Hidden Markov Models in Bioinformatics 14.11 60 min

Definition

Three Key Algorithms

- Summing over Unknown States
- Most Probable Unknown States
- Marginalizing Unknown States

Key Bioinformatic Applications

- Pedigree Analysis
- Isochores in Genomes (CG-rich regions)
- Profile HMM Alignment
- Fast/Slowly Evolving States
- Secondary Structure Elements in Proteins
- Gene Finding
- Statistical Alignment

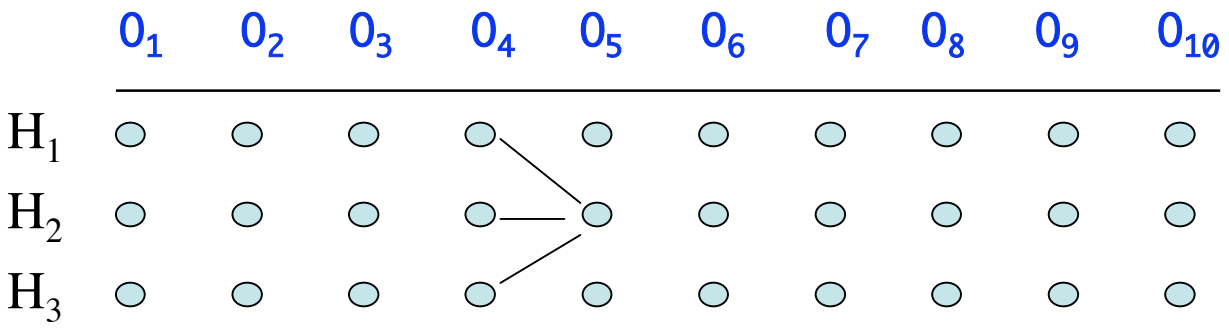
Hidden Markov Models

(O_1,H_1), (O_2,H_2),……. (O_n,H_n) is a sequence of stochastic variables with 2 components - one that is observed (O_i) and one that is hidden (H_i).

The marginal discribution of the H_i's are described by a Homogenous Markov Chain:

- Let $p_{i,j} = P(H_k=i,H_{k+1}=j)$
- Let $\pi_i = P\{H_1=i)$ - often π_i is the equilibrium distribution of the Markov Chain.
- Conditional on H_k (all k), the O_k are independent.
- The distribution of O_k only depends on the value of H_i and is called the emit function

$$e(i,j) = P\{O_k = i \mid H_k = j\}$$



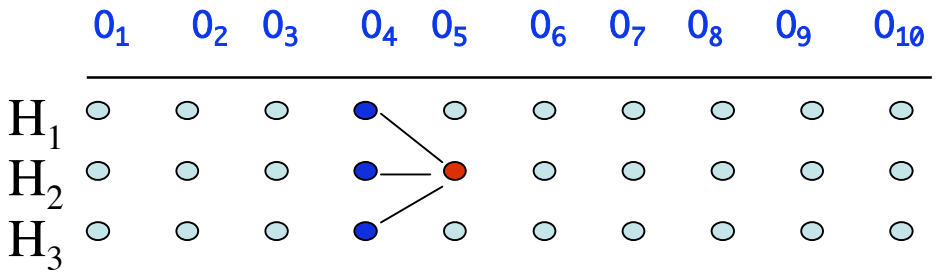
What is the probability of the data?

The probability of the observed is $P(\vec{O}) = \sum_{\vec{H}} P(\vec{O}|\vec{H})P(\vec{H})$, which can be hard to calculate. However, these calculations can be considerably accelerated. Let $P_{O_k=i}^{H_k=j}$ the probability of the observations $(O_1,..O_k)$ conditional on $H_k=j$. Following recursion will be obeyed:

$$i.\quad P_{O_k=i}^{H_k=j} = P(O_k = i|H_k = j)\sum_{H_{k-1}=j} P_{O_{k-1}}^{H_{k-1}=j} p_{j,i}$$

$$ii.\quad P_{O_1=i}^{H_1=j} = P(O_1 = i|H_1 = j)\pi_j \quad \text{(initial condition)}$$

$$iii.\quad P(O) = \sum_{H_n=j} P_{O_n=i}^{H_n=j}$$



$$P_{O_5=i}^{H_5=2} = P(O_5 = i|H_5 = 2)\sum_{H_4=j} P_{O_4}^{H_4=j} p_{j,i}$$

What is the most probable ”hidden” configuration?

