# Applications of Hidden Markov Models

Lecture 7.4

*by Marina Barsky*

# HMM applications

- Robot planning + sensing when there’s uncertainty
- Speech Recognition/Understanding
- Consumer decision modeling
- Economics & Finance
- Human Genome Project
- …

# Classic example: Speech recognition

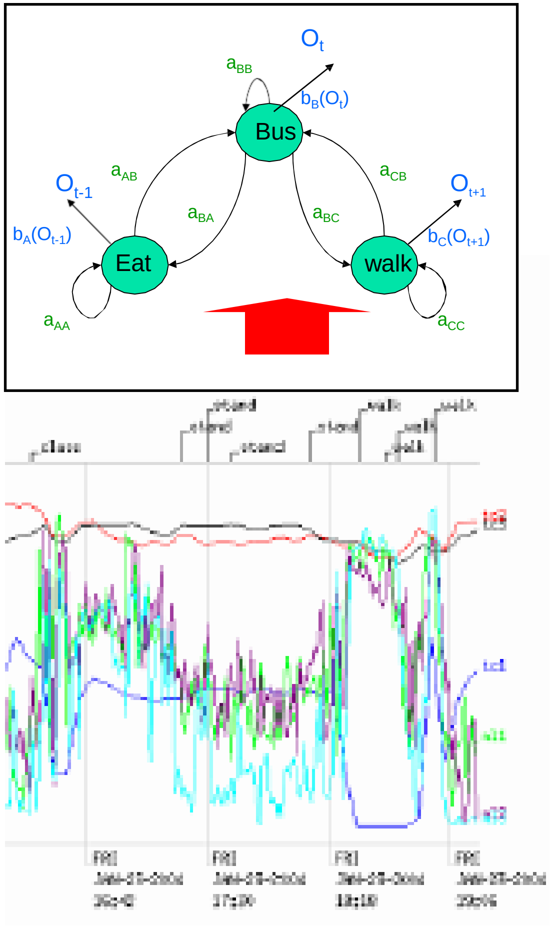
- Signal → words
  - Observable is signal
  - Hidden state is part of word

- Formulation:
  - What is the most probable word given this signal?

**UTTERLY GROSS SIMPLIFICATION**

In practice: many levels of inference – not only HMM

# Example: Human daily activities recognition from wearable sensor signals



# Bio-sequence application: Gene finding

# CpG islands

- C nucleotide followed by G is easily methylated
- Methylated C easily becomes T
- The methylation is suppressed in important regulatory regions – around promoters (starting sites of transcription)
- Thus, an overall low frequency of C->G di-nucleotide is significantly increased in the gene promoter regions

# Biological questions

- Given a short stretch of DNA sequence, determine whether it came from a CpG island or not
- Given a long un-annotated DNA sequence, find CpG islands in it

# Transition probability estimation: from real DNA sequences

From 48 **known** CpG islands of a total length 60,000 nucleotides, and from regular DNA stretches:

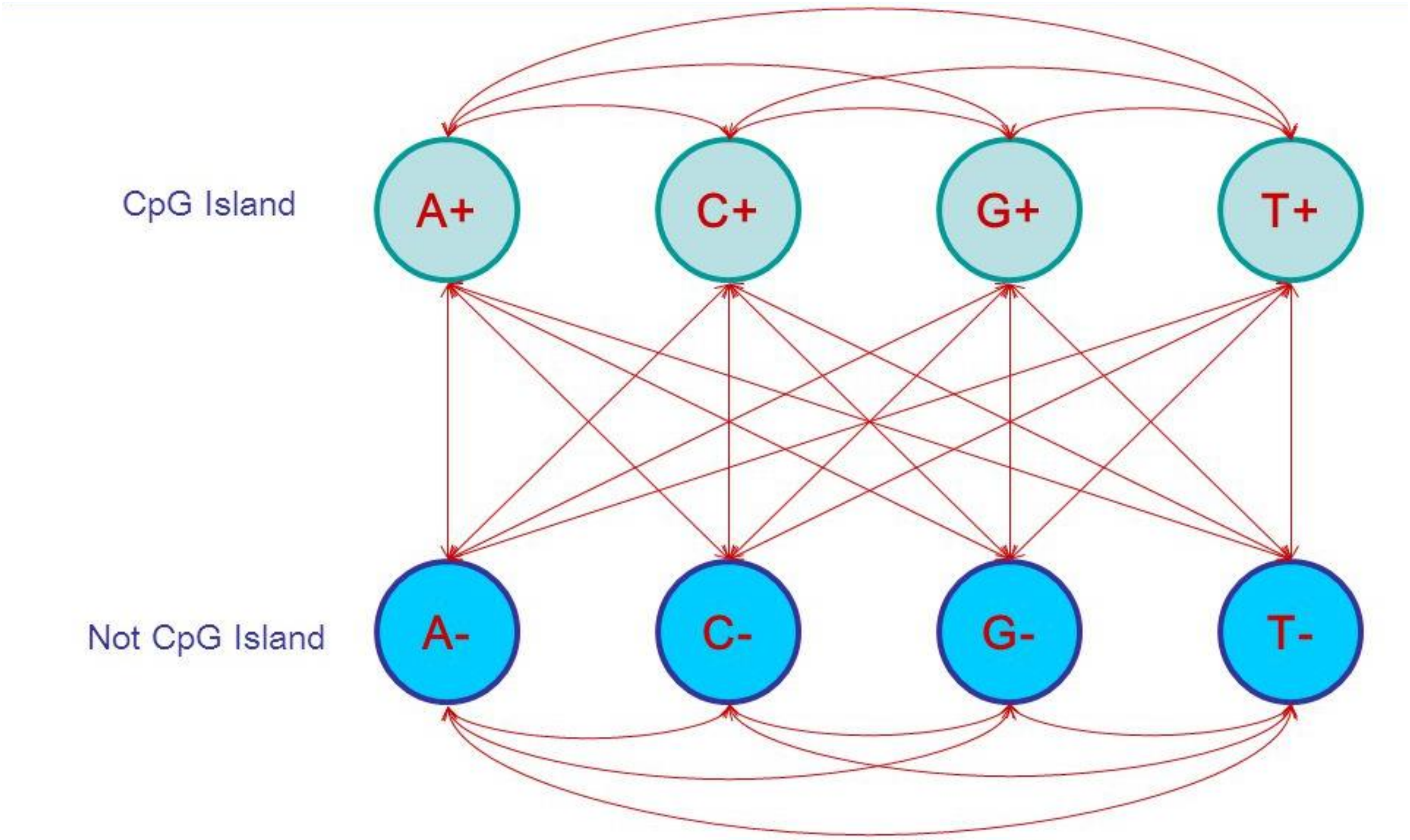
the transition probabilities for each pair of nucleotides were estimated (expected 0.25 if at random)

| + | A | C | G | T |
|---|---|---|---|---|
| A | 0.18 | 0.27 | 0.43 | 0.12 |
| C | 0.17 | 0.37 | 0.27 | 0.19 |
| G | 0.16 | 0.34 | 0.38 | 0.12 |
| T | 0.08 | 0.36 | 0.38 | 0.18 |

| - | A | C | G | T |
|---|---|---|---|---|
| A | 0.30 | 0.20 | 0.29 | 0.21 |
| C | 0.32 | 0.30 | 0.08 | 0.30 |
| G | 0.25 | 0.25 | 0.30 | 0.20 |
| T | 0.18 | 0.24 | 0.29 | 0.29 |

