I529: Machine Learning in Bioinformatics (Spring 2017)

HMM for modeling aligned multiple sequences: phylo-HMM & multivariate HMM

Yuzhen Ye
School of Informatics and Computing
Indiana University, Bloomington
Spring 2017

Content

- Consideration of two aligned sequences
  - TWINSCAN
- Generalization to approaches for modeling multiple, aligned sequences
- Phylo-HMM
  - Definition of phylo-HMM
  - Pruning algorithm for calculating the probability of a column (of the multiple alignment) given a phylogenetic model
- Multivariate HMM

TWINSCAN model for gene finding in human and mouse genomes

- TWINSCAN is an augmented version of the GHMM used in Genscan.
- Input: syntenic regions in human and mouse genome
  - assumption: the gene structure (exon/intron boundaries) is *conserved* in these two genomes, and the conserved boundaries are aligned *precisely* in the pairwise genome alignment
- Output: the annotation of gene structure

TWINSCAN model

```
Human:      ACGGCGACUGUGCACGU
Mouse:      ACUGUGAC GUGCACUU
Alignment:  ||:|:|||-||||||:|
```

$\underbrace{\quad\quad\quad}_{\text{intron}}\ \underbrace{\quad\quad\quad\quad}_{\text{exon}}$

Why using multiple sequences?

1)multiple sequences gives strong signal (e.g. if a sequence profile of a splicing site is preserved, it is more likely to be a true splicing site); and more importantly,
2)the conservation pattern can be used to discriminate exons and introns: exons tend to be more conserved than introns.

TWINSCAN algorithm

The key idea: converting a pairwise alignment into a single observation sequence on an expanded alphabet

1. Align the two sequences (e.g. from human and mouse genomes);
2. Use the similar hidden states as Genscan;
3. Design a **new "alphabet"** for observation symbols:
   4 x 3 = 12 symbols:
   Σ = { A-, A:, A|, C-, C:, C|, G-, G:, G|, U-, U:, U| }
   gap (-), mismatch (:), match (|)

Example

```
Human:      ACGGCGACUGUGCACGU
Mouse:      ACUGUGAC GUGCACUU
Alignment:  ||:|:|||-||||||:|
```

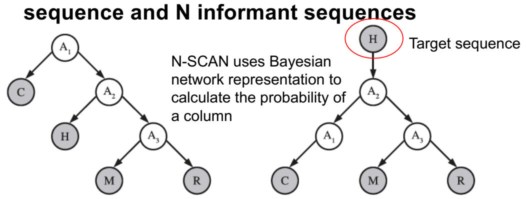
Input to TWINSCAN HMM (observation sequence)

A| C| G: G| C: G| A| C| U- G| U| G| C| A| C| G: U|

Recall, $\mathbf{e_E(A|) > e_I(A|)}$ **and** $\mathbf{e_E(A-) < e_I(A-)}$
Likely exon will be annotated for the entire region

N-SCAN

- GHMM in TWINSCAN outputs a target genomic sequence and a **conservation sequence**
- **GHMM in N-SCAN outputs a target genomic sequence and N informant sequences**



Target sequence

N-SCAN uses Bayesian network representation to calculate the probability of a column

FIG. 1. A phylogenetic tree relating chicken (C), human (H), mouse (M), and rat (R). The graph can also be interpreted as a Bayesian network (**left**). The result of transforming the Bayesian network (**right**).

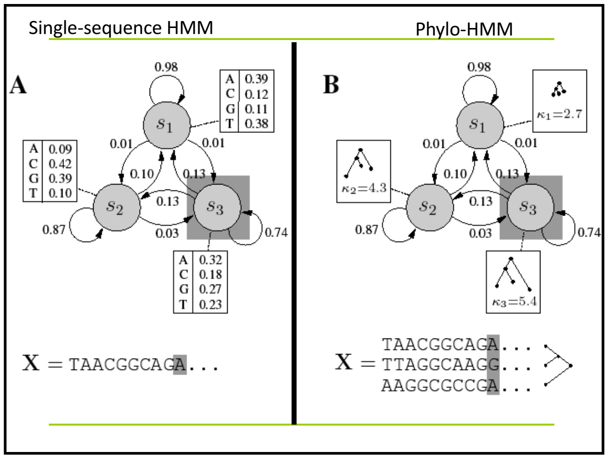
Using multiple alignments to improve gene prediction. JCB, 2006

