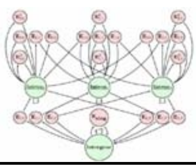
Computational Genomics

10-810/02-710, Spring 2009

Model-based Comparative Genomics

Eric Xing

Lecture 14, March 2, 2009

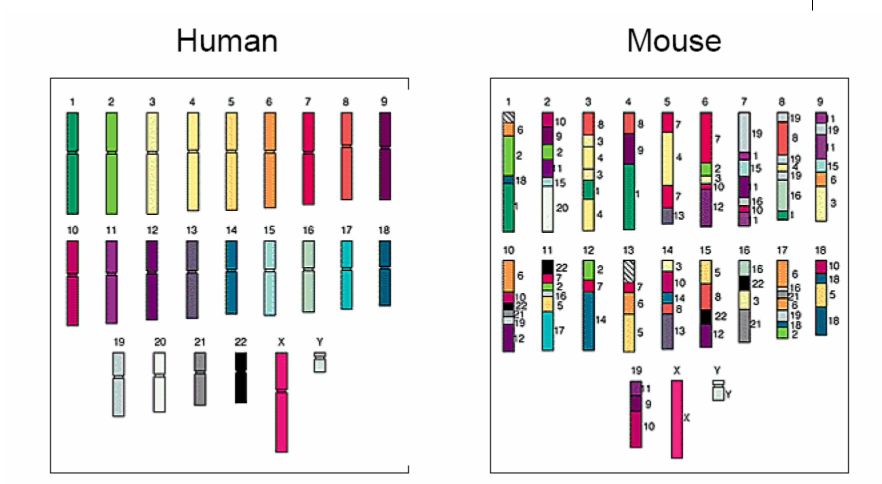


Reading: class assignment

Uses of evolutionary theory

- Comparative genomics (this lecture)
  - Cladistics: figuring out closely related species, proteins, sequences
    - Drug design and testing
    - Building chimera : mixing genetic codes of species, genetic technology
    - Sequence prediction in related unsequenced species : use in sequencing, primer design, etc
  - Phylogenetic footprinting
    - Functional constraints on a genomic region inversely proportional to evolutionary rate, from neutral theory
    - Look at two concrete examples : transcription factor binding site (motif) and gene prediction
- Population genetics (module 4)
  - Population structure
  - Understanding evolutionary driving force underlying genome variation

Comparative Genomics



A pairwise comparison between human and mouse genome

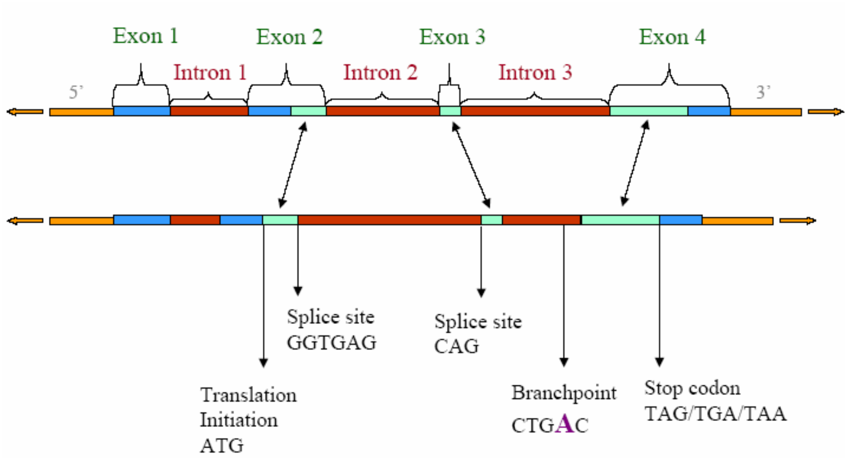
Total Homologies: 4776
Total mapped in both species: 3313

mouse, laboratory

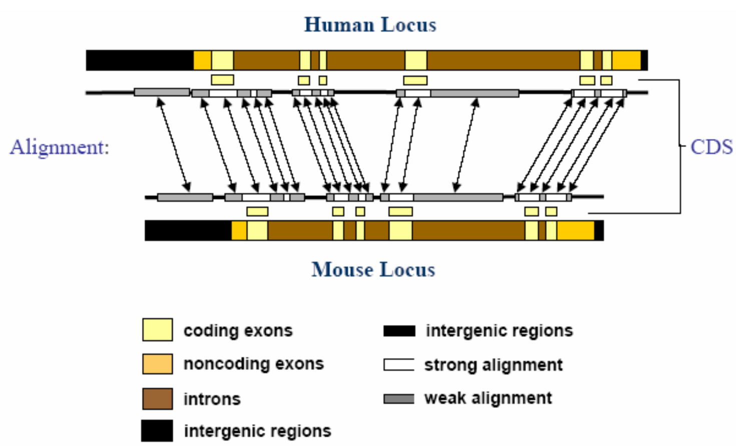
human	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT	human
1	99	1	90	101	2	1		6			2	1	7	1			1	1					111		1
2	77	48			6	29					9	17	1				10	1					59		2
3	1		19	1		27		1	61		1			12		31	1						54		3
4			26		80	4		21					2				1		1				31		4
5	1										47		54		13		4	34					39		5
6	5			6	1		1		7	29			23				110						44		6
7		1			70	68	1				13	13	6				1						40		7
8	6		8	10	1			26						20	33	3				1			30		8
9		45		54	2								10						12				33		9
10		12				3	12			13			1	17		1		3	49				23		10
11		33		1			77		50		1		1						48				66		11
12		1	1		23	67				41					46								53		12
13	3		2		8			14						23									18		13
14			1				1		1	1		51		29									38		14
15		30					33		35		1		1										23		15
16		1					22	58			8				1	17	19						30		16
17			1			1		1			200	2		1									60		17
18	3				2			1								1	4	33					12		18
19						1	95	31	12	29		1					16	1					61		19
20		80																					19		20
21										24						31	6						19		21
22	1				6	3		3		11	9				39	34							21		22
X																				185		1	33		X
Y																					6				Y
XY																									XY
UN	12	22	11	15	8	6	13	5	9	16	12	5	9	1	6	3	10	5	3	4			370		UN
MT																								37	MT
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	X	Y	XY	UN	MT	

