CISC 636 Computational Biology & Bioinformatics (Fall 2016)

Hidden Markov Models (IV)

a. Profile HMMs
b. ipHMMs
c. GeneScan
d. TMMOD
e. ipHMM and ETB-viterbi

Profile HMM for a family of sequences

Applications of HMM's

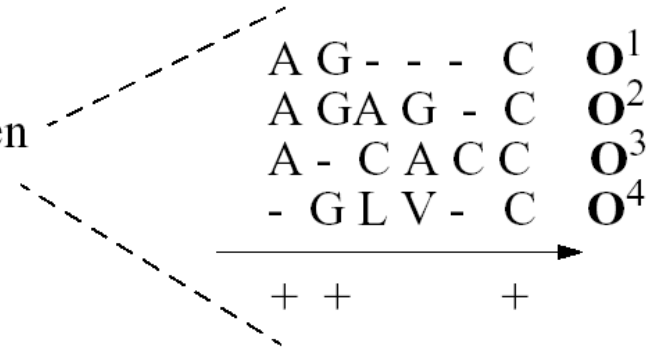
- Given a family of sequences, $\mathbf{O}^l = O^l_1 \ldots O^l_{K^l}$, build a hidden Markov model that best fits to this family-->Problem 3
  - Correct multiple alignment is given--> Problem 3, path known
    - MA built from structural information
    - MA obtained from other sequence based alignment procedures
  - Alignment is not assumed--> Problem 3, path not known (B-W)
- Use the obtained model to:
  - Score potential matches of new sequences-->Problem 1
  - Align new sequences--> Problem 2

Javier Garcia-Frias

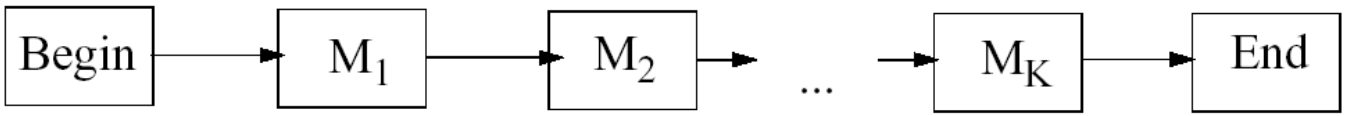
Profile HMM: Correct alignment assumed

HMM construction

Example: Assume MA given
(columns marked with +)



- Segments of family where an alignment exists are produced by MATCH STATES



- Generation probabilities are position dependent!
- In previous example, K=3

Javier Garcia-Frias

Profile HMM: Correct alignment assumed

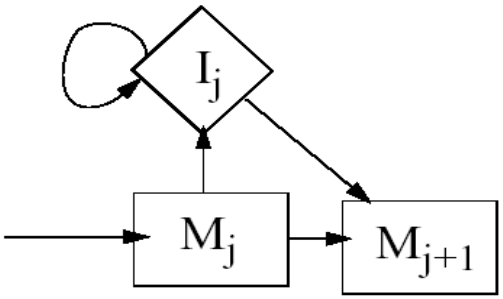
- Handling insertions: Portion of the sequences that are not aligned ---> Add INSERT STATES

Example: Assume MA given (columns marked with +)

A	G	-	-	-	C		$\mathbf{O}^1$
A	G	A	G	-	C		$\mathbf{O}^2$
A	-	C	A	C	C		$\mathbf{O}^3$
-	G	L	V	-	C		$\mathbf{O}^4$
		+	+			+	

- To cope with all possibilities for insertions, an insert state should be added after each match state

State I_k inserts sequence just after match state M_k (i.e., aligned column k)



$\mathbf{O}^1$--> $M_1M_2M_3$

$\mathbf{O}^2$--> $M_1M_2I_2I_2M_3$

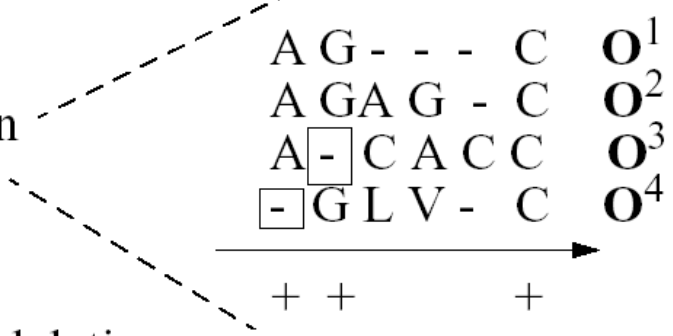
$\mathbf{O}^3$--> M_1? State M_2 is skipped

Javier Garcia-Frias

Profile HMM: Correct alignment assumed

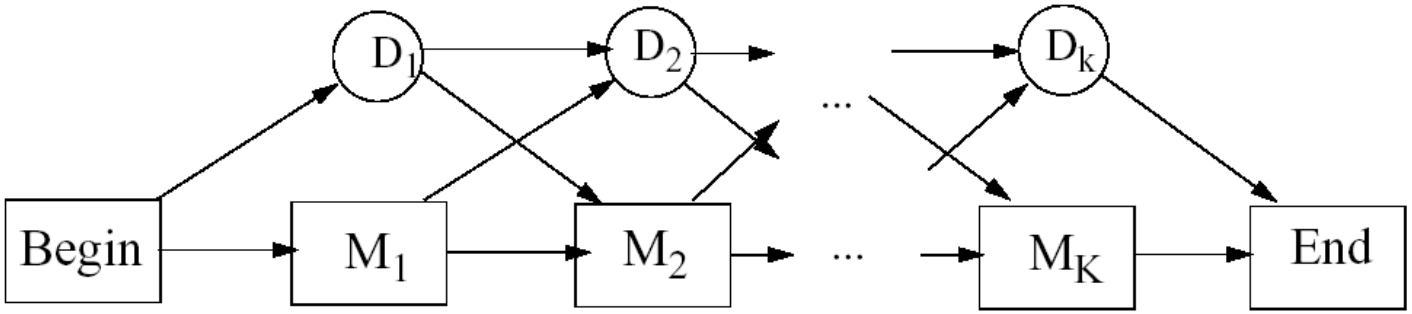
- Handling deletions: Portion of the sequences that "skips" the alignment---> Add SILENT (DELETE) STATES

Example: Assume MA given (columns marked with +)



