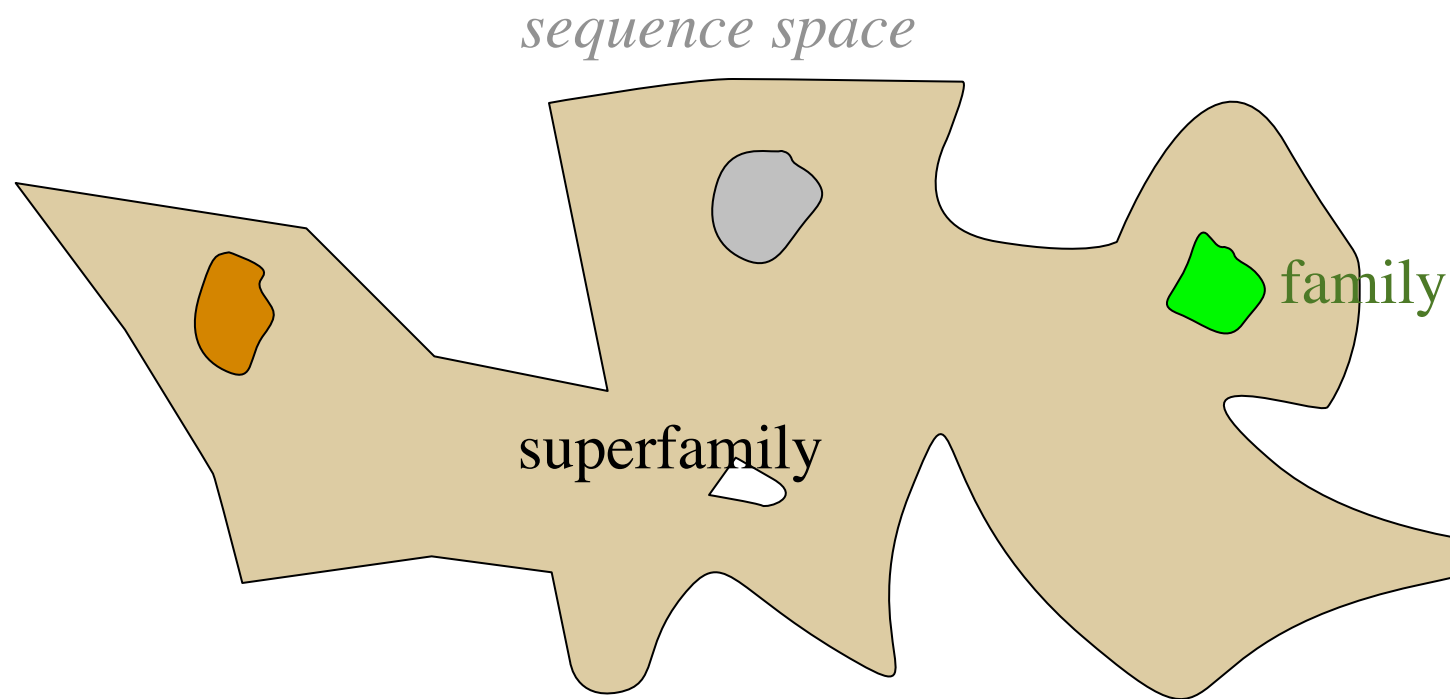


Bioinformatics 1 -- lecture 9

Building a representative model for a gene

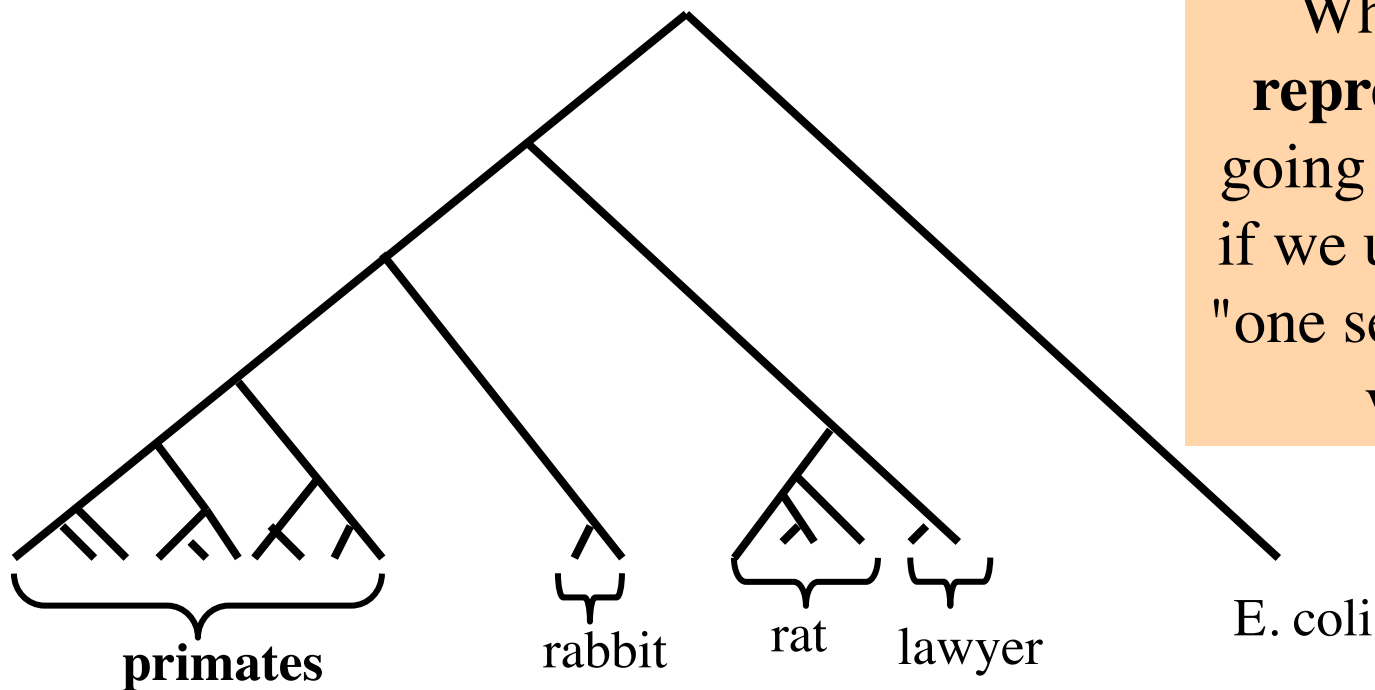
Families and superfamilies



To model a whole family or superfamily, we need to sample sequence space. A family is the set of sequences of true homologs that can “see” each other in a database search. A superfamily is a group of distant homologs that cannot be easily found by sequence searches.

Sampling sequences

If we submit one sequence (for example, citrate synthase from human) to the GenBank database (using BLAST for example), and take 100 results, and we build a cladogram from this, we might get something like this...



What is our **representative** going to look like if we use the rule: "one sequence one vote"?

Sequence weighting corrects for uneven sampling

To build a representative model we can...