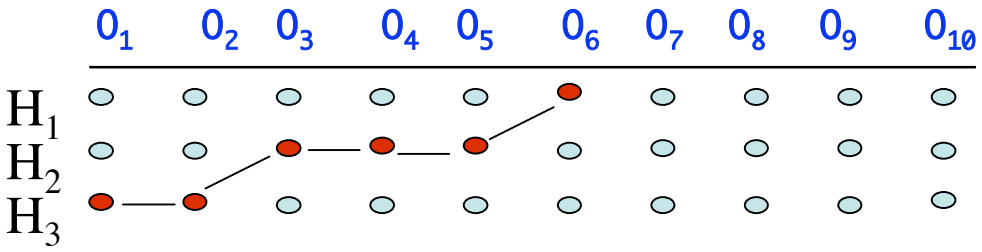
Let H^* be the sequences of hidden states that maximizes the observed sequence O ie $\text{ArgMax}_H[\, P\{O|H\}]$. Let H_k^j probability of the most probability of the most probable path up to k ending in hidden state j.

Again recursions can be found

$$i.\quad H_k^j = \max_i\{H_{k-1}^i p_{i,j}\}e(O_k, j) \qquad ii.\quad H_1^j = \pi_j e(O_1, 1)$$

The actual sequence of hidden states H_k^* can be found recursively by

$$iii.\quad H_{k-1}^* = \{i | H_{k-1}^i p_{i,j} e(O_k, j) = H_k^{H_k^*}\}$$



$$H_6^1 = \max_j\{H_6^1 * p_{j,1} * e(O_6, 1)\}$$

$$H_5^* = \{i | H_5^i * p_{i,1} * e(O_6, 1) = H_6^1\}$$

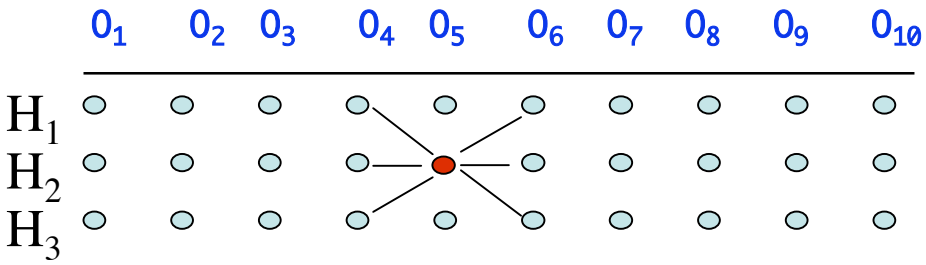
What is the probability of specific ”hidden” state?

Let Q_k^j be the probability of the observations from $j+1$ to n given $H_k=j$. These will also obey recursions:

$$Q_k^j = \sum_{H_{k+1}=i} P(O_k | H_{k+1} = i) p_{j,i} Q_{k+1}^i$$

The probability of the observations and a specific hidden state can found as: $P\{O, H_k = j) = P_k^j Q_k^j$

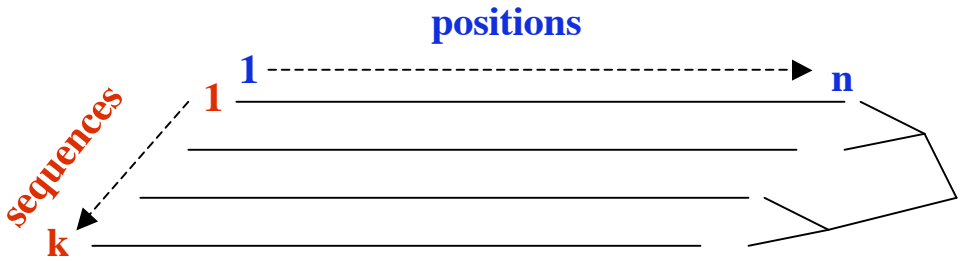
And of a specific hidden state can found as: $P\{H_k = j) = P_k^j Q_k^j / P(O)$



$$P\{H_5 = 2) = P_5^2 Q_5^2 / P(O)$$

Fast/Slowly Evolving States

Felsenstein & Churchill, 1996



HMM:

slow - r_s

fast - r_f

- π_r - equilibrium distribution of hidden states (rates) at first position
- $p_{i,j}$ - transition probabilities between hidden states
- $L_{(j,r)}$ - likelihood for j'th column given rate r.
- $L^{(j,r)}$ - likelihood for first j columns given j'th column has rate r.