$$a_{from,to}=count_{from,to}/\Sigma_x count_{from,x}$$

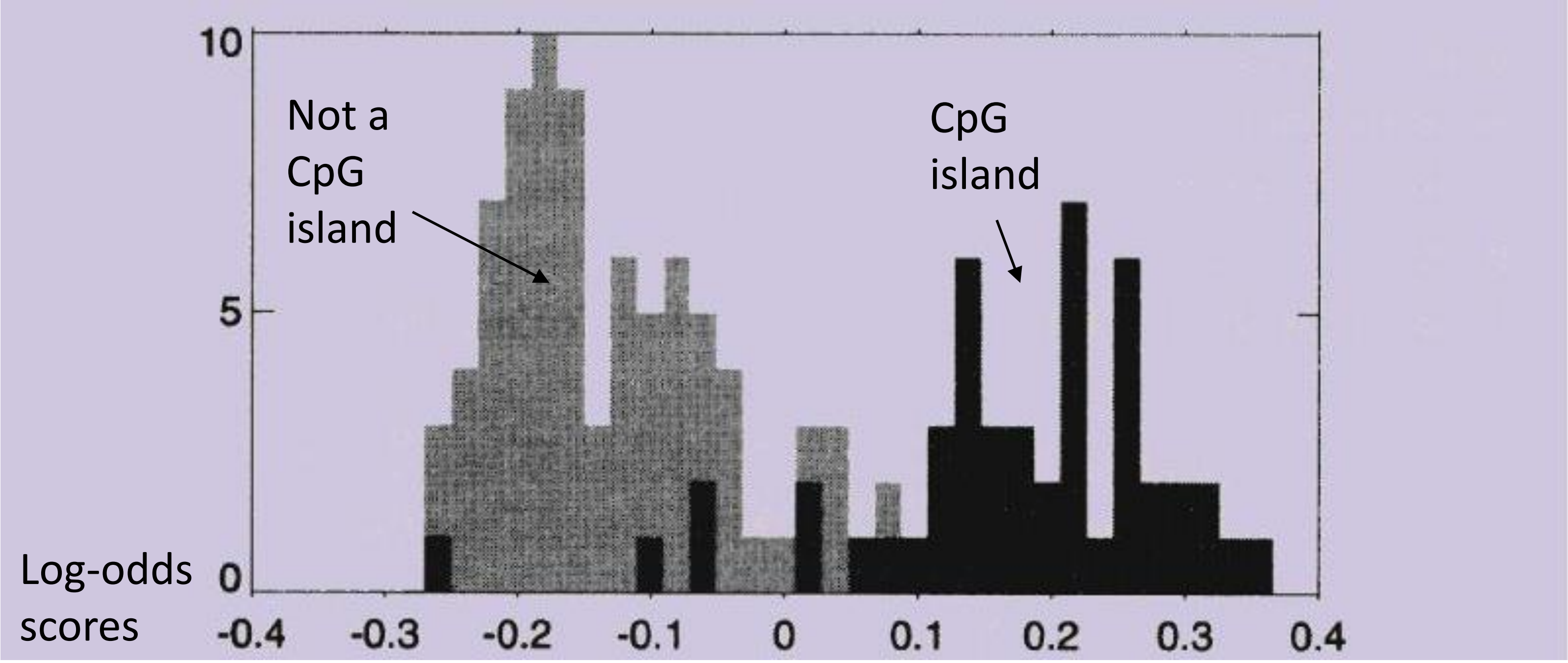
# Markov model for DNA sequence



# Am I inside a CpG island?

To use these (+) and (-) models for discrimination for a given sequence we  calculate the log-odds ratio:

- **Score(M)=log [ P(M|given model +)/P(M|given model -)]**
- If this value is positive, we are in the CpG island, if not, we are not


Not a CpG island

CpG island

Model efficiency: results of tests on another set of labeled DNA sequences

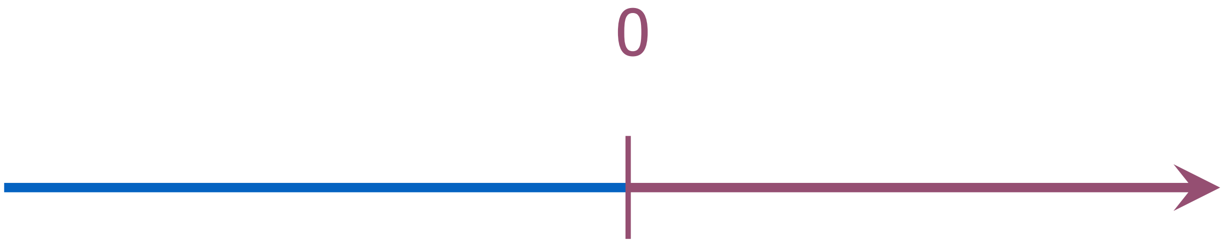
# Finding CpG islands - HMM

- HMM: essential difference from a simple Markov chain is that there is no one-to-one correspondence between the states and the symbols
- By looking at a *single symbol*, there is no way to tell whether it  came from state C+ or C-

# Computing Log-odds Ratios in a sliding window

$$x_1x_2\boxed{x_3x_4x_5x_6x_7}x_8\ldots x_n$$

- Consider a *sliding window* of the outcome sequence
- Find the log-odds for this short window



*Biased* die most likely used (CpG island)