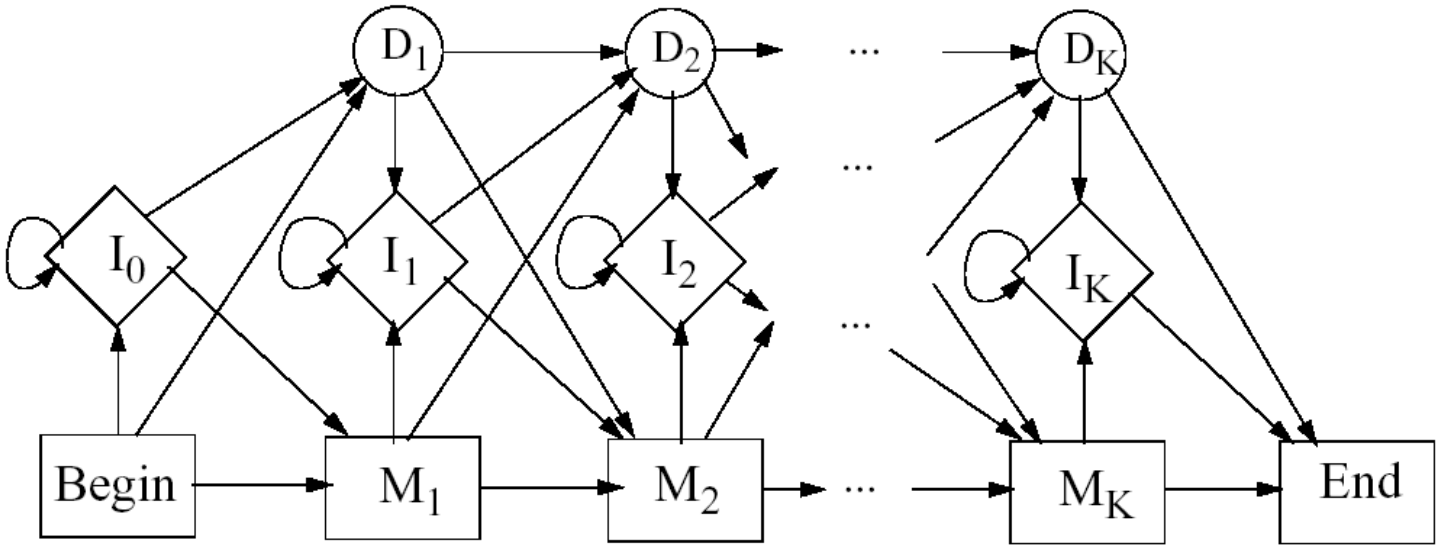
- To cope with all possibilities for deletions
  - Connect all possible match states (big complexity)
  - Add silent states (less complexity, but loss of generality)-->NO EMISSION

State D_k skips match state M_k (i.e., aligned column k)



Profile HMM: Correct alignment assumed

Resulting HMM (Profile HMM)



- Notice we have added transitions between insert and delete states

Example: Assume MA given (columns marked with +)

A G - - - C	$\mathbf{O}^1$	$M_1M_2M_3$
A GA G - C	$\mathbf{O}^2$	$M_1M_2I_2I_2M_3$
A - C A C C	$\mathbf{O}^3$	$M_1D_2I_2I_2I_2M_3$
- G L V - C	$\mathbf{O}^4$	$D_1M_2I_2I_2M_3$
+ + +		

Profile HMM: Correct alignment assumed

Key idea of profile HMM

- Transition and emission probabilities capture specific information about each position in the multiple alignment of the whole family
- Profile HMM=Statistical model representing the family

Questions

- How do we build the profile HMM that best fits to a given family? -->Problem 3 (simplified)
- How do we detect potential membership in this family (for new sequences)? --> Problem 1
- How do we align a new sequence? --> Problem 2

Javier Garcia-Frias

Parameterization of profile HMM's: Correct alignment assumed

Profile HMM parametrization (simplified Problem 3)

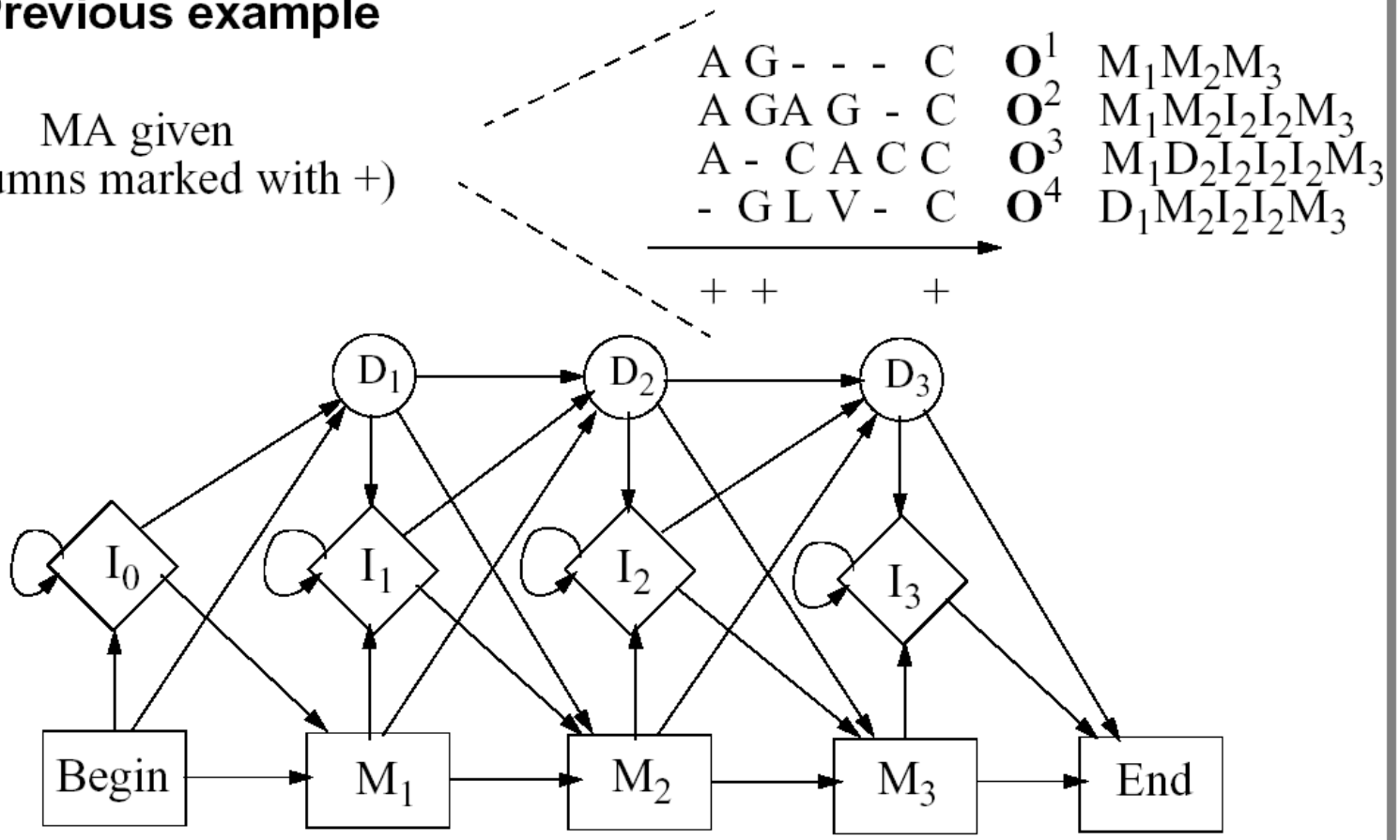
- **Model length**
  - Length (and structure) completely defined when we decide which MA columns should be assigned to match states
    - Manual construction
    - Heuristic construction: e.g., column aligned if proportion of gaps is less than a threshold
    - More sophisticated methods
- **Parameter estimation**
  - Alignment is given-->Path through model is given for any sequence
  - Apply solution to Problem 3 when path is given (just count events)

Javier Garcia-Frias

Parameterization of profile HMM's: Correct alignment assumed

Previous example

MA given
(columns marked with +)



