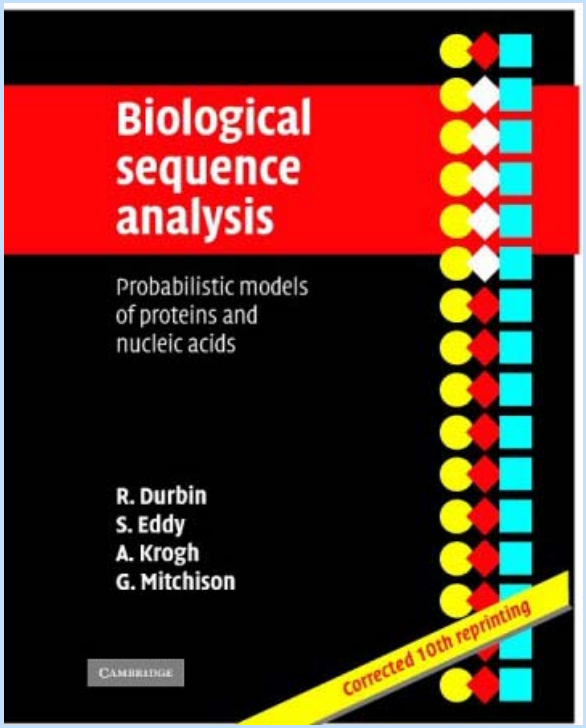
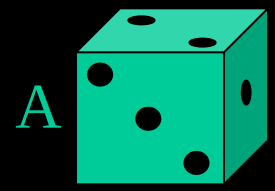
Hidden Markov Models



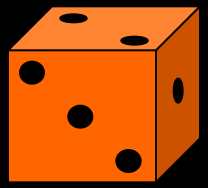
Main source: Durbin et al.,
"Biological Sequence Alignment"
(Cambridge, '98)

The occasionally dishonest casino



A

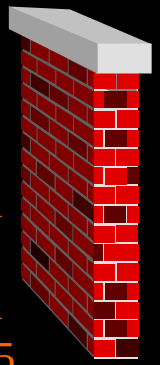
$P_A(1)$ =
$P_A(2)$ =
… = 1/6



$P_{A->B}$ =
$P_{B->A}$ =
1/10

B

$P_B(1)$=0.1
…
$P_B(5)$=0.1
$P_B(6)$ =0.5



13652656643662612564

13652656643662612564

Can we tell when the loaded die is used?

Example - CpG islands

- **CpG islands**:
  - DNA stretches (100~1000bp) with frequent CG pairs (contiguous on same strand).
  - Rare, appear in significant genomic parts.
- **Problem (1)**: Given a short genome sequence, decide if it comes from a CpG island.

Preliminaries: Markov Chains

(S, A, p)

- S: State set
- p: Initial state prob. vector $\{p(x_1=s)\}$ [alternatively, use a begin state]
- A: Transition prob. matrix $a_{st} = P(x_i=t \mid x_{i-1}=s)$

Assumption: $X=x_1 \ldots x_n$ is a random process with *memory length 1*, i.e.: $\forall s_i \in S$

$P(x_i=s_i \mid x_1=s_1,\ldots,x_{i-1}=s_{i-1}) = P(x_i=s_i \mid x_{i-1}=s_{i-1}) = a_{s_{i-1},s_i}$

- Sequence probability:

$$P(X) = p(x_1) \cdot \Pi_{i=2\ldots L} a_{x_{i-1},x_i}$$

Markov Models

- - Transition probs for non-CpG islands
- + Transition probs for CpG islands

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

+	A	C	G	T
A	0.180	0.274	0.425	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

CpG islands: Fixed Window

- **Problem (1):** Given a short genome sequence X, decide if it comes from a CpG island.

- **Solution**: Model by a Markov chain. Let
  - a^+_{st}: transition prob. in CpG islands,
  - a^-_{st}: transition prob. outside CpG islands.

  Decide by log-likelihood ratio score:

$$score(X) = \log \frac{P(X \mid CpG-island)}{P(X \mid non-CpG-island)} = \sum_{i=1}^{n} \log \frac{a^+_{x_{i-1},x_i}}{a^-_{x_{i-1},x_i}}$$

$$bits\_score(X) = \frac{1}{n} \sum_{i=1}^{n} \log_2 \frac{a^+_{x_{i-1},x_i}}{a^-_{x_{i-1},x_i}}$$

Discrimination of sequences via Markov Chains

48 CpG islands, tot length ~60K nt. Similar non-CpG.

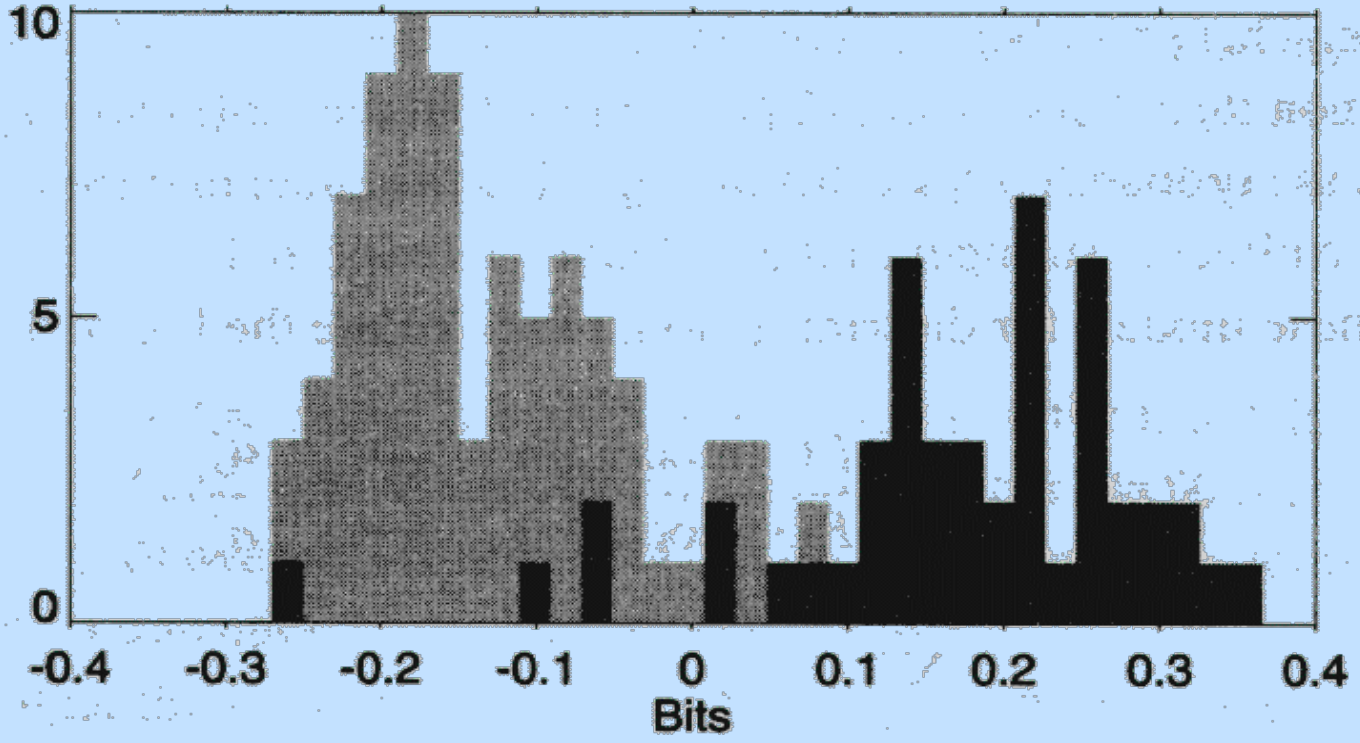


Figure 3.2 *The histogram of the length-normalised scores for all the sequences. CpG islands are shown with dark grey and non-CpG with light grey.*

Durbin et. al, Fig. 3.2

CpG islands – the general case

- **Problem(2):** Detect CpG islands in a long DNA sequence.
- **Naive Solution - Sliding windows**: $\forall 1 \leq k \leq L-l$,
  - window: $X^k = (x_{k+1},\ldots,x_{k+l})$
  - score: $score(X^k)$
  - positive score $\Rightarrow$ potential CpG island

Disadvantage: what is the length of the islands? How do we identify transitions?