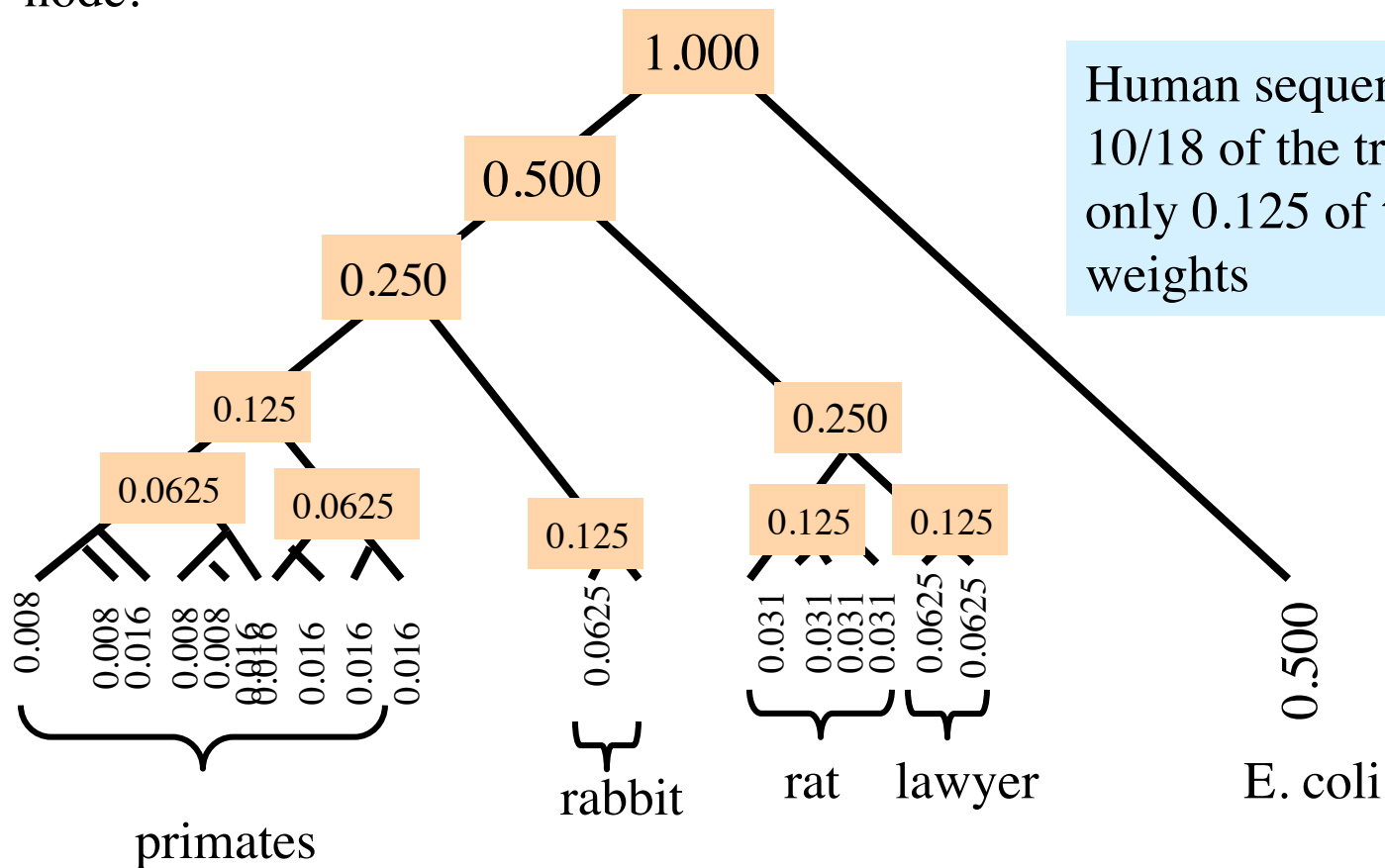
(1) **throw out all redundant sequences** and keep representatives of each clade only, or

(2) **apply a weight** to each sequence reflecting how non-redundant that sequence is.

One measure of **non-redundancy** is sequence-distance, or evolutionary distance.

Crude weights from a cladogram

Simplest weighting scheme: Start with weight = 1.0 at the common ancestor of the tree. Split the weight evenly at each node.



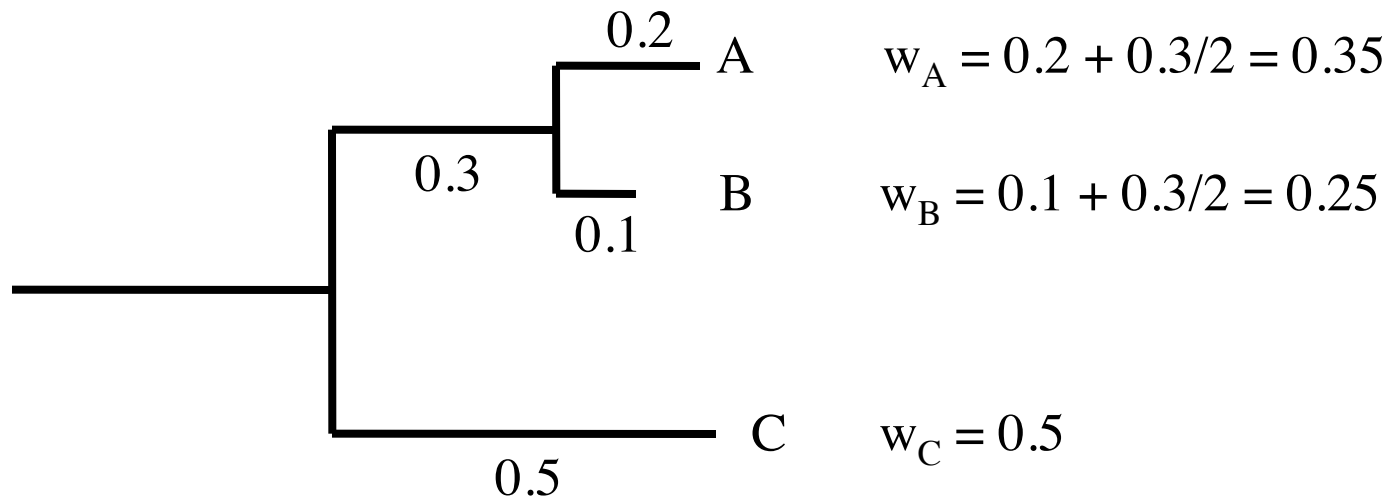
Human sequences are 10/18 of the tree, but only 0.125 of the weights