Javier Garcia-Frias

Parameterization of profile HMM's: Correct alignment assumed

Emission probabilities: Estimate from number of emissions

$N(A\|M_1)=3$	$N(other\|M_1)=0$
$N(A\|M_2)=3$	$N(other\|M_2)=0$
$N(C\|M_3)=4$	$N(other\|M_3)=0$

I_0, I_1, I_3 are not used

$N(A|I_2)=2$ $N(C|I_2)=2$ $N(G|I_2)=1$

$N(L|I_2)=1$ $N(V|I_2)=1$ $N(other|I_2)=0$

Transition probabilities: Estimate from number of transitions

$N(M_1\|B)=3$	$N(D_1\|B)=1$
$N(M_2\|M_1)=3$	$N(D_2\|M_1)=1$
$N(M_3\|M_2)=1$	$N(I_2\|M_2)=2$
$N(E\|M_3)=3$	

$N(I_2|D_2)=1$

$N(I_2|I_2)=4$ $N(M_3|I_2)=3$

- **If number of sequences is not high enough, estimation should be modified**

Javier Garcia-Frias

Membership in a profile HMM

Detection of potential membership, for a new sequence, in family defined by a profile HMM (Problem 1)

- Apply forward equation
- Since P(**O**|M) is length dependent, usually scoring function is modified

$$\text{Scoring} = \log \frac{P(\mathbf{O}|M)}{P(\mathbf{O}|S)}$$

S is called "standard model": Model to use if sequences were independently distributed

- Other statistical approaches can also be used to improve the scoring system

Multiple alignment using profile HMM's

No alignment is assumed

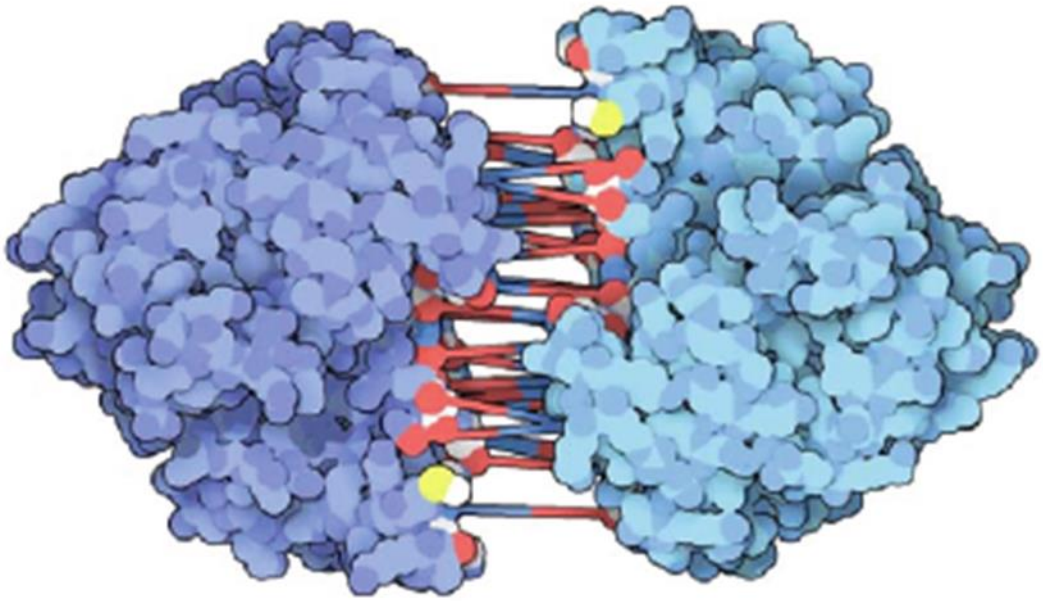
- From an initially unaligned family of sequences, jointly perform:
  - Profile HMM estimation
  - Alignment estimation

1. Initialization
- Choose length of profile HMM and initialize parameters

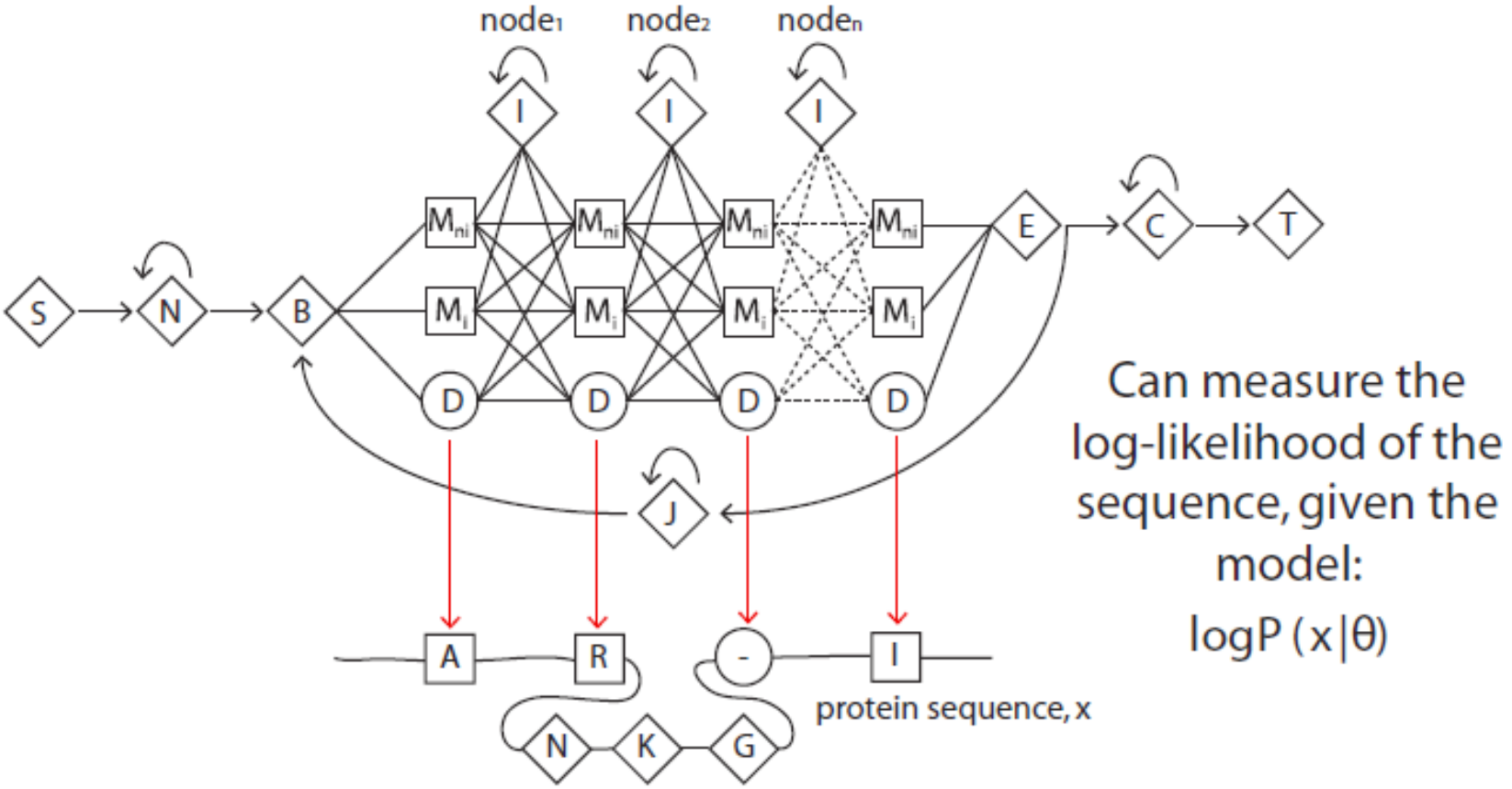
2. Training
- Estimate parameters of the profile HMM
- Path not known (no alignment)--> Problem 3 (Baum-Welch)