Idea: Use Markov chains as before, with additional (hidden) states

Hidden Markov Model (HMM)

Finite set of states, capable of emitting symbols.
Example:
$Q = \{A_+, C_+, G_+, T_+, A_-, C_-, G_-, T_-\}$

Alphabet of symbols
Example: {A, C, G, T}

$M=(\Sigma, Q, \Theta)$

- $\Theta=(A,E)$
- A: Transition prob. a_{kl} $\forall k,l \in Q$
- E: Emission prob. $e_k(b)$ $\forall k \in Q$, $b \in \Sigma$

path $\Pi=\pi_1,\ldots,\pi_n$ (sequence of states - simple Markov chain; convention: π_0 - begin, π_{L+1} - end)

Given sequence $X = (x_1,\ldots,x_L)$:

- $a_{kl} = P(\pi_i=l \mid \pi_{i-1}=k)$,
- $e_k(b) = P(x_i=b \mid \pi_i=k)$

$$P(X,\Pi) = a_{\pi_0,\pi_1} \cdot \Pi_{i=1\ldots L} e_{\pi_i}(x_i) \cdot a_{\pi_i,\pi_{i+1}}$$

Ex.: express $P(X,\Pi)$ in terms of counts

Viterbi's Decoding Algorithm
(finding most probable state path)

Want: path Π maximizing $\mathbf{P(X, \Pi)}$

$v_k(i)$ = prob. of most probable path ending in state k at step i.

Init: $v_0(0) = 1$; $v_k(0)=0 \ \forall k>0$

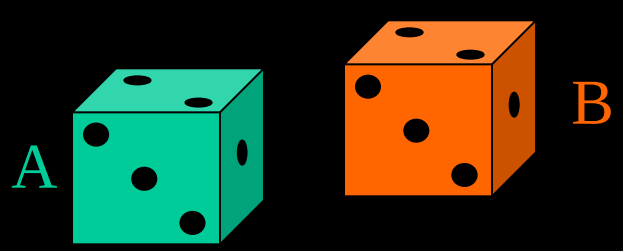
Step: $v_k(i+1)=e_k(x_{i+1})\cdot\max_l\{v_l(i)\cdot a_{lk}\}$

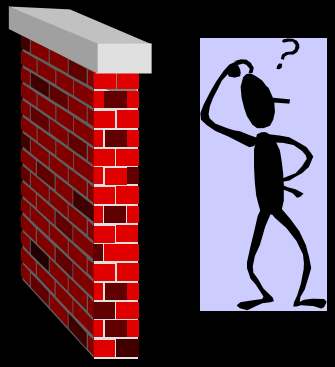
End: $P(X, \Pi^*) = \max_l\{v_l(L) \cdot a_{l0}\}$

Time complexity: $O(Ln^2)$ for n states, L steps

Can find Π^* using back pointers.

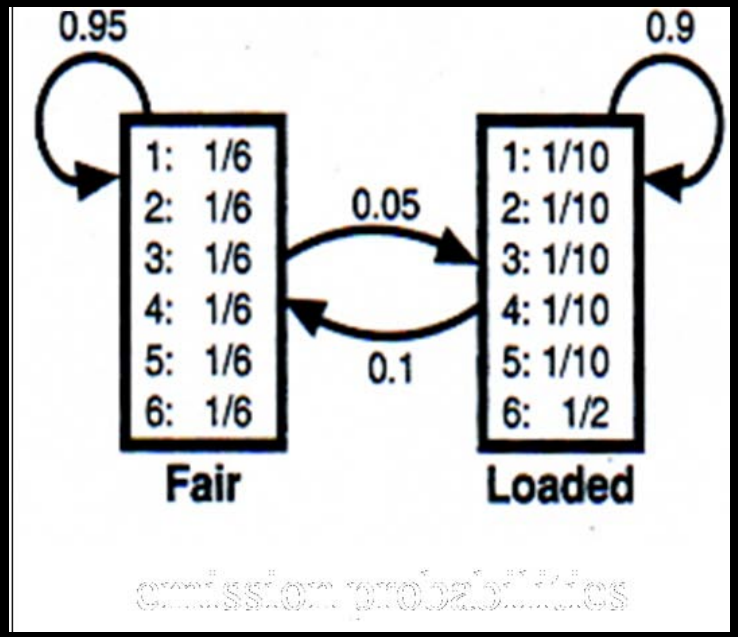
The occasionally dishonest casino (2)





1365265664366261256 4

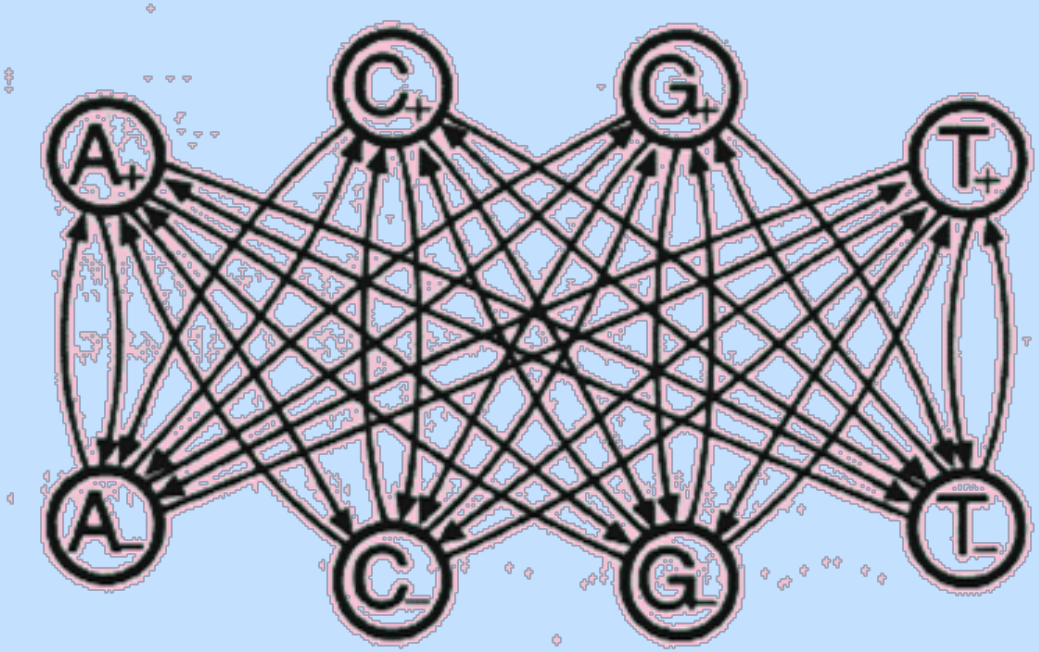
1365265664366261256 4



The occasionally dishonest casino (2)

HMM for CpG Islands

- States: A_+ C_+ G_+ T_+ A_- C_- G_- T_-
- Symbols: A C G T A C G T
- Path $\Pi=\pi_1,\ldots,\pi_n$: sequence of states



+	A	C	G	T
A	0.180	0.274	0.425	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

-	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

transition prob.

http://www.cs.huji.ac.il/~cbio/handouts/class4.ppt

Posterior State Probabilities

Goal: calculate $P(\pi_i=k \mid X)$

- **Our strategy:**
- $P(X, \pi_i=k) =$

  $= P(x_1,\ldots,x_i, \pi_i=k) \cdot P(x_{i+1},\ldots,x_L \mid x_1,\ldots,x_i, \pi_i=k)$

  $= P(x_1,\ldots,x_i, \pi_i=k) \cdot P(x_{i+1},\ldots,x_L \mid \pi_i=k)$
- $P(\pi_i=k \mid X) = P(\pi_i=k, X) / P(X)$

➔Need to compute these two terms - and P(X)

Forward Algorithm

Goal: calculate $P(X) = \Sigma_{\Pi} P(X, \Pi)$

Approximation: take max path Π^* from Viterbi alg.