In class exercise: Edit a MSA using UGENE

- Select a sequence
- Find all near homologs using NCBI BLAST. Keep 200.
- Make a *representative* MSA using MUSCLE as follows (in short-hand):
 - *Align. Trim. Align. Prune. Align. Prune.*
(Calculate distances. *Prune. Align.*)ⁿ
- Calculate a profile (you must trim all gaps first)

This takes ~30 minutes

Better weights from a phylogram



The sequence weight is calculated starting from the distance from the taxon to the first ancestor node, adding half of the distance from the first ancestor to the second ancestor, 1/4th of the distance from the second to third ancestor, and so on.

Finally, the weights are normalized to sum to 1.00

Distance-based weights

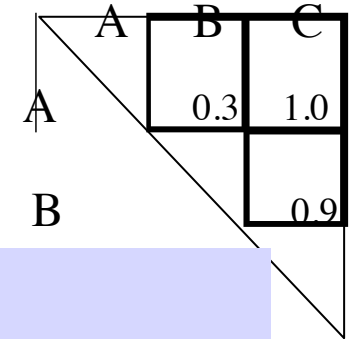
	A	B	C
A		0.3	1.0
B			0.9
C			

- (1) Sum the weighted distances to get new weights.
- (2) Normalize the new weights
- (3) Repeat (1) and (2) until no change.

Pseudocode :

```
all  $w_i$  initialized to 1.  
while ( $w_i \neq w'_i$ ) do  
  for  $i$  from A to C do  
     $w'_i = \sum_j w_j D_{ij}$   
  end do  
  for  $i$  from A to C do  
     $w_i = w'_i / \sum_j w'_j$   
  end do  
end do
```

Distance-based weights



Running the pseudocode :

(1) Sum the weighted distances to get new weights.

$$w'_A = 0.3 + 1.0 = 1.3$$

$$w'_B = 0.3 + 0.9 = 1.2$$

$$w'_C = 1.0 + 0.9 = 1.9$$

(2) Normalize the new weights

$$w_A = 1.3 / (1.3 + 1.2 + 1.9) = 0.30$$

$$w_B = 1.2 / 4.4 = 0.27$$

$$w_C = 1.9 / 4.4 = 0.43$$

(1) Sum the weighted distances to get new weights.

$$w'_A = 0.3 * 0.27 + 1.0 * 0.43 = 0.51$$

$$w'_B = 0.3 * 0.3 + 0.9 * 0.43 = 0.48$$

$$w'_C = 1.0 * 0.3 + 0.9 * 0.27 = 0.54$$

...

$$w_{ABC} = 0.33 \quad 0.31 \quad 0.35$$

$$w_{ABC} = 0.30 \quad 0.28 \quad 0.42$$

$$w_{ABC} = 0.31 \quad 0.29 \quad 0.40$$

$$w_{ABC} = 0.30 \quad 0.28 \quad 0.41$$

$$w_{ABC} = 0.30 \quad 0.28 \quad 0.41 \text{ converged.}$$

(3) Repeat (1) and (2) until no change.

Amino acid probability profiles

An **amino acid profile** is defined as a set of probability distributions over the 20 amino acids, a probability density function (PDF) for each position in the alignment. Gap probabilities are usually not included.

$$P(a|i) = \frac{\sum_{S \forall S_i=a} w_S}{\sum w_S}$$

[The probability of amino acid a at position i is the sum of the sequence weights w_S over all sequences S such that the amino acid at position i of that sequence S_i is a , divided by the sum over the sequence weights w_S for all sequences S .]