HMM for multiple aligned sequences

- Strategy 1: converting the alignment of multiple observation sequences into *one* observation sequence
  - New alphabet representing *convoluted* observation symbols, e.g., the TWINSCAN model
  - Not practical for *n* sequences: the size of the alphabet grows exponentially with O(2^n).
- Strategy 2: employing another probabilistic model to emit multiple aligned observation sequences simultaneously (Phylo-HMM model)
- Strategy 3: emitting multiple aligned observation sequences simultaneously but independently, each following a different emission probability distribution (multivariate HMM)

Phylo-HMMs: model multiple alignments of syntenic sequences

- A phylo-HMM is a probabilistic machine that generates a multiple alignment, column by column, such that each column is defined by a phylogenetic model
- Unlike single-sequence HMMs, the "emission" probabilities of phylo-HMMs are complex distributions defined by **phylogenetic models**
- Molecular evolution can be viewed as a combination of two Markov processes
  - One operates in the dimension of space (along a genome)
  - One operates in the dimension of time (along the branches of a phylogenetic tree)
- Phylo-HMMs combine phylogeny and HMM



Phylo-HMMs: formal definition

A phylo-HMM can be specified as $\theta = (S, \psi, A, b)$,

1) $S = \{S_1, S_2, \cdots, S_M\}$, a set of states

2) $\psi = \{\psi_1, \psi_2, \cdots, \psi_M\}$, a set of associated phylogenetic models

3) $A = \{a_{jk}\}(1 \leq j, k \leq M)$, a matrix of state-transition probabilities

4) $b = (b_1, \cdots, b_M)$, a vector of state-initial probabilities

a_{jk} is the conditional probability of visiting state k at some site i given that state l is visited at site $i-1$. b_j is the probability that state j is visited first.

Questions we can ask using phylo-HMM

A path through the phylo-HMM is a sequence of states $\phi = (\phi_1, \cdots, \phi_M)$, such that $\phi_i \in \{1, \cdots, M\}$ for all $1 \leq i \leq L$.

The joint probability of a path and an alignment is,

$$p(\phi, X|\theta) = b_{\phi_1} P(X_1|\psi_{\phi_1}) \prod_{i=2}^{L} a_{\phi_{i-1},\phi_i} P(X_i|\psi_{\phi_i})$$

The probability of emitting column *i* given a **phylogenetic model**

The probability of the observation (likelihood) is,

$$p(X|\theta) = \sum_{\phi} P(\phi, X|\theta)$$

Forward algorithm

The most probable (maxmum-likelihood) path,

$$\hat{\phi} = \arg\max_{\phi} P(\phi, X|\theta)$$

Viterbi algorithm

Phylogenetic models

- The different phylogenetic models associated with the states of a phylo-HMM may **reflect different overall rates of substitution** (e.g. in conserved and non-conserved regions), different patterns of substitution or background distributions, or even different tree topologies (as with recombination)

Phylogenetic models

$$\psi_j = (Q_j, \pi_j, \tau_j, \beta_j)$$

HKY model

Q_j : substitution rate matrix

π_j : background frequencies

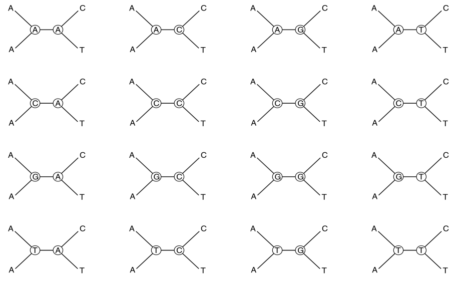
τ_j : binary tree

β_j : branch lengths

$$\mathrm{Q}_j = \begin{pmatrix} - & \pi_{\mathrm{C},j} & \kappa_j\pi_{\mathrm{G},j} & \pi_{\mathrm{T},j} \\ \pi_{\mathrm{A},j} & - & \pi_{\mathrm{G},j} & \kappa_j\pi_{\mathrm{T},j} \\ \kappa_j\pi_{\mathrm{A},j} & \pi_{\mathrm{C},j} & - & \pi_{\mathrm{T},j} \\ \pi_{\mathrm{A},j} & \kappa_j\pi_{\mathrm{C},j} & \pi_{\mathrm{G},j} & - \end{pmatrix}$$