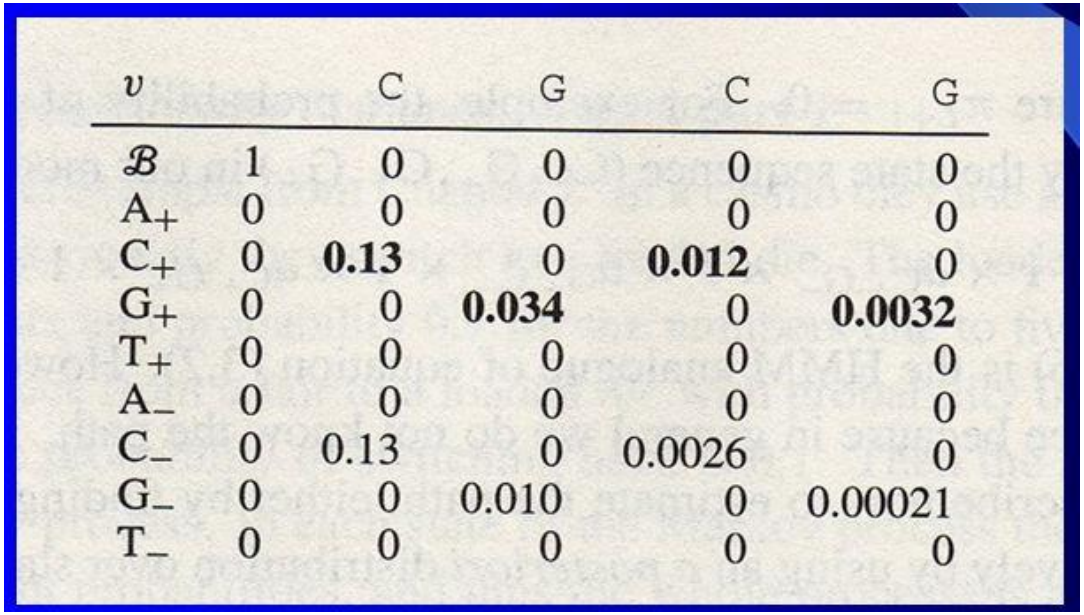
*Fair* die most likely used (non CpG island)

Disadvantages:

- the length of CpG-island is not known in advance
- different windows may classify the same position differently

# The most probable path through the sequence of states

## The most probable path for sequence ***CGCG***



| $v$ | | C | G | C | G |
|---|---|---|---|---|---|
| $\mathcal{B}$ | 1 | 0 | 0 | 0 | 0 |
| $A_+$ | 0 | 0 | 0 | 0 | 0 |
| $C_+$ | 0 | **0.13** | 0 | **0.012** | 0 |
| $G_+$ | 0 | 0 | **0.034** | 0 | **0.0032** |
| $T_+$ | 0 | 0 | 0 | 0 | 0 |
| $A_-$ | 0 | 0 | 0 | 0 | 0 |
| $C_-$ | 0 | 0.13 | 0 | 0.0026 | 0 |
| $G_-$ | 0 | 0 | 0.010 | 0 | 0.00021 |
| $T_-$ | 0 | 0 | 0 | 0 | 0 |

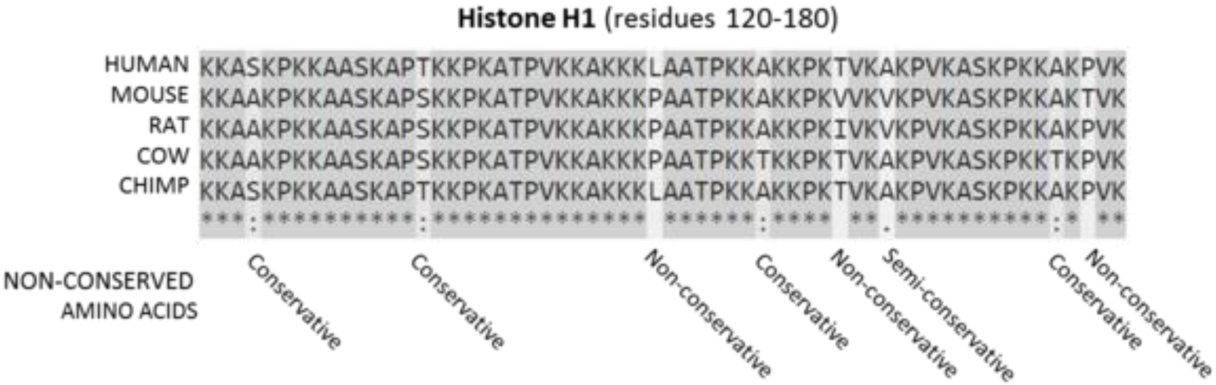
When we apply the Viterbi algorithm to a long un-annotated DNA sequence, the states will switch between + and -, giving suggested boundaries for CpG islands

# Bio-sequence application: Aligning a given sequence to a family of sequences

Profile HMM

# Multiple Alignments and Protein family classification

- Multiple alignment of a protein family shows variations in conservation along the length of a protein
- Example: after aligning many globin proteins, the biologists recognized that the helices region in globins are more conserved than others.