Likelihood Recursions:

$$L^{(j,f)} = (L^{(j-1,f)} p_{f,f} + L^{(j-1,s)} p_{s,f}) L_{(j,f)} \qquad L^{(j,s)} = (L^{(j-1,f)} p_{f,s} + L^{(j-1,s)} p_{s,s}) L_{(j,s)}$$

Likelihood Initialisations:

$$L^{(1,f)} = \pi_f L_{(1,f)} \qquad L^{(1,s)} = \pi_s L_{(1,s)}$$

Statistical Alignment

Steel and Hein,2000 + Holmes and Bruno,2000

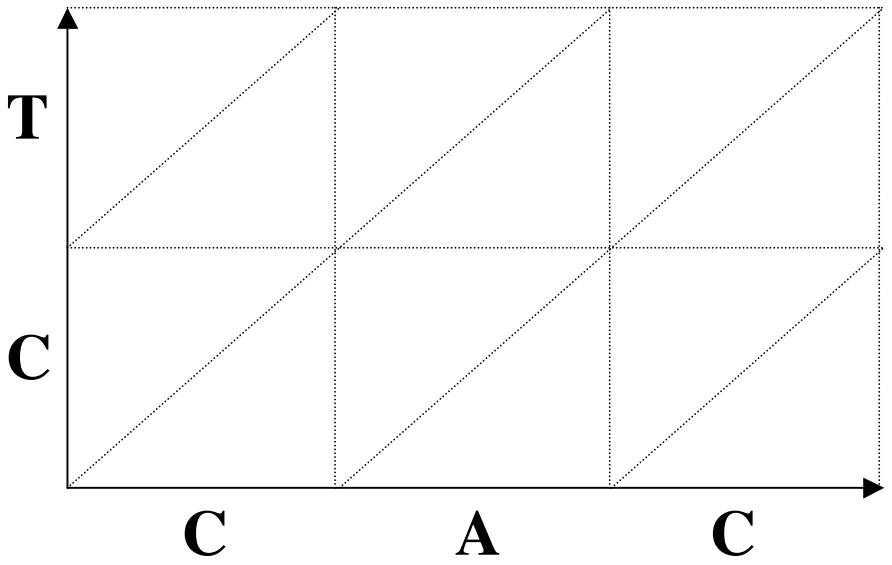
Emit functions:

$e(\#\#) = \pi(N_1) f(N_1, N_2)$

$e(\#-) = \pi(N_1)$, $e(-\#) = \pi(N_2)$

$\pi(N_1)$ - equilibrium prob. of N

$f(N_1, N_2)$ - prob. that N_1 evolves into N_2

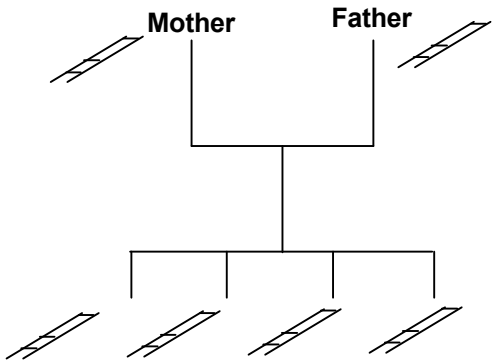


An HMM Generating Alignments

	-	#	#	E
	#	#	-	E
*	$\lambda\beta$	$\lambda/\mu\,(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu\,(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)\,(1-\lambda\beta)$
*	$\lambda\beta$	$\lambda/\mu\,(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu\,(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)\,(1-\lambda\beta)$
-	$\lambda\beta$	$\lambda/\mu\,(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu\,(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)\,(1-\lambda\beta)$
#	$\lambda\beta$	$\lambda/\mu\,(1-\lambda\beta)e^{-\mu}$	$\lambda/\mu\,(1-\lambda\beta)(1-e^{-\mu})$	$(1-\lambda/\mu)\,(1-\lambda\beta)$
#	$\frac{1-\lambda\beta e^{-\mu}}{1-e^{-\mu}}$	$\frac{\lambda\beta e^{-\mu}}{1-e^{-\mu}}$		$\frac{(\mu-\lambda)\beta}{1-e^{-\mu}}$
-			$\lambda\beta$	