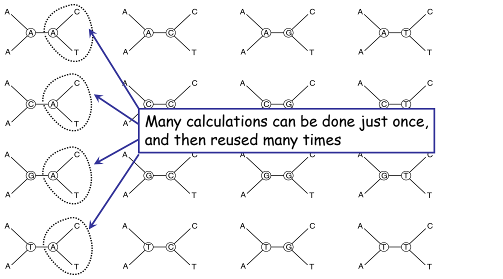
- The model is defined with respect to an alphabet Σ of size *d*
- The substitution rate matrix has dimension *d x d*
- The background frequencies vector has dimension *d*
- The tree has *n* leaves, corresponding to *n* extant taxa in the multiple alignment of observation sequences
- The branch lengths are associated with the tree
- *How to calculate the likelihood of a column given a model?*

Brute force approach to likelihood calculation



Given a column AACT
Brute force strategy: Try all 16 combinations of ancestral states and sum

Pruning algorithm



Many calculations can be done just once, and then reused many times

*The pruning algorithm was introduced by: Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* 17:368-376

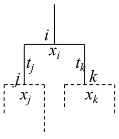
© Paul Lewis

Pruning algorithm

P (X|Ψ) (*X*: a column; *Ψ*: phylogenetic model; subscript not shown for clarity) can be computed recursively from leaves to root.

L_i(x_i): the probability of observing leaves in the subtree *rooted by i*, while the root is assigned to x_i (A, T, C or G),

$$L_i(x_i) = \Big\{\underbrace{\sum_{x_j} p_{x_i x_j}(t_j) L_j(x_j)}\Big\} \times \Big\{\sum_{x_k} p_{x_i x_k}(t_k) L_k(x_k)\Big\}$$



Sum over all possible x_j (A, T, C, or G)

j & *k:* offspring nodes of node *i*

t_j & $t_{k:}$ the branch lengths

$p_{a,b}$(t): the probability to observe a substitution from *a* to *b* within the evolution time of *t*.

The total probability at the root node *r*: $P(X \mid \psi) = \sum_{x_r} L_r(x_r)$

Substitution probabilities

- Pruning algorithm requires the conditional probabilities of substitution *$p_{a,b}$(t)* for all bases *a,b* ∈ Σ and branch lengths *t* ∈ β

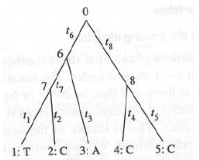
  Another way to denote this probability: $P(b \mid a, t, \psi)$

- It can be computed using a **continuous-time Markov model** of substitution, defined by the rate matrix *Q*

$$P(t) = (e^Q)^t = e^{Qt}$$

(matrix multiplication)

Example



Substitution matrices: P(t=0.1) P(t=0.2)

$$P(0.1) = \begin{array}{cccc} T & C & A & G \\ \hline 0.906563 & 0.045855 & 0.023791 & 0.023791 \\ 0.045855 & 0.906563 & 0.023791 & 0.023791 \\ 0.023791 & 0.023791 & 0.906563 & 0.045855 \\ 0.023791 & 0.023791 & 0.045855 & 0.906563 \end{array}$$

$$P(0.2) = \begin{bmatrix} 0.825092 & 0.084274 & 0.045317 & 0.045317 \\ 0.084274 & 0.825092 & 0.045317 & 0.045317 \\ 0.045317 & 0.045317 & 0.825092 & 0.084274 \\ 0.045317 & 0.045317 & 0.084274 & 0.825092 \end{bmatrix}$$

Branch lengths: $t_1=t_2=t_3=t_4=t_5=0.2$; $t_6=t_7=t_8= 0.1$.

$$P(X_i \mid \psi_j) = P(TCACC \mid T, t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, \kappa)$$
$$= \sum_{x_0}\sum_{x_6}\sum_{x_7}\sum_{x_8}\left[\pi_{x_0} P_{x_0x_6}(t_6) P_{x_0x_8}(t_8) P_{x_6x_7}(t_7) P_{x_7T}(t_1) P_{x_7C}(t_2) P_{x_6A}(t_3) P_{x_8C}(t_4) P_{x_8C}(t_5)\right]$$

Example

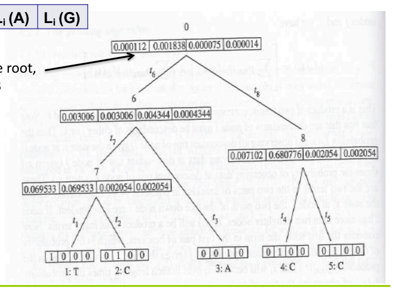
By using pruning algorithm, it can be computed through $L_i(x_i)$, from leaves to root.

L_i(T)	L_i (C)	L_i (A)	L_i (G)

P(TCACC)?

