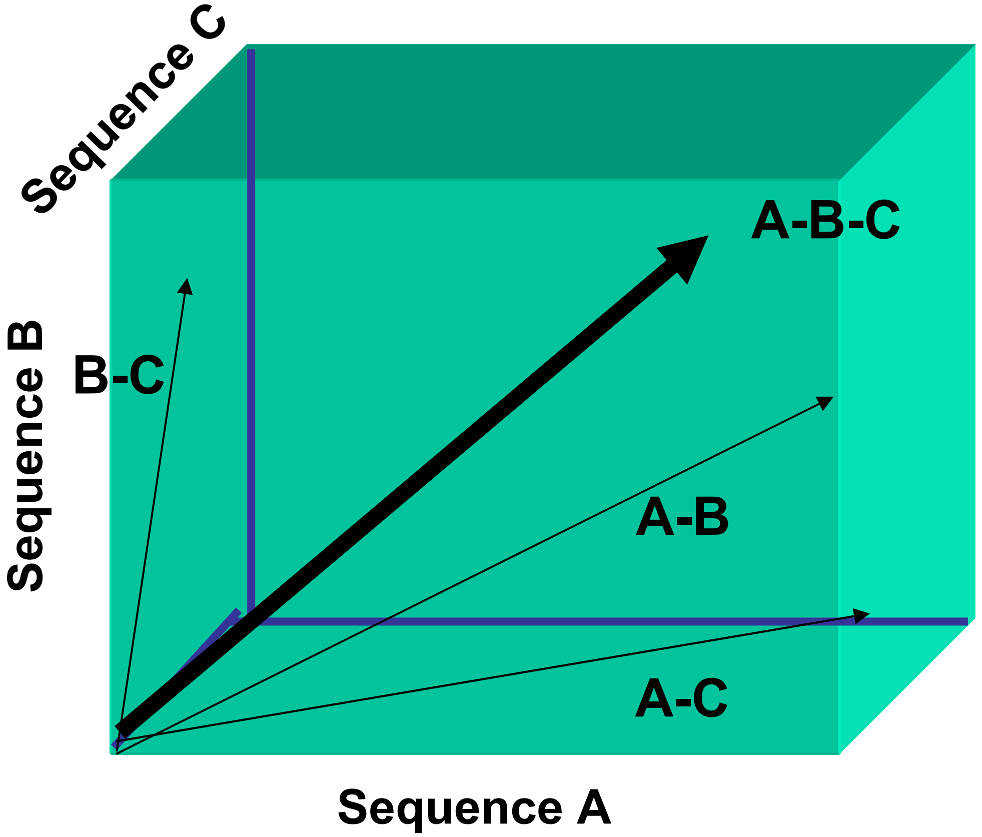
Consider sequences each 300 amino acids

Uh Oh !!!
Our polynomail problem
Just became exponential!

2 sequences – $(300)^2$
3 sequences – $(300)^3$
but for v sequences – $(300)^{v}$

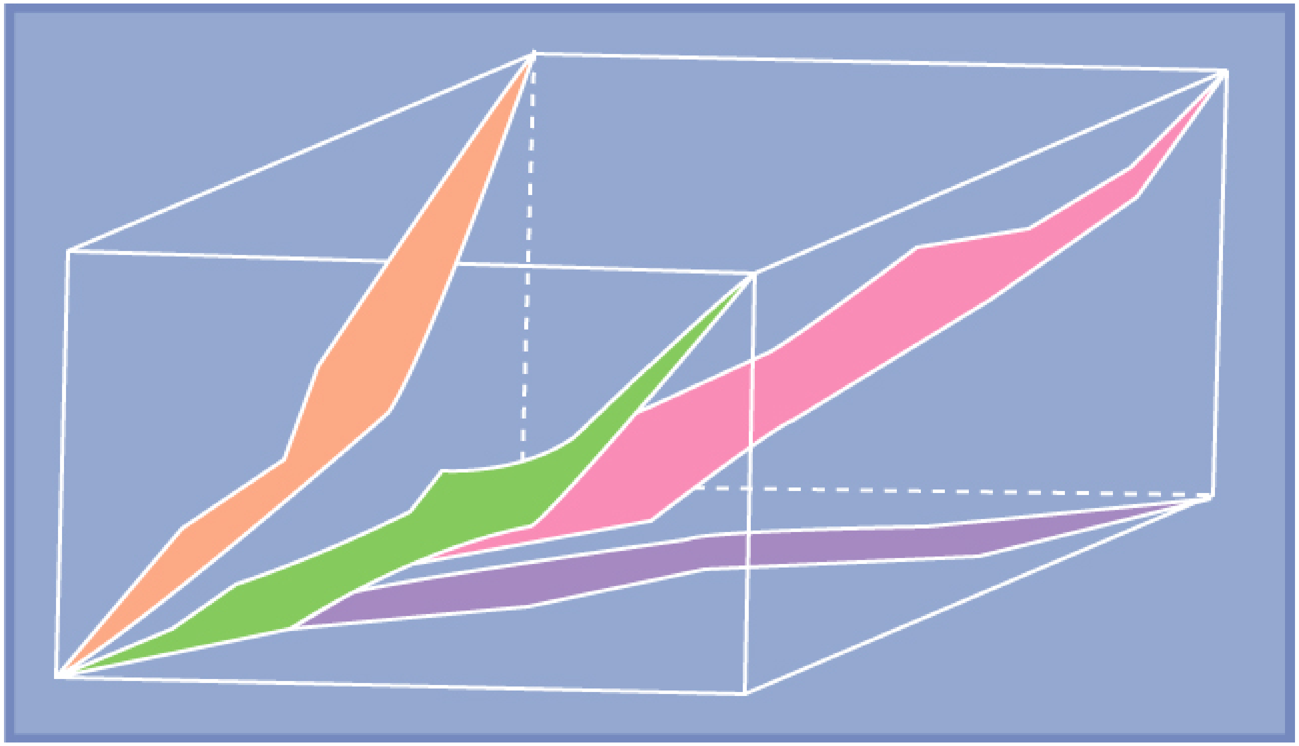
Consider pairwise alignments between 3 sequences

Carillo and Lipman – Sum of Pairs method



Do we need to Score each node?

Get the multiple alignment score within the cubic lattice by Adding together the scores of the pairwise alignments…

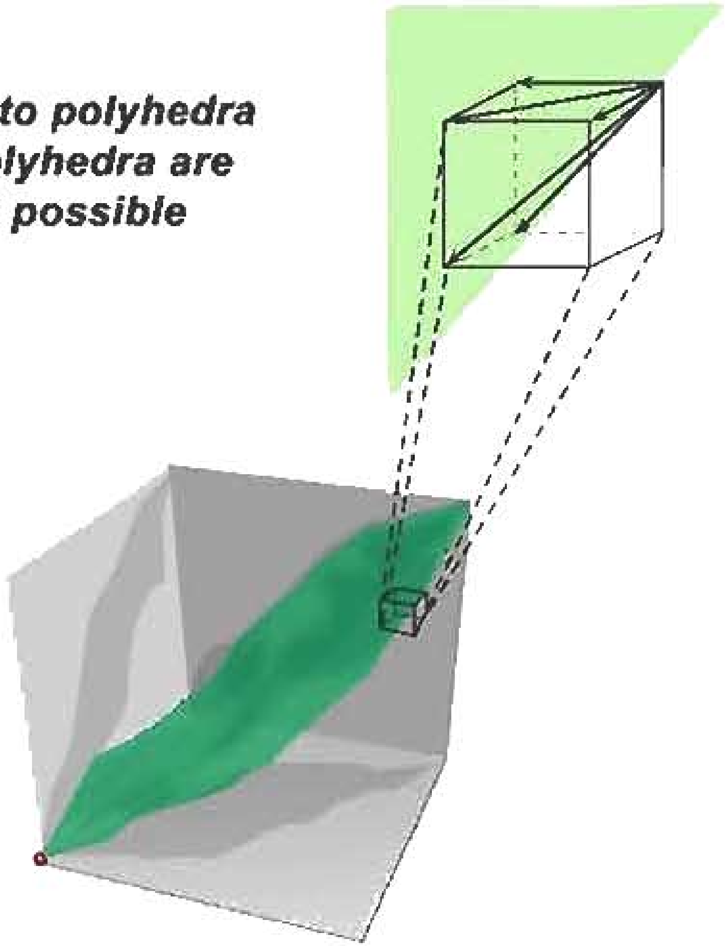


In practice, doesn't give optimal alignment…But we're close!

Seems reasonable that the optimal alignment won't be far from the diagonal we were on…so we just set bounds on the location of the msa within the cube based on each pairwise-alignment.

Then just do dynamic programing within the volume defined by the pre-imposed bounds

...the volume is broken into polyhedra and the borders of the polyhedra are defined by paths through possible alignments



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

- Progressive Methods of Multiple Sequence Alignment

Concept – simple:

1-Use DP to build pairwise alignments of most closely related sequences

2- Then progressively add less related sequences or groups of sequences…

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.

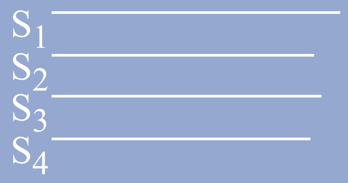
New features: Clustal ⌦ClustalW (allows weights) ⌦ ClustalX (GUI-based

Weighting is important to avoid biasing an alignment by many sequence Members that are closely related to each other evolutionarily!

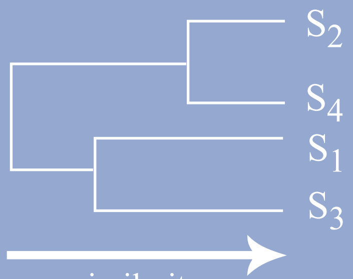
Steps in Multiple Alignment

Pairwise Alignment

Example - 4 sequences S_1 S_2 S_3 S_4

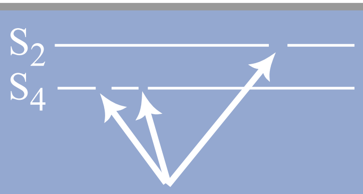


6 pairwise comparisons
then cluster analysis



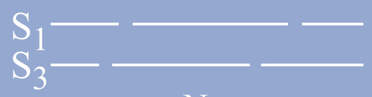
similarity

Multiple alignment following the tree from A



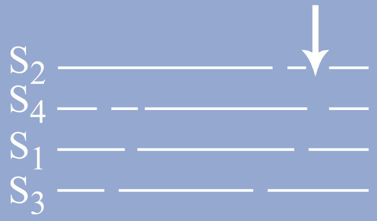
Gaps to optimize alignment

align most similar pair



align next most similar pair

New gap to optimize
alignment of (S_2S_4) with (S_1S_3)



align alignments-preserve gaps

Progressive Alignments

Note that the final msa is EXTREMELY DEPENDENT on the initial pairwise sequence alignments!