3. Alignment
- Align all sequences using Viterbi algorithm (Problem 2)

Javier Garcia-Frias



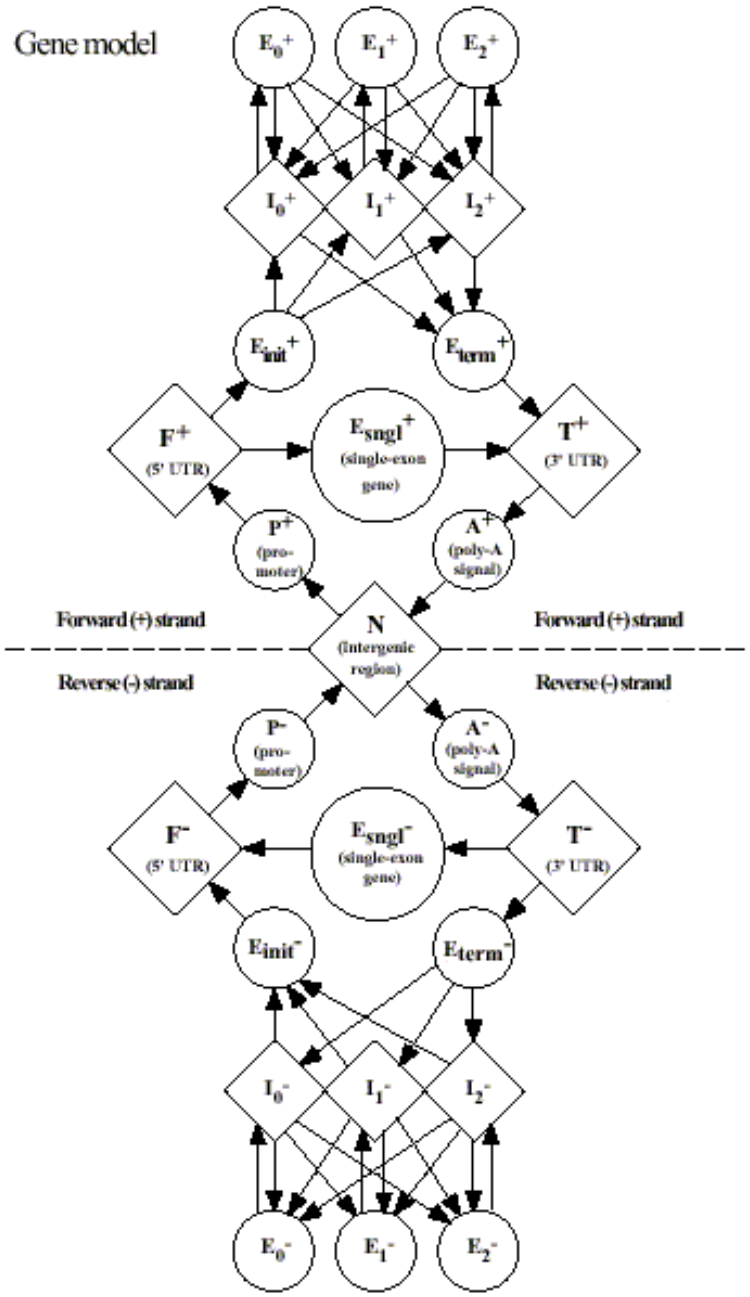
Interaction profile HMM (ipHMM)



Friedrich et al, *Bioinformatics* 2006

GENSCAN (generalized HMMs)

- Chris Burge, PhD Thesis ’97, Stanford
- http://genes.mit.edu/GENSCAN.html
- Four components
  - A vector π of initial probabilities
  - A matrix T of state transition probabilities
  - A set of length distribution f
  - A set of sequence generating models P
- Generalized HMMs:
  - at each state, emission is not symbols (or residues), rather, it is a fragment of sequence.
  - Modified viterbi algorithm


Gene model
E0+
E1+
E2+
I0+
I1+
I2+
Einit+
Eterm+
F+
(5' UTR)
Esngl+
(single-exon gene)
T+
(3' UTR)
P+
(pro-moter)
A+
(poly-A signal)
Forward (+) strand
N
(intergenic region)
Forward (+) strand
Reverse (-) strand
Reverse (-) strand
P-
(pro-moter)
A-
(poly-A signal)
F-
(5' UTR)
Esngl-
(single-exon gene)
T-
(3' UTR)
Einit-
Eterm-
I0-
I1-
I2-
E0-
E1-
E2-

- Initial state probabilities
  - As frequency for each functional unit to occur in actual genomic data. E.g., as ~ 80% portion are non-coding intergenic regions, the initial probability for state N is 0.80
- Transition probabilities
- State length distributions

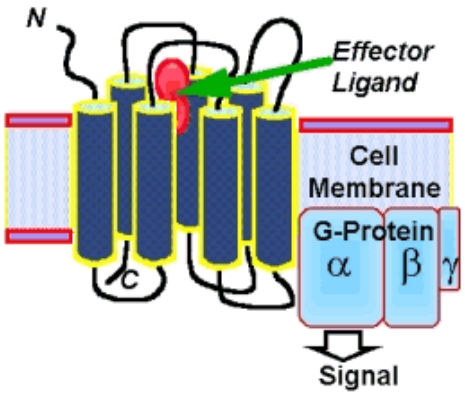
- Training data
  - 2.5 Mb human genomic sequences
  - 380 genes, 142 single-exon genes, 1492 exons and 1254 introns
  - 1619 cDNAs

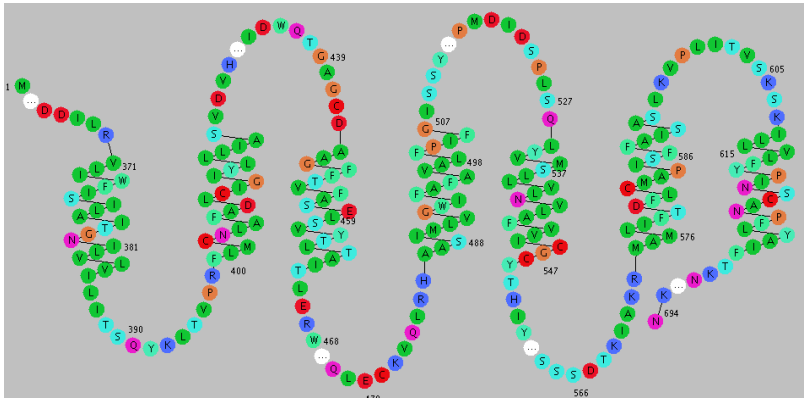
Open areas for research

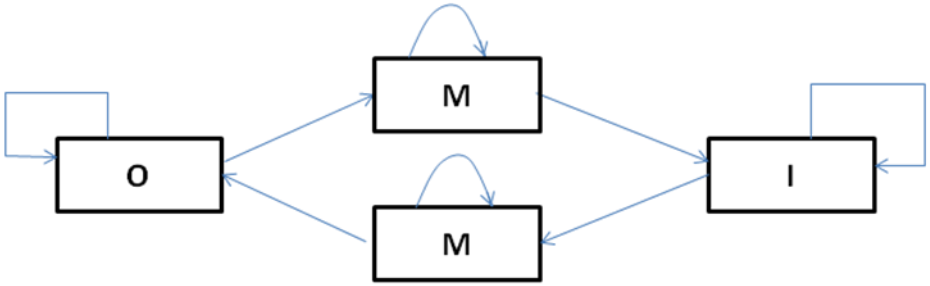
- Model building
  - Integration of domain knowledge, such as structural information, into profile HMMs
  - Meta learning?
- Biological mechanism