PSSMs, Log-likelihood ratios

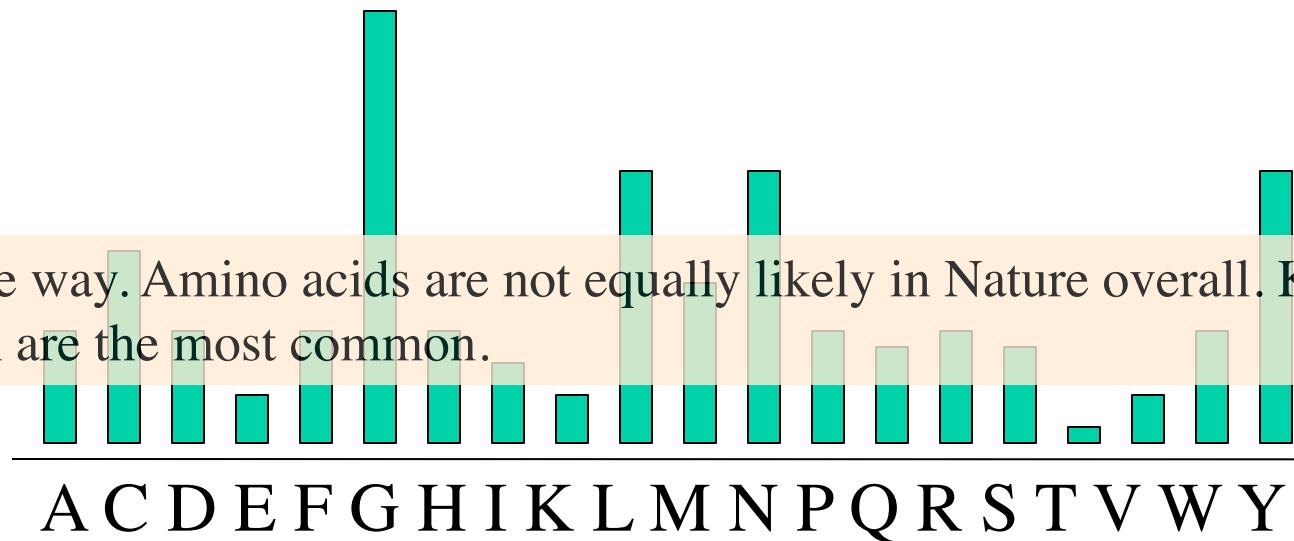
position specific scoring matrix

$$\text{LLR}(a) = \log(P(a|i) / P(a))$$

probability of a in one column

likelihood of a overall
(the whole database)

By the way. Amino acids are not equally likely in Nature overall. K, L and R are the most common.



Pseudocounts, to account for unsampled data...

$$\text{LLR}(a) = \log(P(a|i) / P(a))$$

If $P(a|i)=0.$, you can't take the log

The probability of seeing a in column i of a sequence alignment is never really zero. So we add a small number of 'pseudocounts' ϵ .

$$\text{LLR}(a) = \log((P(a|i)+\epsilon) / P(a))$$

This LLR does not go to negative infinity as $P(a) \rightarrow 0.000$.
Instead it goes to $\log(\epsilon/P(a))$.

Smart Pseudocounts, extrapolating using BLOSUM..

$$\text{LLR}(a) = \log((P(a|i) + \epsilon\omega) / P(a))$$

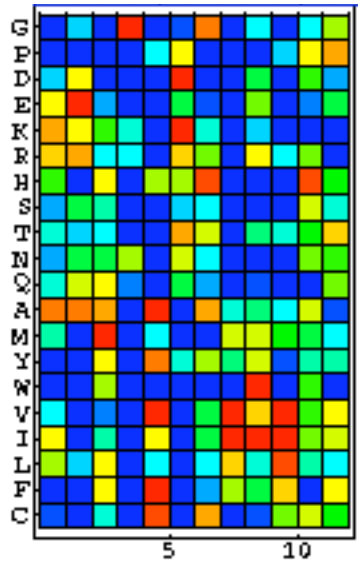
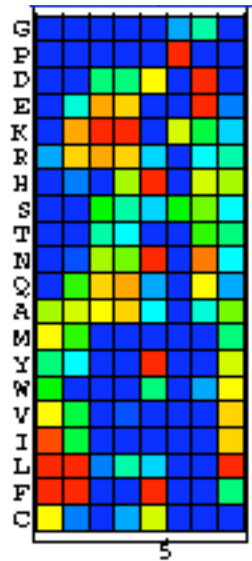
where