If the sequences are close in evolution, and you can see the alignment – GREAT!

If the sequences are NOT close in evolution, and you CANNOT See the alignment – errors will be propogated to the final msa

Has led to other approaches to do msa's that aren't so Dependent on initial states….i.e. genetic algorithm

Finding patterns (i.e. motifs and domains) in Multiple Sequence Analysis

Block Analysis, Position Specific Scoring Matrices (PSSM)

BUILD an msa from groups of related proteins

BLOCKS represent a conserved region in that msa that is LACKING IN GAPS – i.e. no insertions/deletions

The BLOCKS are typically anwhere from 3-60 amino acids long, based on exact amino acid matches – i.e. alignment will tolerate mismatches, but doesn't use any kind of PAM or BLOSUM matrix…in fact they generate the BLOSUM matrix!

A single protein contains numerous such BLOCKS separated by stretches of intervening sequences that can differ in length and composition.

These blocks may be whole domains, short sequence motifs, key parts of enzyme active sites etc, etc.
BLOCKS database….so far exploration limited. Lots of stuff to probe!

Can use these conserved BLOCKS to derive a PSSM

- The dirty secret behind prosite! Scansite! And in a twised way Psi-BLAST!

```
    12345…………..11
….GDSFHYFVSHG…. .
….GDAFHYYISFG…. .
….GDSYHYFLSFG…. .
….SDSFHYFMSFG…. .
….GDSFHFFASFG…. .
```

Now build a matrix with 20 amino acids as the columns, and 11 rows
For the positions in the BLOCK

Each matrix entry is the log(frequency of the amino acid occurance) at that position in the BLOCK.

….GDSFHYFVSHG…. .

….GDAFHYYISFG…. .

….GDSYHYFLSFG…. .

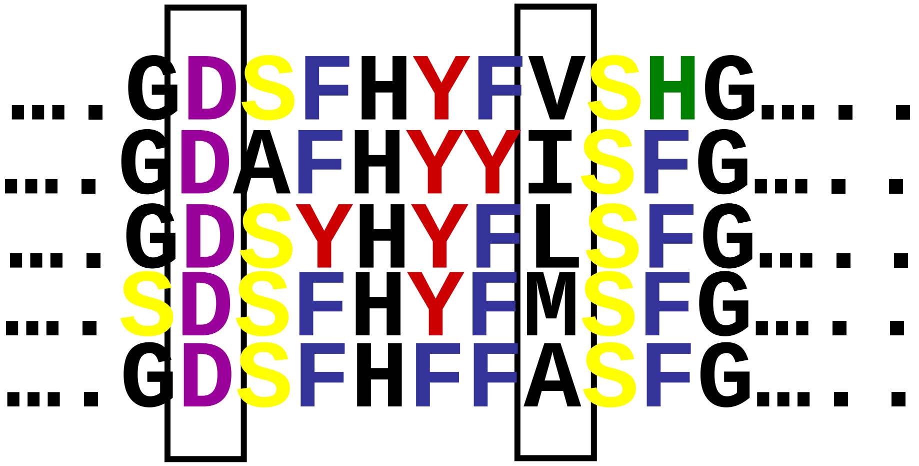
….SDSFHYFMSFG…. .

….GDSFHFFASFG…. .

	A	C	D	E	F	G	H	I	K….
1						Log(4)			
2			Log(5)						
3									
4									
5									

We can now use the PSSM to search a database for other proteins that have the BLOCK (or motif).

Problem 1 – We need to think about what kind of information is Contained within the PSSM.
→Leads to concepts of Information Content & Entropy



```
….GDSFHYFVSHG…. .
….GDAFHYYISFG…. .
….GDSYHYFLSFG…. .
….SDSFHYFMSFG…. .
….GDSFHFFASFG…. .
```

Problem 2 –The PSSM must accurately represent the expected BLOCK Or motif….and we have only limited amounts of data! Is it a good statistical Sampling of the BLOCK/motif? Is it too narrow because of small dataset? Should we broaden it by adding extra amino acids that we choose using Some type of randomization scheme (called adding pseudocounts). If so, How many should we add?

Finding patterns (i.e. motifs and domains) in Multiple Sequence Analysis

Block Analysis, Position Specific Scoring Matrices (PSSM)

BUILD an msa from groups of related proteins

BLOCKS represent a conserved region in that msa
that is LACKING IN GAPS – i.e. no insertions/deletions

The BLOCKS are typically anwhere from 3-60 amino acids long, based on exact amino acid matches – i.e. alignment will tolerate mismatches, but doesn't use any kind of PAM or BLOSUM matrix…in fact they generate the BLOSUM matrix!

These blocks may be whole domains, short sequence motifs, key parts of enzyme active sites etc, etc.

Position Specific Scoring Matrices
PSSM

```
  12345………….11
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….GDSYWNYASFG…. .
```

This BLOCK might represent some small part of a modular protein domain, or might represent a motif for something …..like a phosphorylation site on the S in position 9

Now build a matrix with 20 amino acids as the columns, and 11 rows For the positions in the BLOCK

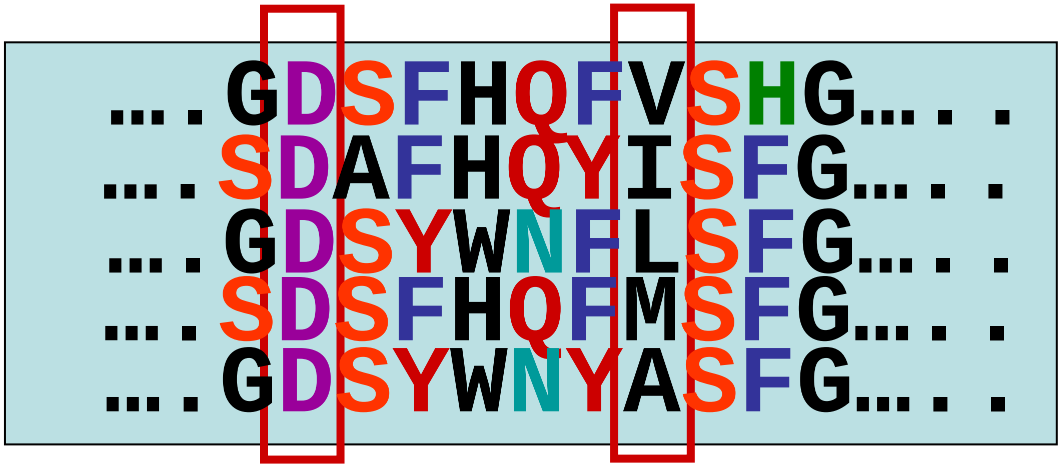
Each matrix entry is the Log(frequency of the amino acid occurance) at that position in the BLOCK.

```
    12345…………..11
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….GDSYWNYASFG…. .
```

Position	A	C	D	E	F	G	H	I	K….		S	T…
1						Log(3)					Log(2)	
2			Log(5)									
3												
4												
5												

We can now use the PSSM to look for the BLOCK (motif) in single proteins
-or-
use the PSSM to search a database for other proteins that have the BLOCK (or motif).

Problem 1 –The PSSM must accurately represent the expected BLOCK Or motif….and we have only limited amounts of data! Is it a good statistical Sampling of the BLOCK/motif? Is it too narrow because of small dataset? Should we broaden it by adding extra amino acids that we choose using Some type of randomization scheme (called adding pseudocounts). If so, How many should we add?



```
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….GDSYWNYASFG…. .
```

Problem 2 –We need to think about what kind of information is Contained within the PSSM.

→Leads to concepts of Information Content & Entropy

Pseudocounts

- If the number of sequences in the training set is both large and diverse, then the sequences in the training set represent a good statistical sampling of the motif….*if not, then we have a sampling error!*

Correct for this by adding pseudocounts. How many to add?

→ ***Too many pseudocounts dominate the frequencies… and the resulting matrix won't work!***

→ ***Too few pseudocounts then we'll miss many amino acid variations, and matrix will only find sequences that produced the motif!***

Add few pseudocounts if sampling is good (robust), and add more pseudocounts if sampling is sparse

One reasonable approach is to add √N pseudocounts, where N is the number of sequences…
As N increases, the influence of pseusocounts decreases since N increases faster than √N, but doesn't add enough at low N

Pseudocounts

How do we choose which amino acids to use for the pseudocounts?

Approach 1 **– Choose the randomly based on f(i), the frequency of With which amino acid I is found in all proteins in GenBank… ……seems wrong somehow, since not all amino acids should substitute for others in an alignment….**

Approach 2 **– Use information from scoring matrices (i.e. PAM250 or BLOSUM 62 matrices, since these give estimates of amino acid substitutions….OK, but may not make optimal use of the data from an aligned sequence block**

Approach 3 **– Use data from some subset of other aligned sequence BLOCKS as an estimate of the amino acid distribution in the block we are working with. These capture some features of protein structure and function. Sjolander et al have prepared such amino acid distributions and these are referred to as Dirichlet mixtures.**

A word on Dirichlet Mixtures

Concept comes from Bayesian statistics….idea of posterior distributions. Obtained by multiplying prior probabilities (i.e. what you know) with likelihood of new data ….OK, maybe more on this later

An easier to understand, almost correct explanation:
(Bayes without tears)

Suppose in one column of an alignment of 27 sequences we find:
12 G's, 6 T's, 6 F's and 3 P's.

The probability of such a distribution of amino acids is:

$$P(12G, 6T, 6F, 3P)=\frac{n!(pG)^{12}(pT)^{6}(pF)^{6}(pP)^{3}}{12!\ 6!\ 6!\ 3!}$$

Where n! = (12+6+6+3)!

This is called a multinomial distribution.

A word on Dirichlet Mixtures

We want to get at the prior probablities that would generate this type of multinomial distribution,
i.e. that would statistically favor the choice amino acids we actually observed in our alignment.

