

## Motifs, profiles and hidden Markov models

CRC-CGF Bioinformatics Course

September 9-13, 2002

Terry Speed, WEHI

1

## The objects of our study

DNA, RNA and proteins: macromolecules which are unbranched polymers built up from smaller units.

**DNA:** units are the nucleotide residues A, C, G and T

**RNA:** units are the nucleotide residues A, C, G and U

**Proteins:** units are the amino acid residues A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y.

To a considerable extent, the chemical properties of DNA, RNA and protein molecules are encoded in the **linear sequence** of these basic units: their primary structure.

2

## The statistics of biological sequences can be global or local

### Base composition of genomes:

*E. coli*: 25% A, 25% C, 25% G, 25% T

*P. falciparum*: 82%A+T

### Translation initiation:

ATG is the near universal **motif** indicating the start of translation in DNA coding sequence.

3

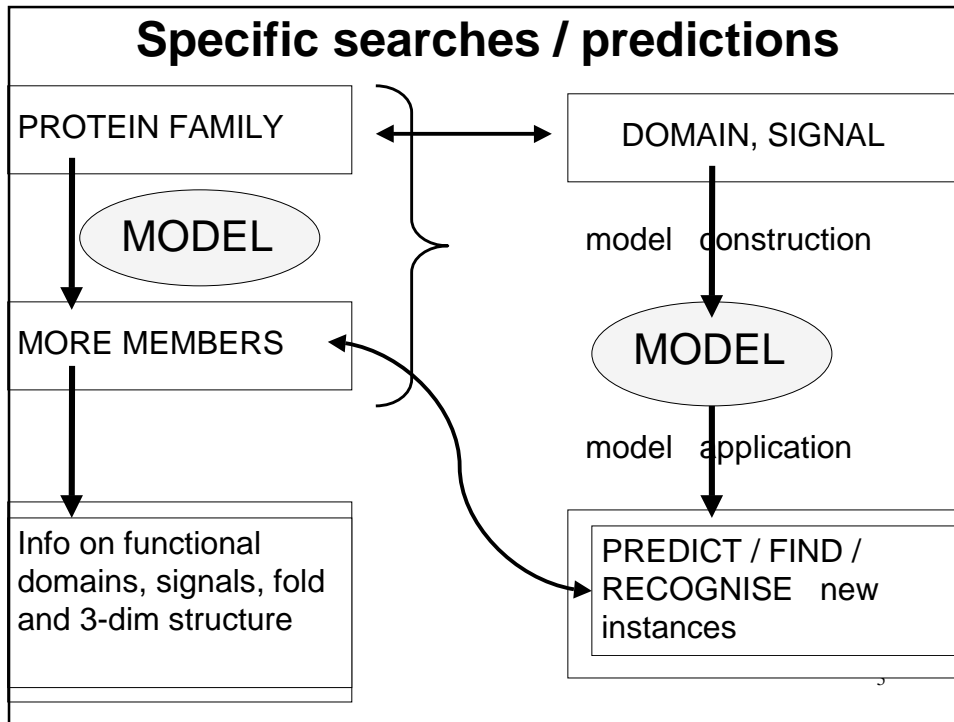
## Motifs - Sites - Signals - Domains

For this talk, I'll use these terms interchangeably to describe **recurring elements** of interest to us.

In **PROTEINS** we have: transmembrane domains, coiled-coil domains, EGF-like domains, signal peptides, phosphorylation sites, antigenic determinants, ..., and **protein families**.

In **DNA / RNA** we have: enhancers, promoters, terminators, splicing signals, translation initiation sites, centromeres, ...

4



## Motifs and models

Motifs typically represent regions of structural significance with specific biological **function**.

Are generalisations from known **examples**.

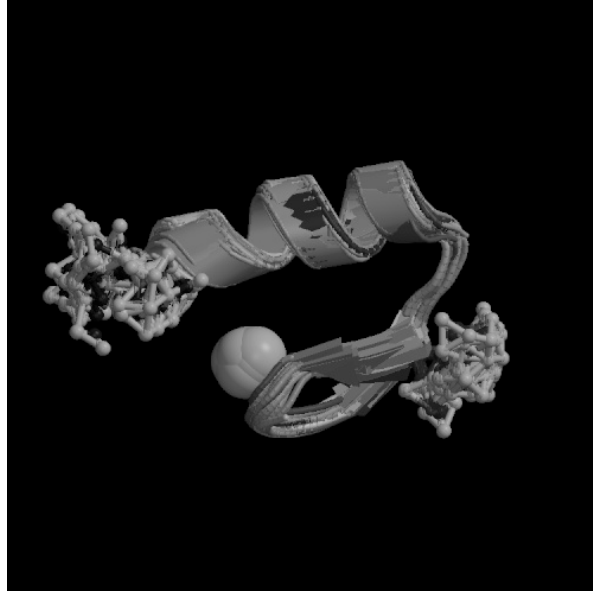
The models can be **highly specific**.

**Multiple models** can be used to give higher sensitivity & specificity in their detection.

Can sometimes be generated **automatically** from examples or multiple alignments.

6

**From certainty to statistical models: a brief case study**



**1 ZNF: Cys-Cys-His-His zinc finger DNA binding domain**

7

**Prosite patterns**

An early effort at collecting descriptors for functionally important protein motifs. They do not attempt to describe a complete domain or protein, but simply try to identify the most important residue combinations, such as the catalytic site of an enzyme. They use regular expression syntax, and focus on the most highly conserved residues in a protein family.