**Histone H1** (residues 120-180)

```
HUMAN KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVKPVKASKPKKAKTVK
  RAT KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVKPVKASKPKKAKPVK
  COW KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
CHIMP KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
      ***:**********:************** ******:**** **.**********:* **
```

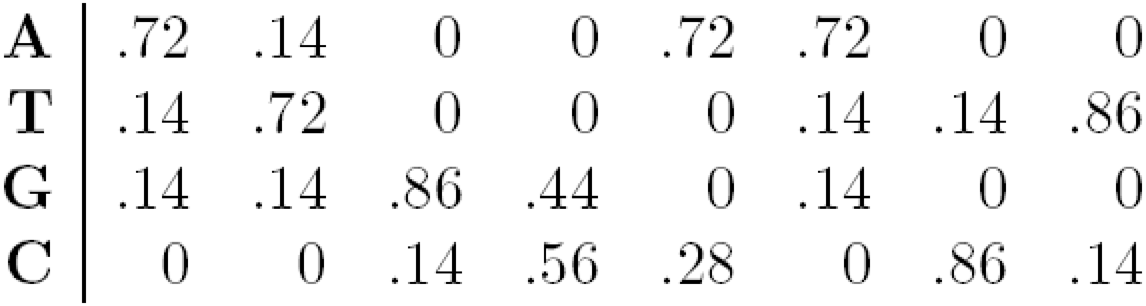
# Finding distant members of a Protein family

- A distant cousin of functionally related sequences in a protein family may have weak pairwise similarities with each member of the family and thus fail significance test
- However, they may have weak similarities with many members of the family

- The goal is to align a sequence to all members of the family at once.
- Family of related proteins can be represented by their multiple alignment and the corresponding profile.

# Profile representation of Protein families

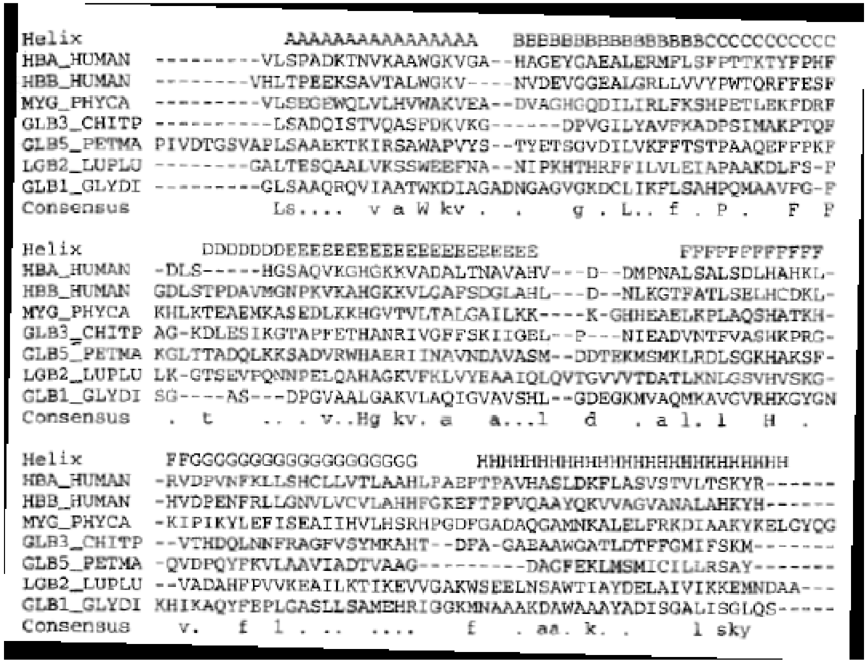
For example, aligned **DNA sequences** can be represented by a
*4 ·n* profile matrix reflecting the frequencies
of nucleotides in every aligned position.



| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **A** | .72 | .14 | 0 | 0 | .72 | .72 | 0 | 0 |
| **T** | .14 | .72 | 0 | 0 | 0 | .14 | .14 | .86 |
| **G** | .14 | .14 | .86 | .44 | 0 | .14 | 0 | 0 |
| **C** | 0 | 0 | .14 | .56 | .28 | 0 | .86 | .14 |

Protein family can be represented by a *20·n* profile representing frequencies of amino acids.