  DNA replication
- Hybrid models
  - Generalized HMM
  - …

TMMOD: An improved hidden Markov model for predicting transmembrane topology

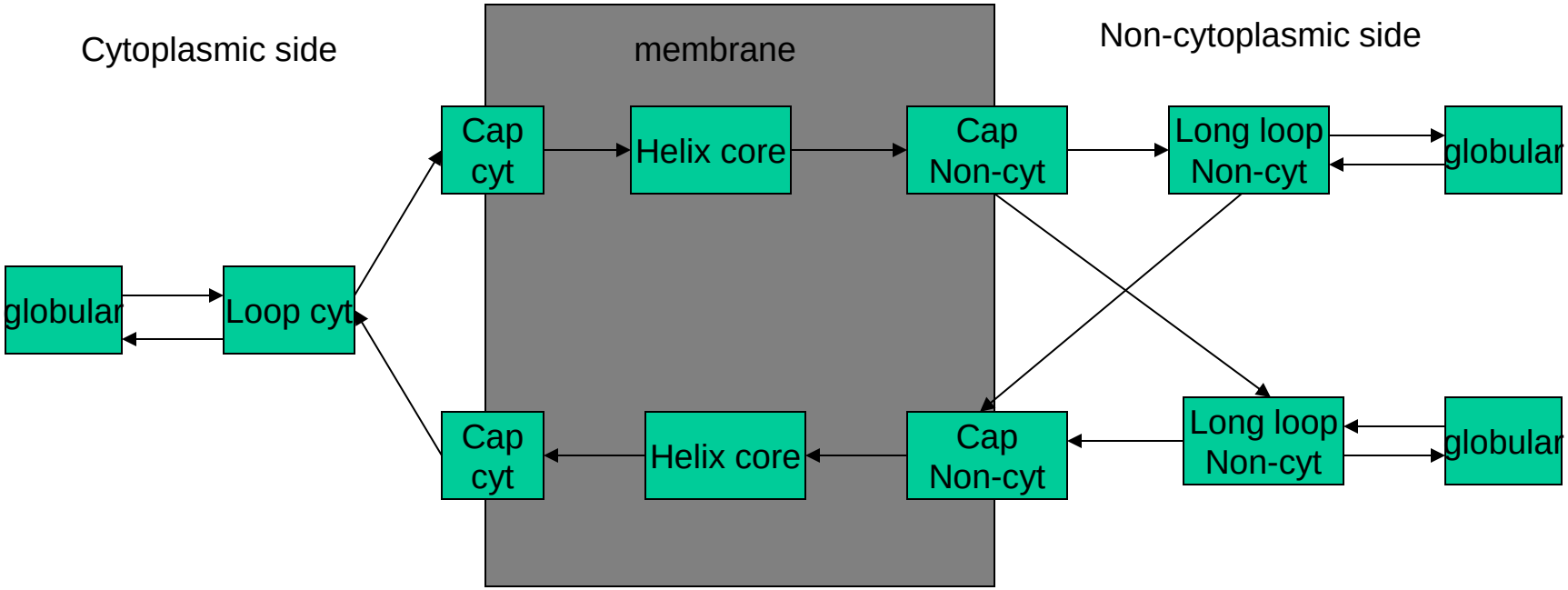




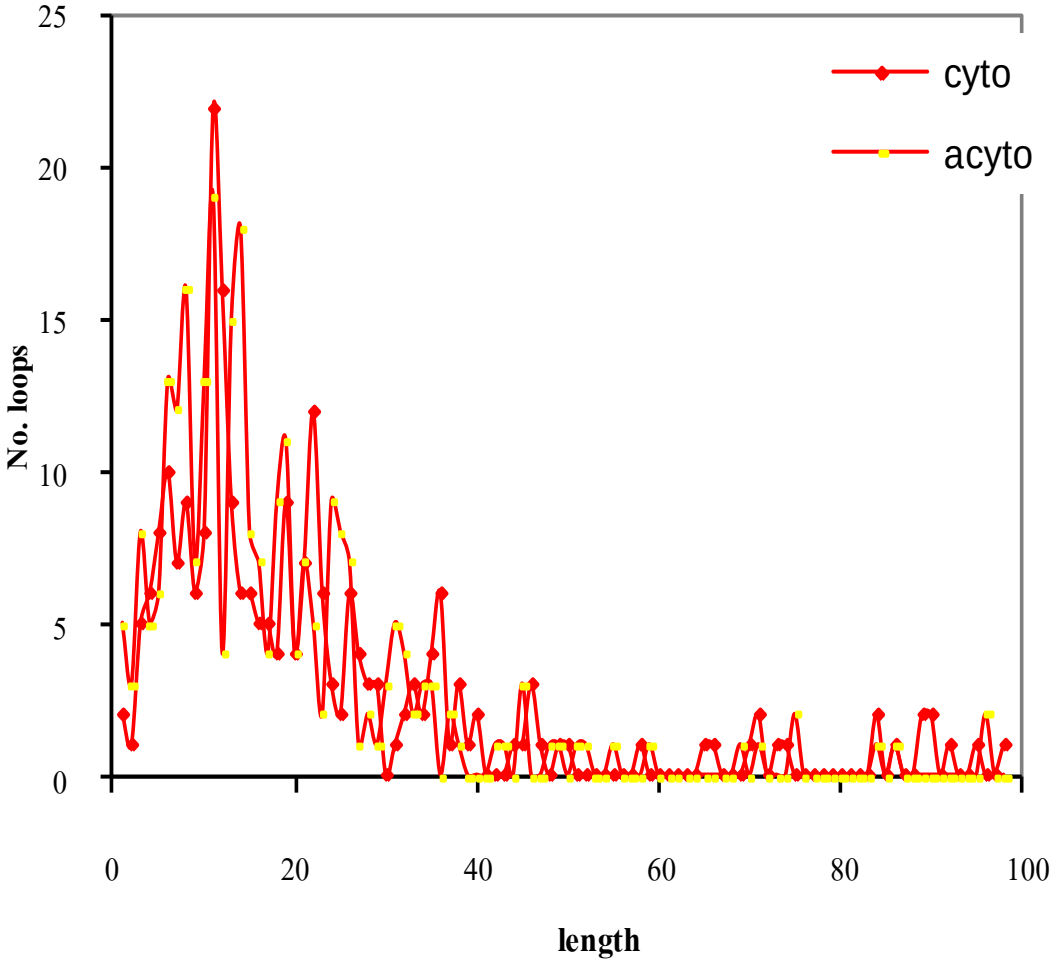


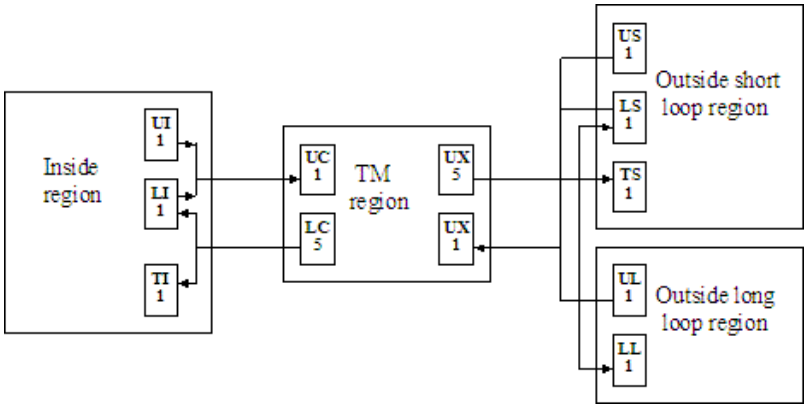

```
MFQLLAGVRMNSTGRPRAKIILLYALLIAFNIGAWLCALAAFRDHPVLLGT
iiiiiiiiiiiiiiiiiiiMMMMMMMMMMMMMMMMMMMMMMoooooooooo
```

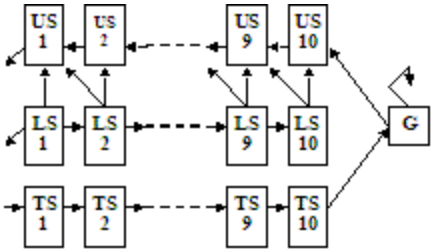
TMHMM by Krogh, A. *et al* *JMB* **305**(2001)567-580

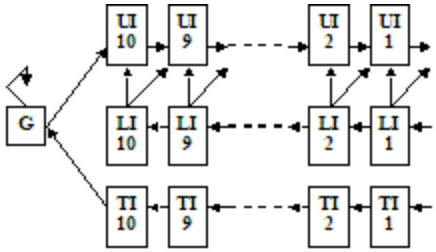


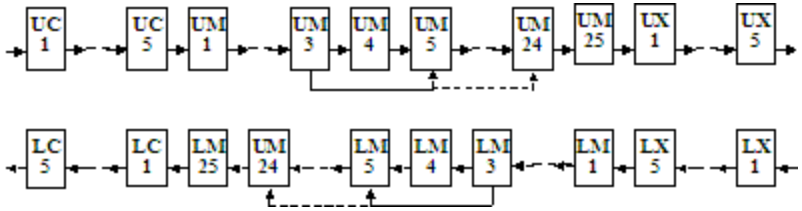
Accuracy of prediction for topology: 78%

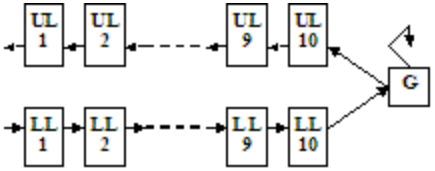

25
20
15
10
5
0
No. loops
0
20
40
60
80
100
length
cyto
acyto


Inside region
UI 1
LI 1
TI 1
UC 1
LC 5
TM region
UX 5
UX 1
US 1
LS 1
TS 1
Outside short loop region
UL 1
LL 1
Outside long loop region


US 1
US 2
US 9
US 10
LS 1
LS 2
LS 9
LS 10
TS 1
TS 2
TS 9
TS 10
G


G
UI 10
UI 9
UI 2
UI 1
LI 10
LI 9
LI 2
LI 1
TI 10
TI 9
TI 2
TI 1

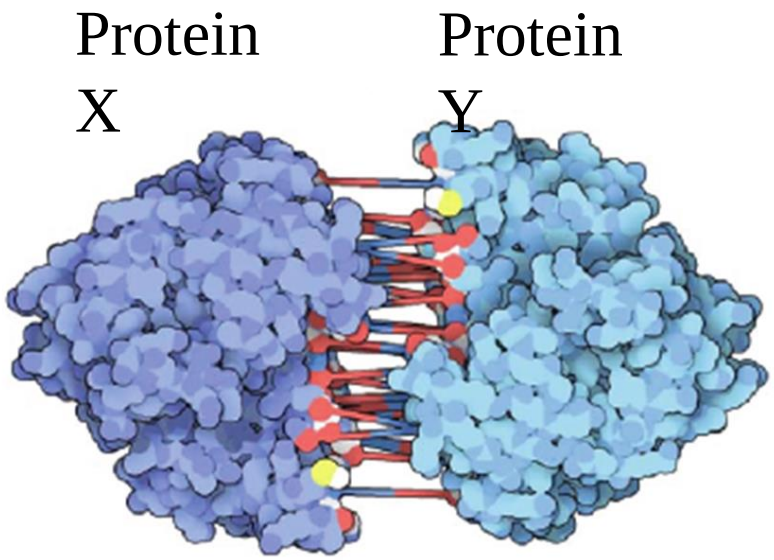

UC 1
UC 5
UM 1
UM 3
UM 4
UM 5
UM 24
UM 25
UX 1
UX 5
LC 5
LC 1
LM 25
UM 24
LM 5
LM 4
LM 3
LM 1
LX 5
LX 1


UL 1
UL 2