We also have to consider several other alternatives that are close to the exact distribution of amino acids we actually observed, but might also be reasonable, i.e.
11 G's, 7 T's, 5 F's and 4 P's.
This would have a slightly different multinomial distribution

So we also need to consider the prior probabilities that would reflect these other similar, but not identical amino acid distributions for the alignment

Finally, we weigh all of these different prior probabilities that would generate these similar but not identical multinomial distributions into a combined frequency distribution…
AND THIS IS THE DIRICHLET MIXTURE

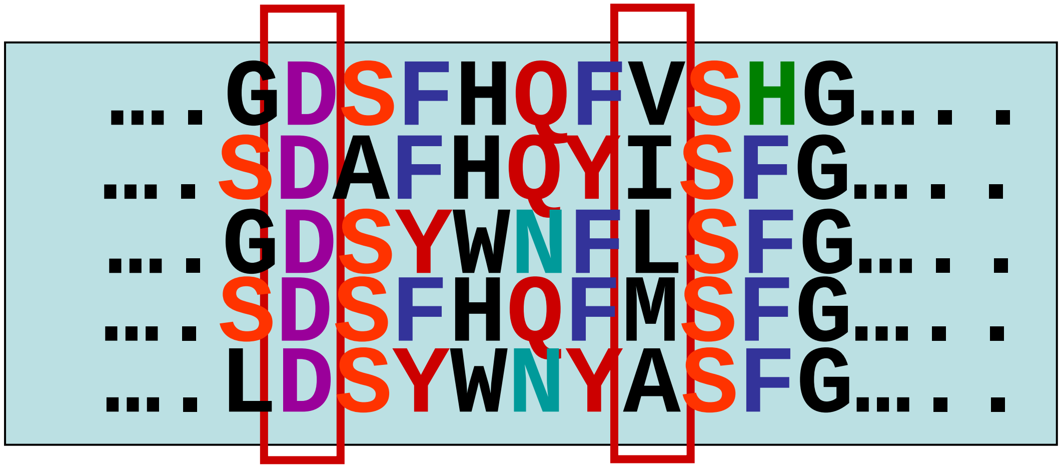
The last word on Dirichlet Mixtures

At the moment, these Dirchelet mixtures provide the best way to analyze amino acid compositions in an alignment that can detect weak but significant sequence similarity.

Dirchelet mixtures are used extensively in several Hidden Markov Models of sequence alignments (later in this lecture).

Why? Because you can use fewer sequences in the training set (i.e. 20 or so for HMMs) and still derive good motifs!

Information Content



```
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….LDSYWNYASFG…. .
```

Derived from the field of Information Theory

Concept is simple – ***some positions in an alignment provide alot more information than others as far as identifying a motif is concerned***

How do we quantitate this?

Information Theory

Photograph courtesy of Lucent Technologies. http://www.bell-labs.com/news/2001/february/26/1.html
Copyright notice: http://www.lucent.com/copyright.html

Originated with Claude Shannon, an MIT PhD who worked for Bell Labs and was interested in how much information you could transmit, how fast you could send it, and how accurately you could transmit it, across telephone lines

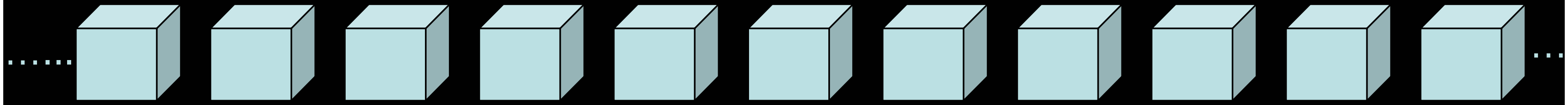
Information Theory

What is the minimum amount of information necessary to convey a message unambiguously?
How do we quantitate the uncertainty in a message?

How much information is present in each position in a sequence motif. How can we quantitate information content in motif sequences? How can this help us understand the biological meaning of motifs?

Complex Problem

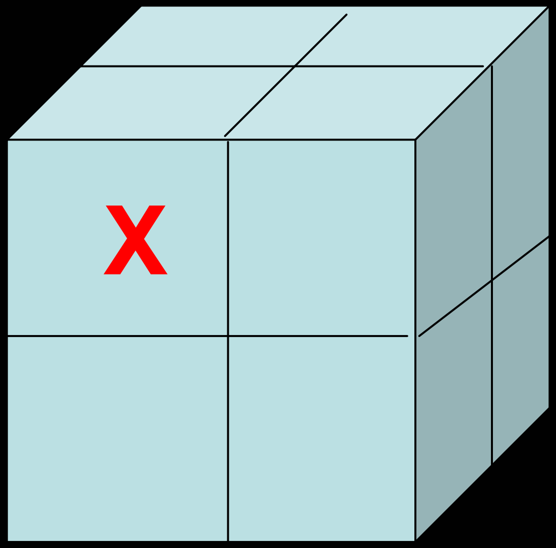
I put 1,048,576 identical boxes in a straight line and I hide $500 into one of them at random



Now let's play 20 questions….(yes/no)

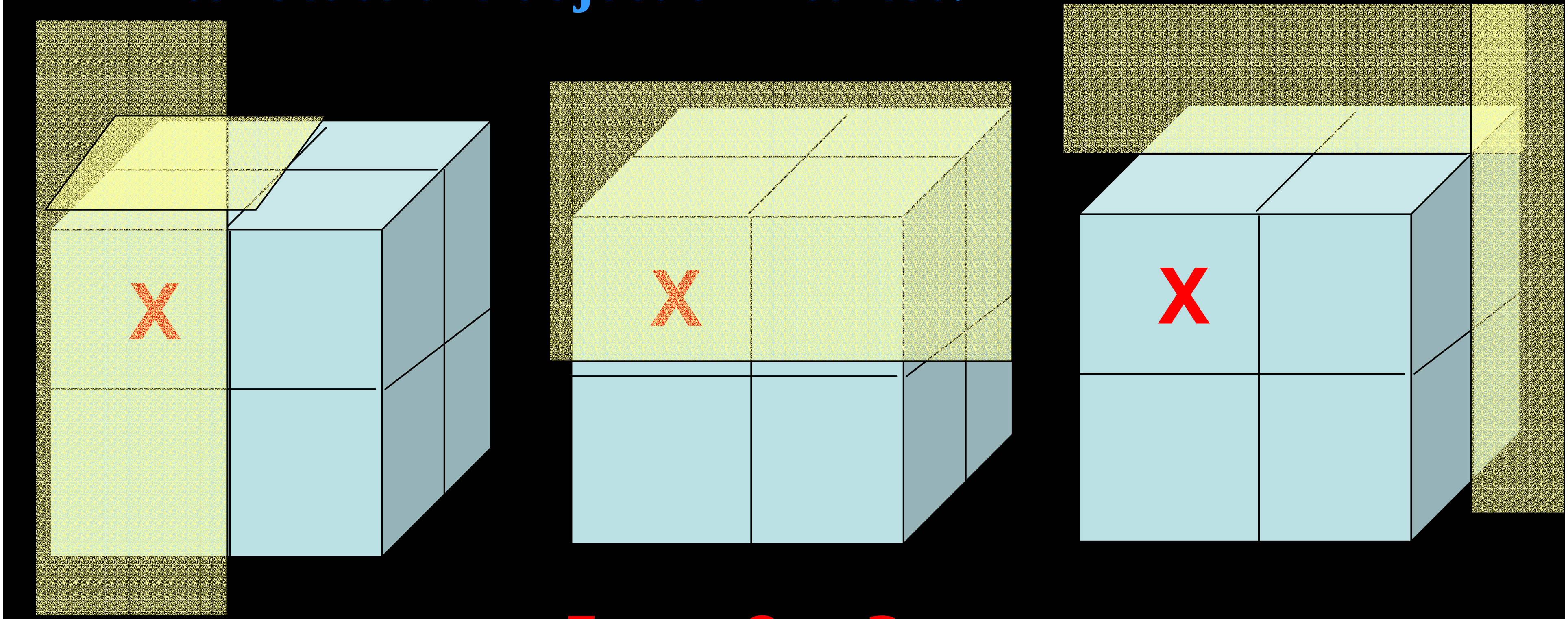
Simpler Example

Minimum number of yes-no questions to locate an object of interest?



Simple Example

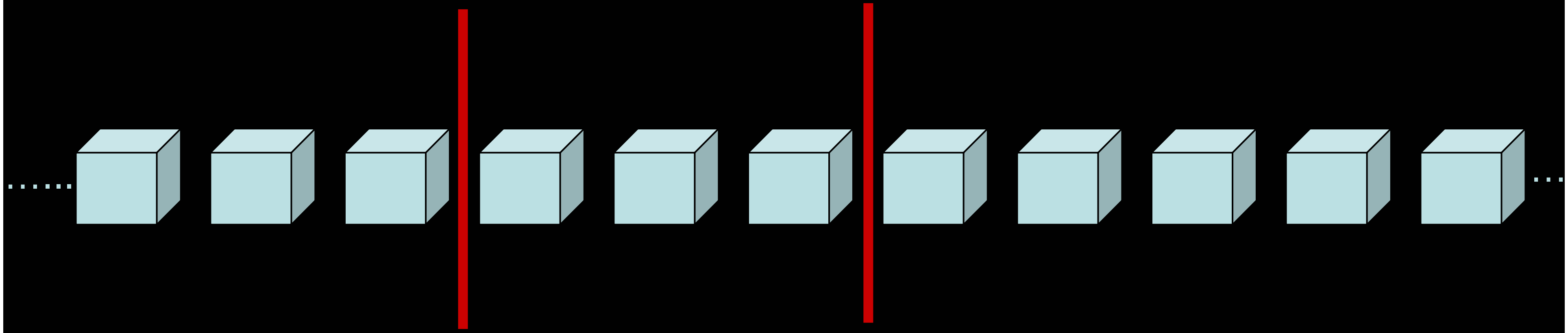
Minimum number of questions to locate the object of interest?



$$\textbf{Log}_2\ \mathbf{8 = 3}$$

Complex Problem

I put 1,048,576 identical boxes in a straight line and I hide $500 into one of them at random



Now let's play 20 questions….(yes/no)

2^{20}=1,048,576

Information Theory

Since the information was derived using base2 of the total number of probabilities (i.e. computers), the resulting information is said to be coded in units of bits (binary digits)

Could have used base *e* for the logs of the probabilities, (i.e. natural Logs) – then the resulting information is said to be coded in units of nats (natural digits)

Could have used base 10 (i.e.) for the logs of the probabilities – then the resulting information is said to be coded in units of dits
(decimal digits)

Information Theory

So we have the relationship that : $2^b=M$
where M = number of possibilities.

or

$$b= \log_2(M)$$

where b=number of bits required to uniquely determine which, of all possibilties M, a particular value is.

b tells us something about how many bits of information we need to specify all the possibilities for a position.