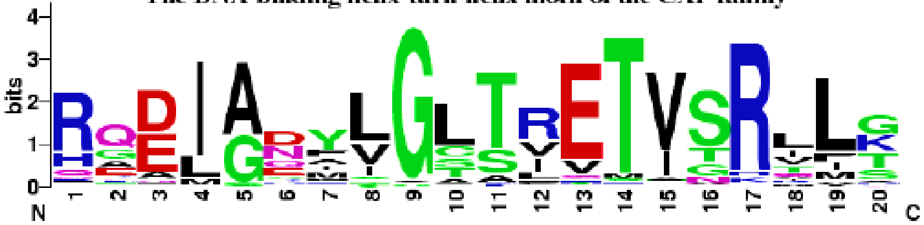
# Multiple alignment and symbol probabilities



```
Helix                 AAAAAAAAAAAAAAAA    BBBBBBBBBBBBBBBBCCCCCCCCCCC
HBA_HUMAN   ---------VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF
HBB_HUMAN   --------VHLTPEEKSAVTALWGKV----NVDEVGGEALGRLLVVYPWTQRFFESF
MYG_PHYCA   ---------VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRF
GLB3_CHITP  ----------LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKPTQF
GLB5_PETMA  PIVDTGSVAPLSAAEKTKIRSAWAPVYS--TYETSGVDILVKFFTSTPAAQEFFPKF
LGB2_LUPLU  --------GALTESQAALVKSSWEEFNA--NIPKHTHRFFILVLEIAPAAKDLFS-F
GLB1_GLYDI  ---------GLSAAQRQVIAATWKDIAGADNGAGVGKDCLIKFLSAHPQMAAVFG-F
Consensus             Ls....  v a W kv .  .    g . L.. f . P .   F  F

Helix           DDDDDDDEEEEEEEEEEEEEEEEEEEEEE               FFFFFFFFFFFF
HBA_HUMAN   -DLS-----HGSAQVKGHGKKVADALTNAVAHV---D--DMPNALSALSDLHAHKL-
HBB_HUMAN   GDLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHL---D--NLKGTFATLSELHCDKL-
MYG_PHYCA   KHLKTEAEMKASEDLKKHGVTVLTALGAILKK----K-GHHEAELKPLAQSHATKH-
GLB3_CHITP  AG-KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASHKPRG-
GLB5_PETMA  KGLTTADQLKKSADVRWHAERIINAVNDAVASM--DDTEKMSMKLRDLSGKHAKSF-
LGB2_LUPLU  LK-GTSEVPQNNPELQAHAGKVFKLVYEAAIQLQVTGVVVTDATLKNLGSVHVSKG-
GLB1_GLYDI  SG----AS---DPGVAALGAKVLAQIGVAVSHL--GDEGKMVAQMKAVGVRHKGYGN
Consensus    . t   ... v..Hg kv. a   a...l   d   . a l. l   H  .

Helix        FFGGGGGGGGGGGGGGGGGG      HHHHHHHHHHHHHHHHHHHHHHHHHHH
HBA_HUMAN   -RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR------
HBB_HUMAN   -HVDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH------
MYG_PHYCA   -KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKDIAAKYKELGYQG
GLB3_CHITP  --VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGMIFSKM-------
GLB5_PETMA  -QVDPQYFKVLAAVIADTVAAG---------DAGFEKLMSMICILLRSAY-------
LGB2_LUPLU  --VADAHFPVVKEAILKTIKEVVGAKWSEELNSAWTIAYDELAIVIKKEMNDAA---
GLB1_GLYDI  KHIKAQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGALISGLQS-----
Consensus     v.    f  l . ..  ....    f   . aa. k. .     l sky
```

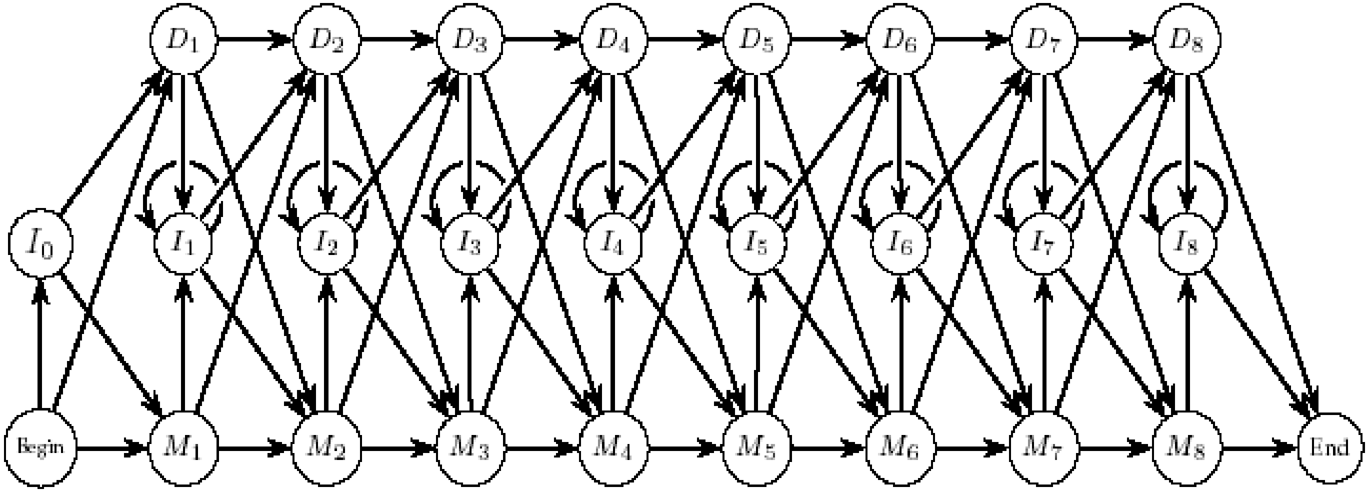


The DNA-binding helix-turn-helix motif of the CAP family

# What are Profile HMMs?

- A Profile HMM is a probabilistic representation of a multiple alignment
- A given multiple alignment (of a protein family) is used to build a profile HMM
- This model then may be used to find and score less obvious potential matches of new protein sequences