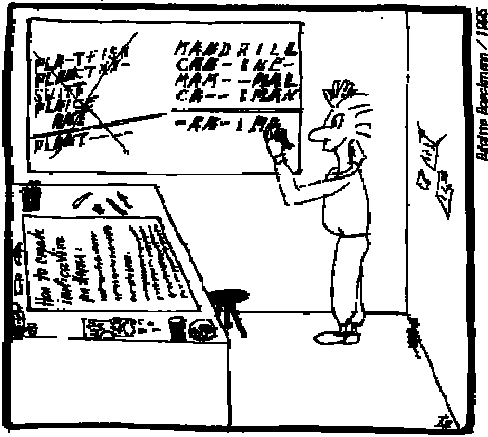
<http://au.expasy.org>

8

### Consensus in sequences

This pattern, which must be in the N-terminal of the sequence ('<'), means:  
 $\langle A - x - [ST] (2) - x(0,1) - V - \{L\}$   
 Ala-any- [Ser or Thr]-[Ser or Thr] - (any or none)-Val-(any but Leu, Ile)

**How we develop Prosite patterns!**



9

### Example: C<sub>2</sub>H<sub>2</sub> zinc finger DNA binding domain

The characteristic motif of a Cys-Cys-His-His zinc finger DNA binding domain has **regular expression**

**C-X(2,4)-C-X(3)-[LIVMFYWC]-X(8)-H-X(3,5)-H**

Here, as in algebra, **X** is unknown. The sequence of our example domain 1ZNF is as follows, clearly fitting the model.

**XYKCGLCERSFVEKSALSRHQRVHKNX**

10

## Searching with regular expressions

[http://www.isrec.isb-sib.ch/software/PATFND\\_form.html](http://www.isrec.isb-sib.ch/software/PATFND_form.html)  
c.{2,4}c...[livmfywc].....h.{3,5}h

### PatternFind output

[ISREC-Server] Date: Wed Aug 22 13:00:41 MET 2001

...

gp|AF234161|7188808|01AEB01ABAC4F945 nuclear  
protein NP94b [Homo sapiens] Occurences: 2

Position : 514 CYICKASCSSQQEFQDHMSEPQH

Position : 606 CTVCNRYFKTPRKFVEHVKSQGH

.....

11

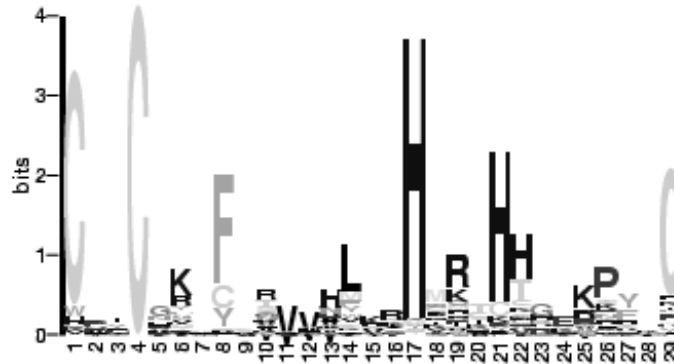
## Regular expressions can be limiting

The regular expression syntax is too rigid to represent many highly divergent protein motifs.

Also, short patterns are sometimes insufficient with today's large databases. Even requiring perfect matches you might find many false positives. On the other site some real sites might not be perfect matches.

We need to go beyond apparently equally likely alternatives, and ranges for gaps. We deal with the former first, having a distribution at each position.

### Cys-Cys-His-His profile: sequence logo form



A sequence logo is a scaled position-specific a.a.distribution.  
 Scaling is by a measure of a position's information content.

### Calculation of a position-specific scoring matrix (PSSM) from counts

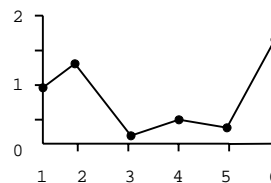
|   |     |     |     |     |     |     |
|---|-----|-----|-----|-----|-----|-----|
| A | 9   | 214 | 63  | 142 | 118 | 8   |
| C | 22  | 7   | 26  | 31  | 52  | 13  |
| G | 18  | 2   | 29  | 38  | 29  | 5   |
| T | 193 | 19  | 124 | 31  | 43  | 216 |

|   |      |      |      |      |      |      |
|---|------|------|------|------|------|------|
| A | 0.04 | 0.88 | 0.26 | 0.59 | 0.49 | 0.03 |
| C | 0.09 | 0.03 | 0.11 | 0.13 | 0.22 | 0.05 |
| G | 0.07 | 0.01 | 0.12 | 0.16 | 0.12 | 0.02 |
| T | 0.80 | 0.08 | 0.51 | 0.13 | 0.18 | 0.89 |

Counts from 242 known sites

Relative frequencies:  $f_{b1}$

|   |     |     |    |     |     |     |
|---|-----|-----|----|-----|-----|-----|
| A | -38 | 19  | 1  | 12  | 10  | -48 |
| C | -15 | -38 | -8 | -10 | -3  | -32 |
| G | -13 | -48 | -6 | -7  | -10 | -40 |
| T | 17  | -32 | 8  | -9  | -6  | 19  |



$$\text{PSSM} : 10 \times \log \frac{f_{b1}}{p_b}$$

$$\text{Information} : 2 + \sum_b p_{b1} \log_2 p_{b1}$$

## Derivation of PSSM entries

Suppose that we have aligned sequence data on a number of instances of a given type of site.

candidate sequence           CTATAATC....

aligned position             123456

S=site (and independence)

**Hypotheses:**

R=random (equiprobable, independence)

$$\begin{aligned} \log_2 \left| \frac{\text{pr}(\text{CTATAA}|S)}{\text{pr}(\text{CTATAA}|R)} \right| &= \log_2 \frac{.09 \times .03 \times .26 \times .13 \times .51 \times .01}{.25 \times .25 \times .25 \times .25 \times .25 \times .25} \\ &= (2 + \log_2 .09) + \dots + (2 + \log_2 .01) \\ &= \frac{1}{10} \{-15 - 32 + 1 - 9 + 10 - 48\} \end{aligned}$$

Generally, PSSM score  $s_{bl} = \log f_{bl}/p_b$            l=position, b=base  
 $p_b$ =background frequency

## Use of a PSSM to find sites

Move the matrix along the sequence and score each window.

|   | C   | T   | A  | T   | A   | A   | T | C |     |
|---|-----|-----|----|-----|-----|-----|---|---|-----|
| A | -38 | 19  | 1  | 12  | 10  | -48 |   |   | sum |
| C | -15 | -38 | -8 | -10 | -3  | -32 |   |   |     |
| G | -13 | -48 | -6 | -7  | -10 | -48 |   |   | -93 |
| T | 17  | -32 | 8  | -9  | -6  | 19  |   |   |     |