mouse, laboratory

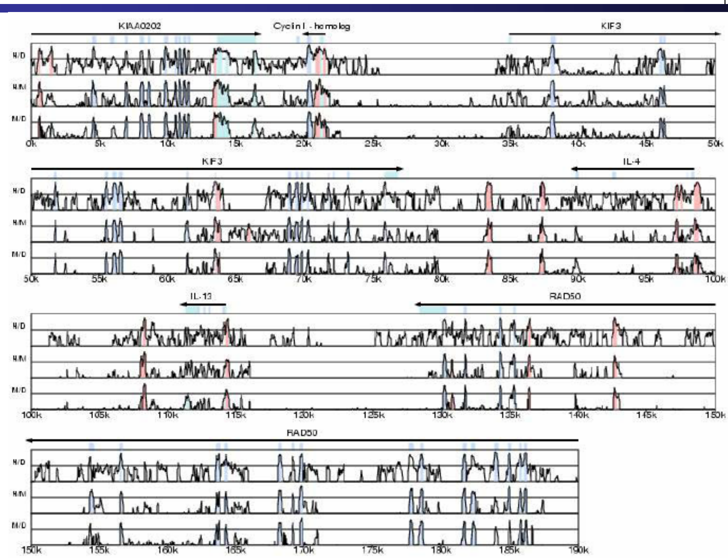
Aligning One Locus



Example: a human/mouse ortholog

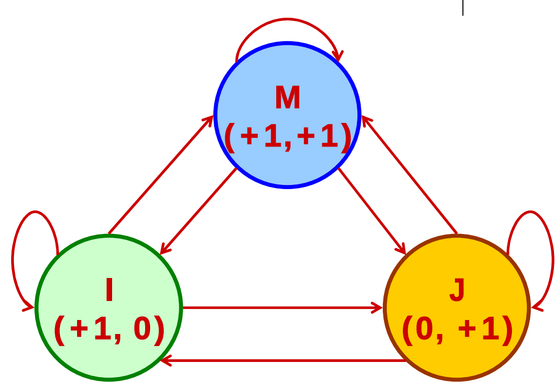


Three Pairwise Alignments



Paired HMM

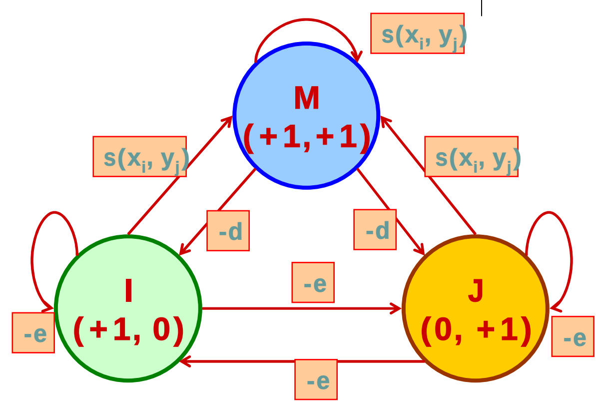
Alignments correspond 1-to-1 with sequences of states M, I, J




```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
IMMJMMMMMMMJJMMMMMMJMMMMMMMIIMMMMMIII
```

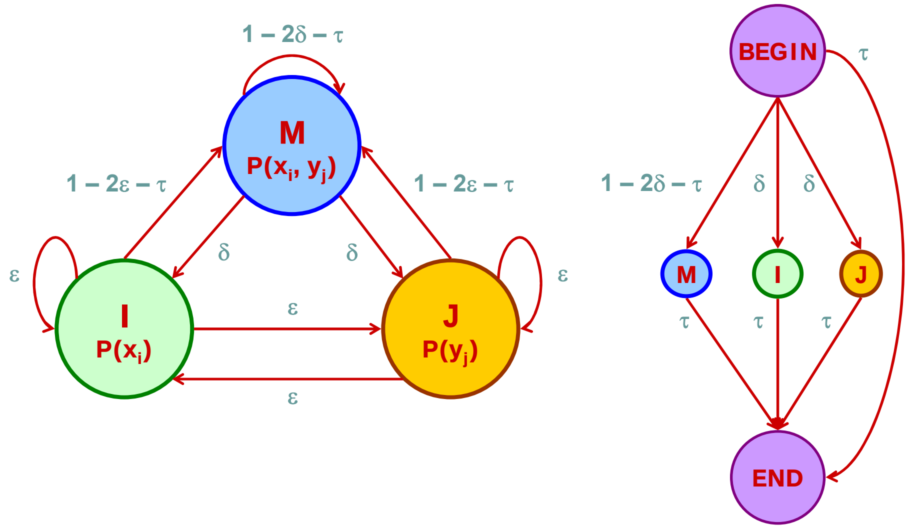
Let's score the transitions

Alignments correspond 1-to-1 with sequences of states M, I, J

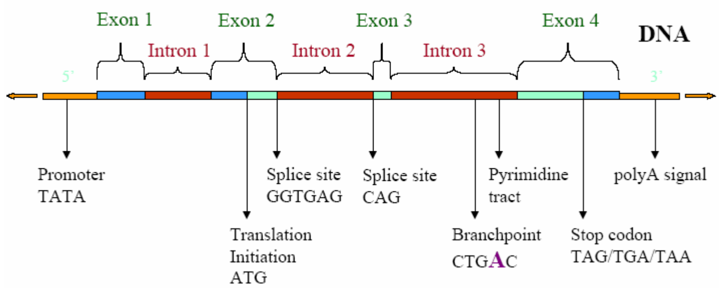



```
-AGGCTATCACCTGACCTCCAGGCCGA--TGCCC---
TAG-CTATCAC--GACCGC-GGTCGATTTGCCCGACC
IMMJMMMMMMMJJMMMMMMJMMMMMMMIIMMMMMIII
```