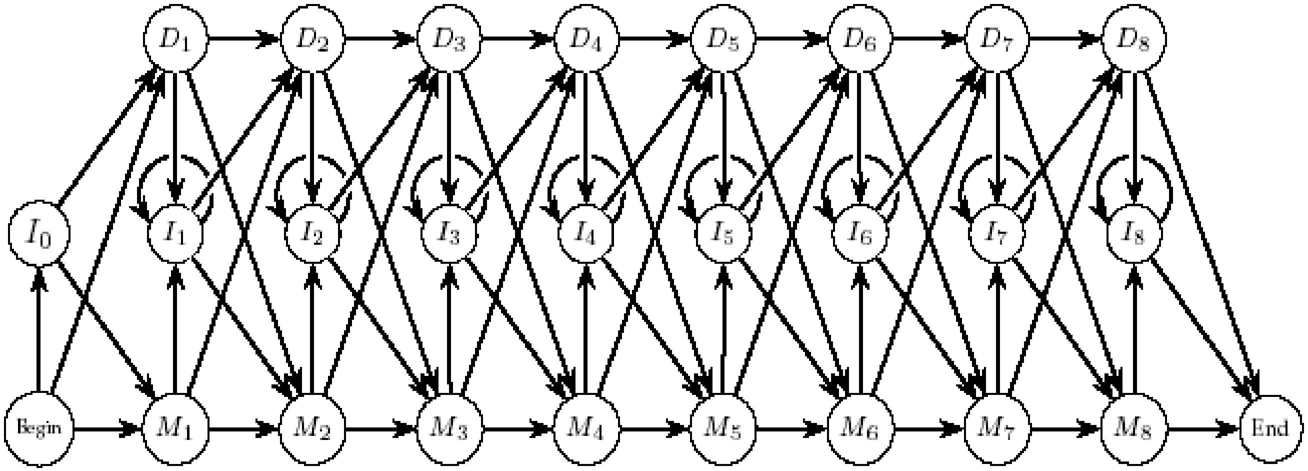
# Building a profile HMM



- Assign each column (sequence position) to a *Match* state in HMM. Add I*nsertion* and *Deletion* state.
- Estimate the emission probabilities according to amino acid counts in column from the multiple alignment. Different positions in the protein will have different emission probabilities.
- Estimate the transition probabilities between *Match, Deletion* and *Insertion* states
- The HMM model gets **trained** to derive the optimal parameters

# States of Profile HMM

- Match states $M_1 \ldots M_n$ (plus *begin/end* states)
- Insertion states $I_0 I_1 \ldots I_n$
- Deletion states $D_1 \ldots D_n$



# Aligning new sequence to a profile

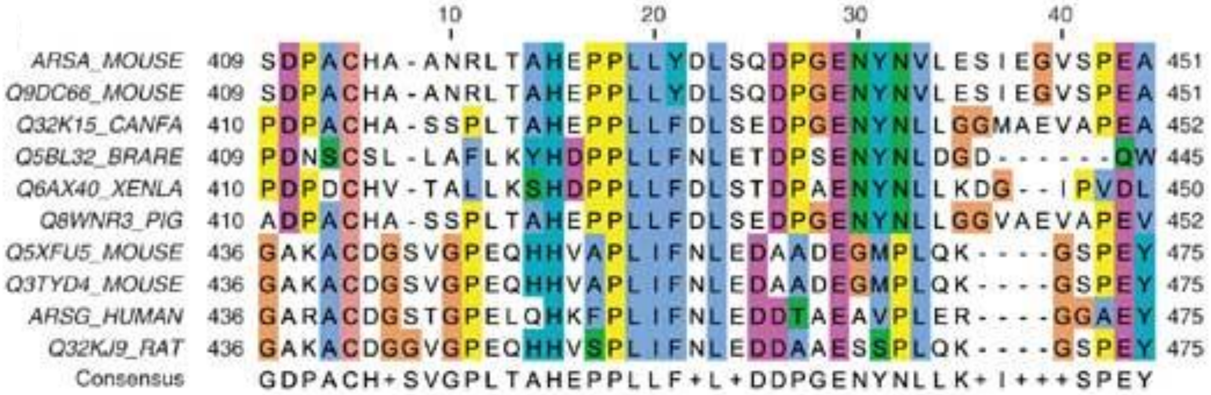
- HMMs can be used for aligning a sequence against a profile representing protein family
- A $20 \cdot n$ profile $P$ corresponds to $n$ sequentially linked *match* states $M_1,\ldots,M_n$ in the profile HMM of $P$