Not justified when $\exists$ several near maximal paths

Exact alg : "Forward Algorithm"

$f_k(i) = P(x_1,\ldots,x_i, \pi_i=k)$

- Init: $f_0(0) = 1;$ $f_k(0)=0 \ \forall k>0$
- Step: $f_k(i+1) = e_k(x_{i+1}) \cdot \Sigma_l f_l(i) \cdot a_{lk}$
- End: $P(X) = \Sigma_l f_l(L) \cdot a_{l0}$

Backward Algorithm

- $b_k(i) = P(x_{i+1},\ldots x_L \mid \pi_i=k)$
- <u>init</u>: $\forall k,\ b_k(L) = a_{k0}$
- <u>step</u>: $b_k(i) = \sum_l a_{kl}\cdot e_l(x_{i+1})\cdot b_l(i+1)$
- <u>End</u>: $P(X) = \sum_k a_{0k}\cdot e_k(x_1)\cdot b_k(1)$

Posterior State Probabilities (2)

Goal: calculate $P(\pi_i=k \mid X)$

- Recall:
  - $f_k(i) = P(x_1,\ldots,x_i\ ,\ \pi_i=k)$
  - $b_k(i) = P(x_{i+1},\ldots x_L \mid \pi_i=k)$
  - Each can be used to compute P(X)
- $P(X,\ \pi_i=k)$ =

  $= P(x_1,\ldots,x_i,\ \pi_i=k) \cdot P(x_{i+1},\ldots,x_L \mid x_1,\ldots,x_i,\ \pi_i=k)$

  $= P(x_1,\ldots,x_i,\ \pi_i=k) \cdot P(x_{i+1},\ldots,x_L \mid \pi_i=k)$

  $= f_k(i) \cdot b_k(i)$
- $P(\pi_i=k \mid X) = P(\pi_i=k, X) / P(X)$

Dishonest Casino (3)

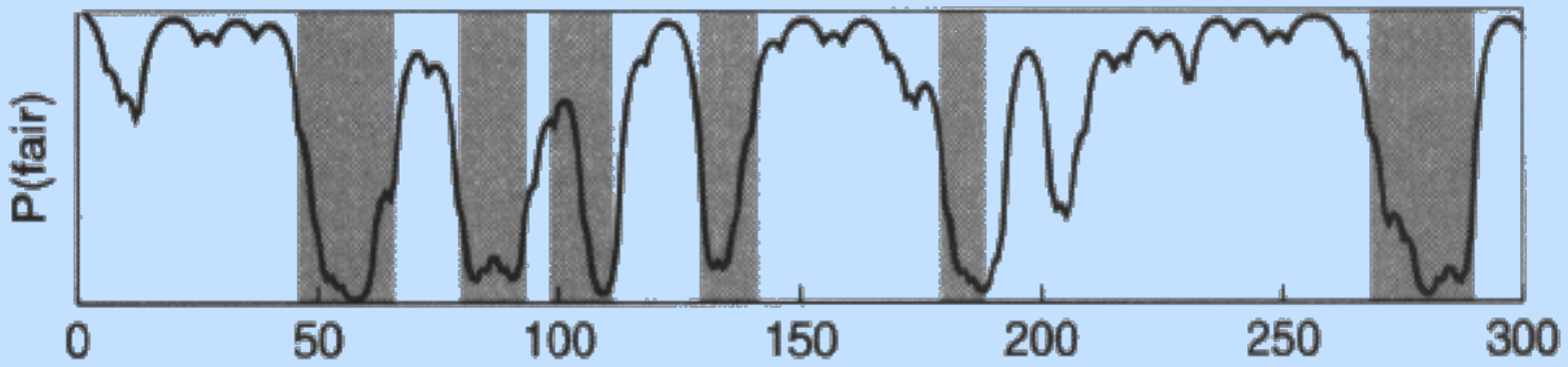


Figure 3.6 *The posterior probability of being in the state corresponding to the fair die in the casino example. The x axis shows the number of the roll. The shaded areas show when the roll was generated by the loaded die.*

Durbin et al. pp. 60

Posterior Decoding

- Now we have $P(\pi_i=k \mid X)$. How do we decode?

1. $\pi_i^* = \text{argmax}_k P(\pi_i=k \mid X)$
   - Good when interested in state at particular point
   - path of states $\pi_1^*,\ldots,\pi_L^*$ may not be legal
2. Define a function of interest $g(i)$ on the states. Compute $G(i|X) = \Sigma_k P(\pi_i=k \mid X) \cdot g(k)$
   - E.g.: $g(i)$ =1 for states in S, 0 on the rest: $G(i|X)$ is posterior prob of symbol i coming from S

e.g., CpG island
$S=\{A_+,C_+,G_+,T_+\}$

Parameter Estimation for HMMs

Log likelihood of model

Input: $X^1,\ldots,X^n$ independent training sequences

Goal: estimation of $\Theta = (A,E)$ (model parameters)

Note: $P(X^1,\ldots,X^n \mid \Theta) = \Pi_{i=1\ldots n} P(X^i \mid \Theta)$ (indep.)

$l(x^1,\ldots,x^n \mid \Theta) = \log P(X^1,\ldots,X^n \mid \Theta) = \sum_{i=1\ldots n} \log P(X^i \mid \Theta)$

Case 1 - Estimation When State Sequence is Known:

A_{kl} = #(occurred $k \to l$ transitions)

$E_k(b)$ = #(emissions of symbol b that occurred in state k)

Max. Likelihood Estimators:

- $a_{kl} = A_{kl} / \sum_{l'} A_{kl'}$
- $e_k(b) = E_k(b) / \sum_{b'} E_k(b')$

small sample, or
prior knowledge correction:

$A'_{kl} = A_{kl} + r_{kl}$

$E'k(b) = E_k(b) + r_k(b)$

- "Dirichlet priors"

Parameter Estimation in HMM

Case 2: -Estimation When States are Unknown

Input: $X^1,\ldots,X^n$ indep training sequences

- guarantees convergence
- monotone
- many local optima
- special case of EM

Baum-Welch alg. (1972):

✶ **Expectation**:

- compute expected no. of k→l state transitions: *(ex.)*

$$P(\pi_i=k,\ \pi_{i+1}=l \mid X, \Theta) = [1/P(x)] \cdot f_k(i) \cdot a_{kl} \cdot e_l(x_{i+1}) \cdot b_l(i+1)$$

$$\Rightarrow A_{kl} = \sum_j [1/P(X^j)] \cdot \sum_i f_k^j(i) \cdot a_{kl} \cdot e_l(x^j_{i+1}) \cdot b_l^j(i+1)$$

- compute expected no. of symbol b appearances in state k

$$E_k(b) = \sum_j [1/P(X^j)] \cdot \sum_{\{i \mid x^j_i=b\}} f_k^j(i) \cdot b_k^j(i)$$ *(ex.)*

✸ **Maximization**:

- re-compute new parameters from A, E using max. likelihood.

repeat (1)+(2) until improvement $\leq \varepsilon$

Profile HMMs (Haussler et al, 1993)

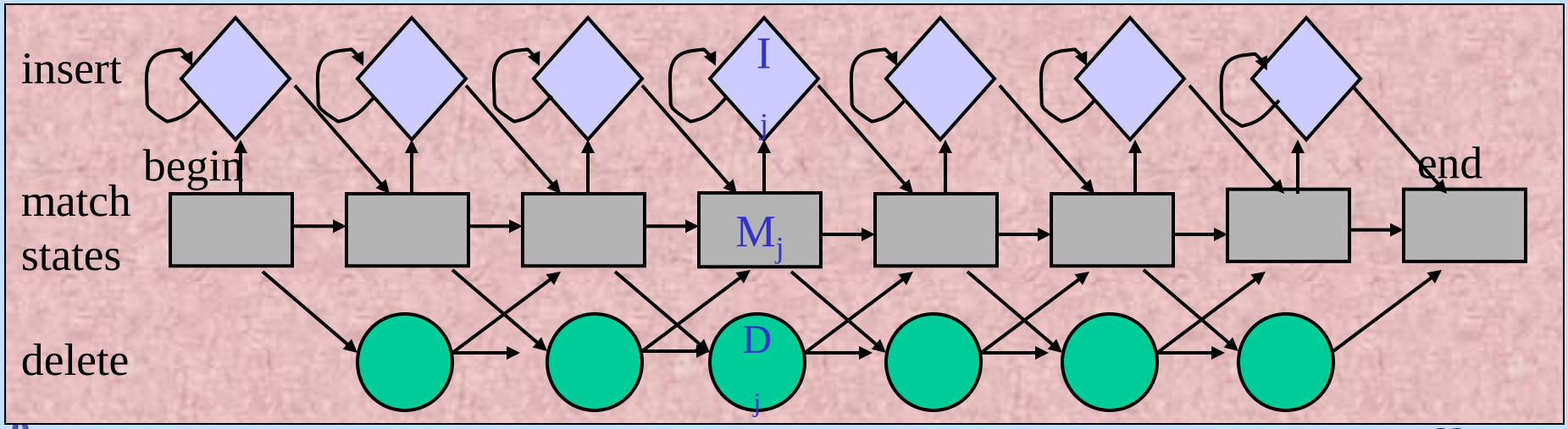
- Ungapped alignment of X against a profile M:
  - $e_i(a)$ = prob. of observing a at position i.
  - $P(X \mid M) = \Pi_{i=1\ldots L} e_i(x_i)$, or
  - $Score(X \mid M) = \sum_{i=1\ldots L} \log [e_i(x_i) / q_{x_i}]$

AG--C
A-AT-C
AG-T-C

background prob.