UL 9
UL 10
LL 1
LL 2
LL 9
LL 10
G

Mod.	Reg.	Data set	Correct topology	Correct location	Sens-itivity	Speci-ficity
TMMOD 1	(a)	S-83	65 (78.3%)	67 (80.7%)	97.4%	97.4%
	(b)		51 (61.4%)	52 (62.7%)	71.3%	71.3%
	(c)		64 (77.1%)	65 (78.3%)	97.1%	97.1%
TMMOD 2	(a)	S-83	61 (73.5%)	65 (78.3%)	99.4%	97.4%
	(b)		54 (65.1%)	61 (73.5%)	93.8%	71.3%
	(c)		54 (65.1%)	66 (79.5%)	99.7%	97.1%
TMMOD 3	(a)	S-83	70 (84.3%)	71 (85.5%)	98.2%	97.4%
	(b)		64 (77.1%)	65 (78.3%)	95.3%	71.3%
	(c)		**74 (89.2%)**	**74 (89.2%)**	**99.1%**	**97.1%**
TMHMM		S-83	**64 (77.1%)**	**69 (83.1%)**	**96.2%**	**96.2%**
PHDtm		S-83	**(85.5%)**	**(88.0%)**	**98.8%**	**95.2%**
TMMOD 1	(a)	S-160	117 (73.1%)	128 (80.0%)	97.4%	97.0%
	(b)		92 (57.5%)	103 (64.4%)	77.4%	80.8%
	(c)		117 (73.1%)	126 (78.8%)	96.1%	96.7%
TMMOD 2	(a)	S-160	120 (75.0%)	132 (82.5%)	98.4%	97.2%
	(b)		97 (60.6%)	121 (75.6%)	97.7%	95.6%
	(c)		118 (73.8%)	135 (84.4%)	98.4%	97.2%
TMMOD 3	(a)	S-160	120 (75.0%)	133 (83.1%)	97.8%	97.6%
	(b)		110 (68.8%)	124 (77.5%)	94.5%	98.1%
	(c)		**135 (84.4%)**	**143 (89.4%)**	**98.3%**	**98.1%**
TMHMM		S-160	**123 (76.9%)**	**134 (83.8%)**	**97.1%**	**97.7%**