A Pair HMM for alignments

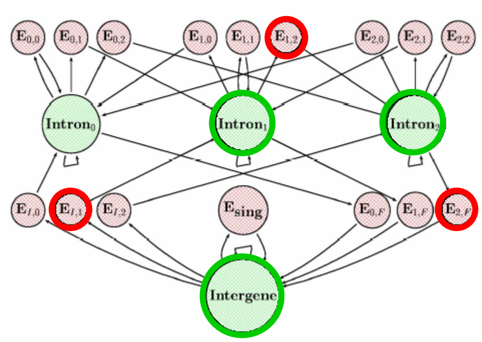


Gene Finding

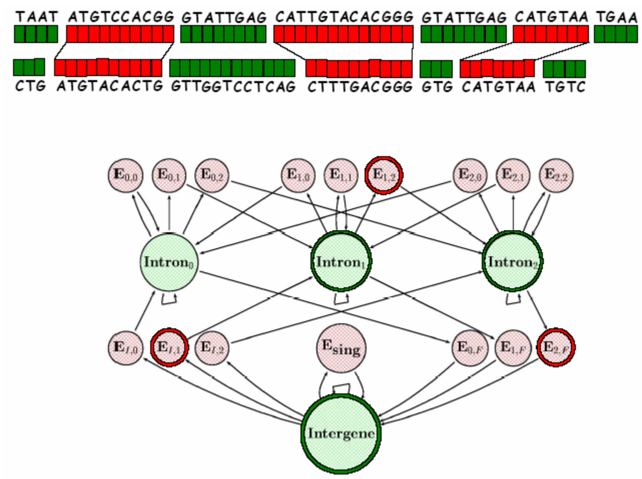


Recall generalized HMM gene finder

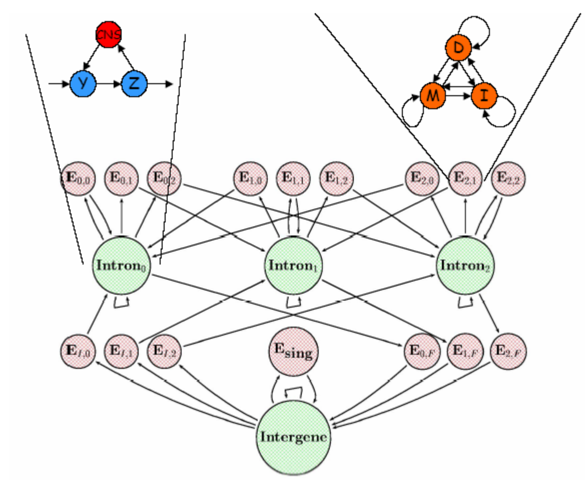
TAAT ATGTCCACGG GTATTGAG CATTGTACACGGG GTATTGAG CATGTAA TGAA



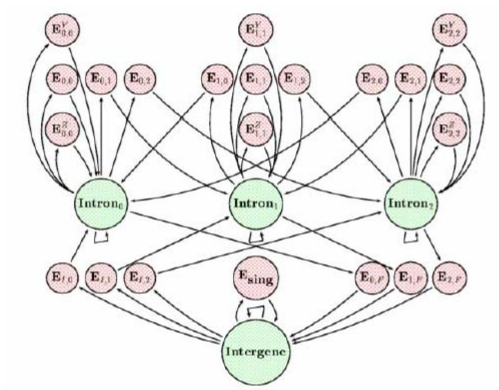
Generalized Pair-HMM gene finder



Hierarchical state transition in pHMM : knowledge of structure



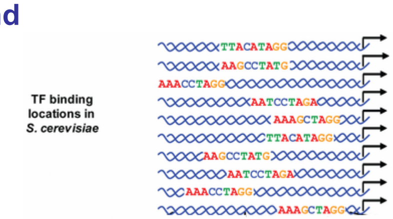
Allowing for inserted exons: knowledge of structure



Motif finding: recap

Recap:

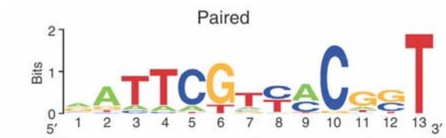
- **Functional regions in sequences often occur as small, noisy, repeating subsequences**
  - **In DNA : transcription factor binding sites, transcription start sites, splicing signals, etc**
  - **In proteins : transmembrane domains, phosphorylation sites, signal peptides. etc**
- **Subsequence similarity and functional significance go hand in hand**



Habib et al, PLoS CompBiol Feb 08

Problem formulation

- Models : Consensus, RegEx, Weight Matrix



- Supervised motif detection : Given a set of sequences, and a PWM *w* of length *k*, find the maximum likelihood set of *k*-mers which correspond to the WMM.
  - Given TF binding specificities, can we find the TF binding sites ?
- De novo (unsupervised) motif detection : Given a set of sequences, find the most overrepresented set of *k*-mers and the corresponding WMM *w*
  - Given a set of genes with similar expression (putatively co-regulated), can we find the TF binding sites common to them and the specificity of the corresponding TF?

Multi-species data pooling

- Simply pool together regulatory regions in related species
  - Bacterial DNA motifs, McGuire *et al* **Gen Res** 2000 & Gelfand *et al* **Nuc Ac Res** 2000
  - Hunchback TFBS in Drosophila species demonstration :

CACCACTTTTTATGCCGAGTTAAT **D. melanogaster**

GGTTTTTTCGATTCAATCGGTATA **D. yakuba**

AGTTAGCGTTTACCCTATTTTTTAC **D. persimilis**

GCATTTATCCTCTTTTTATAAGCTT **D. mojavensis**

- What could be problematic with this approach ?

Multi-species data pooling

- Biases analyses towards motifs in a bunch of closely related species – no explicit phylogenetic information used
- No distinction between paralogs and orthologs
- Variation in number of binding sites in orthologous CRMs much less than in CRMs of coregulated genes in same species
  - Signals in one species may be drowned out by cross species signals, or vice versa

Orthologous sequence analysis

- What if the sequences are orthologous ?

```
D. melanogaster     CTTTACGTATTTTAGTTATCGAGTTTATCTTCTGCTTGCTATCTCGCGC
D. yakuba           T--TACGTATTTTAGTTATCGAGTTTATCTTCTGCTTGCTATCTCGCGC
D. persimilis       GTTTACGTATTTTAGTTATCGAGTTTATCTTTCGCTT------TCTCGC
D. mojavensis       CTTTACGTATTTGAGTTATCAACTTTGT--TTTGCTT--TGCTTTTCGC
```

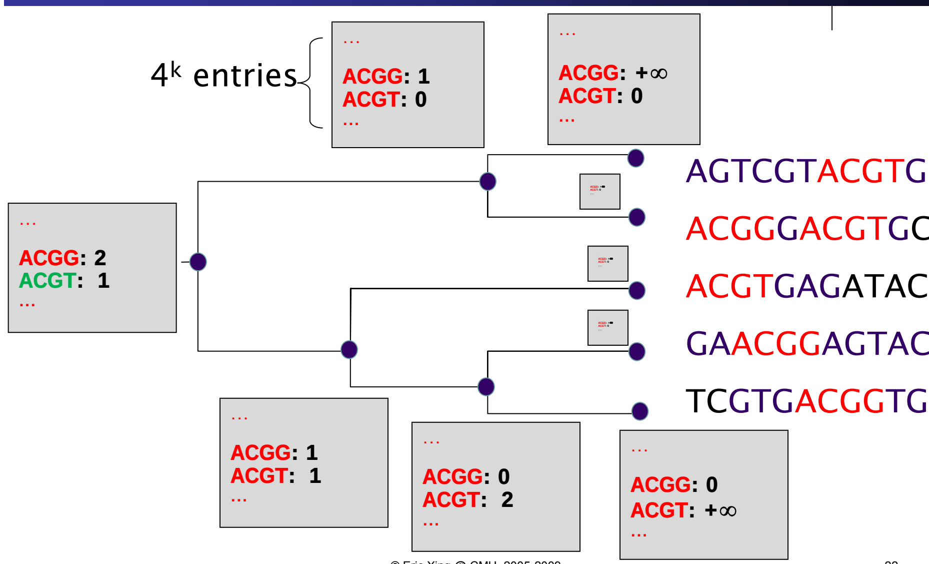
- Functional regions like TFBSs are more conserved than background
- Phylogenetic dependencies between orthologs may be modelled to get more accurate scores for P(data | model)
- Paralogous sequence analysis also possible

Chronology : from one to many

Method	Single species	Multispecies
Combinatorial	Waterman (1985)	FootPrinter (2002)
LRT-like score threshold	Staden-PWM (1989)	rVista (2002) PhyloCon (2003)
Explicit mixture models	MEME (1994)	EMnEM (2004) PhyME (2004) CSMET (2008)
Gibbs Sampling	GMS (1993/1995) BioProspector (2001) AlignACE (2000)	Motif Sampler+ (2000) CompareProspector (2004) PhyloGibbs (2005)
HMM+	Cister (2001) HMDM (2002) LOGOS (2003) BayCis (2008)	PhyloHMM (2004 –gene / 2008 - motif) PhyME (2004) MORPH (2008) CSMET (2008)
Ensemble models	EMD (2006)	-