Peaks should occur at the true sites.

|   |     |     |    |     |     |     |  |  |     |
|---|-----|-----|----|-----|-----|-----|--|--|-----|
| A | -38 | 19  | 1  | 12  | 10  | -48 |  |  |     |
| C | -15 | -38 | -8 | -10 | -3  | -32 |  |  |     |
| G | -13 | -48 | -6 | -7  | -10 | -48 |  |  | +85 |
| T | 17  | -32 | 8  | -9  | -6  | 19  |  |  |     |

Of course in general any threshold will have some **false positive** and **false negative** rate.

|   |     |     |    |     |     |     |  |  |     |
|---|-----|-----|----|-----|-----|-----|--|--|-----|
| A | -38 | 19  | 1  | 12  | 10  | -48 |  |  |     |
| C | -15 | -38 | -8 | -10 | -3  | -32 |  |  |     |
| G | -13 | -48 | -6 | -7  | -10 | -48 |  |  | -95 |
| T | 17  | -32 | 8  | -9  | -6  | 19  |  |  | 16  |

## Representation of motifs: the next steps

Missing from the position-specific distribution representation of motifs are good ways of dealing with:

- **Length distributions for insertions/deletions**
- **Non-local association of amino acids**

**Profiles**, and then **Hidden Markov models** help with the first. The second remains a hard unsolved problem.

17

## Profiles

Are a variation of the position specific scoring matrix approach just described. Profiles are calculated slightly differently to reflect amino acid substitutions, and the possibility of gaps, but are used in the same way.

In general a profile entry  $M_{la}$  for location  $l$  and amino acid  $a$  is calculated by

$$M_{la} = \sum_b S_b w_{lb} S_{ab}$$

where  $b$  ranges over amino acids,  $w_{lb}$  is a **weight** (e.g. the observed frequency of a.a.  $b$  in position  $l$ ) and  $S_{ab}$  is the  $(a,b)$ -entry of a **substitution matrix** (e.g. PAM or BLOSUM).

**Position specific gap penalties** can also be included.

18

# Derivation of a profile for Ig domains

FileUp of: @/home/ucsb00/George/.WAG/pileup-8391.8424

Symbol comparison table: GenRunData:pileuppep.cmp CompCheck: 1254

GapWeight: 3.000  
GapLengthWeight: 0.100

pileup.msf MSF: 49 Type: P May 27, 1999 13:18 Check: 9167 ..

|          |         |             |              |
|----------|---------|-------------|--------------|
| Name: g1 | Len: 49 | Check: 2179 | Weight: 1.00 |
| Name: m3 | Len: 49 | Check: 0    | Weight: 1.00 |
| Name: k2 | Len: 49 | Check: 6739 | Weight: 1.00 |
| Name: l3 | Len: 49 | Check: 249  | Weight: 1.00 |

```
1 49
g1 ELVKAGSSVK MSCKATGYTF SSYE...LY WVRQAPGQGL EDLGYISS
m3 GLVEPGGSLR LSCSASGFTF SAND...MN WVRQAPGKGL EWLSFIGGS
k2 LPVTPGEPAS ISCRSSQSL L DSGDGNTYLN WYLQKAGQSP QLLIYTLSY
l3 VSVALGQTVR ITCQ.GDSLR GYDAA..... WYQQKPGQAP LLVIYGRNN
```

19

## (Peptide) PROFILEMAKE v4.40 of:

/home/ucsb00/George/biochem/1999/immunoglob.msf[\*] Length: 49  
Sequences: 4 MaxScore: 27.72 May 27, 1999 13:24

Gap: 1.00 Len: 1.00  
GapRatio: 0.33 LenRatio: 0.10

|                    |         |        |              |
|--------------------|---------|--------|--------------|
| immunoglob.msf[g1] | From: 1 | To: 49 | Weight: 1.00 |
| immunoglob.msf[m3] | From: 1 | To: 49 | Weight: 1.00 |
| immunoglob.msf[k2] | From: 1 | To: 49 | Weight: 1.00 |
| immunoglob.msf[l3] | From: 1 | To: 49 | Weight: 1.00 |