$$\omega = \sum_b P(b|i) S(ab)$$

where

$S(ab)$ = probability of substitution $b \rightarrow a$

One way to visualize profiles

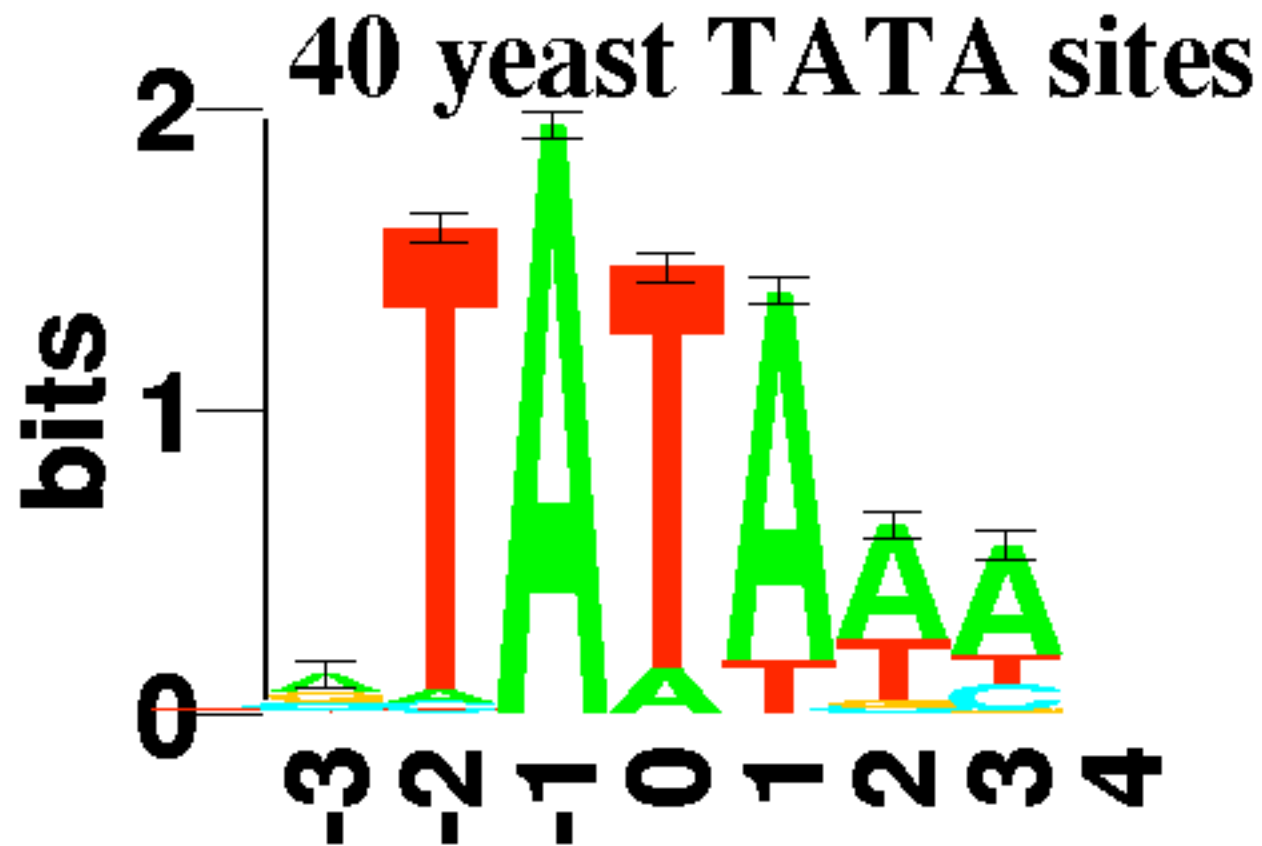


Color matrix

Color = LLR.