Can also write $b= - \log_2 (1/M)$

For equivalent probabilities P = 1/M

$$b= - \log_2 (P)$$

For 20 amino acids in one position in a sequence motif
If all 20 had an equal chance, P = 1/20

$b= - \log_2 (1/20) =$ ***4.32 bits of information***

Information Theory

What happens when the probabilities aren't the same?

Concept of surprisals (Tribus)

Watching TV analogy….I flash up one of the amino acids, and you guess Which one it is…..

Define u_i = - log2 (P_i)

u_i = the surprisal….a measure of how surprised you are by the amino acid that appears on the screen….

In a sequence alignment, we want to know the average surprisal in each column of the alignment.

Information Theory

```
….GDSFHQFVSHG….  .
….SDAFHQYISFG….  .
….GDSYWNFLSFG….  .
….SDSFHQFMSFG….  .
….LDSYWNYASFG….  .
```

Average surprisal = $$\frac{u_V+u_I+u_L+u_M+u_A}{N}$$ ← *total # of sequences*

In general case: $$\frac{\sum_{i=1}^{20} N_i u_i}{N}$$ ← *Ni = # of sequences with Amino acid i in that position N =total # of sequences*

Bring N inside the Summation sign $\sum_{i=1}^{20} \frac{N_i u_i}{N}$ ← *Ni = # of sequences with Amino acid i in that position*
N =total # of sequences

But Ni/N is just Pi, the probability of finding amino acid i in that position of the alignment.

So we get…

$$\sum_{i=1}^{20} P_i u_i$$

We defined u_i = - log2 (P_i)

So we end up with Shannon's famous formula:

$$H = - \sum_{i=1}^{20} Pi(\log_2 Pi)$$

Where H = the "Shannon Entropy"
In bits per position in the alignment

Information Theory

So we end up with Shannon's famous formula:

$$H = - \sum_{i=1}^{20} \mathit{P}\text{i}(\log_2 \mathit{P}\text{i})$$

Where H = the "Shannon Entropy"
In bits per position in the alignment

What does this mean???

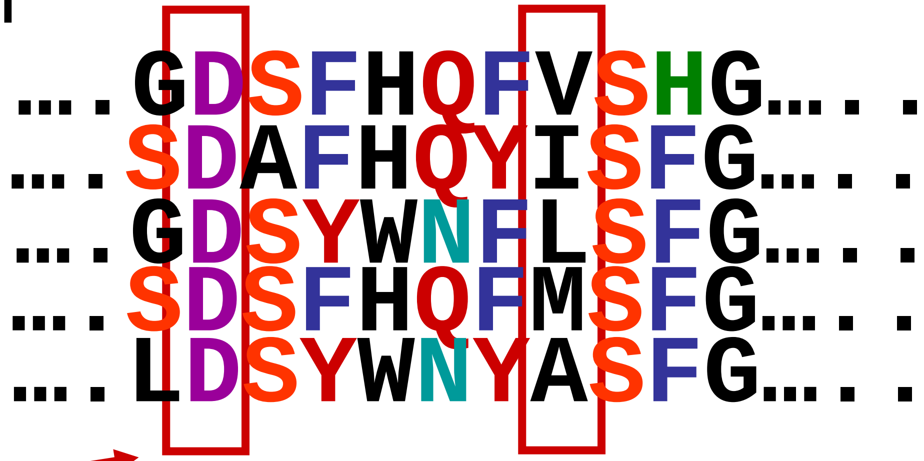
H is a measure of entropy or randomness or disorder ….it tells us how much uncertainty there is for the different amino acid abundances at one position in a sequence motif

Information Theory

Shannon's famous formula:

$$H = -\sum_{i=1}^{20} Pi(\log_2 Pi)$$

Where H = the "Shannon Entropy"
In bits per position in the alignment



```
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….LDSYWNYASFG…. .
```

Here P(D)=1 so H = -(1 * $\log_2$(1)) = -(1*0) = 0 !

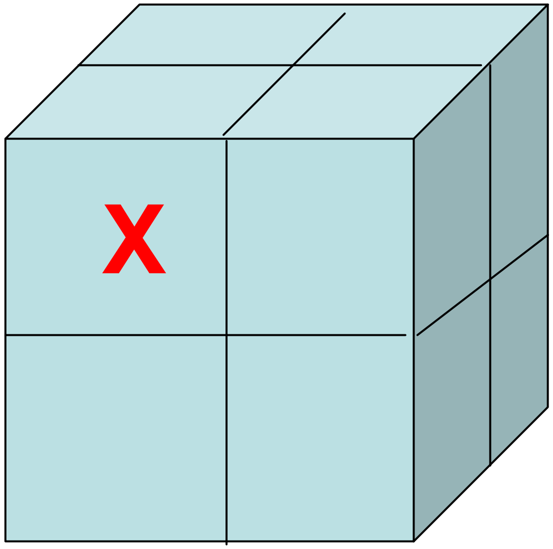
No uncertainty in this position!!

Here H = -(1/5 * $\log_2$(1/5)) **-(1/5 * $\log_2$(1/5)) -(1/5 * $\log_2$(1/5)) -(1/5 * $\log_2$(1/5)) -(1/5 * $\log_2$(1/5))** = 2.32

Great uncertainty in this position!!

Information Theory

From Uncertainty to Information



Back to our cube

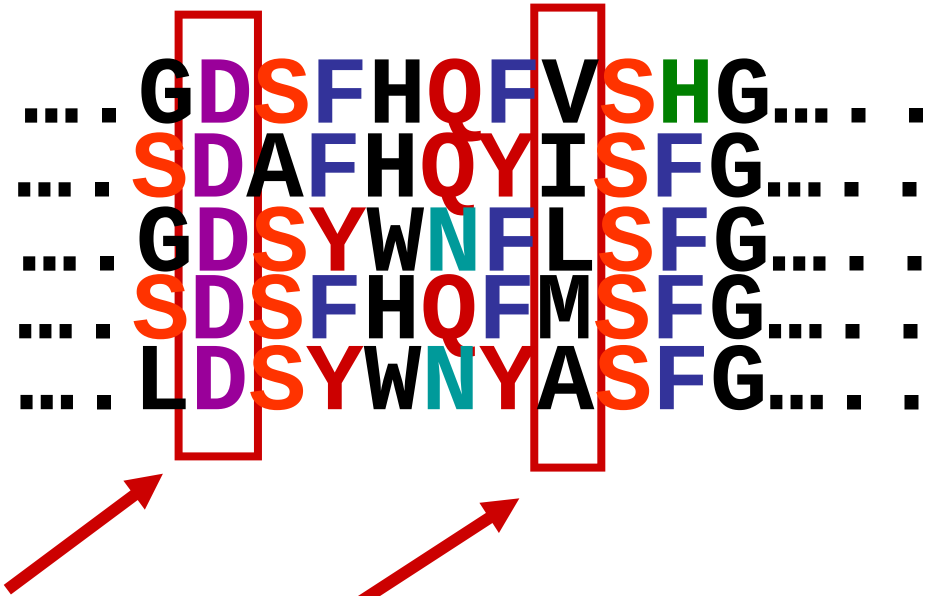
Suppose I would only answer 2 questions…

Now the uncertainty before would have been 3 bits

Uncertainty after is 1 bit

Therefore, the information content you have gained, R, is
$R = H_{before} – H_{after}$ = 3 bits -1 bit = 2 bits!

Information Theory



```
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….LDSYWNYASFG…. .
```

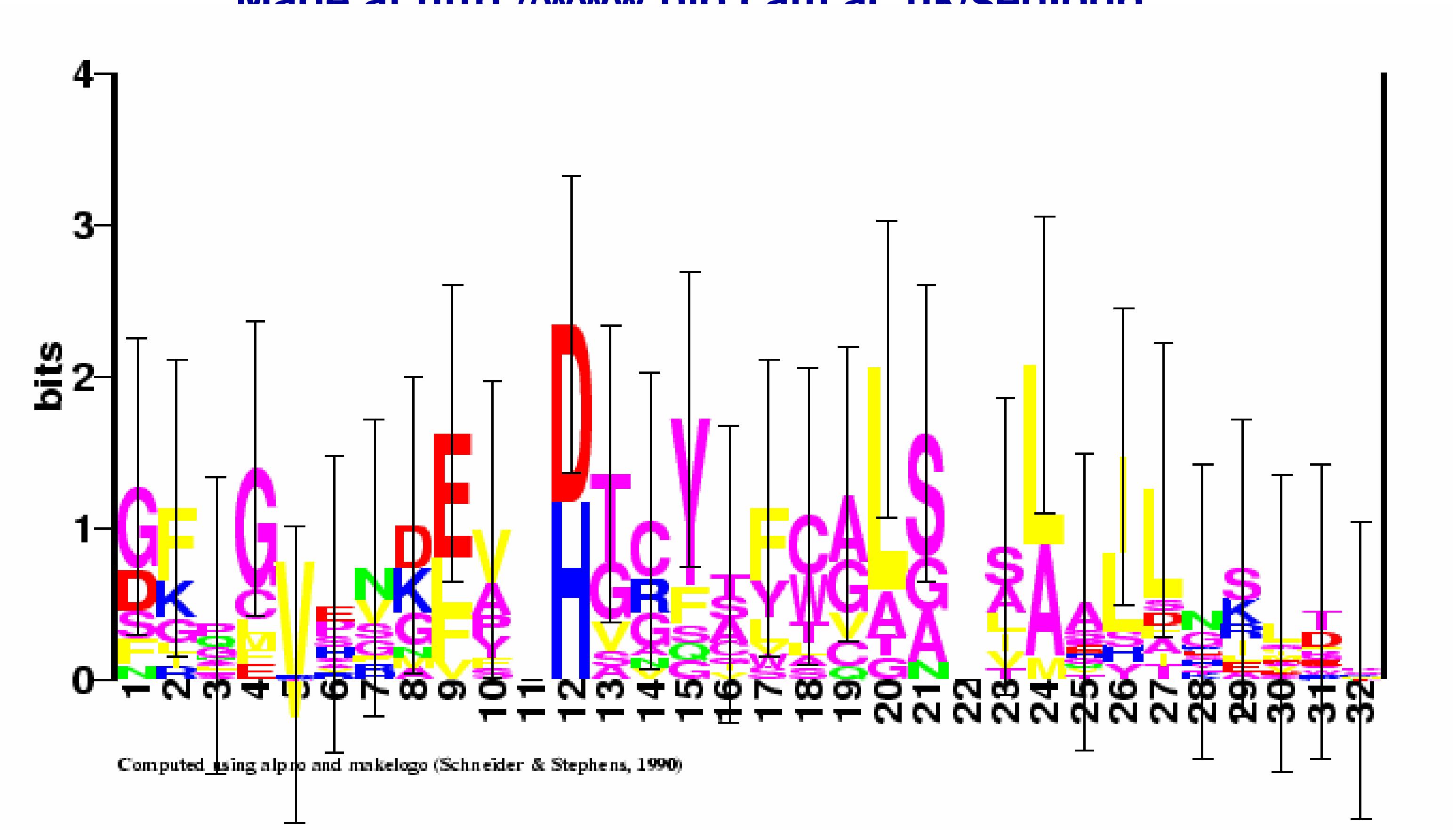
Assuming all 20 amino acids equally possible:
H_{before} = 4.32, H_{after}=0
Therefore, this position encodes 4.32-0 =4.32 bits of information!