- indels:

AG---C



Profile HMMs

- Gapped alignment of X against a profile M:

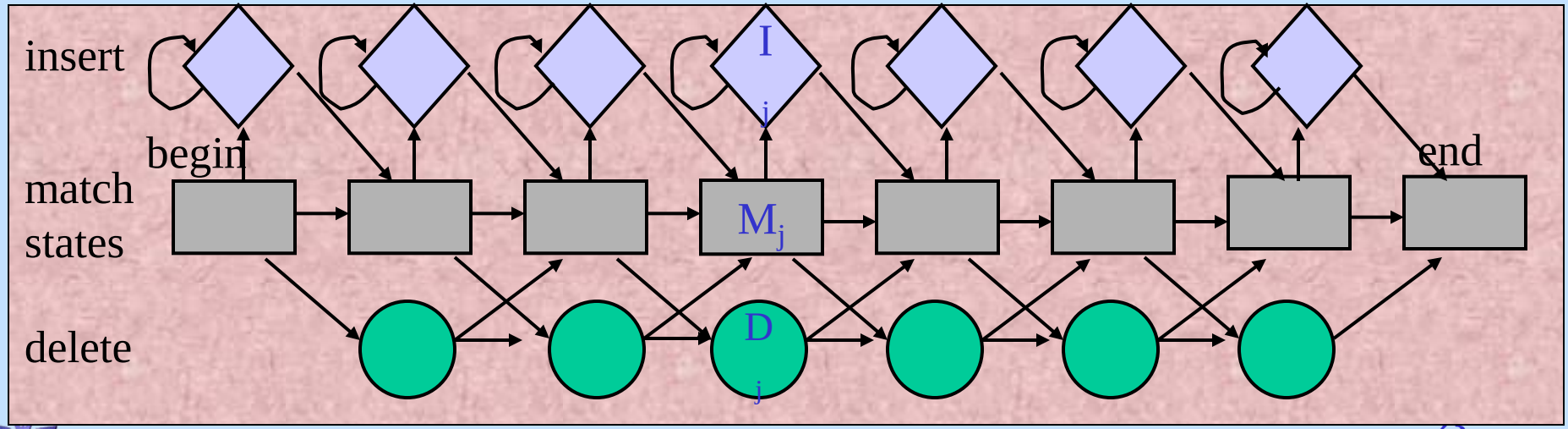
assume $e_{I_j}(a) = q_a$ (backgroud prob.)

$\Rightarrow$ gap of length k contributes to log-odds:

$$\log(a_{M_j,I_j}) + \log(a_{I_j,M_{j+1}}) + (k-1) \cdot \log(a_{I_j,I_j})$$

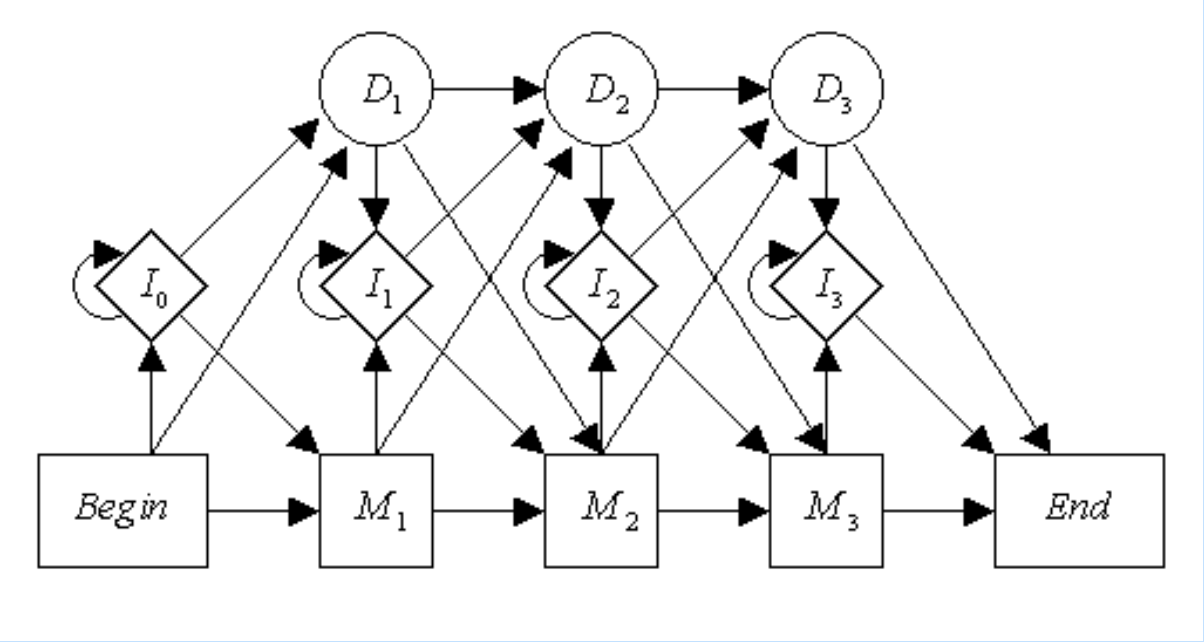
{ gap open } (gap extension)

No log odds contribution from the emission



Profile HMM

- Transition Probabilities
  - *Mi → Mi+1*
  - *Mi → Di+1*
  - *Mi → Ii*
  - *Ii → Mi+1*
  - *Ii → Ii*
  - *Ii → Di+1*
  - *Di → Di+1*
  - *Di → Mi+1*
  - *Di → Ii*
- Emission probabilities
  - $M_i \rightarrow a$
  - $I_i \rightarrow a$



http://www.cs.huji.ac.il/~cbio/handouts/Class6.ppt

Example

- Suppose we are given the aligned sequences

```
**---*
AG---C
A-AT-C
AG-AA-
--AAAC
AG---C
```

  - Suppose also that the "match" positions are marked…

Calculating A, E

count transitions and emissions:

transitions

	0	1	2	3
M-M				
M-D				
M-I				
I-M				
I-D				
I-I				
D-M				
D-D				
D-I				

```
**---*
AG---C
A-AT-C
AG-AA-
--AAAC
AG---C
```

emissions

	0	1	2	3
A				
C				
G				
T				
A				
C				
T				
G				

Calculating A, E

count transitions and emissions:

transitions

	0	1	2	3
M-M	4	3	2	4
M-D	1	1	0	0
M-I	0	0	1	0
I-M	0	0	2	0
I-D	0	0	1	0
I-I	0	0	4	0
D-M	-	0	0	1
D-D	-	1	0	0
D-I	-	0	2	0

```
**---*
AG---C
A-AT-C
AG-AA-
--AAAC
AG---C
```

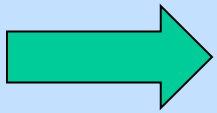
emissions

	0	1	2	3
A	-	4	0	0
C	-	0	0	4
G	-	0	3	0
T	-	0	0	0
A	0	0	6	0
C	0	0	0	0
T	0	0	1	0
G	0	0	0	0

Estimating **Maximum Likelihood** probabilities using fractions

emissions

	0	1	2	3
A	-	4	0	0
C	-	0	0	4
G	-	0	3	0
T	-	0	0	0
A	0	0	6	0
C	0	0	0	0
T	0	0	1	0
G	0	0	0	0

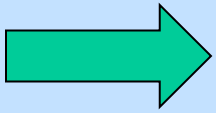


	0	1	2	3
A	-	1	0	0
C	-	0	0	1
G	-	0	1	0
T	-	0	0	0
A	.25	.25	.86	.25
C	.25	.25	0	.25
T	.25	.25	.14	.25
G	.25	.25	0	.25

Estimating ML probabilities (contd)