Symbol comparison table: profilepep.cmp FileCheck: 1254

Stringent treatment of non-observed characters

Exponential weighting of characters

| Cons | A   | B   | C   | D    | E    | F    | G   | H    | I   | K   | L   | M   | N   | P   | Q   | R   | S   | T   | V   | W    | Y   | Z   | Gap | Len |     |
|------|-----|-----|-----|------|------|------|-----|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|-----|-----|-----|-----|-----|
| V    | 15  | 8   | -14 | 14   | 21   | 1    | 24  | -4   | 19  | -4  | 21  | 19  | 3   | 3   | 8   | -14 | 4   | 10  | 32  | -33  | -14 | 14  | 100 | 100 |     |
| L    | 9   | -10 | -14 | -12  | -5   | 24   | -3  | -6   | 21  | -5  | 38  | 34  | -9  | 17  | 0   | -7  | 14  | 5   | 25  | 9    | -7  | -4  | 100 | 100 |     |
| V    | 20  | -20 | 20  | -20  | 20   | 20   | 20  | 110  | 20  | 80  | 60  | -30 | 10  | -20 | -30 | -10 | 20  | 150 | -80 | -10  | -20 | 100 | 100 |     |     |
| T    | 31  | 21  | -10 | 25   | 32   | -31  | 21  | 4    | -3  | 28  | -11 | 0   | 18  | 14  | 17  | 6   | 15  | 32  | 0   | -33  | -24 | 25  | 100 | 100 |     |
| P    | 35  | -1  | -4  | 0    | 3    | -13  | 12  | 2    | 5   | -1  | 10  | 12  | -3  | 50  | 11  | 0   | 13  | 14  | 17  | -30  | -25 | 6   | 100 | 100 |     |
| G    | 70  | 60  | 20  | 70   | 50   | -60  | 150 | -20  | -30 | 110 | -30 | 40  | 30  | 20  | -30 | 60  | 40  | 20  | 100 | -70  | 30  | 100 | 100 |     |     |
| E    | 22  | 29  | -4  | 36   | 40   | -33  | 39  | 10   | -12 | 11  | -18 | -11 | 22  | 15  | 32  | 3   | 31  | 11  | -4  | -32  | -31 | 35  | 100 | 100 |     |
| S    | 25  | 14  | 26  | 11   | 11   | -23  | 29  | -5   | -3  | 11  | -18 | -12 | 12  | 38  | 0   | 6   | 57  | 24  | 1   | -10  | -28 | 4   | 100 | 100 |     |
| V    | 26  | -11 | -1  | -9   | -6   | 16   | 9   | -14  | 46  | -11 | 45  | 37  | -12 | 6   | -5  | -19 | -3  | 11  | 61  | -30  | -3  | -6  | 100 | 100 |     |
| R    | -4  | 13  | -8  | 7    | 7    | -30  | -3  | 14   | -14 | 49  | -22 | 5   | 13  | 16  | 17  | 60  | 27  | 4   | -14 | 50   | -33 | 12  | 100 | 100 |     |
| I    | -1  | -17 | -13 | -19  | -13  | 66   | -21 | -16  | 67  | -8  | 64  | 58  | -19 | -13 | -11 | -12 | -13 | 5   | 54  | -13  | 6   | -11 | 100 | 100 |     |
| S    | 28  | 20  | 43  | 14   | 14   | -21  | 40  | -13  | -3  | 14  | -24 | -17 | 20  | 27  | -7  | 4   | 50  | 38  | -3  | 9    | -27 | 1   | 100 | 100 |     |
| C    | 30  | -40 | 150 | -50  | -60  | -10  | 20  | -10  | 20  | -60 | -80 | -60 | -30 | 10  | -60 | -30 | 70  | 20  | 20  | -120 | 100 | -60 | 100 | 100 |     |
| K    | 4   | 18  | -11 | 17   | 17   | -32  | 6   | 15   | -12 | 40  | -17 | 1   | 17  | 15  | 31  | 39  | 24  | 4   | -11 | 18   | -31 | 24  | 100 | 100 |     |
| A    | 53  | 11  | 19  | 12   | 12   | -200 | 31  | -6   | -1  | 3   | -9  | -4  | 11  | 21  | 5   | -8  | 33  | 17  | 5   | -21  | -15 | 6   | 30  | 30  |     |
| S    | 28  | 21  | 28  | 19   | 16   | -22  | 45  | -11  | -5  | 8   | -21 | -14 | 18  | 21  | -2  | -2  | 60  | 36  | 2   | -13  | -27 | 6   | 100 | 100 |     |
| G    | 29  | 41  | -9  | 53   | 39   | -44  | 60  | 9    | -16 | 7   | -24 | -15 | 28  | 15  | 37  | -4  | 20  | 14  | 1   | -54  | -37 | 37  | 100 | 100 |     |
| M    | 2   | -4  | 35  | -14  | -10  | 31   | 1   | -4   | 8   | -12 | 8   | -4  | 1   | -8  | -23 | -12 | 38  | 1   | -2  | 43   | -28 | -18 | 100 | 100 |     |
| L    | 10  | -10 | -19 | -10  | -3   | 29   | -3  | -10  | 32  | -3  | 44  | 41  | -6  | 0   | -6  | -16 | -3  | 44  | 32  | -3   | 0   | -3  | 100 | 100 |     |
| W    | -1  | -28 | -18 | -19  | -26  | 57   | -30 | 1    | 29  | -15 | 53  | 37  | -20 | -22 | -21 | -1  | -14 | -12 | 13  | 68   | 40  | -22 | 100 | 100 |     |
| I    | 21  | 27  | 33  | 18   | 37   | 27   | -32 | 50   | -4  | -10 | 9   | -27 | -19 | 25  | 18  | 9   | -1  | 59  | 18  | -3   | -20 | -29 | 17  | 100 | 100 |
| S    | 29  | 8   | 40  | 4    | 4    | 3    | 19  | -4   | -2  | -2  | -10 | -11 | 11  | 9   | -9  | -9  | 48  | 11  | -2  | 14   | 4   | -6  | 100 | 100 |     |
| B    | 12  | 35  | 6   | 33   | 21   | -10  | 26  | 14   | -10 | 0   | -15 | -15 | 35  | -6  | 10  | -11 | 10  | 7   | -6  | -18  | 3   | 14  | 100 | 100 |     |
| D    | 34  | 47  | -20 | 66   | 57   | -48  | 39  | 37   | -9  | 14  | -21 | -15 | 32  | 11  | 35  | -4  | 15  | 15  | -6  | -61  | -27 | 47  | 100 | 100 |     |
| A    | 31  | 11  | 7   | 14   | 11   | -15  | 31  | -4   | -4  | -1  | -8  | -4  | 8   | 11  | 6   | -8  | 14  | 11  | 6   | -25  | -14 | 7   | 21  | 21  |     |
| N    | 3   | 15  | -4  | 10   | 7    | -7   | 6   | 7    | -4  | 6   | -6  | -4  | 21  | 0   | 6   | 1   | 4   | 3   | -4  | -4   | -1  | 6   | 21  | 21  |     |
| T    | 6   | 3   | 3   | 3    | 3    | -4   | 6   | -1   | 3   | 3   | -1  | 0   | 3   | 4   | -1  | -1  | 4   | 21  | 3   | -8   | -4  | 6   | 21  | 21  |     |
| Y    | -4  | -4  | 14  | -7   | -7   | 19   | -10 | 4    | 1   | -8  | 4   | -1  | -11 | -8  | -8  | -6  | -4  | -1  | 15  | 21   | -8  | 21  | 21  |     |     |
| L    | -3  | -20 | -34 | -21  | -12  | 65   | -20 | -11  | 34  | -7  | 66  | 62  | -17 | -12 | -3  | -10 | -17 | -3  | 14  | 12   | 8   | -8  | 21  | 21  |     |
| N    | 2   | 31  | 4   | 15   | 9    | 4    | 3   | 20   | -8  | 4   | -9  | -11 | 46  | -11 | 4   | -5  | 4   | 2   | -11 | 6    | 18  | 4   | 21  | 21  |     |
| I    | 31  | -80 | -70 | -120 | -110 | -110 | 130 | -100 | -10 | -50 | 10  | 50  | -30 | -30 | -80 | -50 | 140 | 30  | -60 | -80  | 150 | 110 | -80 | 100 | 100 |
| F    | -3  | -16 | 38  | -22  | -22  | 51   | -16 | 0    | 38  | -25 | 35  | 16  | -13 | -22 | -25 | -29 | -16 | -3  | 44  | 10   | 44  | -25 | 100 | 100 |     |
| R    | -5  | 3   | -25 | 3    | 5    | -10  | -14 | 23   | -3  | 27  | 7   | 24  | 3   | 10  | 32  | 48  | -4  | -6  | -1  | 44   | -33 | 15  | 100 | 100 |     |
| Q    | 20  | 50  | -60 | 70   | 70   | -80  | 20  | 70   | -30 | 40  | -10 | 0   | 40  | 30  | 150 | 40  | -10 | -10 | -20 | -50  | -60 | 110 | 100 | 100 |     |
| A    | 48  | 19  | -10 | 19   | 19   | -38  | 19  | 0    | -8  | 48  | -13 | 8   | 19  | 19  | 19  | 18  | 19  | 0   | -22 | -19  | 19  | 100 | 100 | 100 |     |
| P    | 40  | 8   | 10  | 10   | 10   | 47   | 27  | 10   | -11 | 6   | -18 | -11 | 3   | 92  | 20  | 13  | 28  | 23  | 8   | -57  | -50 | 14  | 100 | 100 |     |
| Q    | 10  | 60  | 20  | 70   | 50   | -60  | 150 | -20  | -30 | 110 | -30 | 40  | 30  | 20  | -30 | 60  | 40  | 20  | 100 | -70  | 30  | 100 | 100 | 100 |     |
| Q    | 11  | 14  | -42 | 44   | 44   | 65   | 10  | 41   | -20 | 44  | -10 | 3   | 28  | 18  | 91  | 34  | -3  | -3  | -14 | -27  | -42 | 68  | 100 | 100 |     |
| C    | 49  | 26  | 20  | 29   | 23   | -30  | 66  | -11  | -11 | 0   | -23 | -14 | 20  | 22  | 8   | -12 | 45  | 22  | 8   | -39  | -32 | 12  | 100 | 100 |     |
| L    | 13  | -13 | -22 | -13  | -6   | 16   | -6  | 0    | 19  | -6  | 38  | 35  | -13 | 38  | 6   | -3  | 0   | 6   | 29  | -10  | -16 | 0   | 100 | 100 |     |
| I    | 41  | 11  | 22  | -38  | 35   | 53   | -17 | 12   | 20  | 1   | 11  | 10  | 12  | 16  | 3   | 42  | 0   | -1  | 4   | 2    | -35 | -20 | 47  | 100 | 100 |
| L    | -10 | -10 | -40 | -10  | -11  | 42   | -20 | -2   | 16  | -4  | 48  | 32  | -7  | -19 | 0   | 7   | -6  | -9  | 12  | 21   | 18  | -5  | 100 | 100 |     |
| L    | -3  | -31 | -43 | -31  | -20  | 71   | -26 | -16  | 61  | -20 | 96  | 82  | -27 | -16 | -8  | -27 | -24 | -3  | 66  | 17   | 16  | -14 | 100 | 100 |     |
| T    | 16  | 6   | 16  | 6    | 3    | 10   | 26  | 16   | 40  | 6   | 11  | 11  | 0   | 3   | 8   | -13 | 26  | 16  | 26  | 26   | -13 | 5   | 100 | 100 |     |