Another position in the motif that contains all 20 amino acids…
H_{before} = 4.32, H_{after}=4.32
Therefore, this position encodes 4.32-4.32 =0 bits of information!

Sequence Logos

Motif “A” in prenyltransferases,

Made at http://www.bio.cam.ac.uk/seqlogo



Computed using alpro and makelogo (Schneider & Stephens, 1990)

For the compulsively nerdy – error bars represent possible error in the stack due to limited number of sequences

Sequence Logos

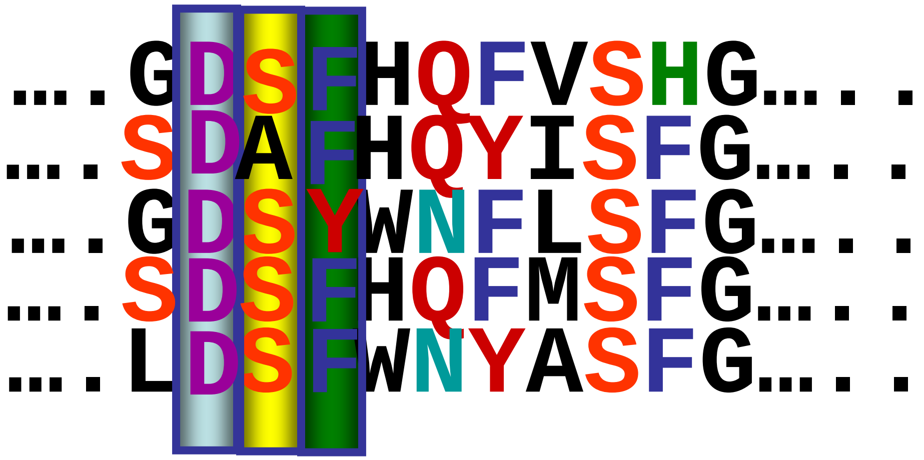
1- Horizontal axis is position within a sequence motif

2 - Vertical axis is amount of information in bits that the Position holds. The height of the individual letters is H_{before}-H_{letter}, i.e. how much of that information that particular letter (i.e. amino acid) contributes

3- Very useful graphical representations that convey relative information content at various positions in a motif and contributions from relative amino acids.

Markov Chains & Models

In PSSMs, each position is evaluated independently.



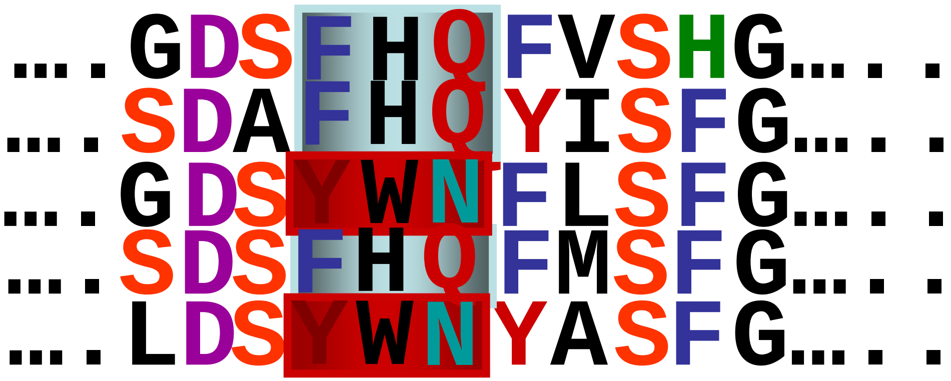
```
….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….LDSFWNYASFG…. .
```

Note – we can use PSSMs for aligning sequences, finding domains, mapping motif sites, and even secondary structure prediction….

Are we capturing all the information?

Markov Chains & Models

What other information is there?

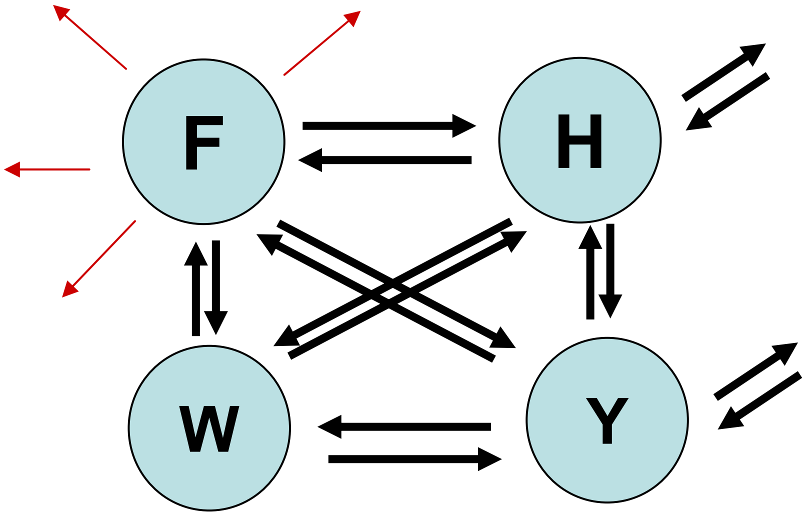


….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….LDSYWNYASFG…. .

How do we model this? …A Markov model!

A Markov Chain

….GDSFHQFVSHG…. .
….SDAFHQYISFG…. .
….GDSYWNFLSFG…. .
….SDSFHQFMSFG…. .
….LDSFWNYASFG…. .



A collection of states:

Probability parameters =transition probabilities

Net probabilities for the entire sequence involve conditional probabilities – Bayes!

The probability of any amino acid depends ONLY on the value of the preceding amino acid.

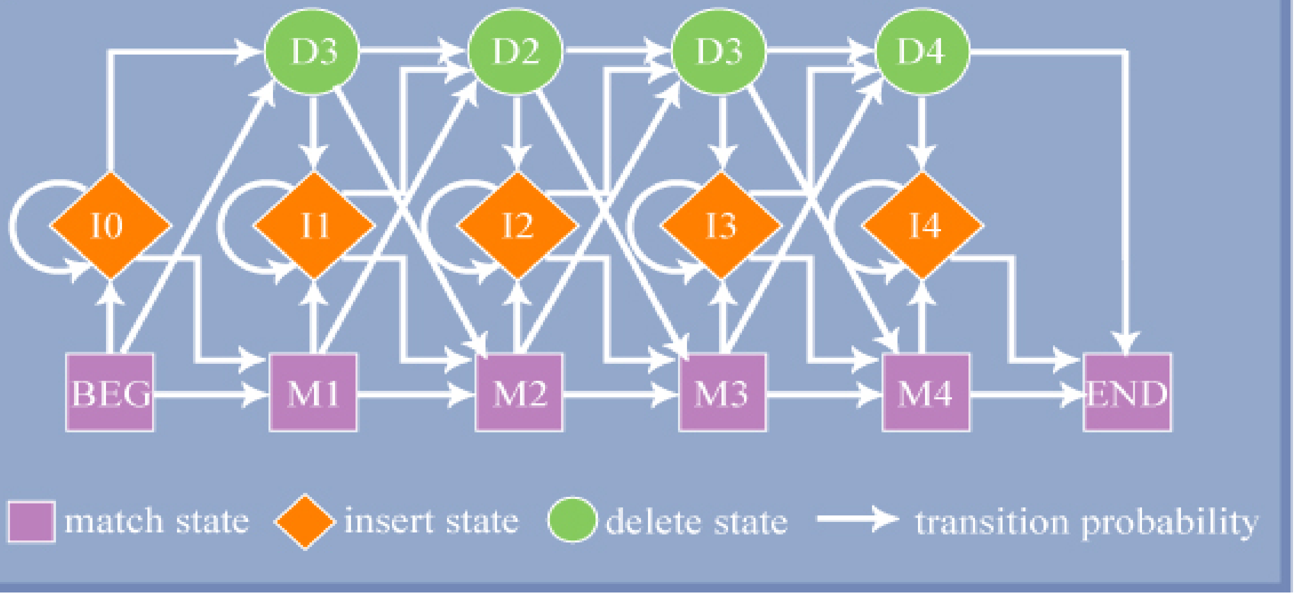
Markov Models

Sequence alignment

N	•	F	L	S
N	•	F	L	S
N	K	Y	L	T
Q	•	W	-	T

PURPLE POSITION REPRESENTS ALIGNMENT IN COLUMN
ORANGE POSITION REPRESENTS INSERT IN COLUMN
GREEN POSITION REPRESENTS DELETE IN COLUMN

Hidden Markov model for sequence alignment



Hidden Markov Models

Statistical model that includes all combinations of matches, mismatches and gaps (prior slide)

Very useful to obtain best multiple sequence alignment or find new members of the alignment!

Make a starting model, and use 20-100 aligned sequences to train it (i.e. get the transition probabilities tuned).