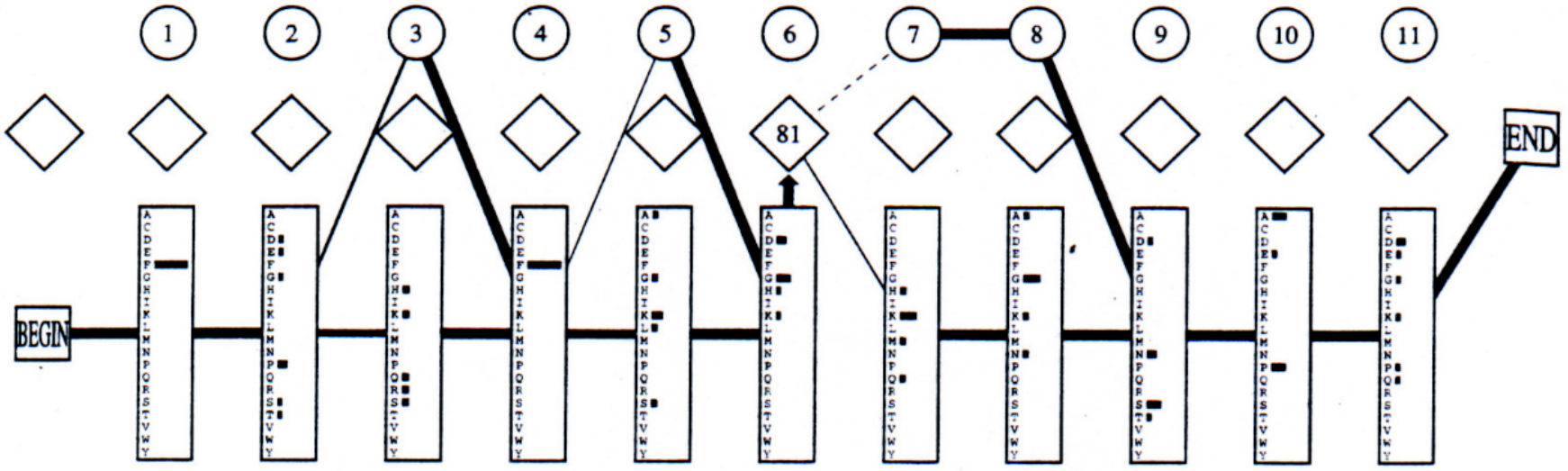
transitions

	0	1	2	3
M-M	4	3	2	4
M-D	1	1	0	0
M-I	0	0	1	0
I-M	0	0	2	0
I-D	0	0	1	0
I-I	0	0	4	0
D-M	-	0	0	1
D-D	-	1	0	0
D-I	-	0	2	0



	0	1	2	3
M-M	.8	.75	.66	1.0
M-D	.2	.25	0	0
M-I	0	0	.33	0
I-M	.33	.33	.28	.33
I-D	.33	.33	.14	.33
I-I	.33	.33	.57	.33
D-M	-	0	0	1
D-D	-	1	0	0
D-I	-	0	1	0

HMM from multiple alignment…

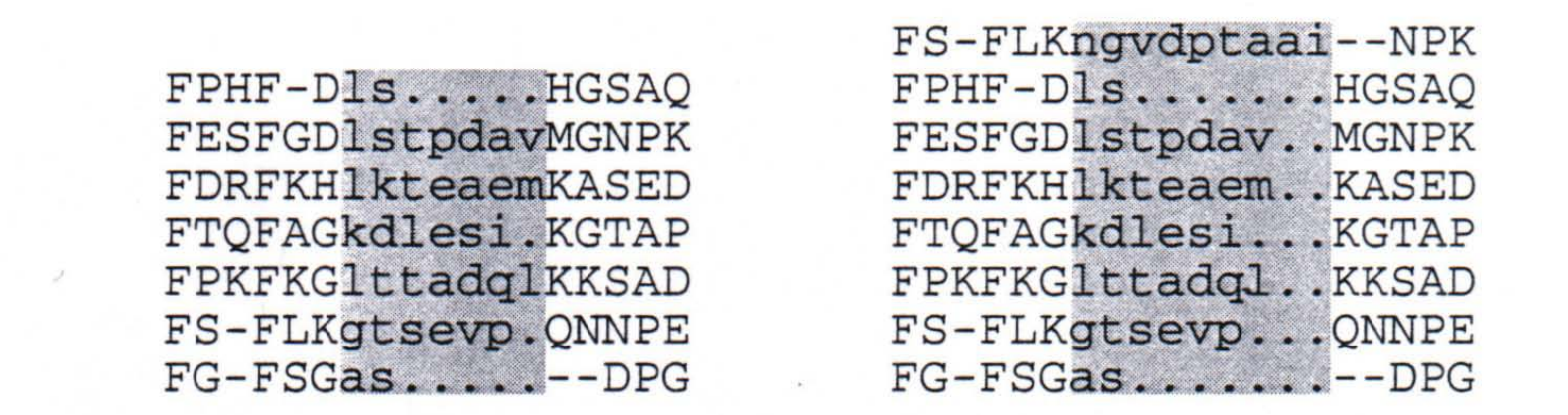



```
FPHF-DLS-----HGSAQ
FESFGDLSTPDAVMGNPK
FDRFKHLKTEAEMKASED
FTQFAG-KDLESIKGTAP
FPKFKGLTTADQLKKSAD
FS-FLK-GTSEVPQNNPE
FG-FSG----AS---DPG
```

Shade: insert

.. and multiple alignment from a given HMM

- Align each sequence to the profile separately



- Right: a new sequence realigend with the model
- Inserts are unaligned.

Searching with Profile HMMs

Compute: $\log \frac{P(X|Model)}{P(X|random)}$

$\Pi_i q_{x_i}$

$V_j^M(i)$: log odds of best path matching $x_1,\ldots,x_i$ to submodel up to level j, ending with x_i emitted by state M_j

$V_j^I(i)$: same, ending with x_i emitted by state I_j

$V_j^D(i)$: score for best path ending in D_j after x_i has been emitted (and x_{i+1} has not been emitted yet)

- $V_j^M(i) = \log [e_{Mj}(x_i) / q_{xi}] + \max \{ V_{j-1}^M(i-1)+\log(a_{Mj-1,Mj}), V_{j-1}^I(i-1)+\log(a_{Ij-1,Mj}), V_{j-1}^D(i-1)+\log(a_{Dj-1,Mj}) \}$

- $V_j^I(i) = \log [e_{Ij}(x_i) / q_{xi}] + \max \{ V_j^M(i-1)+\log(a_{Mj-1,Ij}), V_j^I(i-1)+\log(a_{Ij,Ij}), V_j^D(i-1)+\log(a_{Dj,Ij}) \}$

- $V_j^D(i) = \max \{V_{j-1}^M(i)+\log(a_{Mj-1,Dj}), V_{j-1}^I(i)+\log(a_{Ij-1,Dj}), V_{j-1}^D(i)+\log(a_{Dj-1,Dj}) \}$

Training a profile HMM from unaligned sequences

- choose length of the profile HMM, initialize parameters
- train the model using Baum-Welch
- obtain MA with the resulting profile as before.

(Formulas inlc. forward/backward – in handout)

An Illustrative Study

"HMMs in Computational Biology", Krogh, Brown, Mian, Sjoalnder, Haussler '93

- Globin experiment:
  - Heme-containing proteins, involved in the storage and transport of oxygen
  - 625 globins from Swissprot, ave. legnth 145AA
  - Training set: 400 sequences
  - Built model and ran it against test globins and all Swissprot proteins (25K).

Representative globins

```
Helix                AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCCC    DDDDDDDEE
HBA_HUMAN  ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF-DLS-----HGSA
HBB_HUMAN  --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
MYG_PHYCA  ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE
GLB3_CHITP ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQFAG-KDLESIKGTA
GLB5_PETMA PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSA
LGB2_LUPLU --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-FLK-GTSEVPQNNP
GLB1_GLYDI ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-FSG----AS---DP

Helix      EEEEEEEEEEEEEEEEEE            FFFFFFFFFFFF  FFGGGGGGGGGGGGGGGGGG
HBA_HUMAN  QVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL--RVDPVNFKLLSHCLLVTLAAHLPAE
HBB_HUMAN  KVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL--HVDPENFRLLGNVLVCVLAHHFGKE
MYG_PHYCA  DLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH--KIPIKYLEFISEAIIHVLHSRHPGD
GLB3_CHITP PFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG---VTHDQLNNFRAGFVSYMKAHT--D
GLB5_PETMA DVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF--QVDPQYFKVLAAVIADTVAAG----
LGB2_LUPLU ELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG---VADAHFPVVKEAILKTIKEVVGAK
GLB1_GLYDI GVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGNKHIKAQYFEPLGASLLSAMEHRIGGK

Helix       HHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN  FTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN  FTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA  FGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP FA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA -----DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU WSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI MNAAAKDAWAAAYADISGALISGLQS-----
```