- First usage of term phylogenetic footprinting : Tagle *et al*, J Mol Biol 1988 : Regulatory regions of paralogous gamma and epsilon globin genes in Galago
- Google scholar hits for "phylogenetic footprinting"
  - 1988 – 1990 : 11 hits
  - 1991 – 2000 : 141 hits
  - 2001 – 2009 : 1850 hits
- Gibbs Sampling particularly easily adapted to incorporate phylogenetic footprinting

FootPrinter: going combinatorial



Footprinter

- For each node from the leaf up
  - Fill up a parsimony table W for each k-mer
- If the node n is a leaf
  - If the word is present in corr. string, $W[k]^{(n)} = 0$ else $W[k]^{(n)} = +\infty$
- Else
  - $W[k]^{(n)} = \sum_{V \in child(n)} \min_{t \in k\text{-mer}} (W[t]^{(v)} + d(t,k))$
- Choose most parsimonious k-mer(s) over the tree and alignment from table at root
- Time complexity = $O(n\, 4^{2k})$ for n species for fixed topology

Motif Sampler + footprinting

- A simple idea that works reasonably well
- Wasserman et al, 2000
  - Given an input multiple alignment **A**
  - Compute a score for conservation across the alignment
  - Filter out all regions of the alignment with a score below a threshold *t*
  - Perform Gibbs Motif Sampling on the remaining alignment **A'**

```
D. melanogaster    CTTTACGTATTTTAGTTATCGATTTTATTTTCTGCTTGCTATCTCGCGC
D. yakuba          T--TACGTATTTTAGTTATCGATTTTATTTTCTGCTTGCTATCTCGCGC
D. persimilis      GTTTACGTATTTTAGTTATCGATTTTAGTTTTCGCTT------TCTCGC
D. mojavensis      CTTTACGTATTTGAGTTATCAATTTTGGTTTTTGCTT--TGCTTTTCGC
```

Motif Sampler + footprinting

- A simple idea that works reasonably well
- Wasserman et al, 2000
  - Given an input multiple alignment **A**
  - Compute a score for conservation across the alignment
  - Filter out all regions of the alignment with a score below a threshold *t*
  - Perform Gibbs Motif Sampling on the remaining alignment **A'**

```
D. melanogaster   CTTTACGTATTTTAGTTATCGATTTTATTTTCTGCTTGCTATCTCGCGC
D. yakuba         T--TACGTATTTTAGTTATCGATTTTATTTTCTGCTTGCTATCTCGCGC
D. persimilis     GTTTACGTATTTTAGTTATCGATTTTAGTTTTCGCTT------TCTCGC
D. mojavensis     CTTTACGTATTTGAGTTATCAATTTTGGTTTTTGCTT--TGCTTTTCGC
```


Motif Sampler + footprinting

- A simple idea that works reasonably well
- Wasserman et al, 2000
  - Given an input multiple alignment **A**
  - Compute a score for conservation across the alignment
  - Filter out all regions of the alignment with a score below a threshold *t*
  - Perform Gibbs Motif Sampling on the remaining alignment **A'**

```
D. melanogaster   CTTTACGTATTTTAGTTATCGATTTTATTTTCTGCTTGCTATCTCGCGC
D. yakuba         T--TACGTATTTTAGTTATCGATTTTATTTTCTGCTTGCTATCTCGCGC
D. persimilis     GTTTACGTATTTTAGTTATCGATTTTAGTTTTCGCTT------TCTCGC
D. mojavensis     CTTTACGTATTTGAGTTATCAATTTTGGTTTTTGCTT--TGCTTTTCGC
```


- Motifs sampled according to:

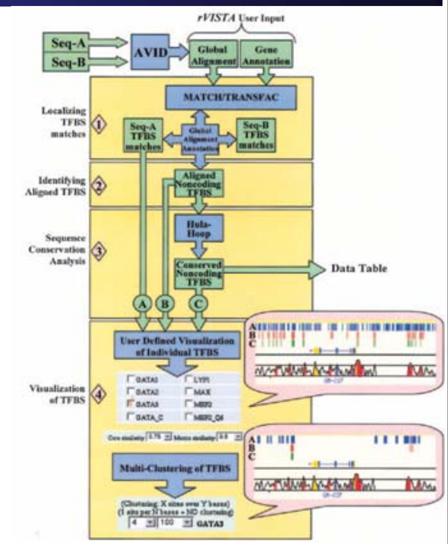
$$\frac{\prod_{k=a}^{a+w} pm_{k-a,R_k}}{\prod_{k=a}^{a+w} p0_{k,R_k}}$$

Score (likelihood) under PWM (motif) scenario

Score (likelihood) under background scenario

rVISTA

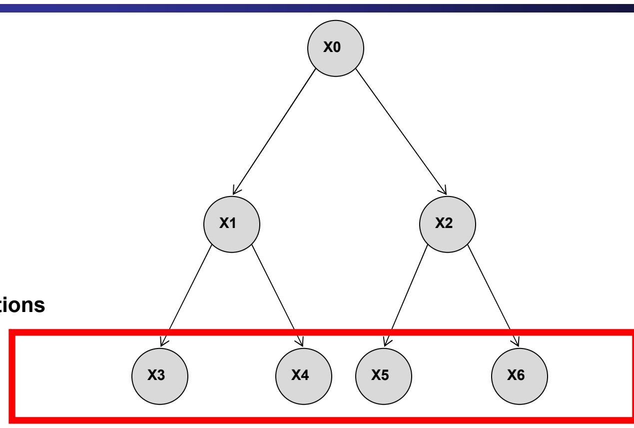
- Select motifs with a PWM score greater than a threshold
- Screens for motifs above a certain threshold for nucleotide conservation
- Two step screening a common way to capture both overrepresentation & conservation
  - Loots et al, Gen Res 2002 [rVISTA]
  - Kellis et al, Nature 2003
  - Wang & Stormo, Bioinf 2003 [PhyloCon]



Loots et el. Gen Res 2002

Phylogenetic model: recap

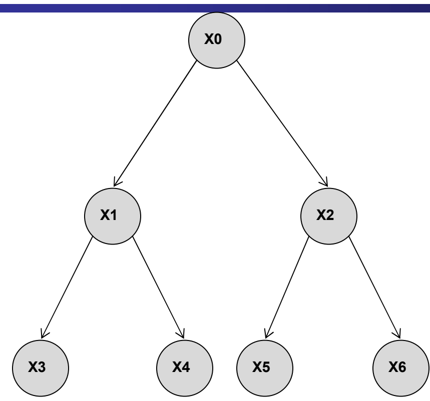
Observations



$P(D|M) = P(X0, X1, X2, X3, X4, X5, X6 \mid \tau, \beta, \theta, \pi)$

$= \sum_{X0, X1, X2} P(X3|X1; tree)\ P(X4|X1; tree)\ P(X5|X2; tree)\ P(X6|X2; tree)\ P(X2|X0; tree)\ P(X1|X0; tree)\ P(X0|\ tree)$