Often use Dirichlet mixtures when deciding which amino acid substitutions to allow in a motif

Try and generate all the sequences in the training set by examining all possible paths through the model. This might be a huge computational problem, but it is reduced using something called the "Forward-Backward" algorithm

Another algorithm, Baum-Welch, counts the number of times each state-to-state transition was used, and that a particular amino acid is required to generate one of the sequences in the training set

Make an improved HMM that uses the results of the Baum-Welch algorithm, and repeat cyclically until the parameters (i.e. the transition probabilities) stabilize.

Now you have a trained model that can provide the most likely path to generate each sequence. There is another algorithm, the Viterbi algorithm, that uses dynamic programming to find the best path to generate each sequence, given the model.

This collection of paths, (i.e. match-match-insertion-match…) defines the final multiple sequence alignment.

Note that in the end, all you have are transition probabilities for each amino acid at each position in an alignment. The model that generated these is "hidden"

…The Hidden Markov Model!

Get multiple sequence alignments as good or better than those
Obtained with other techniques.

Both Pfam and SMART, two software programs that search for domains
In amino acid quieries, use HMMs

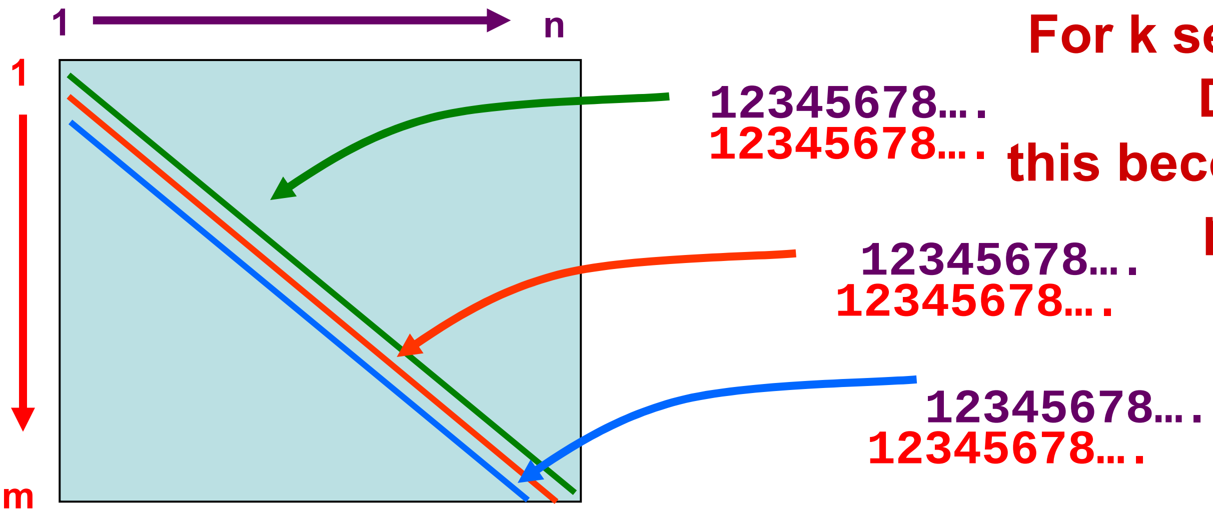
Database Searching

Problem is simple:
I want to find homologues to my protein in the database
How do I do it?

Do the obvious – compare my protein against every other protein in the database and look for local alignments by dynamic programming

Uh Oh!

For k sequences in the Database this becomes an O(mnk) problem!



….essentially an O(mn) problem

Database Searching

Still, this can be done - ~ 50x slower than Blast/FASTA, Smith-Waterman algorithm…
SSEARCH (ftp.virginia.edu/pub/fasta) – do it locally!

But in the old days, needed a faster method…
2 approaches – Blast, FASTA – both heuristic (i.e. tried and true) – almost always finds related Proteins but cannot guarantee optimal solution

FASTA: Basic Idea

1- Search for matching sequence patterns or words

Called k-tuples, which are exact matches of "k" characters between the two sequences

i.e. RW = 2-tuple

Seq 1: AHFYRWNKLCV **Seq 2: DRWNLFCVATYWE**

Database Searching

FASTA: Basic Idea

2- Repeat for all possible k-tuples

i.e. CV = 2-tuple

Seq 1: AHFYRWNKLCV　　　Seq 2: DRWNLFCVATYWE

3- Make a Hash Table (Hashing) that has the position of each k-tuple in each sequence

i.e.

2-tuple	pos. in Seq1	pos in Seq 2	Offset (pos1-pos2)
RW	5	2	3
CV	10	7	3
AH	1	----	----

Database Searching

Seq 1: AHFYRWNKLCV   Seq 2: DRWNLFCVATYWE

3- Make a Hash Table (Hashing) that has the position of each k-tuple in each sequence

i.e.

2-tuple	pos. in Seq1	pos in Seq 2	Offset (pos1-pos2)
RW	5	2	3
CV	10	7	3
AH	1	----	----

4- Look for words (k-tuples) with same offset

These are in-phase and reveal a region of alignment between the two sequences.

5- Build a local alignment based on these, extend it outwards

Seq 1: AHFYRWNKLCV
Seq 2: DRWNLFCVATYWE

Database Searching

With hashing, number of comparisons is proportional To the average sequence length (i.e. an O(n) problem), Not an O(mn) problem as in dynamic programming.

Proteins – ktup = 1-2,
Nucleotides, ktup=4-6

One big problem – low complexity regions.

Seq 1: AHFYPPPPPPPPFSER
Seq 2: DVATPPPPPPPPPPPNLFK

Database Searching

BLAST

Same basic idea as FASTA, but faster and more sensitive!

How?

BLAST searches for common words or k-tuples, but limits the search for k-tuples that are most significant, by using the log-odds values in the Blosum62 amino acid substitution matrix

i.e. look for WHK and might accept WHR but not HFK as a possible match (note 8000 possibilities)

Repeat for all 3-tuples in the query

Search the database for a match to the top 50 3-tuples that match the first query position in the sequence, the second query position, etc.
Use any match to seed an ungapped alignment (old BLAST)

Database Searching

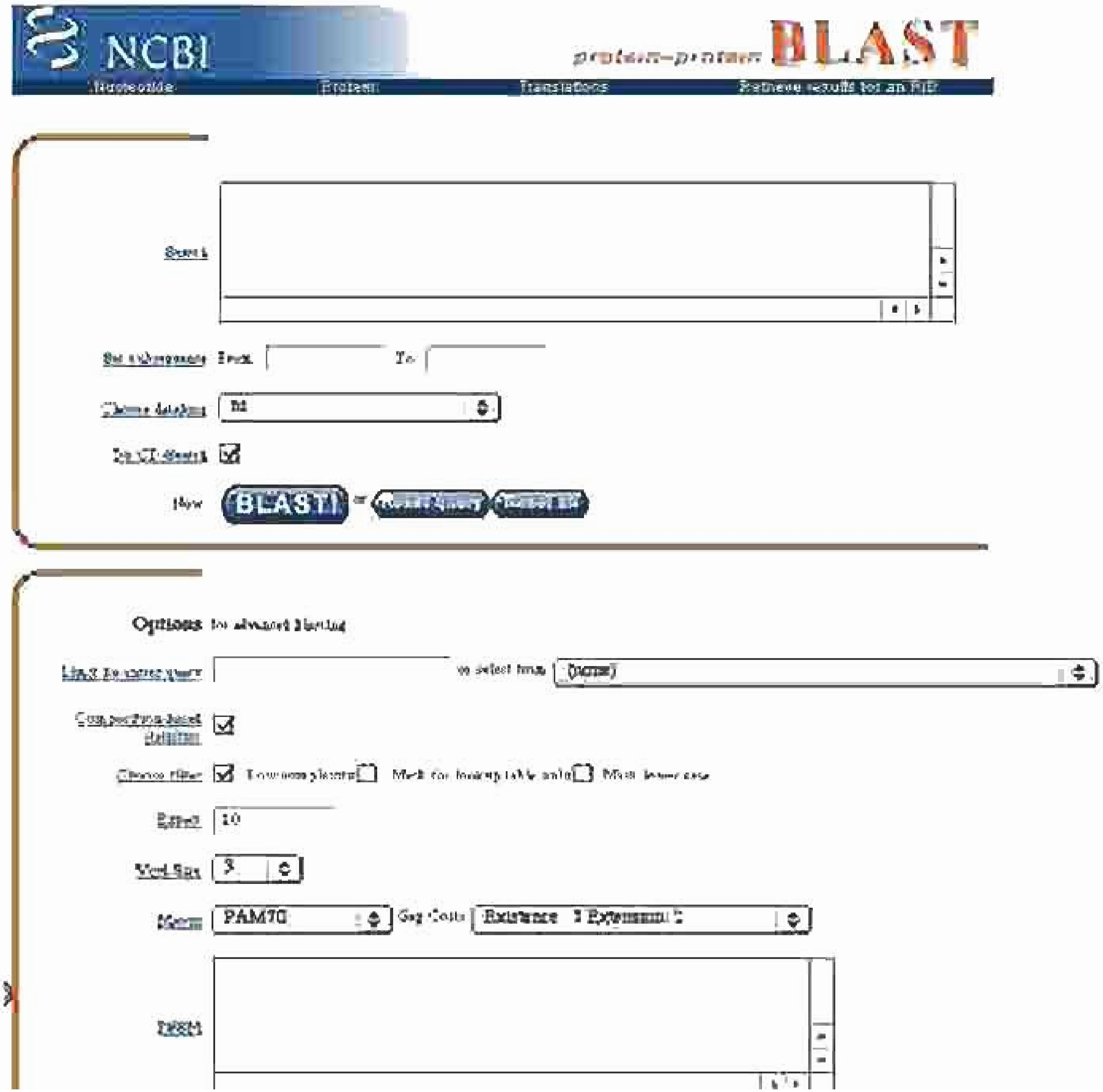
Word length is fixed:	**3-tuple for proteins**
	11-tuple for nucleotides

By default, filters out low complexity regions.

Determine if the alignment is statistically significant. calculates the probability of observing a score greater than or equal to your alignment based on extreme value distribution.
Calculates an E-value = expectation value:

This is the probability of finding an unrelated sequence that shows this good an alignment just by chance.