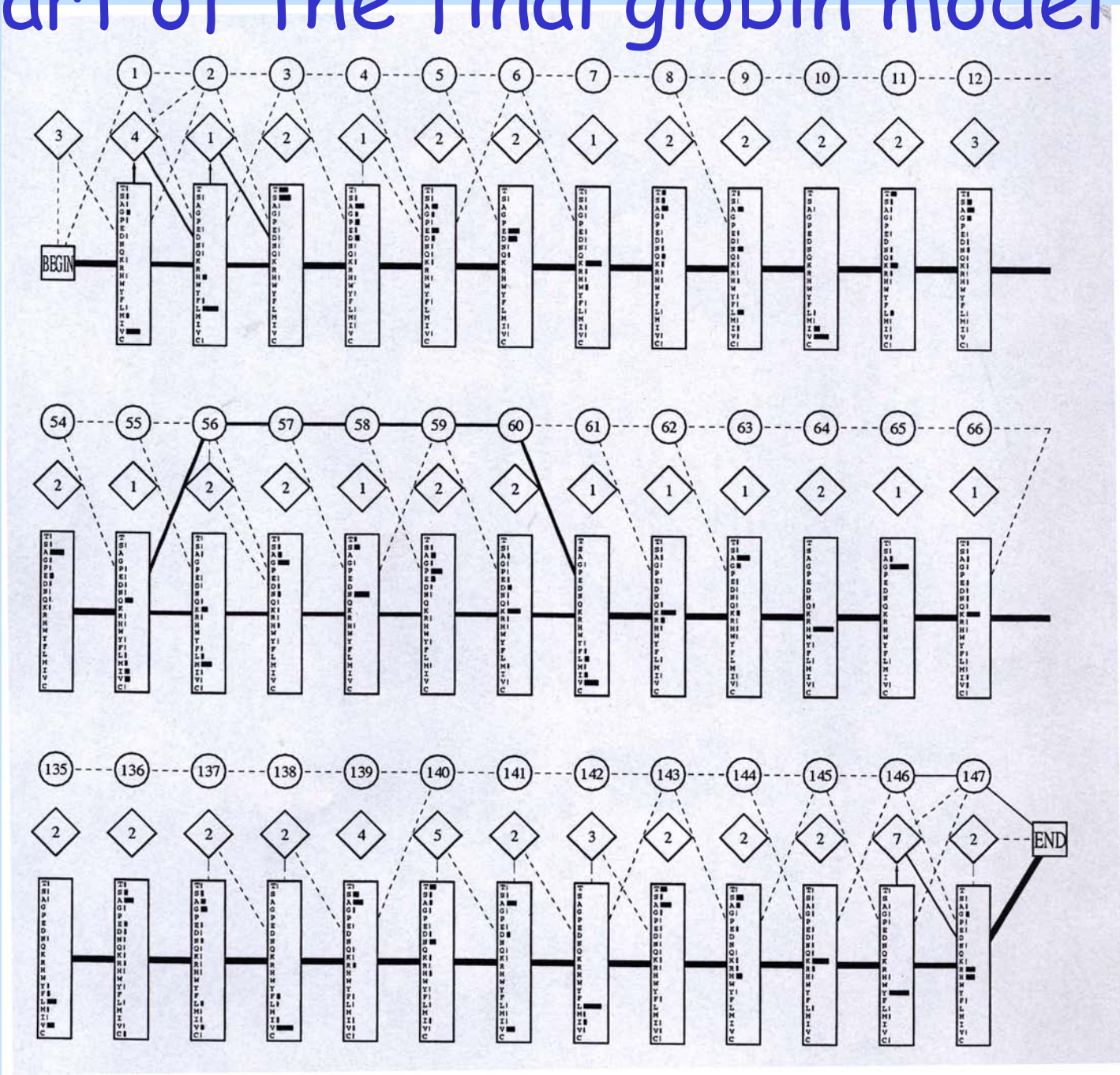
Alignment from Bashford et al 87

A-H: alpha helices

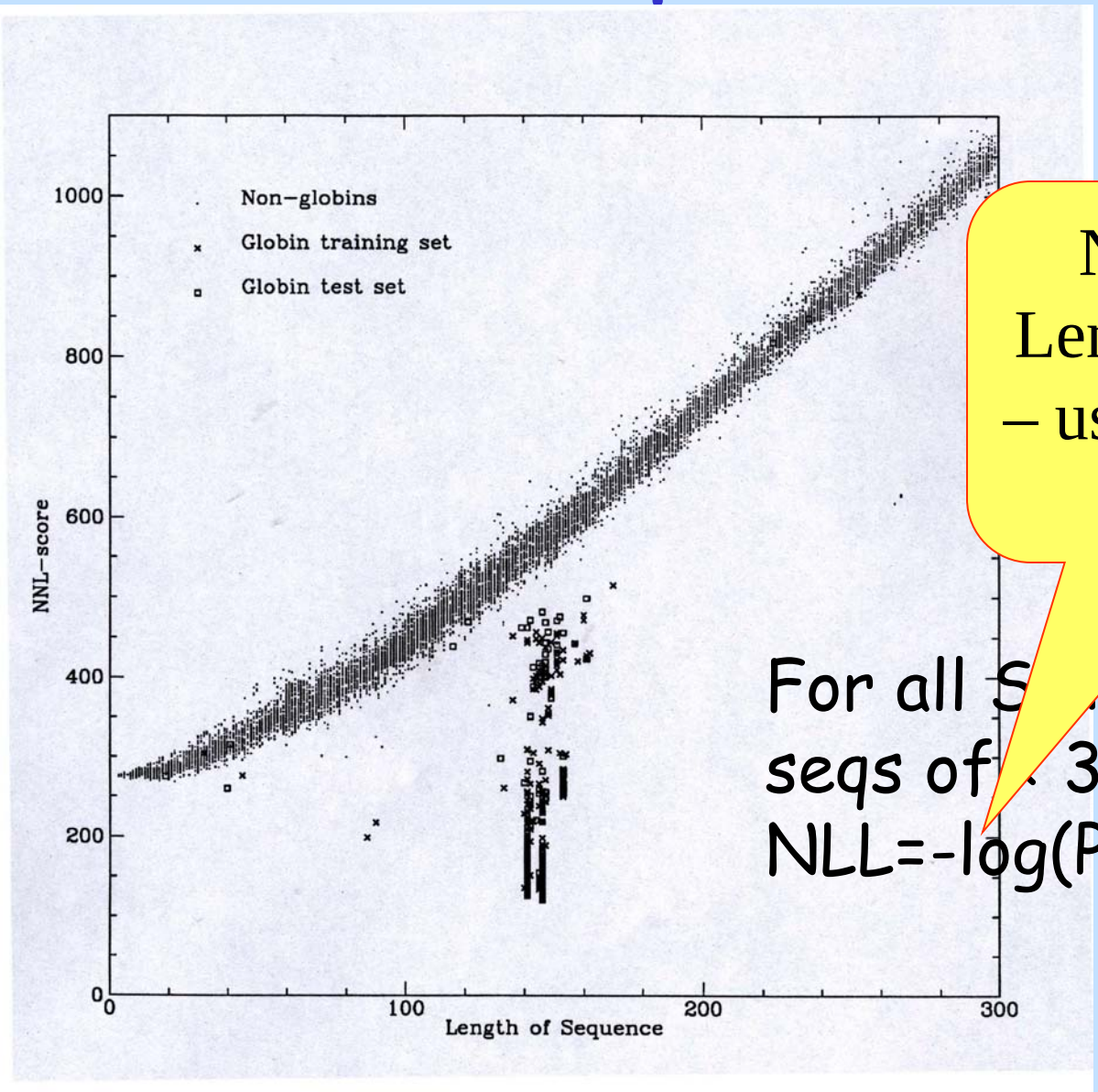
Realignment by trained HMM

```
Helix                AAAAAAAAAAAAAAAA   BBBBBBBBBBBBBBBBCCCCCCCCCCCC    DDDDDDDEE
                     ****************+   ++++++++++++++++++++++++++++           +
HBA_HUMAN  V.........LSPADKTNVKAAWGKVGA..HAGEYGAEALERMFLSFPTTKTYFPHF-DLSHGSAQ----
HBB_HUMAN  Vh........LTPEEKSAVTALWGKV--..NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNP
MYG_PHYCA  V.........LSEGEWQLVLHVWAKVEA..DVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASE
GLB3_CHITP -.........LSADQISTVQASFDKV--..KGDPVG--ILYAVFKADPSIMAKFTQF-AGKDLESIKGTA
GLB5_PETMA PivdtgsvapLSAAEKTKIRSAWAPVYS..TYETSGVDILVKFFTSTPAAQEFFPKFKGLTTADQLKKSA
LGB2_LUPLU Ga........LTESQAALVKSSWEEFNA..NIPKHTHRFFILVLEIAPAAKDLF-SFLKGTSEVPQ-NNP
GLB1_GLYDI G.........LSAAQRQVIAATWKDIAGadNGAGVGKDCLIKFLSAHPQMAAVF-GF----SGASD---P

Helix      EEEEEEEEEEEEEEEEEE            FFFFFFFFF   FFFFFGGG GGGGGGGGGGGGGGGG
           +******************            *********           ***************+
HBA_HUMAN  -VKGHGKKVADALTNAVAHVDD.....MPNALSALSDLHA...HKLRVDPV.NFKLLSHCLLVTLAAHLP
HBB_HUMAN  KVKAHGKKVLGAFSDGLAHLDN.....LKGTFATLSELHC...DKLHVDPE.NFRLLGNVLVCVLAHHFG
MYG_PHYCA  DLKKHGVTVLTALGAILKKKGH.....HEAELKPLAQSHA...TK-HKIPIkYLEFISEAIIHVLHSRHP
GLB3_CHITP PFETHANRIVGFFSKIIGELPN.....IEADVNTFVASHK...PR-GVTHD.QLNNFRAGFVSYMKAH--
GLB5_PETMA DVRWHAERIINAVNDAVASMDDtek..MSMKLRDLSGKHA...KSFQVDPQ.YFKVLAAVIADTVAA---
LGB2_LUPLU ELQAHAGKVFKLVYEAAIQLQVtgvvvTDATLKNLGSVHV...SK-GVADA.HFPVVKEAILKTIKEVVG
GLB1_GLYDI GVAALGAKVLAQIGVAVSHLGDegk..MVAQMKAVGVRHKgygNK-HIKAQ.YFEPLGASLLSAMEHRIG

Helix        HHHHHHHHHHHHHHHHHHHHHHHHH
             ++++++++++++++++++++++++
HBA_HUMAN  AEFTPAVHASLDKFLASVSTVLTSKY......R
HBB_HUMAN  KEFTPPVQAAYQKVVAGVANALAHKY......H
MYG_PHYCA  GDFGADAQGAMNKALELFRKDIAAKYkelgyqG
GLB3_CHITP TDF-AGAEAAWGATLDTFFGMIFSKM......-
GLB5_PETMA GD------AGFEKLMSMICILLRSAY......-
LGB2_LUPLU AKWSEELNSAWTIAYDELAIVIKKEMnda...A
GLB1_GLYDI GKMNAAAKDAWAAAYADISGALISGLq.....S
```

