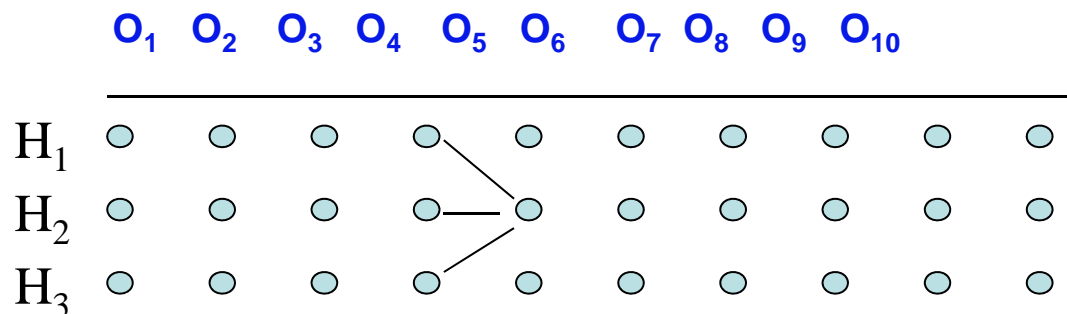


Hidden Markov Models in Bioinformatics



Definition

Three Key Algorithms

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

Key Bioinformatic Applications

- **Pedigree Analysis**
- **Profile HMM Alignment**
- **Fast/Slowly Evolving States**
- **Statistical Alignment**

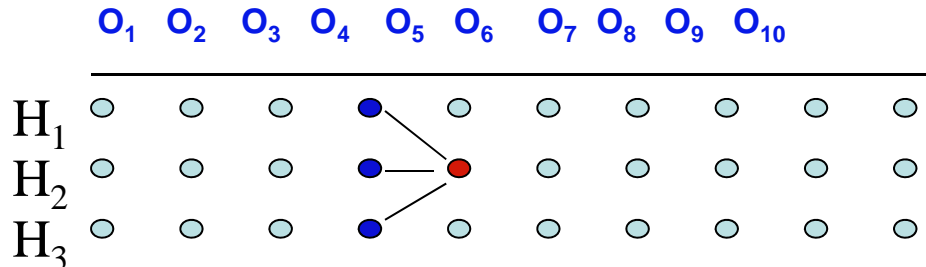
What is the probability of the data?

The probability of the observed is $P(\vec{O}) = \sum_{\vec{H}} P(\vec{O}|\vec{H})P(\vec{H})$, which could 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, \dots, O_k) conditional on $H_k=j$. Following recursion will be obeyed:

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

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

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



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

Example - probability of the data

Observables {0, 1} at times 1, 2, 3. Hidden states {a, b}.

Transition probabilities:

.9	.1
.1	.9

Emission probabilities:

	0	1
a	.7	.3
b	.3	.7

Equilibrium distribution, π , of a b is .5 .5

Example. Observation 0 1 1

Direct calculation: $P(O) = \sum_H P(O|H)P(H)$

$P(aaa) = .5 * .9 * .9$ $P(011|aaa) = .7 * .3 * .3$

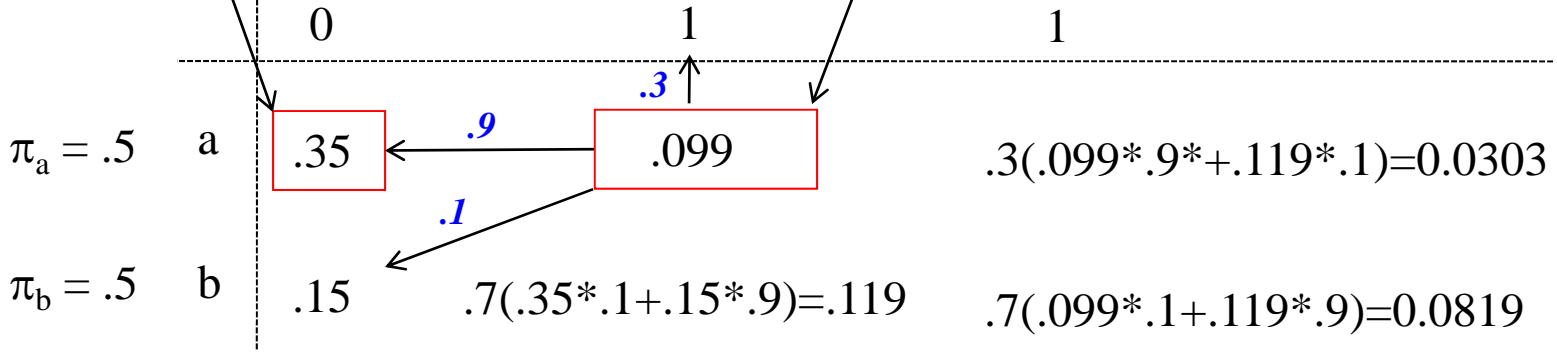
$P(aab) = .5 * .9 * .1$ $P(011|aab) = .7 * .3 * .7$