20

```

g1  ELVKAGSSVK MSCKATGYTF SSYE....LY WV
m3  GLVEPPGSLR LSCSASGFTF SAND....MN WV
k2  LPVTPGEPAS ISCRSSQSLD DSGDGNTYLN WY
l3  VSVALGQTVR ITCQ.GDGLR GYDAA..... WY
    
```

| Cons | A  | B   | C   | D   | E   | ... | Gap | Len |
|------|----|-----|-----|-----|-----|-----|-----|-----|
| 13C  | 30 | -40 | 150 | -50 | -60 |     | 100 | 100 |
| 14K  | 4  | 18  | -11 | 17  | 17  |     | 100 | 100 |
| 15A  | 53 | 11  | 19  | 12  | 12  |     | 30  | 30  |
| 16S  | 28 | 21  | 28  | 19  | 16  |     | 100 | 100 |

These days profiles of this kind have been largely replaced by Hidden Markov Models for sequence searching and alignment. We now turn to HHMs.

21

## Hidden Markov models

Processes  $\{(S_t, O_t), t=1, \dots\}$ , where  $S_t$  is the *hidden state* and  $O_t$  the *observation* at time  $t$ , such that

$$pr(S_t | S_{t-1}, O_{t-1}, S_{t-2}, O_{t-2} \dots) = pr(S_t | S_{t-1})$$

$$pr(O_t | S_{t-1}, O_{t-1}, S_{t-2}, O_{t-2} \dots) = pr(O_t | S_t, S_{t-1})$$

The basics of HHMs were laid bare in a series of beautiful papers by L E Baum and colleagues around 1970, and their formulation has been used almost unchanged to this day.

22

## HMMs: generalities

As the name suggests, the series  $\mathbf{O} = (O_1, O_2, O_3, \dots, O_T)$  is **observed**, while the *hidden states*  $\mathbf{S} = (S_1, S_2, S_3, \dots, S_T)$  are not.