Part of the final globin model



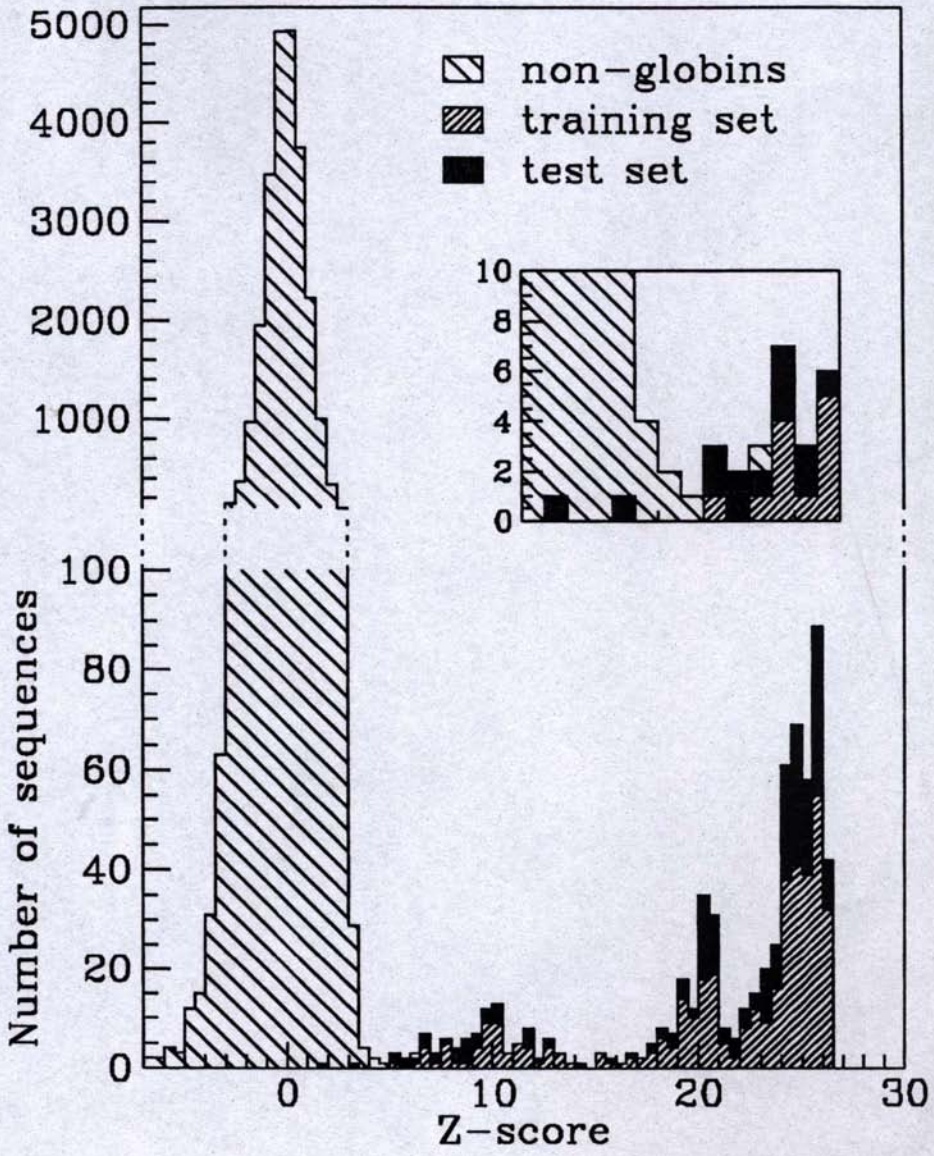
Score vs sequence length



Negative LL.
Length dependent – using log odds is preferable

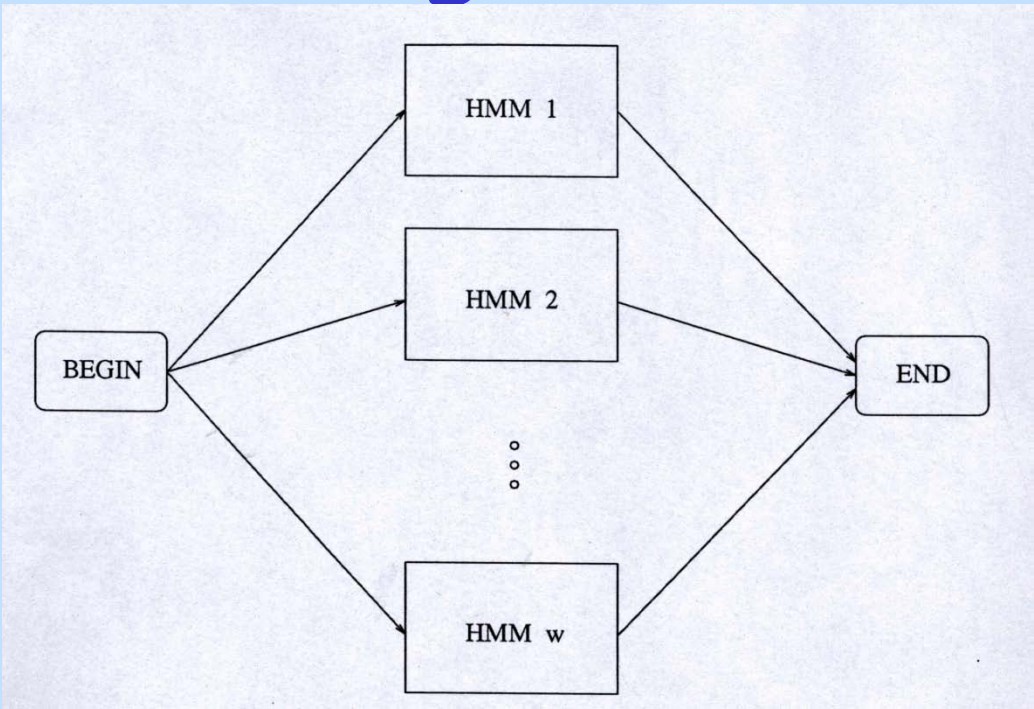
For all Swissprot seqs of ≤ 300AA
NLL=-log(P(seq|model))