Probability of Data given a pedigree.

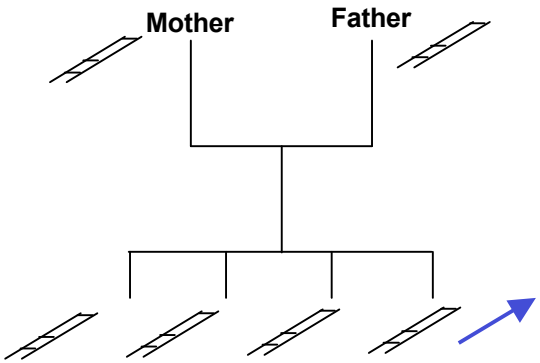
Elston-Stewart (1971) -Temporal Peeling Algorithm:



Condition on parental states

Recombination and mutation are Markovian

Lander-Green (1987) - Genotype Scanning Algorithm:



Condition on paternal/maternal inheritance

Recombination and mutation are Markovian

Comment: Obvious parallel to Wiuf-Hein99 reformulation of Hudson's 1983 algorithm

Further Examples I

Isochore:

Churchill,1989,92

HMM:

poor
rich

$L_p(C)=L_p(G)=0.1,\ L_p(A)=L_p(T)=0.4,$
$L_r(C)=L_r(G)=0.4,\ L_r(A)=L_r(T)=0.1$

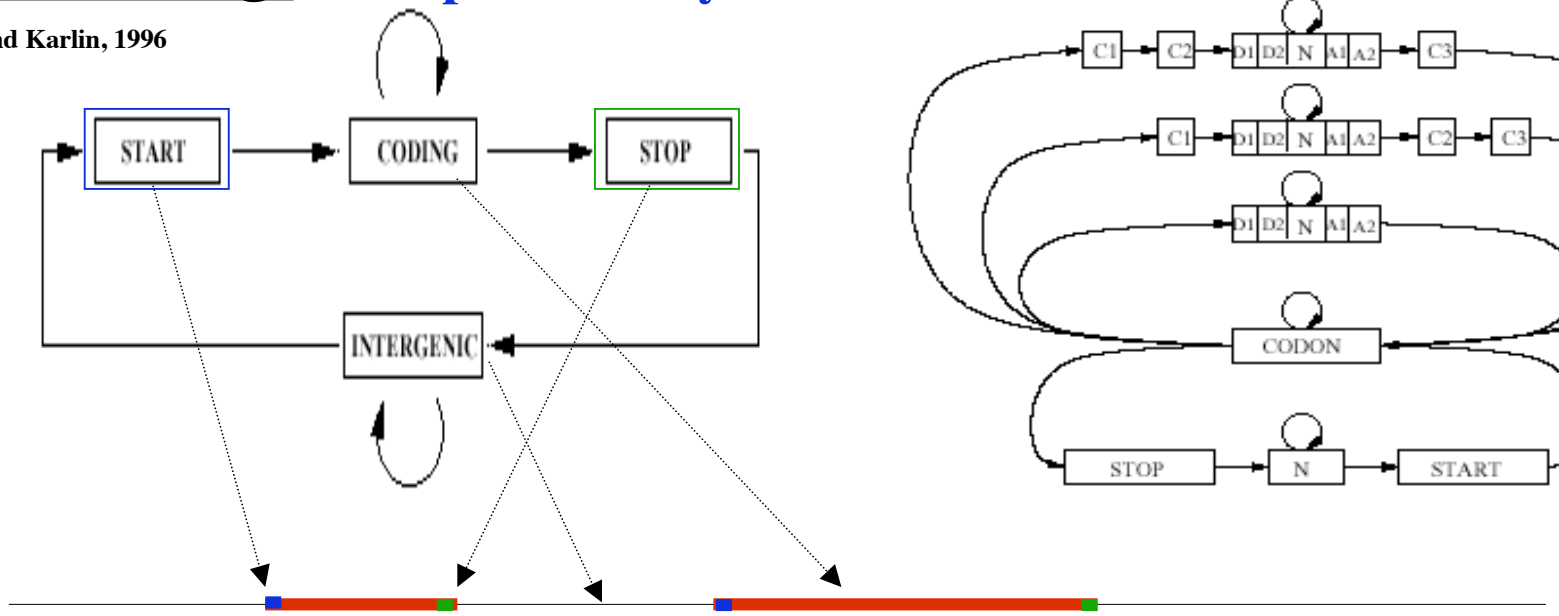
Likelihood Recursions: $L^{(j,p)} = (L^{j-1,p} p_{p,p} + L^{(j-1,s)} p_{s,f}) P_p(S[j]),\ L^{(j,r)} = (L^{(j-1,r)} p_{p,r} + L^{(j-1,r)} p_{r,r}) P_r(S[j])$