Remember if p=.0001 and my database has 500,000 sequences, I will have an E=50! (normal starting E=10)



Search

Set subsequence From To

Choose database nr

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from (none)

Composition-based statistics ☑

Choose filter ☑ Low complexity □ Mask for lookup table only □ Mask lower case

Expect 10

Word Size 3

Matrix PAM70 Gap Costs Existence: 10 Extension: 1

PSSM

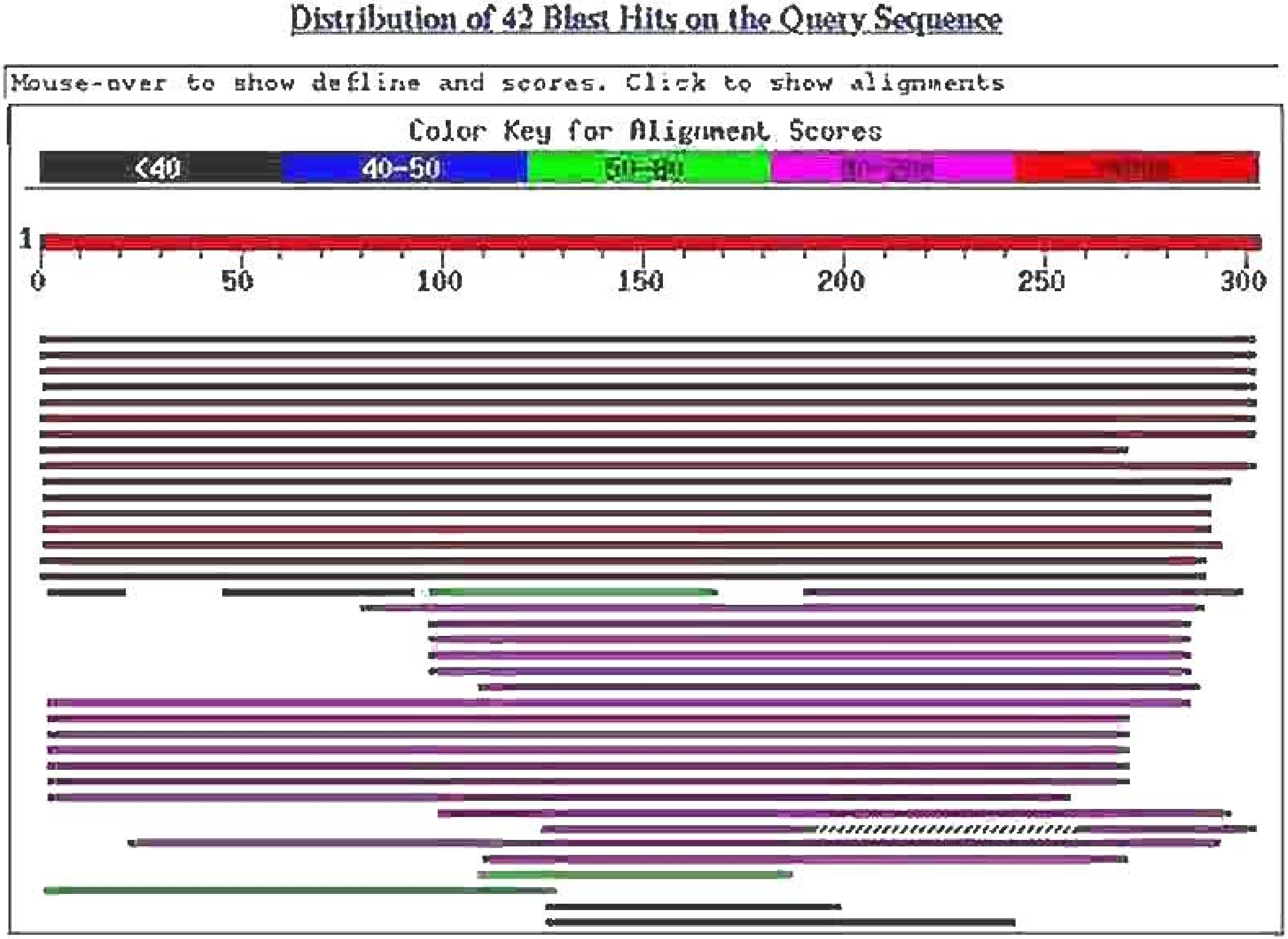
See: http://www.nlm.nih.gov

Psi-BLAST

Position-specific iterative BLAST

Combines BLAST searching with PSSMs!

1- Start with regular BLAST search – look at the results



Distribution of 42 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

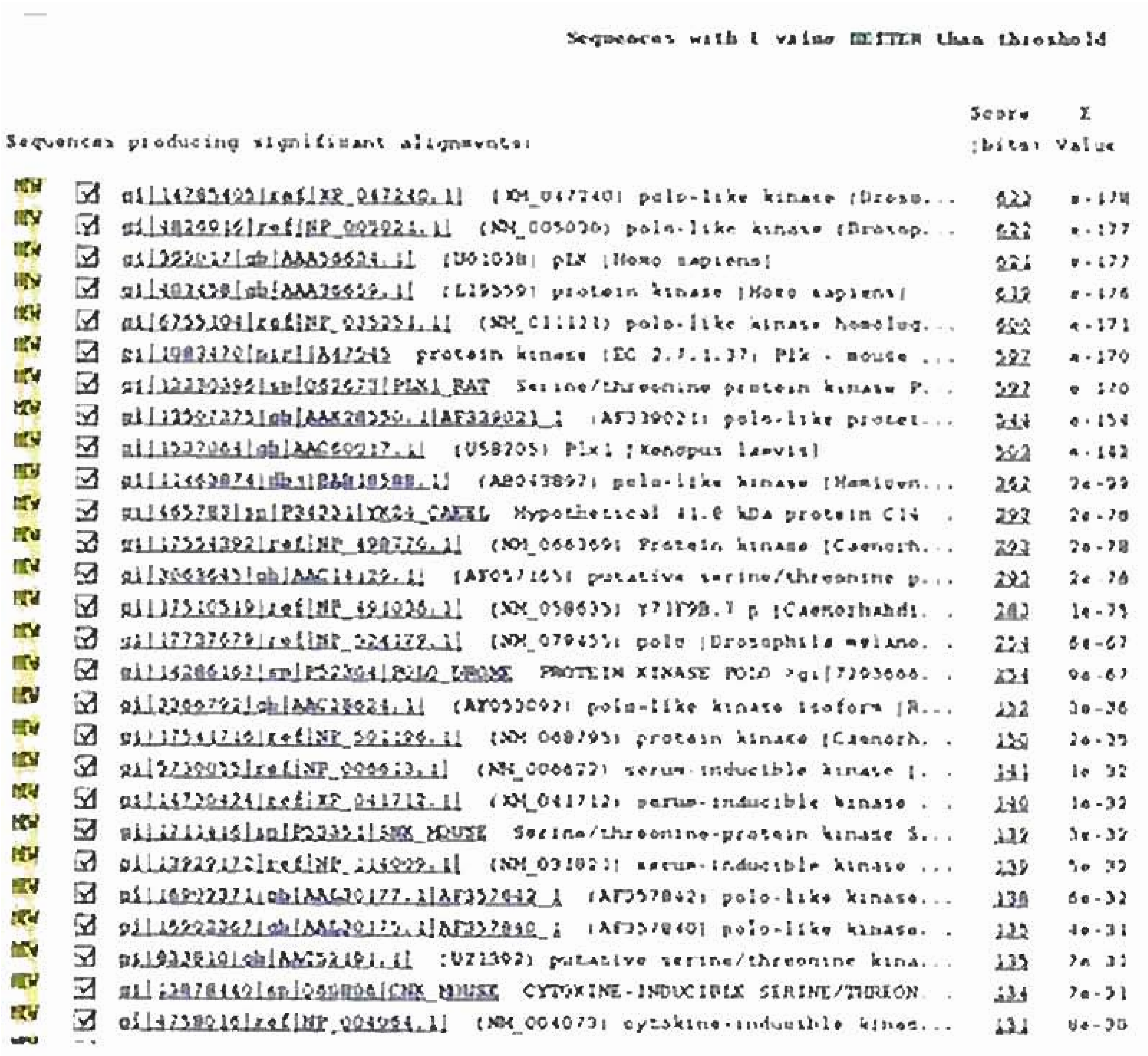
Psi-BLAST

Position-specific iterative BLAST

Combines BLAST searching with PSSMs!

1- Start with regular BLAST search – look at the results

2- Pick the ones you believe are really homologous



```
                                      Sequences with E value BETTER than threshold

                                                                                          Score     E
Sequences producing significant alignments:                                               (bits)  Value

NEW  ☑  gi|14785495|ref|XP_047240.1|  (XM_047240) polo-like kinase (Drosu...              622   e-178
NEW  ☑  gi|4826914|ref|NP_005021.1|  (NM_005030) polo-like kinase (Drosop...              622   e-177
NEW  ☑  gi|322027|gb|AAA36624.1|  (U01038) pLK [Homo sapiens]                             621   e-177
NEW  ☑  gi|402438|gb|AAA36659.1|  (L19559) protein kinase [Homo sapiens]                  619   e-176
NEW  ☑  gi|6755104|ref|NP_035251.1|  (NM_011121) polo-like kinase homolog...              602   e-171
NEW  ☑  gi|1083470|pir||A47245  protein kinase (EC 2.7.1.37) Plk - mouse ...              597   e-170
NEW  ☑  gi|12230392|sp|Q62673|PLK1_RAT  Serine/threonine protein kinase P...              597   e-170
NEW  ☑  gi|13507275|gb|AAK28550.1|AF332021_1  (AF332021) polo-like protei...              544   e-154
NEW  ☑  gi|1532064|gb|AAC60217.1|  (U58205) Plx1 [Xenopus laevis]                         502   e-142
NEW  ☑  gi|11463874|dbj|BAB18588.1|  (AB043897) polo-like kinase [Hemicen...              362   2e-99
NEW  ☑  gi|465783|sp|P34331|YK24_CAEEL  Hypothetical 41.6 kDa protein C14 ...             293   2e-78
NEW  ☑  gi|17554392|ref|NP_498776.1|  (NM_066369) Protein kinase [Caenorh...              293   2e-78
NEW  ☑  gi|3063643|gb|AAC14429.1|  (AF057165) putative serine/threonine p...              292   2e-78
NEW  ☑  gi|17510519|ref|NP_491026.1|  (NM_058633) Y71F9B.7 p [Caenorhabdi...              283   1e-75
NEW  ☑  gi|17737679|ref|NP_524179.1|  (NM_079455) polo [Drosophila melano...              254   6e-67
NEW  ☑  gi|1428616|sp|P52304|POLO_DROME  PROTEIN KINASE POLO >gi|7303666...               254   9e-67
NEW  ☑  gi|2266792|gb|AAC28624.1|  (AF033092) polo-like kinase isoform [R...              152   3e-36
NEW  ☑  gi|17541716|ref|NP_501196.1|  (NM_068795) protein kinase [Caenorh...              150   2e-35
NEW  ☑  gi|5730055|ref|NP_006613.1|  (NM_006622) serum-inducible kinase [...              141   1e-32
NEW  ☑  gi|14720424|ref|XP_041712.1|  (XM_041712) serum-inducible kinase ...              140   1e-32
NEW  ☑  gi|2121416|sp|P53351|SNK_MOUSE  Serine/threonine-protein kinase S...              139   3e-32
NEW  ☑  gi|13929176|ref|NP_113929.1|  (NM_031821) serum-inducible kinase ...              139   5e-32
NEW  ☑  gi|16923371|gb|AAL30177.1|AF357842_1  (AF357842) polo-like kinase...              138   6e-32
NEW  ☑  gi|16923367|gb|AAL30175.1|AF357840_1  (AF357840) polo-like kinase...              135   4e-31
NEW  ☑  gi|1932810|gb|AAC52181.1|  (U71392) putative serine/threonine kina...              135   7e-31
NEW  ☑  gi|12878440|sp|Q9R012|CNK_MOUSE  CYTOKINE-INDUCIBLE SERINE/THREON...              134   7e-31
NEW  ☑  gi|4758016|ref|NP_004064.1|  (NM_004073) cytokine-inducible kinas...              131   8e-30
```