Forward recursion:

$\pi_a * P(0|a) = .5 * .7$

$P^{H_2=a} = P(1|a) \sum_{H_1=j} P_{O_1}^{H_1=j} p_{j,i} = .3[.35 * .9 + .15 * .1]$

Observations:



Hence $P(O)$ up to the 3rd state is $0.0303 + 0.0819 = 0.1122$

What is the most probable "hidden" configuration?

This algorithm is also called Viterby.

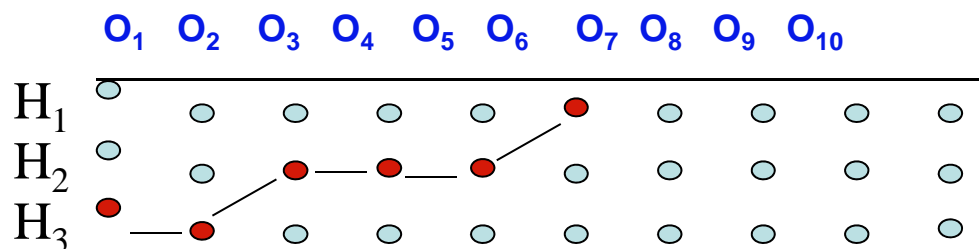
Let H^* be the sequences of hidden states in the most probably hidden path ie $\text{ArgMax}_H[P\{O|H\}]$. Let H_k^j be the probability of the most probable path up to k ending in hidden state j .

Again recursions can be found:

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

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

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



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

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

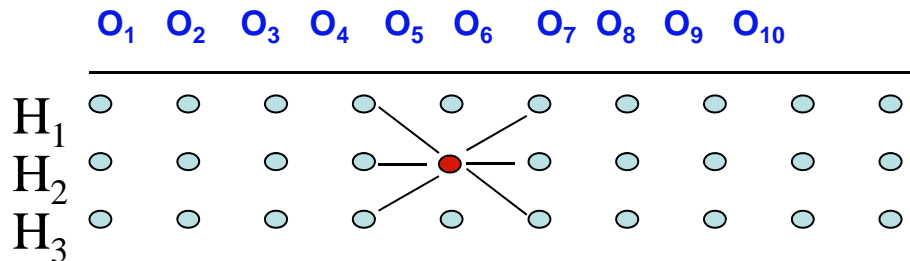
What is the probability of specific "hidden" state?

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

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

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

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



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

Example continued - best path, single hidden state

Best path: $P(H|O)P(O) = P(O|H)P(H)$ or $P(H|O) = P(O|H)P(H)/P(O)$

Observations:		0	1	1
$\pi_a = .5$	a	.7	.189 = $\text{Max}\{.7 * .9 * .3, .3 * .1 * .3\}$.051
$\pi_b = .5$	b	.3	.189	.1191

Transitions: a to a (0) is .9, a to b (0) is .1, b to a (1) is .3, b to b (1) is .7.

Single hidden state:

Forward:

Observations:	0	1	1
$\pi_a = .5$	a	<input type="text"/>	
$\pi_b = .5$	b		

Forward - Backward:

Observations:	0	1	1
$\pi_a = .5$	a	<input type="text"/>	
$\pi_b = .5$	b		

Backward:

Observations:	0	1	1
$\pi_a = .5$	a	<input type="text"/>	
$\pi_b = .5$	b		

Baum-Welch, Parameter Estimation or Training

	O ₁	O ₂	O ₃	O ₄	O ₅	O ₆	O ₇	O ₈	O ₉	O ₁₀
H ₁	○	○	○	○	○	○	○	○	○	○
H ₂	○	○	○	○	●	○	○	○	○	○
H ₃	○	○	○	○	○	○	○	○	○	○

Objective: Evaluate Transition and Emission Probabilities

Set p_{ij} and $e(\cdot)$ arbitrarily to non-zero values

- Use forward-backward to re-evaluate p_{ij} and $e(\cdot)$

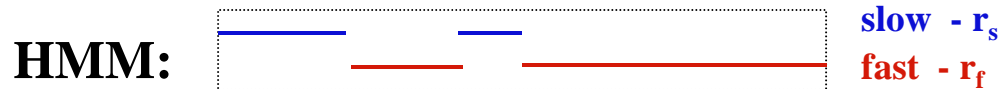
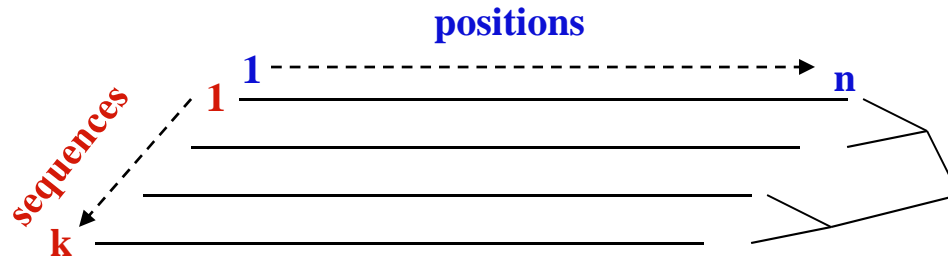
Do this until no significant increase in probability of data

To avoid zero probabilities, add pseudo-counts.

Other numerical optimization algorithms can be applied.

Fast/Slowly Evolving States

Felsenstein & Churchill, 1996



- 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:

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

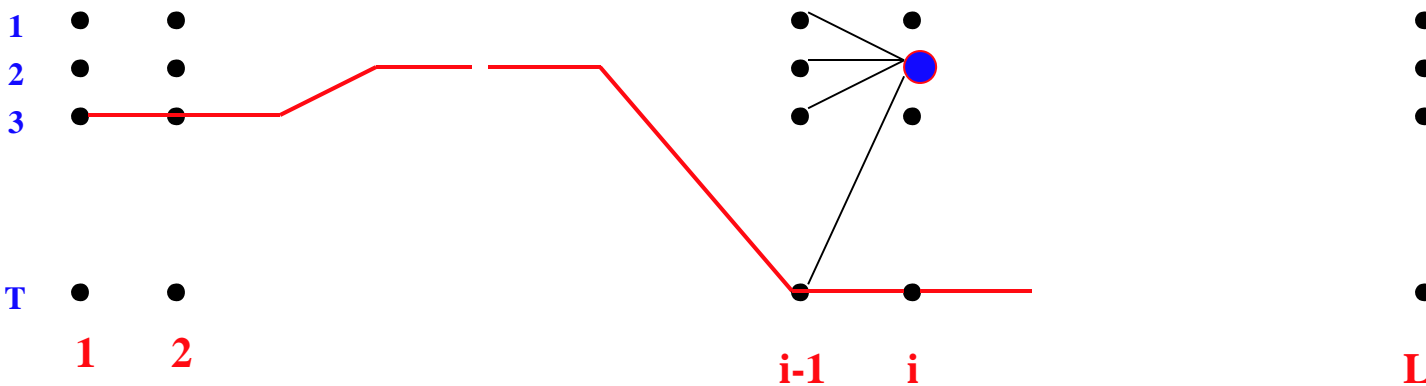
Likelihood Initialisations:

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

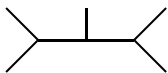
Recombination HMMs

1	0	1	1	1	1	0	0	0	0	0	0	0	0
1	0	1	1	1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	0	1	0	1	0	1	1	1
0	0	0	0	0	0	1	1	1	1	1	1	0	1
0	1	0	0	0	0	1	1	1	1	1	1	0	1

Data

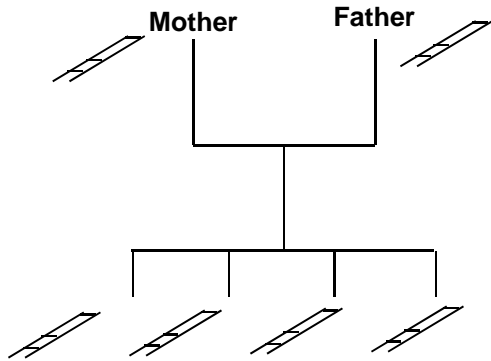


Trees



Probability of Data given a pedigree.