Phylogenetic model: recap



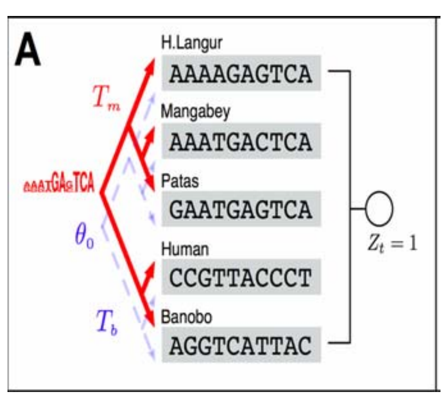
- **Topology** – how the observations are "tied together": τ
- **Branch lengths** – the length for which the CTMP runs: β
- **Parameters of CTMP** – characterizing the substitution model: θ
- **Distribution at root** - maybe stationary dist of CTMP : π

P(D|M) = P(X0, X1, X2, X3, X4, X5, X6 | τ, β, θ, π)

$= \sum_{X0, X1, X2}$ P(X3|X1; tree) P(X4|X1; tree) P(X5|X2; tree) P(X6|X2; tree) P(X2|X0; tree) P(X1|X0; tree) P(X0| tree)

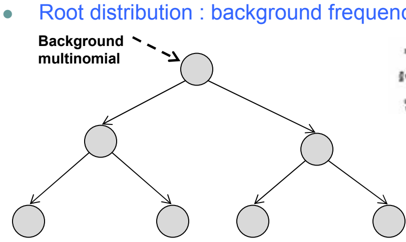
EMnEM: model based approaches

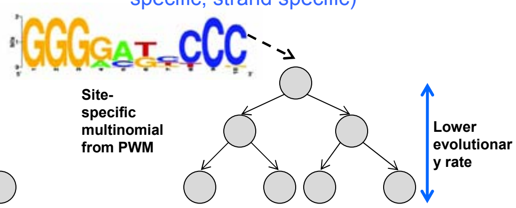
- Mixture model
- Each block of k-mer could be generated from a background model with probability **1-**π_m or from a motif model with probability π_m
- Bernoulli draw for the mixture indicator
- In the spirit of MEME



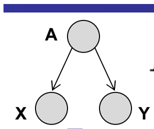
Function specific phylogenetic models

- Background model T_b
  - Topology invariant unless evidence otherwise
  - Substitution matrix invariant unless evidence otherwise
  - Branch lengths longer than functional sites
  - Root distribution : background frequency
- Motif site-specific, strand-specific model $T_{m,k,+/-}$
  - Topology invariant unless evidence otherwise
  - Substitution matrix invariant unless evidence otherwise
  - Branch lengths shorter than background sites
  - Root distribution : from PWM (site specific, strand specific)





EMnEM: Expectation maximization on mixtures of phylogenies



$$L = \prod_{i=0}^{N-w} \sum_{m_i} p(m_i) \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(X_k, Y_k \mid A_{kb}, m_i) p(A_{kb} \mid m_i)$$

Mixture parameter ($p(m_i)$); Ancestral base (A_{kb}); PWM ($p(A_{kb} \mid m_i)$)

- E-step :
  - Mixture parameter:

$$\langle m_i \rangle = p(m_i | X, Y) = \frac{p(m_i) p(X, Y | m_i)}{p(X, Y)}$$

$$p(X, Y | m_i) = \prod_{k=i}^{i+w-1} \sum_{b=0}^{3} p(X_k, Y_k | A_{kb}, m_i) p(A_{kb} | m_i)$$

$$p(X, Y) = \sum_{m_i} p(X, Y | m_i) p(m_i)$$

  - Ancestral nucleotide:

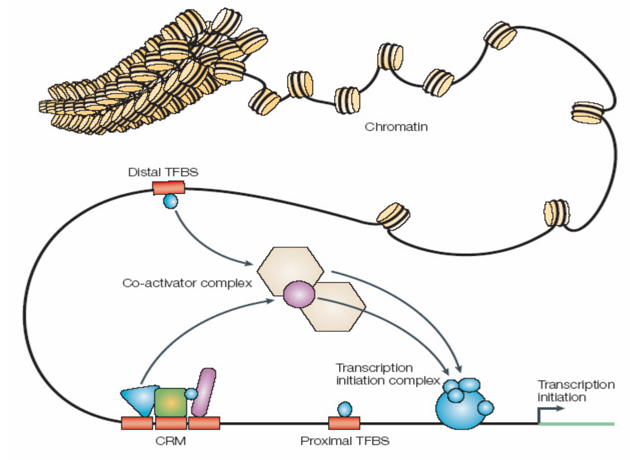
$$\langle A_{ib} \rangle = p(A_{ib} | X_i, Y_i) = \sum_{m_i} p(A_{ib} | X_i, Y_i, m_i) p(m_i) = \sum_{m_i} \frac{p(A_{ib}) p(X_i, Y_i | A_{ib}, m_i)}{p(X_i, Y_i | m_i)} p(m_i)$$

- M-step : $\langle \ln L_c \rangle = \sum_{i=0}^{N-w} \sum_{m_i} \langle m_i \rangle \left[\ln \pi_m + \sum_{k=i}^{i+w-1} \sum_{b=0}^{3} \langle A_{kb} \rangle (\ln p(X_k, Y_k \mid A_{kb}, m_i) + \ln f_{mkb}) \right]$

$$\frac{\partial \langle \ln L_c \rangle}{\partial \pi_m} = 0, \quad \frac{\partial \langle \ln L_c \rangle}{\partial f_{mkb}} = 0 \text{ and } \frac{\partial \langle \ln L_c \rangle}{\partial \alpha_{mk}} = 0$$

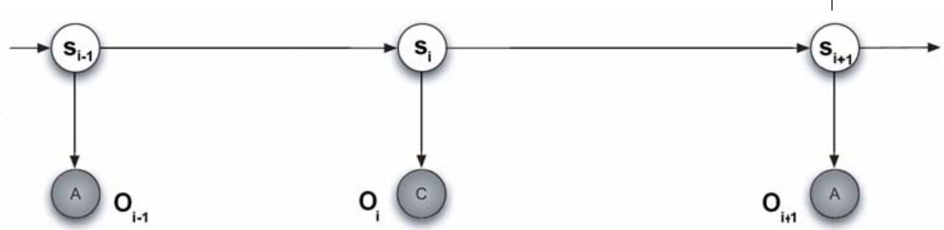
CRM: putting the pieces together

- HMMs !