Z-score distribution



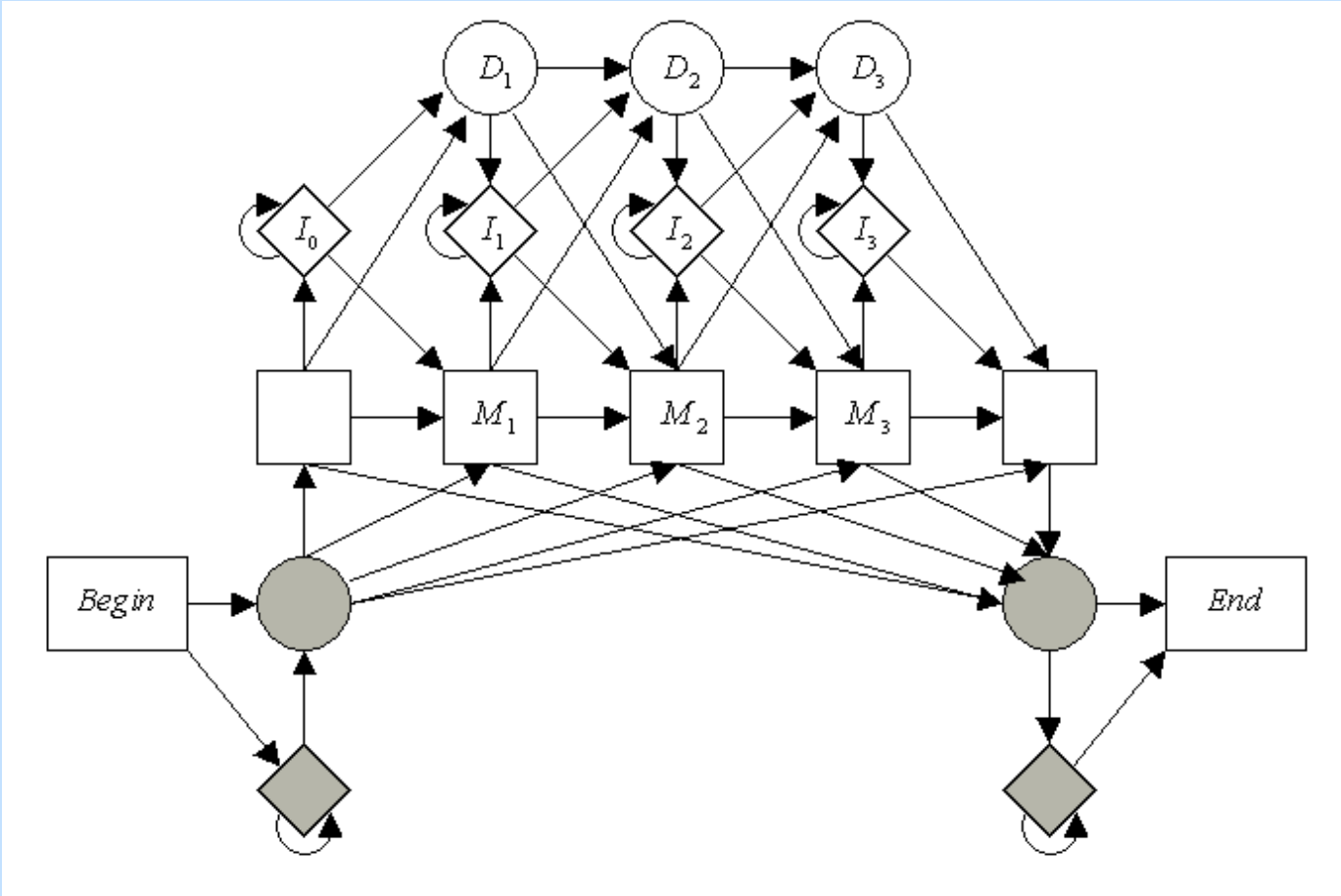
- Z-score: (S-E(S)) / sd(S)
- On ~25K proteins, cutoff 5 misses 2/628 globins with no fp

Discovering subfamilies



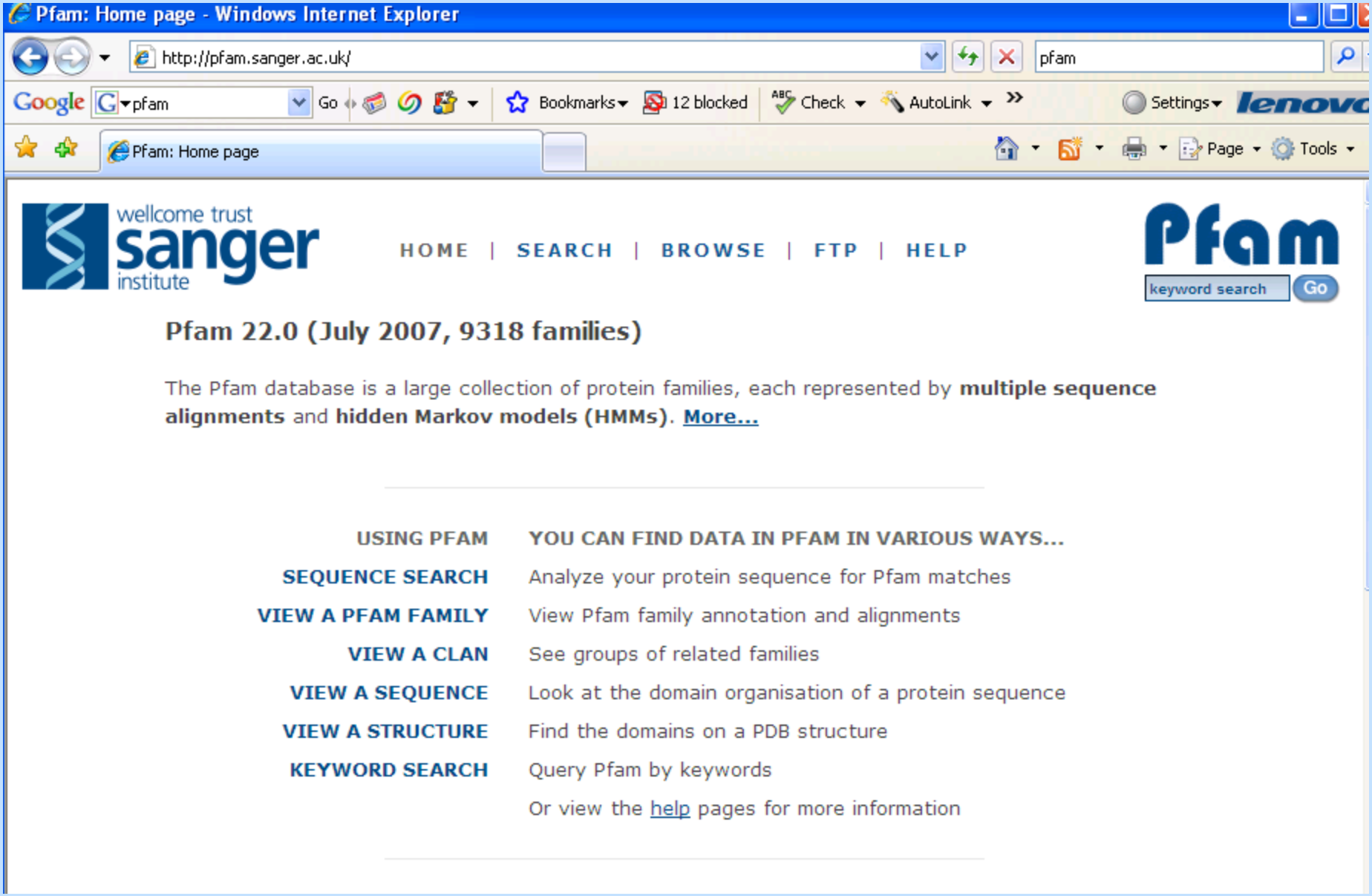
- 10 component HMM. Training set: 628 globins. Classified each sequence using the model.
- Generated 7 nonempty clusters, with ~560 falling into 4 "biologically meaningful" clusters.

Local alignment in HMM



Add flanking states for regions of unaligned seq.

Durbin et al. Ch. 5.5


Pfam: Home page - Windows Internet Explorer
http://pfam.sanger.ac.uk/
pfam
Google
pfam
Go
Bookmarks
12 blocked
Check
AutoLink
Settings
Pfam: Home page
Page
Tools
wellcome trust sanger institute
HOME | SEARCH | BROWSE | FTP | HELP
Pfam
keyword search
Go
Pfam 22.0 (July 2007, 9318 families)
The Pfam database is a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). More...
USING PFAM
YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...
SEQUENCE SEARCH
Analyze your protein sequence for Pfam matches
VIEW A PFAM FAMILY
View Pfam family annotation and alignments
VIEW A CLAN
See groups of related families
VIEW A SEQUENCE
Look at the domain organisation of a protein sequence
VIEW A STRUCTURE
Find the domains on a PDB structure
KEYWORD SEARCH
Query Pfam by keywords
Or view the help pages for more information