Proteins Interact via Domains

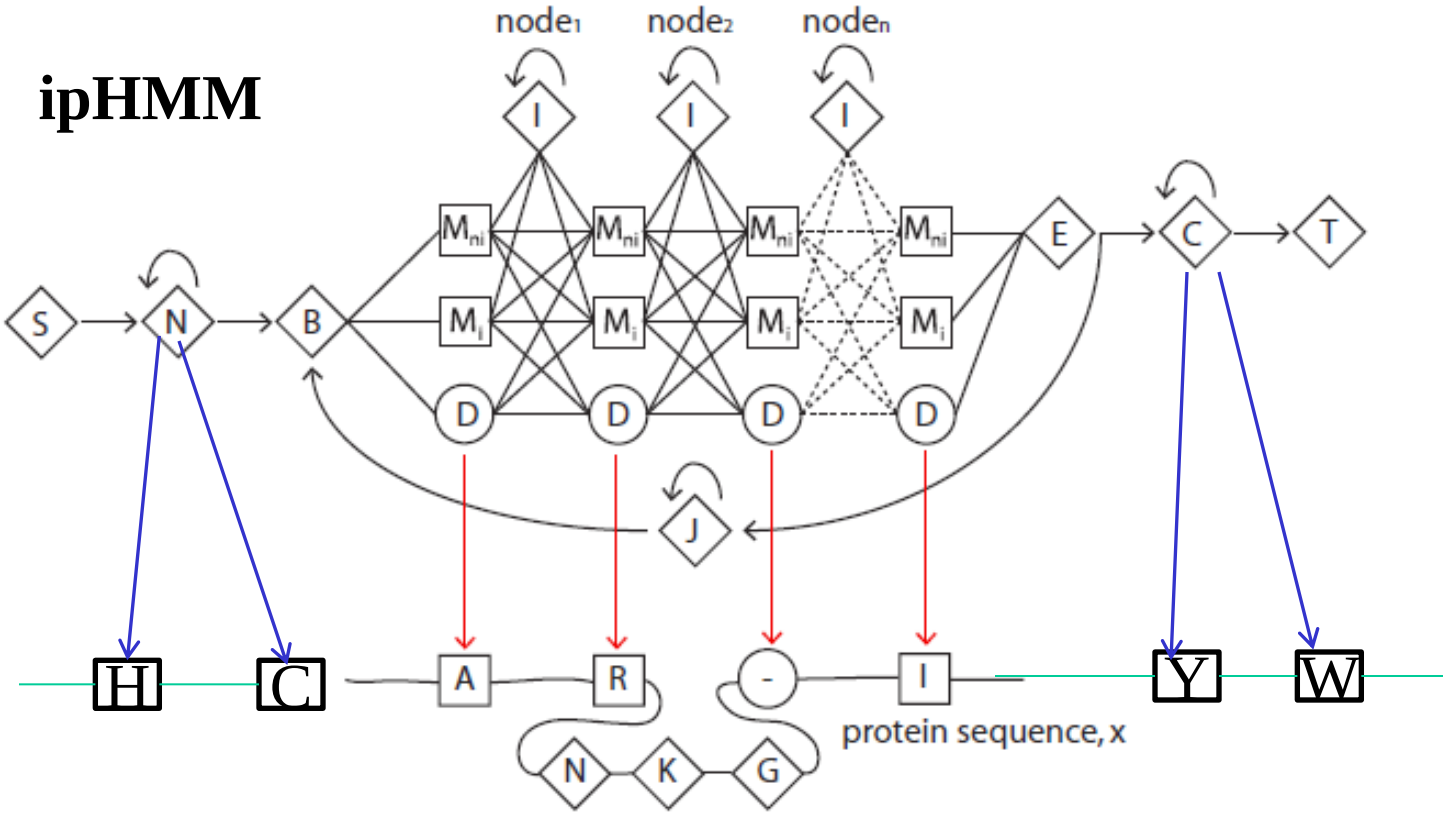
Chemical bonds are formed between amino acids across interface at two interacting proteins.



Residues at interface tend to be more conserved due to selection pressure during evolution.

Profile Hidden Markov Models capturing interaction

- P(x|θ) : probability that sequence x contains a domain described by the model θ.
- Viterbi algorithm can align x against the model to annotate interacting residues.



Friedrich et al, *Bioinformatics* 2006

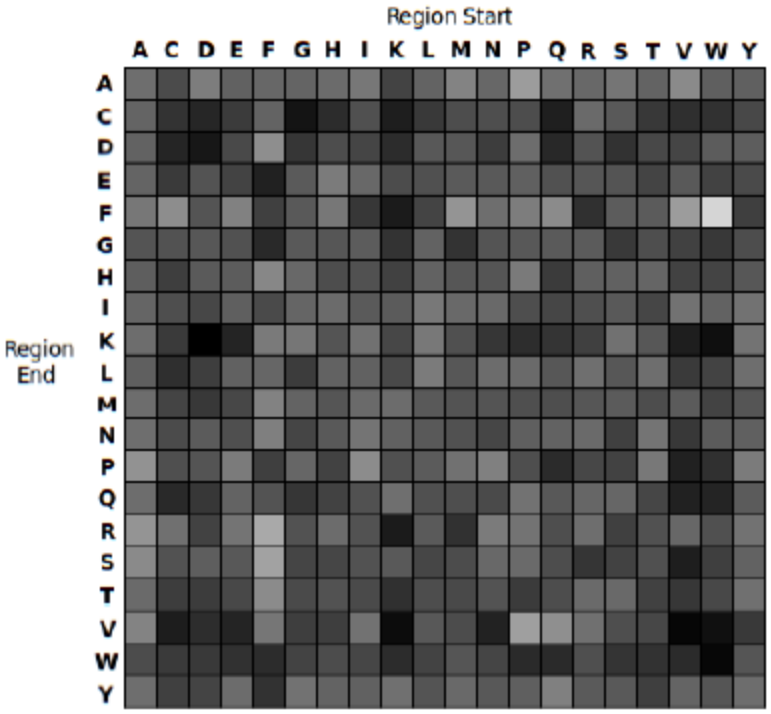
Capture long range correlation with HMM

(Kern, Gonzalez, Liao, Shanker, 2013)

hotspots

..ACHGLCVMPDILY….

$$S(x, y|x@h_i, y@h_{i+1}) = \log\left(\frac{p(y|x)}{p(x)\,p(y)}\right)$$



Initialization: $v_0(0) = 1, v_k(0) = 0 for k > 0$

Recursion:
$$v_j(i) = \begin{cases} e_j(x_i) max_k(v_k(i-1)a_{kj}), \\ \quad j \text{ is not a hotspot state} \\ e_j(x_i) max_k(v_k(i-1)a_{kj}) \\ \quad + \mathbf{C} * \mathbf{S}(\mathbf{x_{i'}}, \mathbf{x_i}), j \text{ is a hotspot state} \end{cases}$$