Psi-BLAST

Position-specific iterative BLAST

Combines BLAST searching with PSSMs!

1- Start with regular BLAST search – look at the results

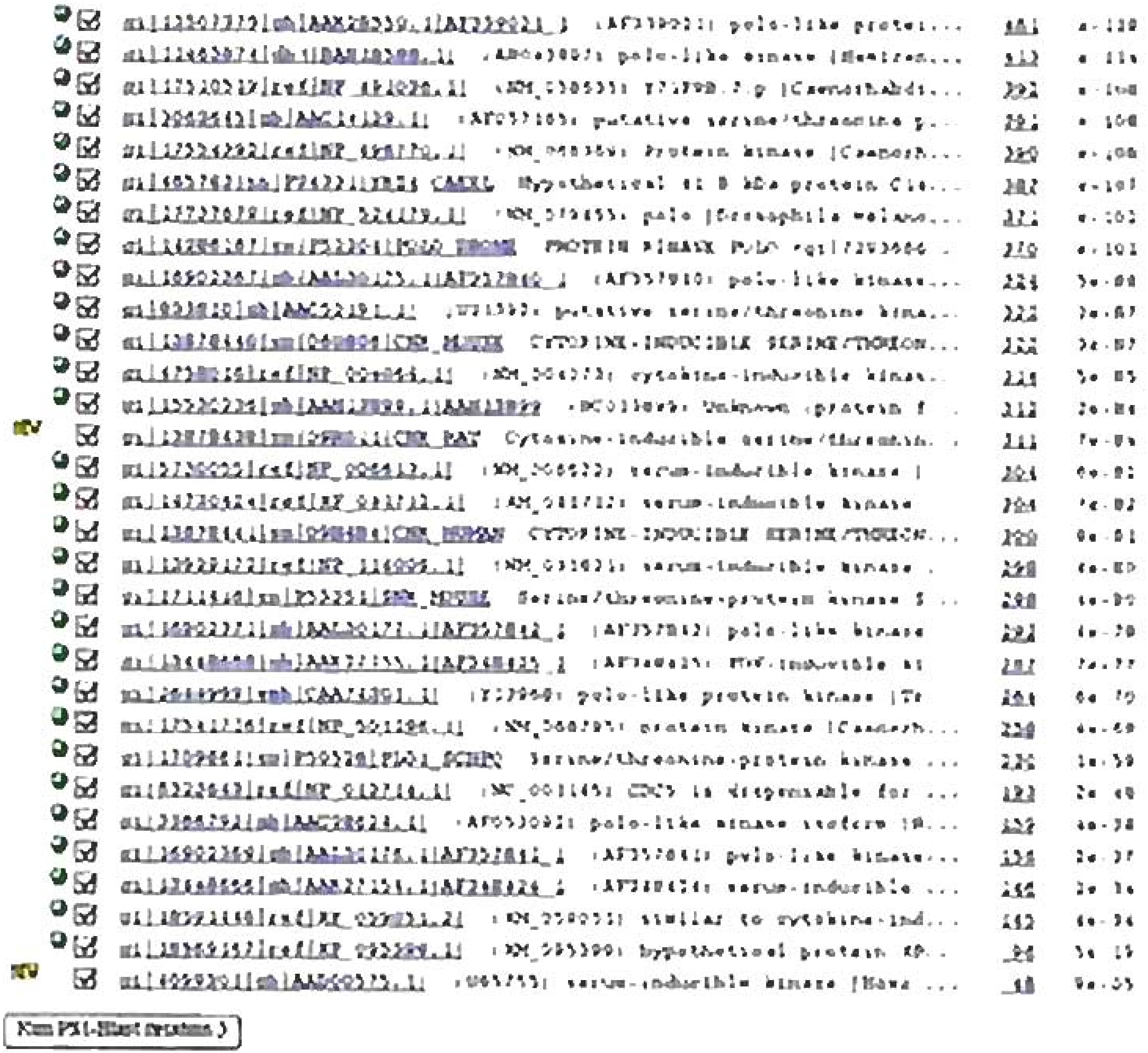
2- Pick the ones you believe are really homologous

3- Now align these sequences to the query sequence and make up a PSSM that tells how much to weigh each amino acid in each position in the alignment

4- Use this PSSM to do another BLAST search

5- Add any new sequences that come up to the old ones if you believe they are really homologous

6- Repeat the alignment to make a new and improved PSSM that tells how much to weigh each amino acid in each position in the alignment



```
● ☑ gi|13507775|gb|AAK28539.1|AF339021_1  (AF339021) polo-like protei...   461   e-128
● ☑ gi|12463074|dbj|BAB18500.1|  (AB046098) polo-like kinase [Hemicen...   413   e-114
● ☑ gi|17510237|ref|NP_491026.1|  (NM_058625) Y71F9B.7.p [Caenorhabdi...   392   e-108
● ☑ gi|3068645|gb|AAC16129.1|  (AF037165) putative serine/threonine p...   391   e-108
● ☑ gi|17554292|ref|NP_498770.1|  (NM_066369) Protein kinase [Caenorh...   388   e-106
● ☑ gi|465762|sp|P34331|YR25_CAEEL  Hypothetical 81.8 kDa protein C14...   382   e-105
● ☑ gi|17737678|ref|NP_524179.1|  (NM_079455) polo [Drosophila melano...   371   e-102
● ☑ gi|1428616|sp|P52304|POLO_DROME  PROTEIN KINASE POLO <gi|7293466 ..   370   e-101
● ☑ gi|16902265|gb|AAL30125.1|AF237910_1  (AF237910) polo-like kinase...   324   3e-88
● ☑ gi|3023819|gb|AAC12219.1|  (U71359) putative serine/threonine kina...   322   3e-87
● ☑ gi|12878446|sp|Q60866|CNK_MOUSE  CYTOKINE-INDUCIBLE SERINE/THREON...   322   3e-87
● ☑ gi|4759116|ref|NP_004064.1|  (NM_004073) cytokine-inducible kinas..   321   5e-85
● ☑ gi|13529234|gb|AAH12909.1|AAH12909  (BC012909) Unknown (protein f ..   312   2e-84
NEW ☑ gi|12878438|sp|Q9R012|CNK_RAT  Cytokine-inducible serine/threonin.. .   311   7e-84
● ☑ gi|5730055|ref|NP_006613.1|  (NM_006622) serum-inducible kinase [   304   6e-82
● ☑ gi|14730424|ref|NP_037172.1|  (NM_013040) serum-inducible kinase   304   7e-82
● ☑ gi|12878441|sp|Q9H4B4|CNK_HUMAN  CYTOKINE-INDUCIBLE SERINE/THREON...   300   9e-81
● ☑ gi|13929122|ref|NP_116099.1|  (NM_032610) serum-inducible kinase .   298   6e-80
● ☑ gi|1711468|sp|P53351|SNK_MOUSE  Serine/threonine-protein kinase S ..   298   6e-80
● ☑ gi|16902221|gb|AAL30171.1|AF237842_1  (AF237842) polo-like kinase   292   4e-78
● ☑ gi|13448668|gb|AAK27155.1|AF348425_1  (AF348425) FNK-inducible ki   287   7e-77
● ☑ gi|2644992|emb|CAA74301.1|  (Y13960) polo-like protein kinase [Tr   264   6e-70
● ☑ gi|17541226|ref|NP_501195.1|  (NM_068795) protein kinase [Caenorh...   226   4e-59
● ☑ gi|1709881|sp|P50528|PLO1_SCHPO  Serine/threonine-protein kinase ...   225   1e-58
● ☑ gi|6322643|ref|NP_012716.1|  (NC_001145) CDC5 is dispensable for ...   183   2e-46
● ☑ gi|3306292|gb|AAC78523.1|  (AF053309) polo-like kinase isoform 1B...   157   4e-38
● ☑ gi|16902269|gb|AAL30176.1|AF237841_1  (AF237841) polo-like kinase...   156   2e-37
● ☑ gi|13448666|gb|AAK27154.1|AF348424_1  (AF348424) serum-inducible ...   152   3e-36
● ☑ gi|18592150|ref|XM_070921.2|  (XM_070921) similar to cytokine-ind...   152   4e-36
● ☑ gi|18569157|ref|XP_095390.1|  (XM_095390) hypothetical protein XP..    86   3e-15
NEW ☑ gi|4092921|gb|AAC60375.1|  (U05373) serum-inducible kinase [Homo ...    48   9e-05
```

Psi-BLAST

7-- Use this PSSM to do another BLAST search

8– Keep iterating until no new sequences are found

Very good for finding weakly related sequences

The End !!!!!