# Emission Probabilities in Profile HMM

- Probability of emitting a symbol *a* at an insertion state $I_j$:

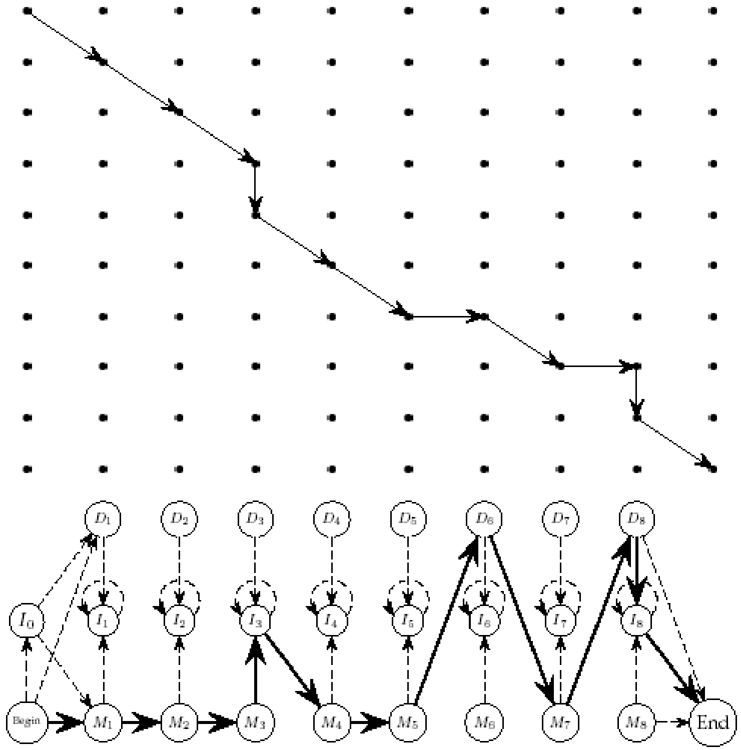
$$e_{I_j}(a) = p(a)$$

where *p(a)* is the frequency of the occurrence of the symbol *a* in all the sequences.



```
                               10        20        30        40
                                |         |         |         |
ARSA_MOUSE    409  SDPACHA-ANRLTAHEPPLLYDLSQDPGENYNVLESIEGVSPEA  451
Q9DC66_MOUSE  409  SDPACHA-ANRLTAHEPPLLYDLSQDPGENYNVLESIEGVSPEA  451
Q32K15_CANFA  410  PDPACHA-SSPLTAHEPPLLFDLSEDPGENYNLLGGMAEVAPEA  452
Q5BL32_BRARE  409  PDNSCSL-LAFLKYHDPPLLFNLETDPSENYNLDGD------QW  445
Q6AX40_XENLA  410  PDPDCHV-TALLKSHDPPLLFDLSTDPAENYNLLKDG--IPVDL  450
Q8WNR3_PIG    410  ADPACHA-SSPLTAHEPPLLFDLSEDPGENYNLLGGVAEVAPEV  452
Q5XFU5_MOUSE  436  GAKACDGSVGPEQHHVAPLIFNLEDAADEGMPLQK----GSPEY  475
Q3TYD4_MOUSE  436  GAKACDGSVGPEQHHVAPLIFNLEDAADEGMPLQK----GSPEY  475
ARSG_HUMAN    436  GARACDGSTGPELQHKFPLIFNLEDDTAEAVPLER----GGAEY  475
Q32KJ9_RAT    436  GAKACDGGVGPEQHHVSPLIFNLEDDAAESSPLQK----GSPEY  475
Consensus          GDPACH+SVGPLTAHEPPLLF+L+DDPGENYNLLK+I+++SPEY
```

# Paths in Edit Graph and Profile HMM



A path through an edit graph and the corresponding path through a profile HMM

# Most used tool: *PFAM*

- PFAM describes ***protein domains***
- Each protein domain family in Pfam has:
  - *Seed alignment*: manually verified multiple alignment of a representative set of sequences.
  - *HMM* built from the seed alignment for further database searches.
  - *Full alignment* generated automatically from the  HMM

- The distinction between seed and full alignments facilitates Pfam updates.
  - Seed alignments are stable resources.
  - HMM profiles and full alignments can be updated with newly found amino acid sequences