$ptr_i(j) = argmax_k(v_k(i-1)a_{kj})$

Termination: $P(x, \pi*) = max_k(v_k(L)a_{k0})$;

$\pi_L* = argmax_k(v_k(L)a_{k0})$

Traceback: $\pi_{i-1}* = ptr_i(\pi_i*)$

Fig. 3. The ETB-Viterbi algorithm

for $i =: 0$ **to length of protein** x**:**
  for $j =: 0$ **to number of states in HMM:**
    $max =: 0$
    $ptr =: null$
    for $k =: 0$ **to number of states in HMM:**
      if $v_k(i-1) > max$
        $max =: v_k(i-1)a_{kj}$
        $ptr =: k$
      End if
    End for
    $v_j(i) =: max$
    $ptr_j(i) =: ptr$
    if j **is a hotspot state**
      traceback following $ptr_j(i)$ until a hotspot (j', i')
      $v_j(i) = v_j(i) + C * S(x'_i, x_i)$
    End if
  End for
End for

Fig. 4. Pseudocode of the recursive stage of the ETB-Viterbi algorithm.

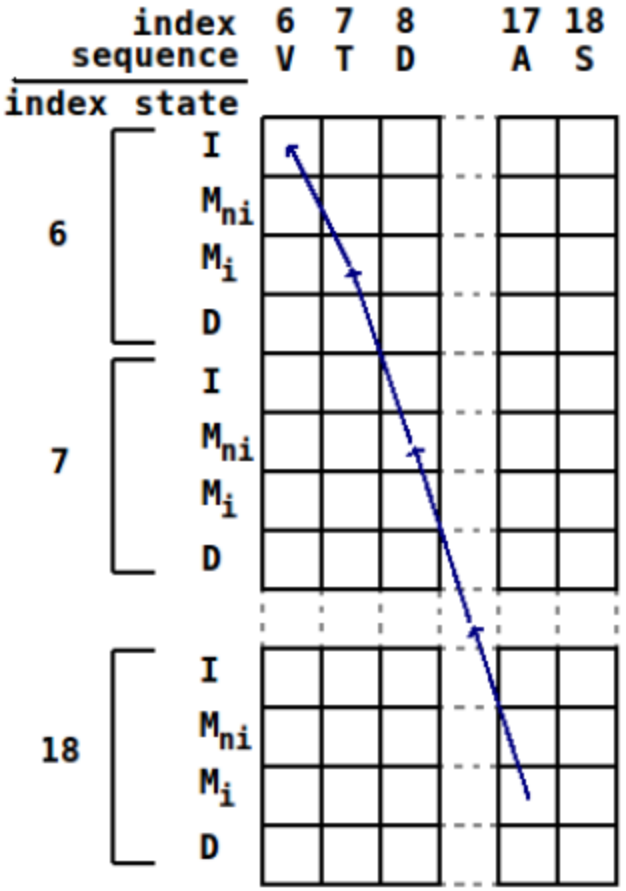


Fig. 5. A segment of the dynamic programming table for ipHMM. The lines show the path of an early traceback, with M_i states defined as hotspots. When hotspot (18, 17) is handled on the bottom right, traceback is performed until the previous hotspot, (6, 7), is reached.

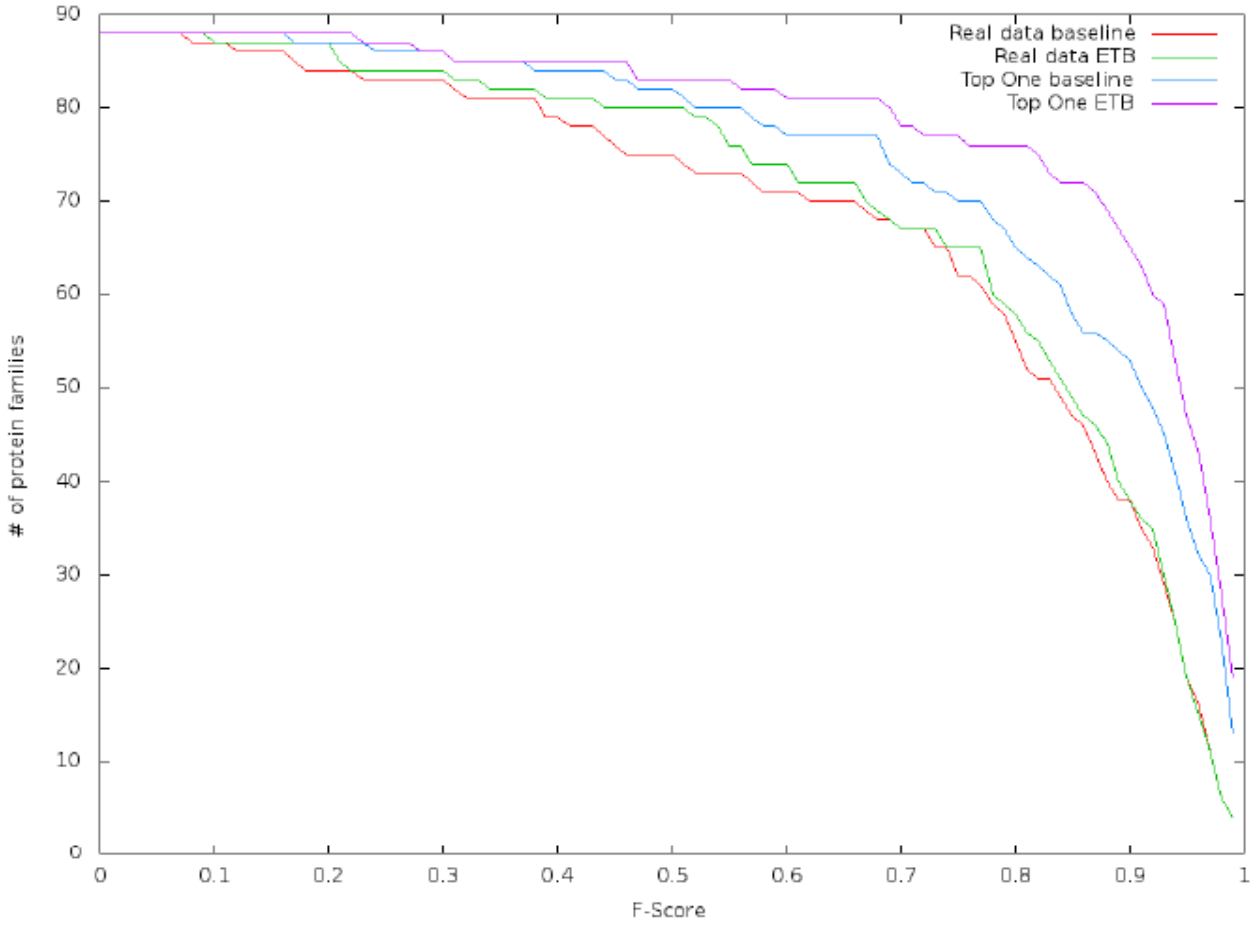


Fig. 6. The improvement made by the ETB-Viterbi algorithm on the real data and the Top 1 simulated dataset.