Courtesy: Simone Scalabria

A vanilla HMM …

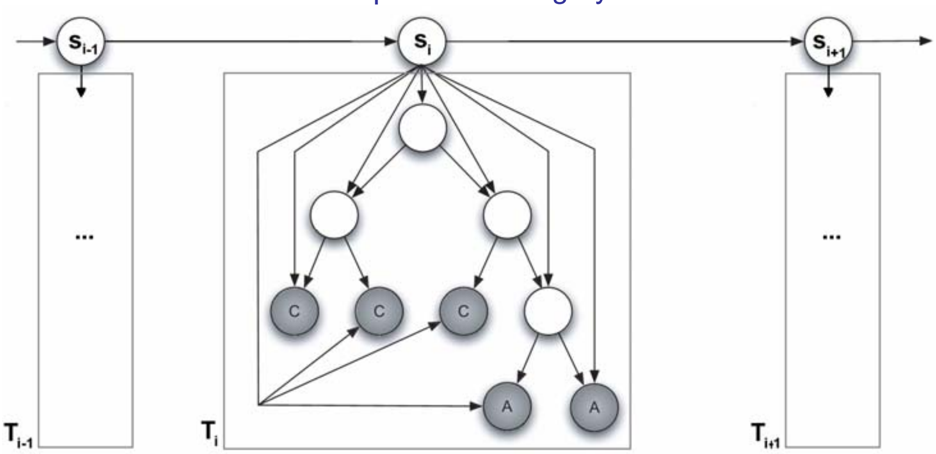


ACATTGCCATACCATAATCCTTAATT…

- Emits a symbol at every discrete step
- A run of the HMM outputs a sequence
- PhyloHMM outputs a vector of characters
- A run of the PhyloHMM outputs a multiple sequence alignment

Phylo-HMM

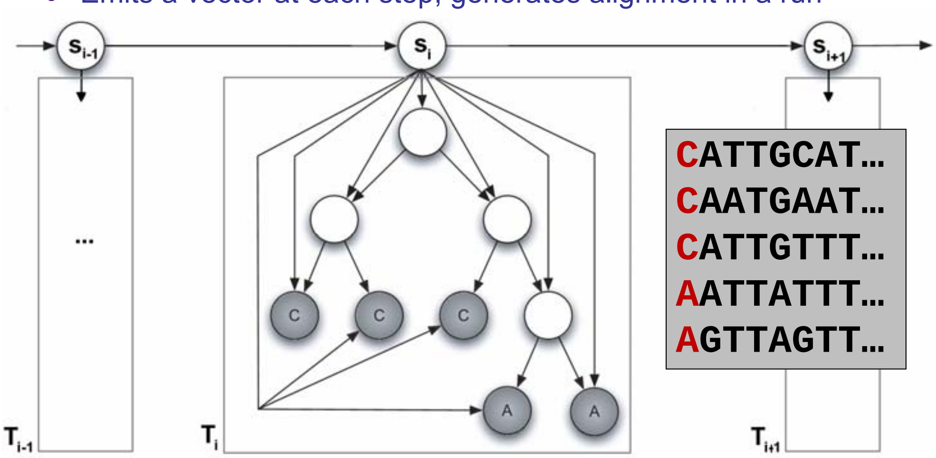
- The emission vector O_i is shaded in gray



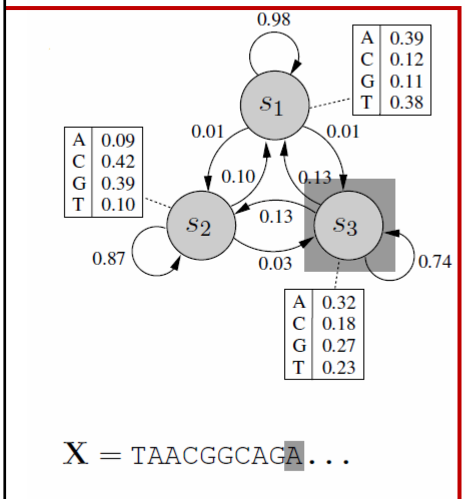
Courtesy: McAuliffe

Phylo-HMM

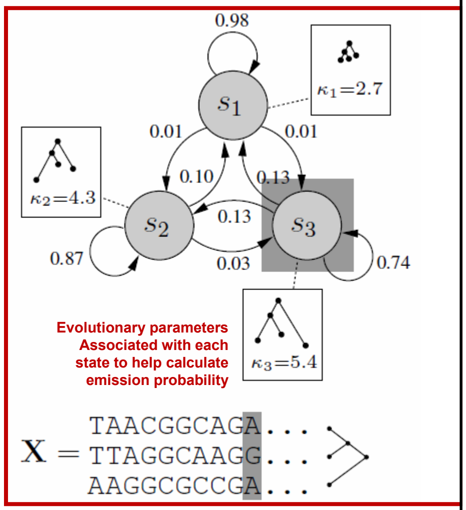
- Emits a vector at each step, generates alignment in a run