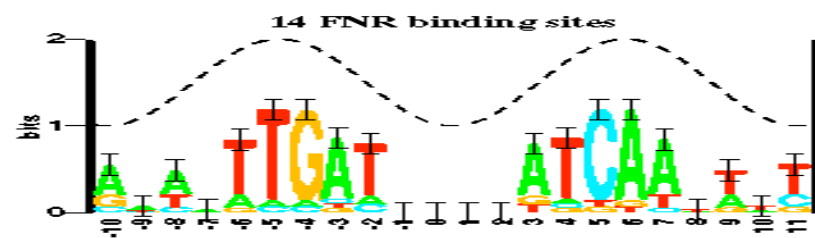
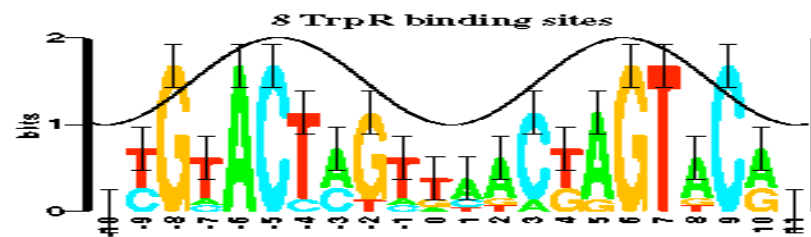
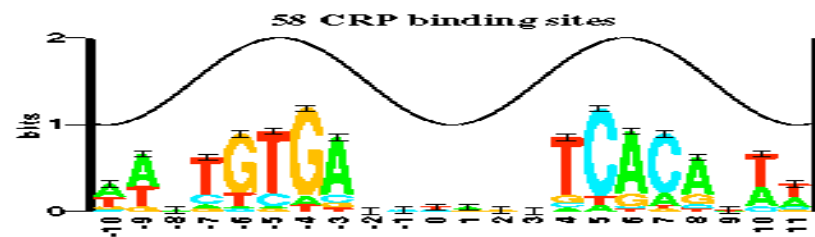
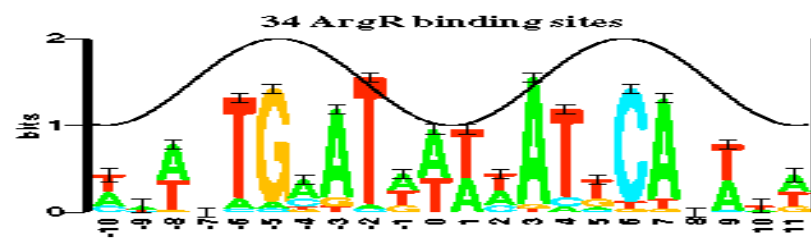
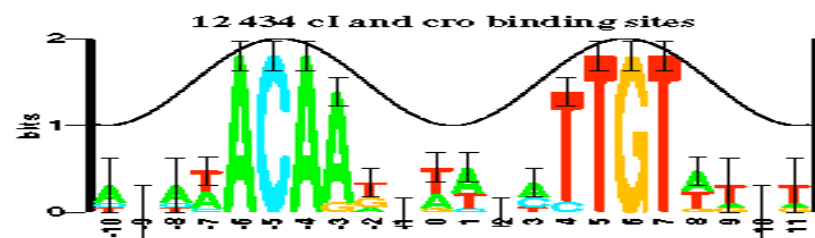
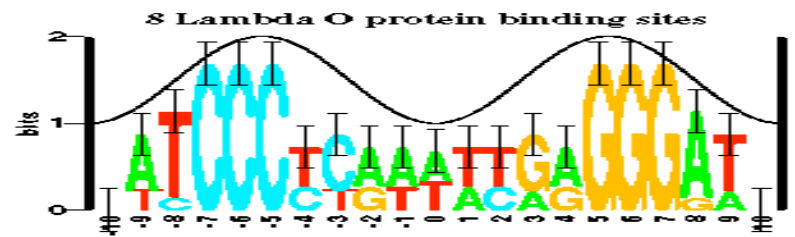
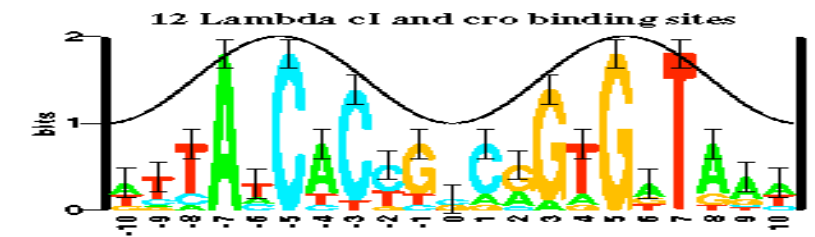
Blue = high negative values. Green = zero. Red = high positive values.