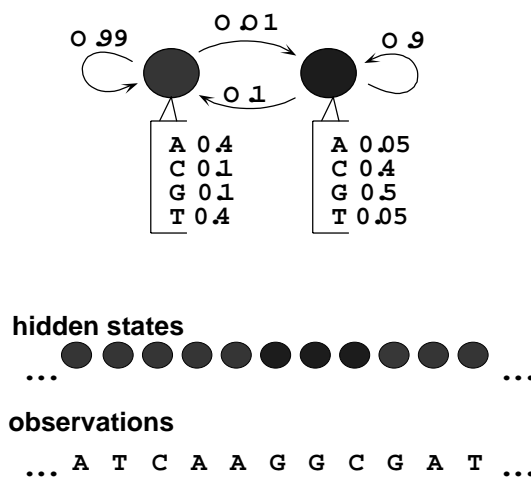
$S = \text{States } \{s_0, s_1, \dots, s_n\}$ ;  $V = \text{Output alphabet } \{v_0, v_1, \dots, v_m\}$   
 $A = \{a_{ij}\} = \text{transition probability from } s_i \rightarrow s_j$   
 $B = \{b_i(j)\} = \text{probability outputting } v_j \text{ in state } s_i$  } parameters  $\theta$

What is the most likely sequence of states that produced a given sequence of observations? What is the likelihood of a sequence? Estimate the parameters.

There are elegant algorithms for calculating  $pr(\mathbf{O}|\theta)$ ,  $arg \max_{\theta} pr(\mathbf{O}|\theta)$  in certain special cases, and  $arg \max_S pr(\mathbf{S}|\mathbf{O}, \theta)$ .

23

## A simple HMM (Churchill, 1989)



24

## Hidden Markov models: extensions

Many variants are now used. For example, the distribution of  $O$  may not depend on previous  $S$  but on previous  $O$  values,

$$pr(O_t | S_t, S_{t-1}, O_{t-1}, \dots) = pr(O_t | S_t), \text{ or}$$

$$pr(O_t | S_t, S_{t-1}, O_{t-1}, \dots) = pr(O_t | S_t, S_{t-1}, O_{t-1}).$$

Most importantly for us, the times of  $S$  and  $O$  may be decoupled, permitting the *Observation* corresponding to *State time*  $t$  to be a string whose length and composition depends on  $S_t$  (and possibly  $S_{t-1}$  and part or all of the previous *Observations*). This is called a hidden semi-Markov or generalized hidden Markov model.

25

### Some early applications of HMMs

- finance, but we never saw them
- speech recognition
- modelling ion channels

In the mid-late 1980s HMMs entered genetics and molecular biology, and they are now firmly entrenched.

### Some current applications of HMMs to biology

- mapping chromosomes
- aligning biological sequences
- predicting sequence structure
- inferring evolutionary relationships
- finding genes in DNA sequence

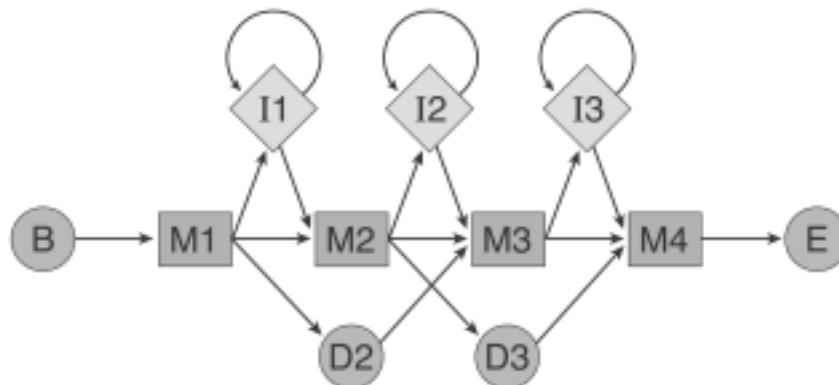
## Alignment using profile Hidden Markov models

There are now many HHMs for protein families such as globins, and these models can be used to infer alignments of new globin sequences to other members of the family.

Such models can also be used to determine whether a given sequence is or is not a member of a specified family.

27

## A very short profile HMM



**M = Match state, I = Insert state, D = Delete state.**  
**To operate, go from left to right. I and M states output amino acids; B, D and E states are “silent”.**

## How profile HMMs work: in brief

**Instances** of the motif are identified by calculating  
 $\log\{pr(sequence | M)/pr(sequence | B)\},$   
where  $M$  and  $B$  are the motif and background HMMs.

**Alignments** of instances of the motif to the HMM are found by calculating

$$\arg \max_{states} pr(states | instance, M).$$

**Estimation** of HMM parameters is by calculating

$$\arg \max_{parameters} pr(sequences | M, parameters).$$

In all cases, we use the **efficient HMM algorithms**.

29

## Pfam domain-HMMs

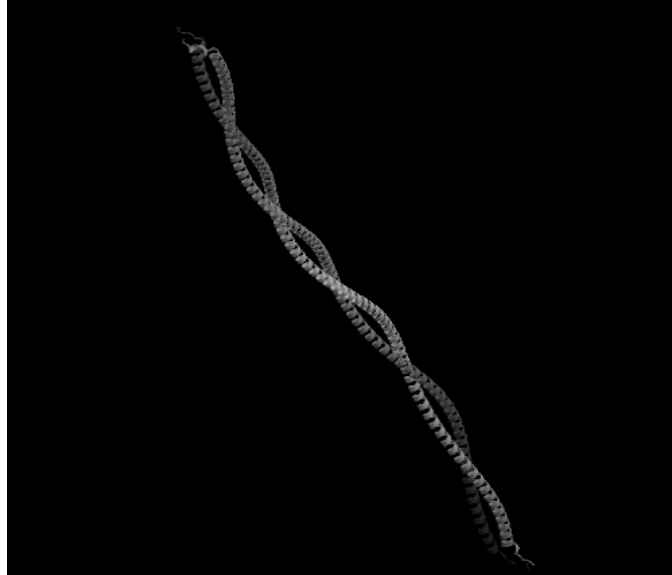
**Pfam** is a library of models of recurrent protein domains. They are constructed semi-automatically using hidden Markov models (HMMs).

Pfam families have permanent accession numbers and contain functional annotation and cross-references to other databases, while Pfam-B families are re-generated at each release and are unannotated.