Likelihood Initialisations: $L^{(1,p)} = \pi_p P_p(S[1]),\ L^{(1,r)} = \pi_r P_r(S[1])$

Gene Finding:

Burge and Karlin, 1996

Simple Prokaryotic



Simple Eukaryotic

Secondary Structure Elements:

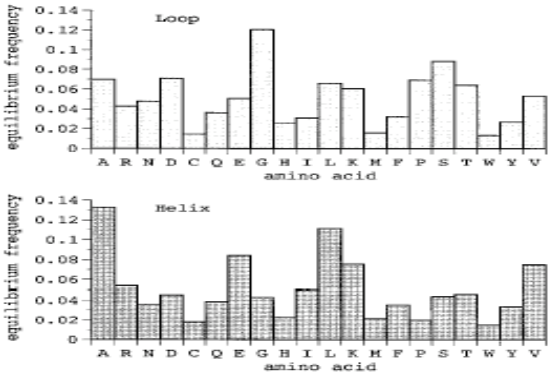
Further Examples II

Goldman, 1996

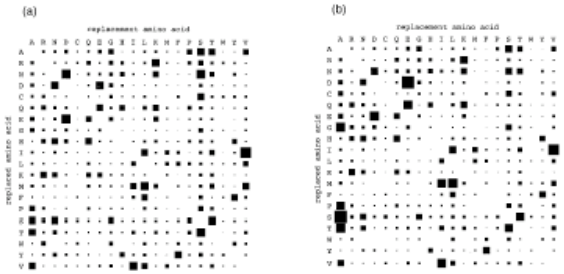
HMM for SSEs:

α L α L β L β

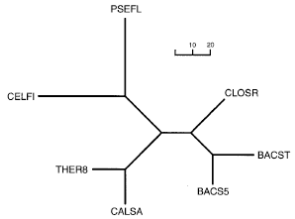
	α	β	L
α	.909	.0005	.091
β	.005	.881	.184
L	.062	.086	.852
	.325	.212	.462

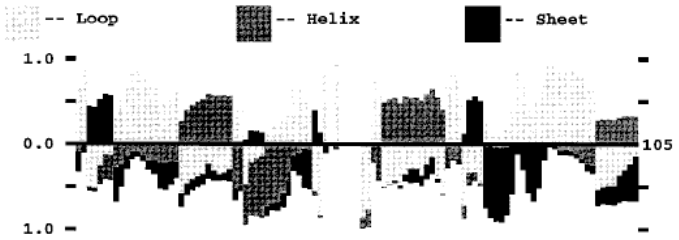


Adding Evolution:



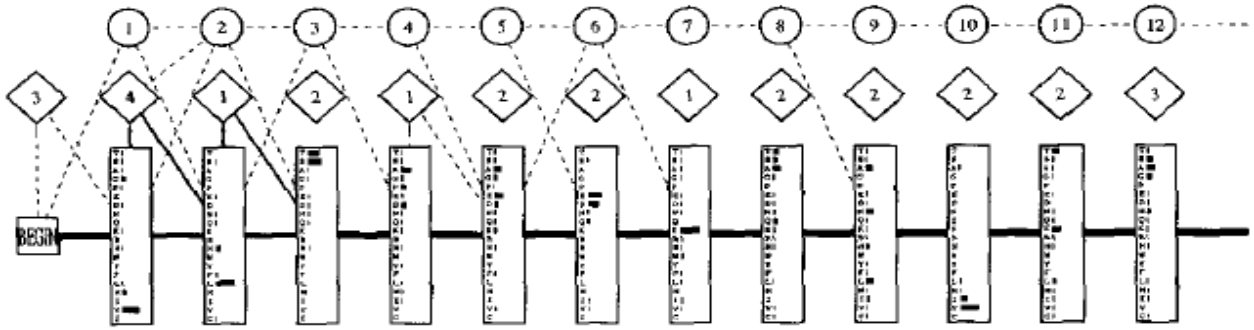
SSE Prediction:



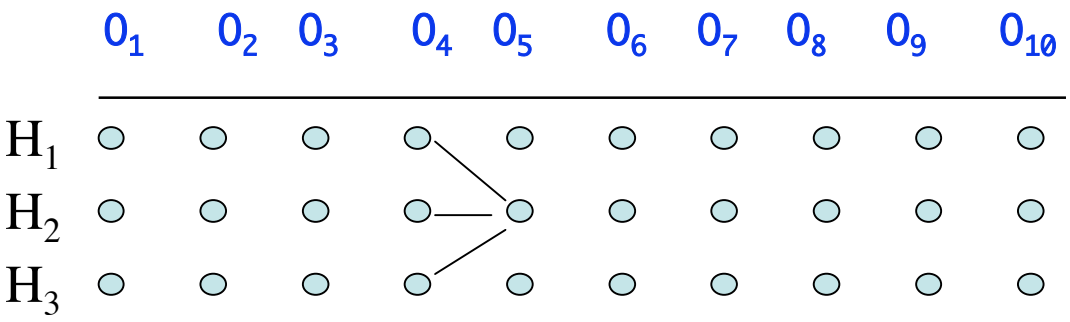


Profile HMM Alignment:

Krogh et al.,1994



Summary



Definition

Three Key Algorithms

- **Summing over Unknown States**
- **Most Probable Unknown States**
- **Marginalizing Unknown States**

Key Bioinformatic Applications

- **Pedigree Analysis**
- **Isochores in Genomes (CG-rich regions)**
- **Profile HMM Alignment**
- **Fast/Slowly Evolving States**
- **Secondary Structure Elements in Proteins**
- **Gene Finding**
- **Statistical Alignment**

Recommended Literature

Vineet Bafna and Daniel H. Huson (2000) The Conserved Exon Method for Gene Finding ISMB 2000 pp. 3-12

S.Batzoglou et al.(2000) Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Genome Research. 10.950-58.

Blayo, Rouze & Sagot (2002) "Orphan Gene Finding - An exon assembly approach" J.Comp.Biol.

Delcher, AL et al.(1998) Alignment of Whole Genomes Nuc.Ac.Res. 27.11.2369-76.

Gravely, BR (2001) Alternative Splicing: increasing diversity in the proteomic world. TIGS 17.2.100-

Guigo, R.et al.(2000) An Assesment of Gene Prediction Accuracy in Large DNA Sequences. Genome Research 10.1631-42

Kan, Z. Et al. (2001) Gene Structure Prediction and Alternative Splicing Using Genomically Aligned ESTs Genome Research 11.889-900.

Ian Korf et al.(2001) Integrating genomic homology into gene structure prediction. Bioinformatics vol17.Suppl.1 pages 140-148

Tejs Scharling (2001) Gene-identification using sequence comparison. Aarhus University

JS Pedersen (2001) Progress Report: Comparative Gene Finding. Aarhus University

Reese,MG et al.(2000) Genome Annotation Assessment in Drosophila melanogaster Genome Research 10.483-501.

Stein,L.(2001) Genome Annotation: From Sequence to Biology. Nature Reviews Genetics 2.493-