Summing at the root,
P=0.000509843

Initialization



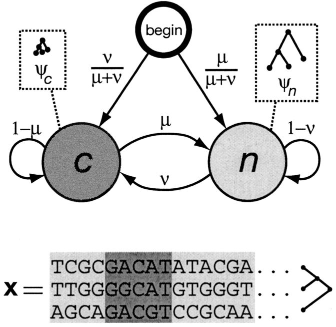
Applications of Phylo-HMMs

- Improving phylogenetic modeling that allow for variation among sites in the rate of substitution (Felsenstein & Churchill, 1996; Yang, 1995)
- Protein secondary structure prediction (Goldman et al., 1996; Thorne et al., 1996)
- Detection of recombination from DNA multiple alignments (Husmeier & Wright, 2001)
- **Comparative genomics (Siepel, et. al. Haussler, 2005)--phastCons**
- Inferring sequence regions under functional divergence in duplicate genes (Huang & Golding, Bioinformatics, 2012)

phastCons

- phastCons is based on a two-state phylogenetic hidden Markov model (phylo-HMM), with a state for conserved regions and a state for nonconserved regions; the free parameters of the model were estimated from a multiple alignment by maximum likelihood, using an EM algorithm.
- Developed for searching for conserved elements in vertebrate genomes, using genome-wide multiple alignments of five vertebrate species (human, mouse, rat, chicken, and *Fugu rubripes*)
  - The predicted (conserved) elements cover roughly 3%–8% of the human genome (depending on the details of the calibration procedure)
  - HCEs (highly conserved elements) are associated with the 3′ UTRs of regulatory genes, stable gene deserts, and megabase-sized regions rich in moderately conserved noncoding sequences. Noncoding HCEs also show strong statistical evidence of an enrichment for RNA secondary structure.

State-transition diagram for the phylo-HMM used by phastCons, which consists of a state for conserved regions (c) and a state for nonconserved regions (n).



Siepel A et al. Genome Res. 2005;15:1034-1050

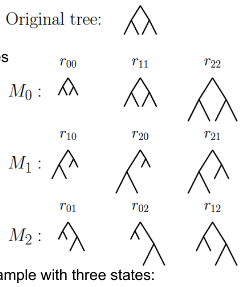
Cold Spring Harbor Laboratory Press

PHAST & RPHAST

- **PHAST and RPHAST: phylogenetic analysis with space/time models (**Brief Bioinform. 2011)
  - http://compgen.bscb.cornell.edu/phast/
  - http://compgen.bscb.cornell.edu/rphast/
- Include phastCons, phastOdds, phyloP, dbless, etc

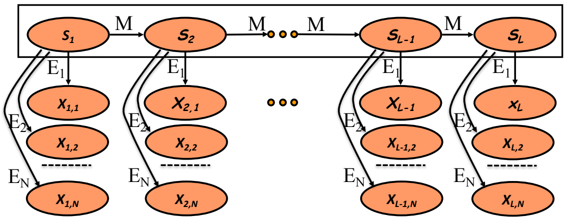
HMMDiverge

Inferring Sequence Regions under Functional Divergence in Duplicate Genes
Bioinformatics (2011)



An example with three states:
M0 (no functional divergence)
M1 & M2 (with functional divergence)

Multivariate HMM



Multivariate HMM (formal definition)

- A multivariate HMM θ has
  - **N** sets of observation symbols, each for one given observation sequence n (n=1, 2, …, N)
  - A set of hidden states
  - Transition probabilities a_{ij}, for any pair of hidden states i and j
  - Initial probabilities $B_j=a_{0j}$ for any hidden states j
  - **N** sets of emission probabilities $e^n{}_s(x_n)$ for the observation symbol being emitted in the *n*th observation sequence from the hidden state *s*.

Multivariate HMM

- Given N observation sequences of the same length L, $X=\{(x_{1,1}\ldots x_{1,L}), \ldots,(x_{N,1}\ldots x_{N,L})\}$ and the hidden state sequence $S=(s_1\ldots s_L)$, the full probability from the multivariate HMM is,

$$P(S,X\mid\theta)=\prod_{j=1}^{L}\left[a_{s_{j-1}s_j}\prod_{n=1}^{N}e_{s_j}\left(x_{n,j}\right)\right]$$

Let $e_{s_i}\left(x_{n,1},\ldots,x_{n,j}\right)=\prod_{n=1}^{N}e_{s_i}\left(x_{n,j}\right)$, the multivariate HMM can be reduced to conventional HMM, except the observation symbol becomes a vector $(x_{n,1}\ldots x_{n,j})$ at position j. The same algorithms for model inference (Viterbi and forward/backward) and learning can be used.