See <http://pfam.wustl.edu/>

30

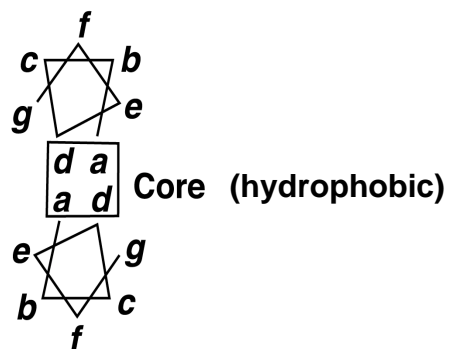
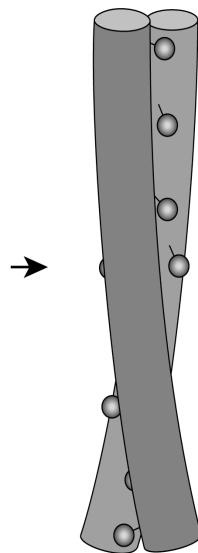
### HMM for coiled-coil domains

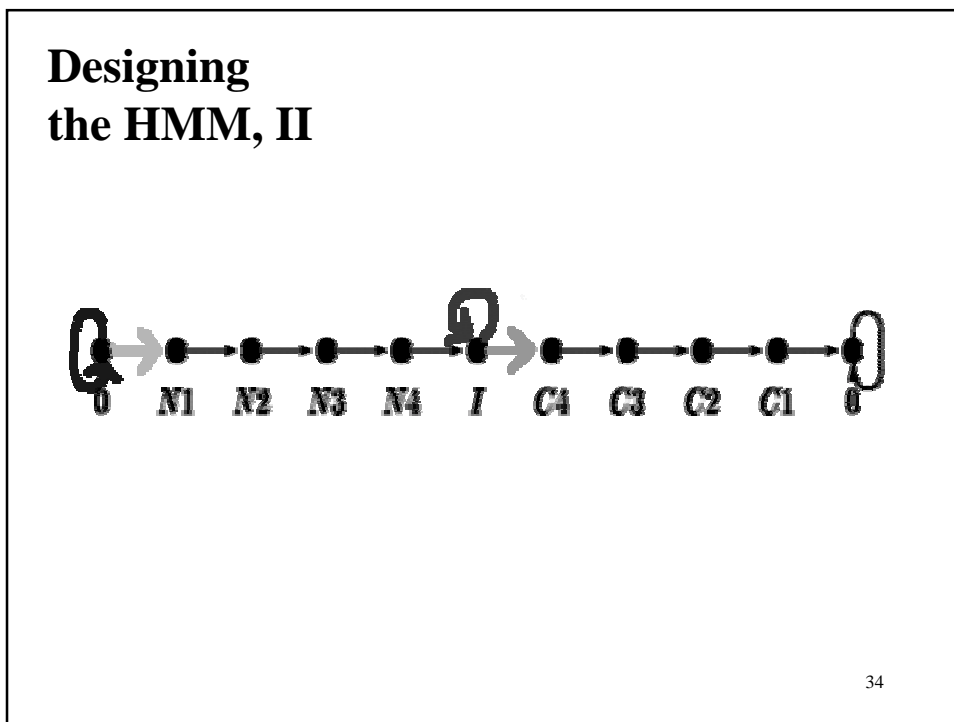
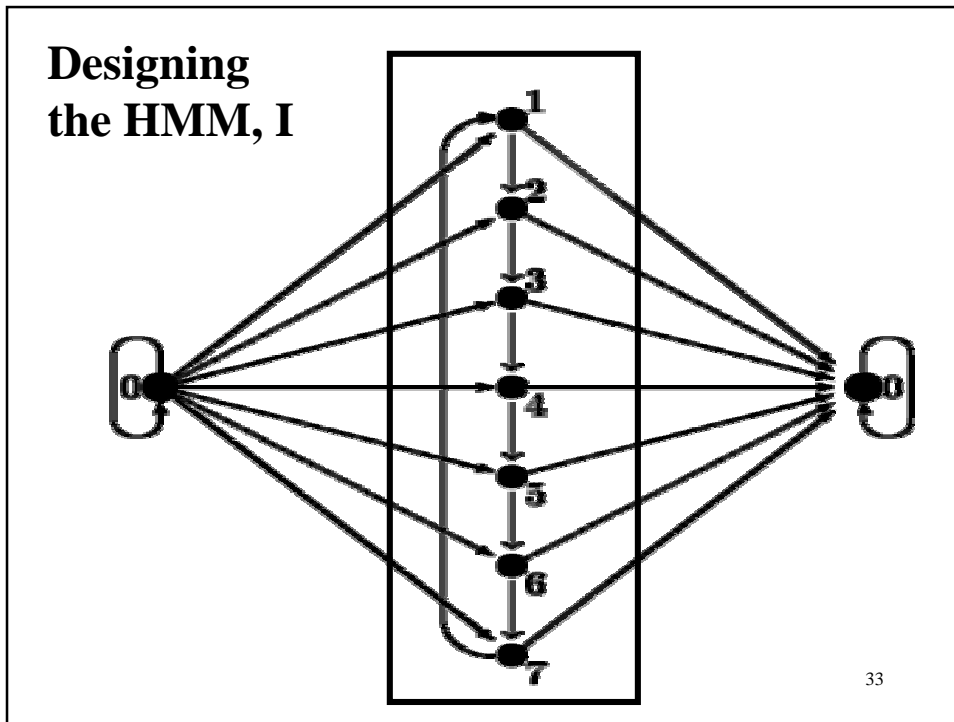