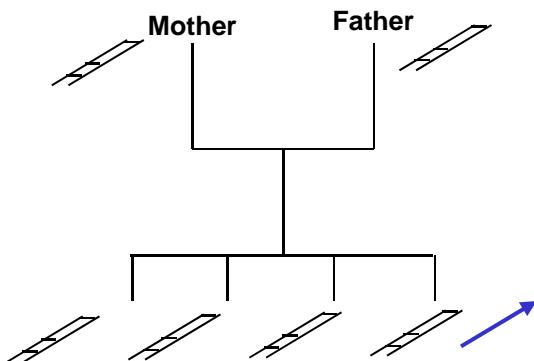
Elston-Stewart (1971) - Temporal Peeling Algorithm:



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

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:

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

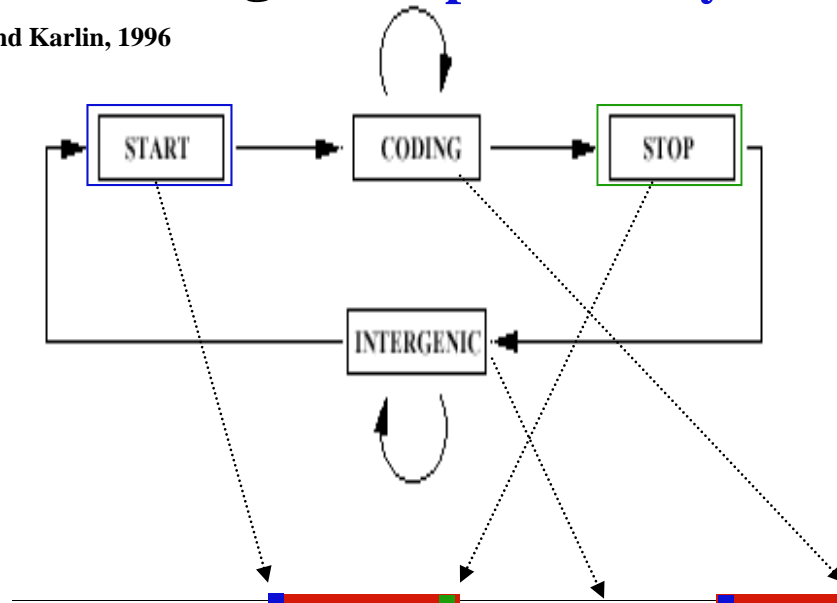
Likelihood Initialisations:

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

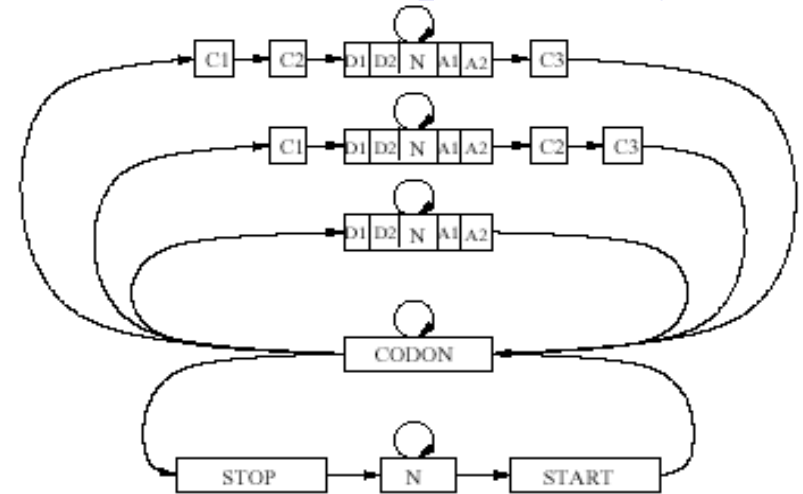
Gene Finding:

Simple Prokaryotic

Burge and Karlin, 1996



Simple Eukaryotic



Secondary Structure Elements:

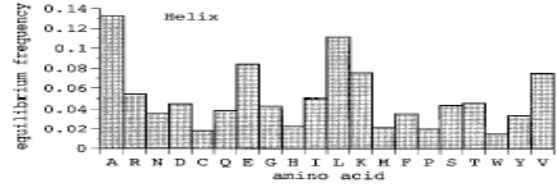
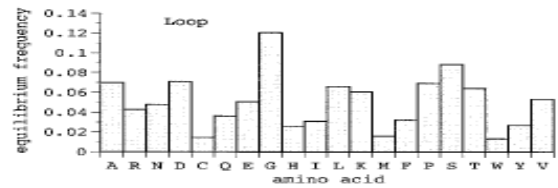
Goldman, 1996

Further Examples

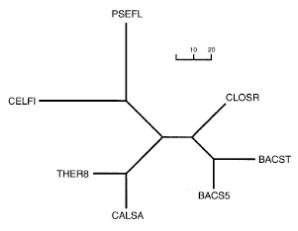
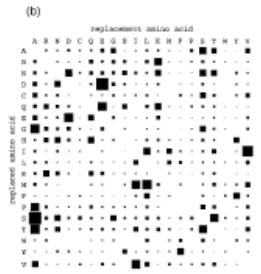
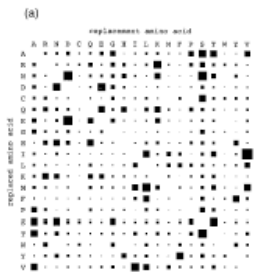
HMM for SSEs:



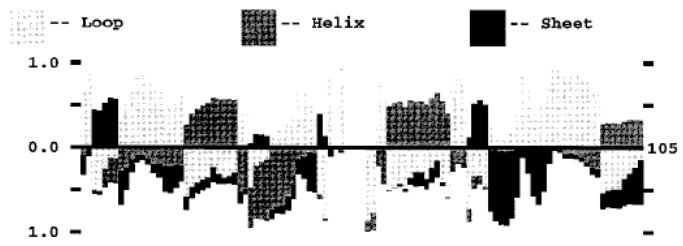
	α	β	L
α	.909	.0005	.091
β	.005	.881	.184
L	.062	.086	.852
	.325	.212	.462



Adding Evolution:



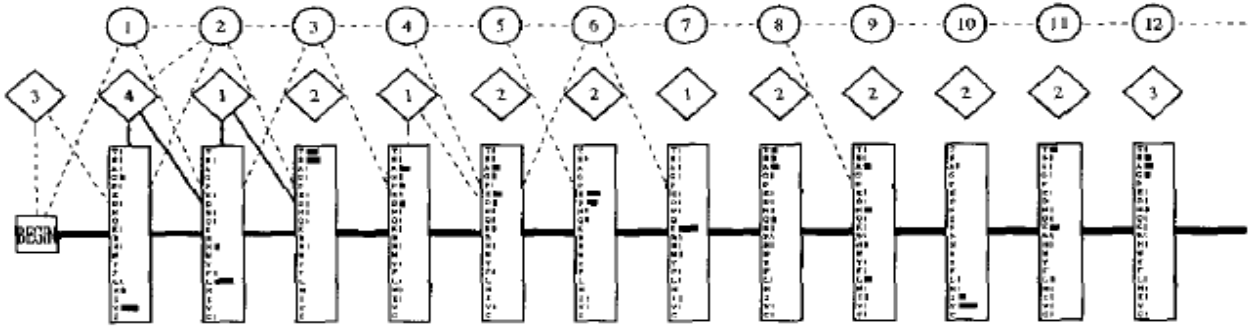
SSE Prediction:



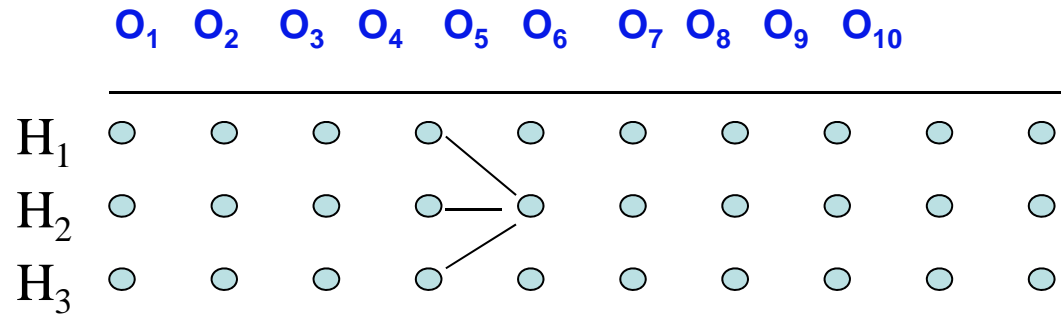
Profile HMM

Alignment:

Krogh et al., 1994



Summary



Definition

Three Key Algorithms

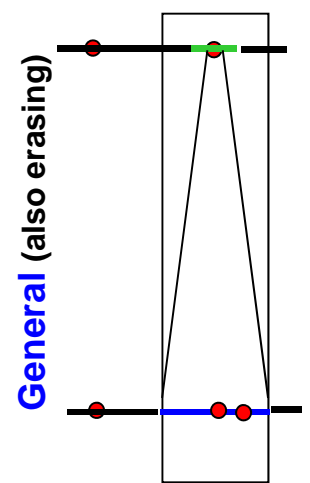
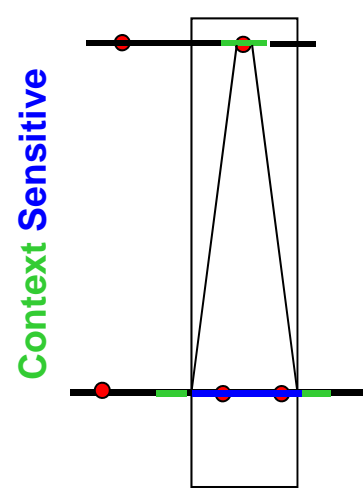
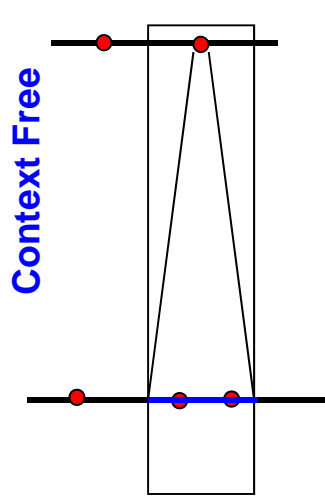
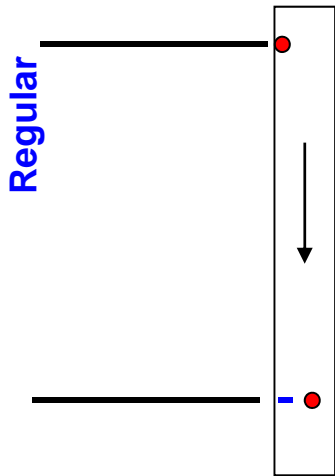
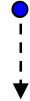
- **Summing over Unknown States**
- **Most Probable Unknown States**
- **Marginalizing Unknown States**

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**

Grammars: Finite Set of Rules for Generating Strings

- i. A starting symbol: • Ordinary letters: — & Variables: •
- ii. A set of substitution rules applied to variables • in the present string: —•—•—



— finished – no variables

Simple String Generators

Terminals (capital) --- Non-Terminals (small)

i. Start with **S** **S** \rightarrow **aT bS**
 T \rightarrow **aS bT** ϵ

One sentence – odd # of **a**'s:

S \rightarrow **aT** \rightarrow **aaS** \rightarrow **aabS** \rightarrow **aabaT** \rightarrow **aaba**

ii. **S** \rightarrow **aSa bSb aa bb**

One sentence (even length palindromes):

S \rightarrow **aSa** \rightarrow **abSba** \rightarrow **abaaba**

Stochastic Grammars

The grammars above classify all string as belonging to the language or not.

All variables has a finite set of substitution rules. Assigning probabilities to the use of each rule will assign probabilities to the strings in the language.

If there is a 1-1 derivation (creation) of a string, the probability of a string can be obtained as the product probability of the applied rules.

i. Start with **S**. **S** \rightarrow (0.3)a**T** (0.7)b**S**
T \rightarrow (0.2)a**S** (0.4)b**T** (0.2) ϵ

S $\xrightarrow{*0.3}$ a**T** \rightarrow aa**S** $\xrightarrow{*0.2}$ aab**S** $\xrightarrow{*0.7}$ aaba**T** $\xrightarrow{*0.3}$ aaba $\xrightarrow{*0.2}$

ii. **S** \rightarrow (0.3)a**S**a (0.5)b**S**b (0.1)aa (0.1)bb

S $\xrightarrow{*0.3}$ a**S**a $\xrightarrow{*0.5}$ ab**S**ba $\xrightarrow{*0.1}$ abaaba

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-

Example continued - parameter optimisation