Another way: Logos



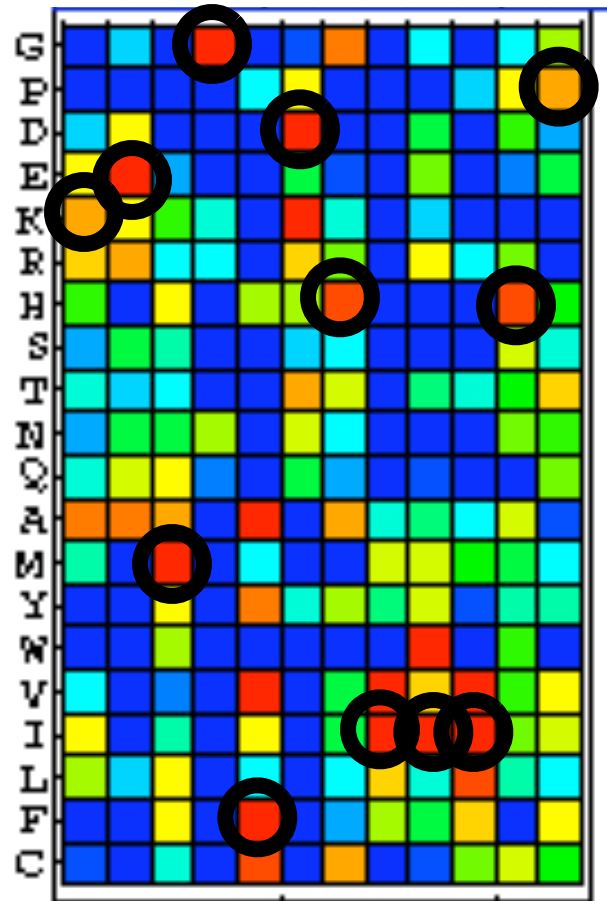
Height of letter is the LLR in bits, half-bits or nats.

Alignments of transcription factor footprint sites



Scoring a sequence versus a profile

The score is the **sum of the log-likelihood ratios** of the amino acid in the sequence.



Sequence=

KEMGF'DHIIHP

$$\text{score} = \sum_i \text{LLR}(a_i)$$

Aligning sequence to profile

```
S(i,j) = 0
```

```
do aa=1,20
```

```
    S(i,j) = S(i,j) + P(aa,i)*B(aa,s(j))
```

```
enddo
```

profile1@i P(aali)

sequence2@j

BLOSUM
score

Aligning profile to profile

```
S(i,j) = 0
```

```
do aai=1,20
```

```
    do aaj=1,20
```

```
        S(i,j) = S(i,j) + P(aai,i)*P(aaj,j)*B(aai,aaj)
```

```
    enddo
```

```
enddo
```

No need to normalize, since $\sum_{aa_i} \sum_{aa_j} P(aa_i|i) * P(aa_j|j) = 1$ 19

Psi-BLAST: Blast with profiles

Psi-BLAST searches the database *iteratively*.

(Cycle 1) Normal BLAST (with gaps)

(Cycle 2) (a) Construct a **profile** from the results of **Cycle 1**.

(b) Search the database using the profile.

(Cycle 3) (a) Construct a **profile** from the results of **Cycle 2**.

(b) Search the database using the profile.

And So On... (user sets the number of cycles)

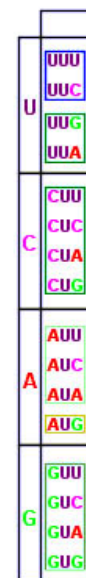
Psi-BLAST is much more *sensitive* than BLAST.

Also more vulnerable to *low-complexity*.

Other forms of BLAST

BLAST	query	database
blastn	nucleotide	nucleotide
blastp	protein	protein
tblastn	protein	translated DNA
blastx	translated DNA	protein
tblastx	translated DNA	translated DNA
psi-blast	protein, profile	protein
phi-blast	pattern	protein
transitive blast*	any	any

*not really a blast. Just a way of using blast.



IUPAC nuc
A
C
G
T (or U)
R
Y
S
W
K
M
B
D
H
V
N
. or -

PHI-BLAST -- Patterned Hit Initiated BLAST

Table I. Detection of subtle protein sequence relationships using PHI-BLAST

Conserved domain or motif under investigation	Pattern ^a	GenBank (30) accession no. of query	Top non-trivial relevant hit found by PHI-BLAST		Top non-trivial relevant hit found by BLAST	
			Accession no.	<i>E</i> -value	Accession no.	<i>E</i> -value
A. P-loop ATPase domain in apoptosis regulators and plant stress response proteins	[GA]xxxxGK[ST]	231729	2213598	0.038	2961373	4.7
B. ATPase domain in mismatch repair protein MutL, type II topoisomerases, histidine kinases, and HS90 molecular chaperones	hxhxDxGxG	127552	488200	0.017	2495364	1.8
C. Nucleotidyltransferase domain in archaeal tRNA nucleotidyltransferases	DhDhhh	2826366	2650333	0.061	2650333	8.6
D. Motif VI of superfamily II helicases in archaeal homologs of bacterial DNA primases	QxxGRx[GA]R	2128723	2499099	0.54		

Transitive Blast searches

- If A detects B, but B detects C (but A does not detect C). A is homologous to C *.
- Transitive-BLAST = using the hits of a BLAST search to do additional BLAST searches.

*exceptions exist!