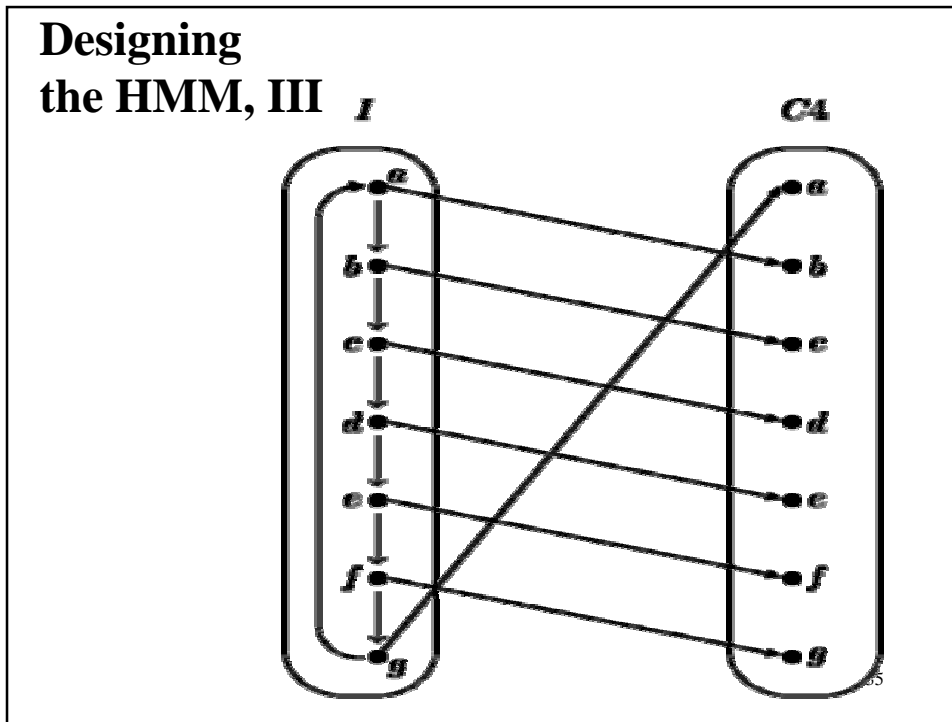
2TMA: Tropomyosin

31

### Left-handed coiled coil (all known structures)





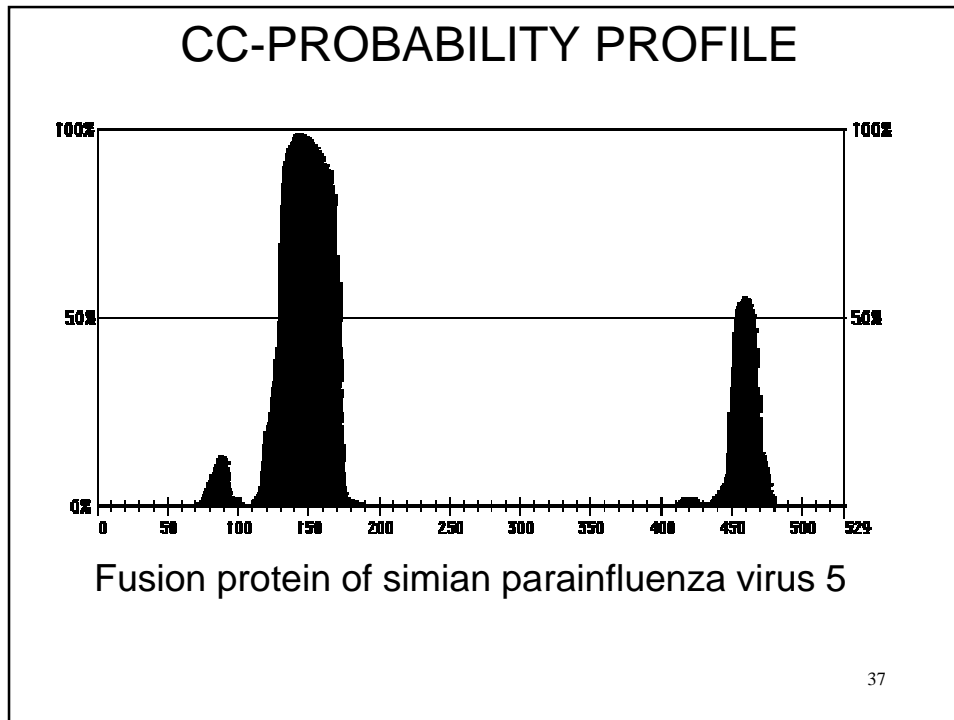


### HMM: decoding

|                 |                            |
|-----------------|----------------------------|
| <b>Sequence</b> | WGP ARQLNES VKDINKM LER HP |
| <b>Labels</b>   | BBB CCCCCC CCCCCC CCC BB   |
| <b>Path1</b>    | 000 abcdefg abcdefg abc 00 |
| <b>Path2</b>    | 00c defgabc defgabc def g0 |

VITERBI decoding: of all possible state-paths, we determine the maximum probability one given the amino acid sequence **O**;

POSTERIOR decoding: at each position, we determine the state with the highest probability given **O**.



## References

***Biological Sequence Analysis*** R Durbin, S Eddy, A Krogh and G Mitchison. Cambridge University Press, 1998.

***Bioinformatics The machine learning approach*** P Baldi and S Brunak. The MIT Press, 1998.

***Post-Genome Informatics*** M Kanehisa  
Oxford University Press, 2000

***Pattern discovery in biomolecular data***, by  
JTL Wang, BA Shapiro and D Shasha

***Computational methods in molecular biology***  
eds SL Salzberg, DB Searls and S Kasif, ch 4

38

## HMM-type software available

SAM oldest profile HMM software.

HMMER profile HMM to do sensitive database searching.

PFTOOLS “generalized profiles” similar to profile HHMs,  
used in Prosite Profiles.

PROBE multiple ungapped HMM motifs including Gibbs  
sampling.

META-MEME motif models similar to PROBE.

PSI-BLAST stripped down ultra-fast iterative profile HMM  
server from the NCBI.

39

## References

### ***Biological Sequence Analysis***

R Durbin, S Eddy, A Krogh and G Mitchison  
Cambridge University Press, 1998.

### ***Bioinformatics The machine learning approach***

P Baldi and S Brunak  
The MIT Press, 1998

### ***Post-Genome Informatics***

M Kanehisa  
Oxford University Press, 2000

40