State space comparisons



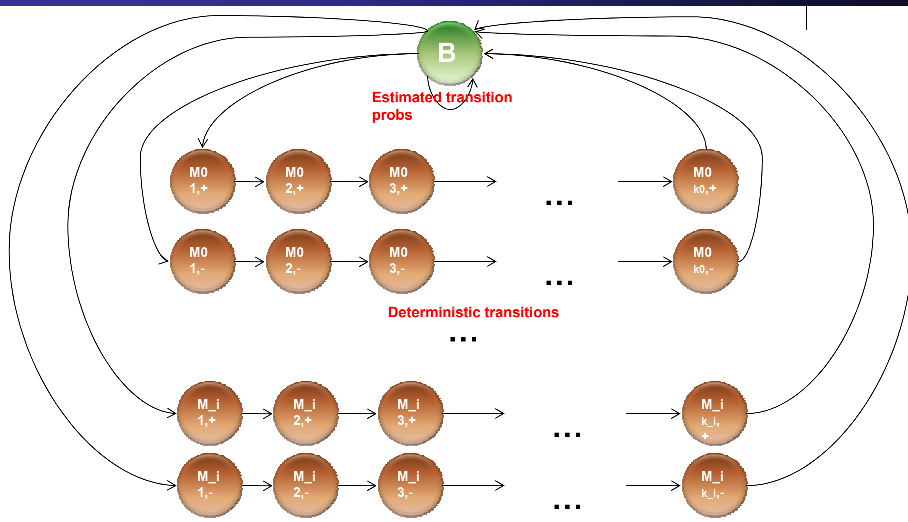
Standard HMM



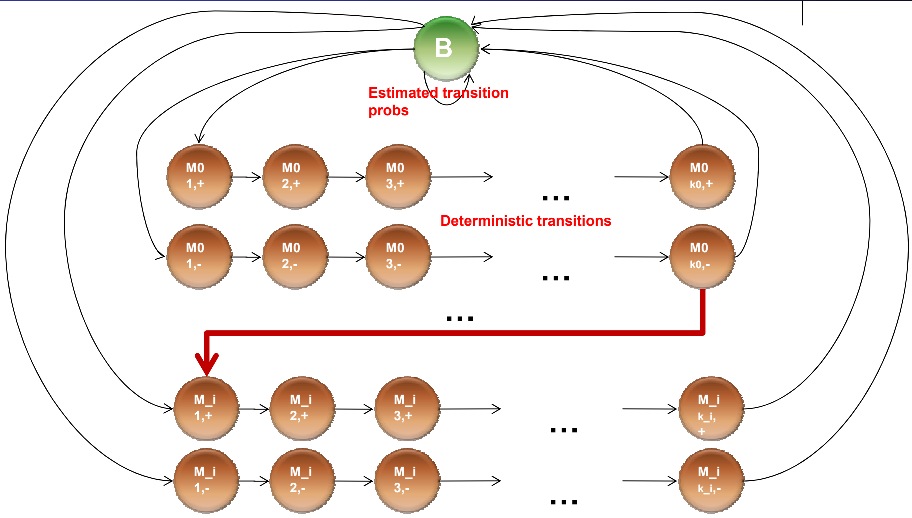
Phylo HMM

Courtesy: Siepel

HMM state space for motif finding



More realism, more parameters



Phylo-HMM

- A normal HMM, except the emission probabilities are a multinomial distribution over the space of **[ATGC]^n**, n being the number of sequences in the alignment
- 4^n emission probabilities can be pre computed
- But usually calculated on the fly using Felsenstein's Pruning Algorithm - a special case of the GM Belief Propagation Algorithm on trees
  - Siepel & Haussler, RECOMB 2004, for gene finding
  - Ray et al. PLoS CompBio 2008, adapted for motif finding

Analogy with HMM

(S, ψ, N, n, T, b)

Set of states → S

Set of associated phylogenetic models → ψ

Set of alphabets → N

No of species → n

Transition probs → T

Initial probs → b

- Emission probability = $P(O_i \mid S_i = s)$

  = $P(O_i \mid \text{phylogenetic model}_s)$

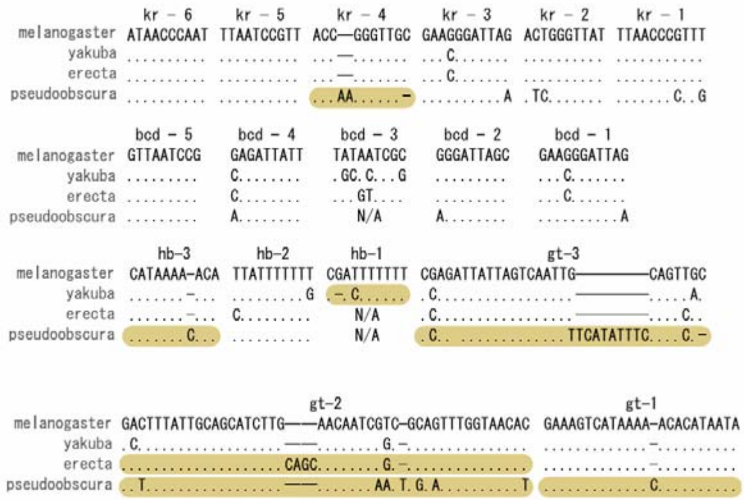
  = $P(\text{Alignment column}_i \mid \psi_s)$

  => Calculate using the Pruning Algorithm

- Apply standard Viterbi (maximize joint) or posterior decoding on the Forward-Backward matrix
- Baum-Welch algorithm (E-M) for unsupervised settings
- Exactly analogous to single species motif finding

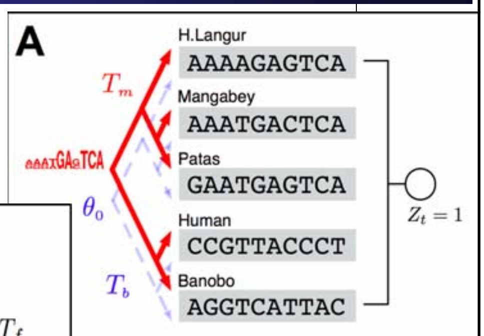
Missing motifs

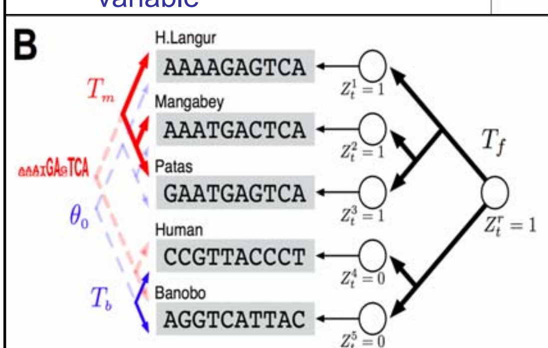
Functional turnover



CSMET : Phylogenetic mixtures of phylogenies

- Mixture models for evolutionary model selection (EMnEM) →
- Bernoulli draw for mixture variable





- What if the mixture variables are phylogenetically related ?
- Output of a "functional" phylogeny

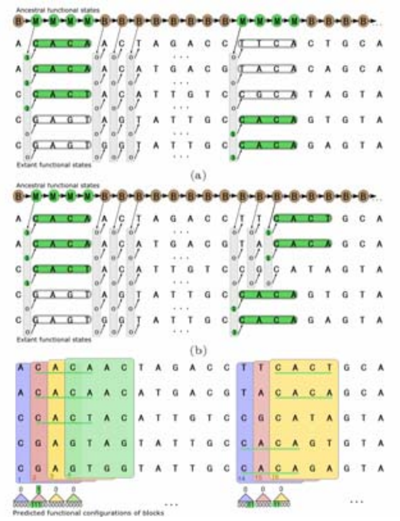
CSMET

- CSMET-HMM :
  - An HMM with emission vector [A,T,G,C]*
  - Each vector is the output of a generative process involving a mixture of trees
  - Mixture indicator variables themselves generated by a phylogeny
  - Similar scheme to PhyloHMM, except for calculating emission probs
- Schematic of generation and ML inference
- CSMET-HMM : (S, TN, TF, N, F, n, T, b)

Corr character sets

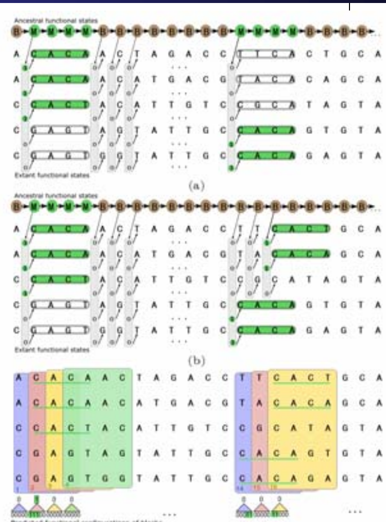
Set of nucleotide phylogenetic models corr to each annotation

Set of functional phylogenetic models corr to each state in S



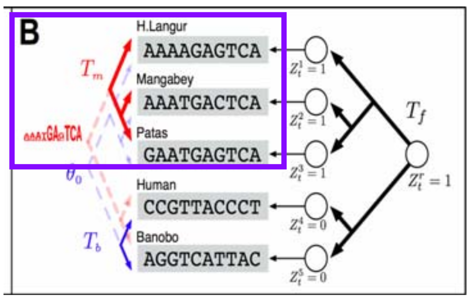
CSMET

- To calculate emission probabilities:
  - Calculate likelihoods of nucleotide data for each subtree of the nucleotide phylogeny
  - Calculate likelihood of functional indicators for the functional phylogeny
  - Putting the likelihoods together using conditional independences
  - Marginalize out hidden variables
- The rest would be analogous to an HMM !



Likelihoods on partial phylogenies

- Marginalize out observed nucleotides present in parts of the phylogeny we are not interested in
- Turns out to be equivalent to calculating the likelihood of the data on the subtree !



$$P(A'_l \mid T'^{(l)}) = \sum_{A''_l} P(A'_l, A''_l \mid T^{(l)}) =$$

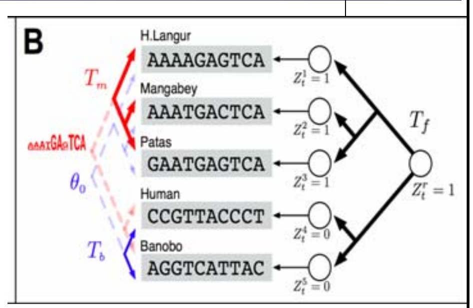
$$\sum_{A''_l} \sum_{\mathbf{v}_{1:K'}} P(\mathbf{V}_{1:K'} = \mathbf{v}_{1:K'}, \mathbf{V} = A''_l, \mathbf{V}' = A'_l)$$

CSMET: toolkit for calculations

- Calculating likelihoods on the nucleotide phylogeny and functional phylogeny

Nucleotide phylogeny :
F84 model – simplest arbitrary stationary distribution

$$Q_N = \begin{pmatrix} * & (1+\kappa/\pi_Y)\iota\pi_C & \iota\pi_A & \iota\pi_G \\ (1+\kappa/\pi_Y)\iota\pi_T & * & \iota\pi_A & \iota\pi_G \\ \iota\pi_T & \iota\pi_C & * & (1+\kappa/\pi_R)\iota\pi_G \\ \iota\pi_T & \iota\pi_C & (1+\kappa/\pi_R)\iota\pi_A & * \end{pmatrix}$$



Functional phylogeny
Jukes Cantor model

$$P_F = \begin{pmatrix} \frac{1}{2}+\frac{1}{2}e^{-2\beta} & \frac{1}{2}-\frac{1}{2}e^{-2\beta} \\ \frac{1}{2}-\frac{1}{2}e^{-2\beta} & \frac{1}{2}+\frac{1}{2}e^{-2\beta} \end{pmatrix}$$

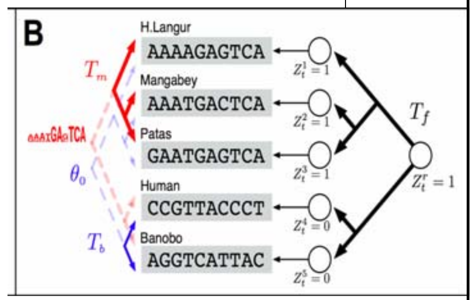
Likelihoods on partial phylogenies

$$P(A'_l \mid T'^{(l)}) = \sum_{A''_l} P(A'_l, A''_l \mid T^{(l)}) =$$

$$\sum_{A''_l} \sum_{\mathbf{v}_{1:K'}} P(\mathbf{V}_{1:K'} = \mathbf{v}_{1:K'}, \mathbf{V} = A''_l, \mathbf{V}' = A'_l)$$

CSMET : emission probabilities

- Emission prob : Prob of block surrounding particular aligned site
- Again, analogous to an HMM, with one twist : Z_i s not observed



Joint Probability for an instantiated block

$$P(\mathbf{A}_t, \mathbf{z}_t, z_t^r) = P(\mathbf{A}_t | Z_t = \mathbf{z}_t, T_m, T_b) P(Z_t = \mathbf{z}_t | Z_t^r = z_t^r, T_f)$$
$$P(Z_t^r = z_t^r) = P(\mathbf{A}'_t | T'_m) P(\mathbf{A}''_t | T'_b) P(\mathbf{z}_t | z_t^r, T_a) P(z_t^r).$$

Conditional probability for the block

$$P(\mathbf{A}_t | Z_t = \mathbf{z}_t, T_m, T_b) = P(\mathbf{A}'_t | T'_m) P(\mathbf{A}''_t | T'_b) = \prod_{l=1}^{L} P(A'_l(t) | T'_m(l)) P(A''_l(t) | T'_b).$$

Emission probability for the block (marginalized)

$$P(\mathbf{A}_t | z_t^r) = \sum_{\mathbf{z}_t} P(\mathbf{A}_t, \mathbf{z}_t | z_t^r) = \sum_{\mathbf{z}_t} P(\mathbf{A}'_t(\mathbf{z}_t) | T'_m(\mathbf{z}_t)) P(\mathbf{A}''_t(\mathbf{z}_t) | T'_b(\mathbf{z}_t)) P(\mathbf{z}_t | T_a, z_t^r),$$

Chronology : aspects of footprinting

- Footprinting + Gibbs Sampling
  - **Motif Sampler+ : 2 species alignment**
  - **CompareProspector : Pairwise alignment**
  - **PhyloGibbs : Multiple alignment**
- Footprinting + HMM
  - **PhyloHMM : Emission of HMM generated by a CTMP phylogenetic tree, no tolerance for functional turnover**
  - **PhyME**
  - **CSMET : Emission of HMM generated by a mixture of CTMP phylogenetic trees, explicit tolerance for functional turnover**
- **Footprinting + alignment**
  - **OrthoMEME**
  - **MORPH**

Can we do even better ?

- Footprinting improves with
  - More knowledge about the functional component we are searching for : what to look for in a single species
  - More knowledge about how it evolves : what to look for in related species
- We know a lot about both aspects for protein coding regions, or genes
- Initial footprinting algorithms on genes and proteins

Evolution of codons

- Genes evolve at a level of higher granularity
  - Nucleotide
  - Codon
- HMM states corresponding to codons
- How to choose priors for transition probabilities ?

Seconed Position

First Position	U code	U Amino Acid	C code	C Amino Acid	A code	A Amino Acid	G code	G Amino Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Courtesy: Bioephemera.com

Incorporating evolutionary processes

- Selection
- Transition probabilities can reflect
  - Synonymous transitions more frequent than non synonymous ones
  - How much more frequent ?
  - Selection parameters estimated from data

Seconed Position

First Position	U code	U Amino Acid	C code	C Amino Acid	A code	A Amino Acid	G code	G Amino Acid	Third Position
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

Courtesy: Bioephemera.com

Summary

- Use genomic representation of functional component
- Use evolutionary models of functional component
- Can be used for non-sequence data too :
  - Gene regulatory network
  - Expression levels : microarray data

Acknowledgments

- **Serafim Batzoglou**: for some of the slides adapted or modified from his lecture slides at Stanford University
- **Lior Pachter'**: for some of the slides modified from his lectures at UC Berkeley
- Acknowledgements for some graphics on corresponding slides