BIOC 218 – Autumn 2009

The Phylo-HMM approach to problems in comparative genomics, with examples.

Keith Bettinger

Introduction

The theory of evolution explains the diversity of organisms on Earth by positing that earlier species of organisms evolved by small changes in their genomes into later ones by a process known as selection. The implication of these gradually integrated differences is that the genomes of two modern-day species have areas of similarity that correspond to each other, both in origin and in function, and areas of divergence, indicating which sequences have changed to produce the new species. This heterogeneity can be exploited to compare the genomes, using the areas of similarity to align the genome sequences, then looking at the areas of divergence to see what makes the species different.

Comparative genomics is the investigation of these evolutionary histories and species differentiations to determine aspects of DNA structure and function. This field examines the genome sequences of related species and attempts to use estimates of how these species evolved from common ancestors to detect important characteristics of the DNA sequences. The overarching principle in comparative genomics is the assumption that selection reduces the ability of evolutionary processes to modify subsequences that are important to the organism's survival.

This paper will outline and explore the foundations and uses of an important tool in comparative genomics known as a phylogenetic hidden Markov model, or phylo-HMM. It will describe how phylo-HMMs relate to other tools used in comparative genomics, define their base characteristics, and contrast three examples of how phylo-HMMs are used. Some work about the statistical power of the phylo-HMM will be described, and the frontier for work on this analytical tool will be outlined.

Related Tools

The phylo-HMM is a descendent of three tools used in comparative genomics:

- Multiple sequence alignment
- Phylogenetic tree
- Hidden Markov model

Each of these tools will be briefly described before defining the phylo-HMM.

The foundation of any comparative genomics study is the *multiple sequence alignment (MSA)* [1]. A multiple sequence alignment (MSA) represents the correspondences calculated among genomic sequences from three or more species, where the matching subsequences in each genome are assumed to have originated from a common ancestor; these correspondences are called homologies. Homologous sequences can be merely similar as well as identical. Differences between sequences in an MSA can be characterized as *point mutations*, where a single nucleotide differs between/across the aligned genomes, *insertions*, where one genome has had a subsequence added to it, and *deletions*, where one genome has had a subsequence removed from it. These homologies and differences then represent the landscape of evolution across these species, suitable for study.

Another important tool in comparative genomics is the *phylogenetic tree* [1]. Phylogenetic trees are undirected simple graphs showing the evolutionary relationships among a set of species. The leaves of the tree are typically modern-day species, and branching points in the tree represent common ancestors to the species on the branches/leaves below. Phylogenetic trees are computed based on the relative homologies between species: pairs of species with more homologies are placed closer together on the tree. Distances within the tree can represent a variety of differing characteristics of the associations; comparative genomics most often uses phylogenetic trees where the branch distances represent the average number of sequence changes between the genomes of the related species (this type of tree is known as a phylogram).

The final major tool in comparative genomics is the *hidden Markov model (HMM)* [1]. An HMM is a specific instance of a finite-state machine, which is a statistical model composed of a set of states, the transitions between them, and the actions performed when the transitions are made. At any given time point, an HMM is in a particular state, and when the time point advances, the HMM transitions from one state to another. Each of the possible states to which a given state can transition has

a certain probability associated with it, so the next state is randomly determined from the previous state. Each state also has a list of actions associated with it to perform when it becomes the current state; these actions also have probabilities associated with them to determine which is run.

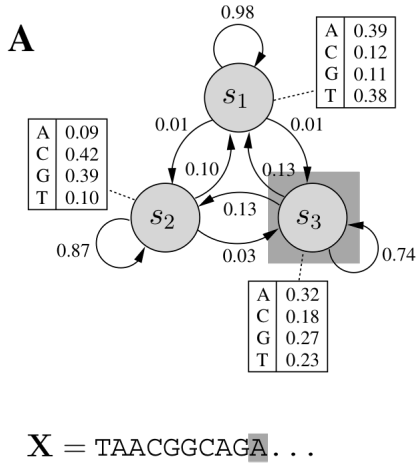

Figure 1: A 3-state HMM representing a genomic sequence.

A example of a simple HMM can be seen in Figure 1 [2], where an HMM which represents a genomic sequence is shown. It has three states, s_1, s_2, and s_3, and three transitions from each state to all the states (including itself). The transition probabilities are above each arc from state to state, and the emission probabilities associated with a state are given in the 4-row table next to the state. The current state is s_3, and it has a 0.13 probability of transitioning to s_1, a 0.13 probability of transitioning to s_2, and a 0.74 probability of transitioning (back) to s_3. Each state emits an nucleotide in the sequence, but each state has a different probability distribution for this emission associated with it. If the current state s_3 transitions to s_1, the HMM will emit an A with a probability of 0.39, a C with a probability of 0.12, a G with a probability of 0.11, and a T with a probability of 0.38.

The random processes represented in an HMM are used in comparative genomics to generalize the properties of DNA sequences. These generalized models can be then applied to sequences to see if they match the model, and to segment the sequences into classes, based on the paths taken through the HMM.

A phylo-HMM combines multiple sequence alignments, phylogenetic trees, and hidden Markov Models into a single framework for representing evolutionary

processes which captures not only the fact that sequences change over time but that the rates of change differ for different subsequences. This integration models evolution of genome sequences across two orthogonal dimension: one dimension along the genome, as its structure changes, and in another dimension across time, as the rate of evolution itself changes [3]. Phylo-HMMs will be described in detail in the next section.

Description of Phylo-HMMs

Phylo-HMMs were first proposed by [3] and [4] to increase the realism of phylogenetic models by allowing for variation across nucleotide sites in the rate of their modification. Soon after, they were applied problems in protein secondary structure prediction [5-7] and the detection of recombination site events [8, 9]. The subsequent proliferation of completed genomes (e.g., [10, 11]) and improved multiple sequence alignment algorithms led to a growth in the application of phylo-HMMs to comparative genomics problems, where they have been used to detect coding sequences [12], conserved sequences [13, 14], gene locations [14-17], and regulatory regions [18, 19].

A phylo-HMM differs from a single-sequence HMM in that the states of the phylo-HMM emit an entire column from a multiple alignment instead of just a single nucleotide. Each row in these columns is no longer a probability distribution over the set of possible nucleotides, but is now the result of executing a separate phylogenetic model for each site in the generated alignment. An example of a phylo-HMM can be seen in Figure 2 [2].

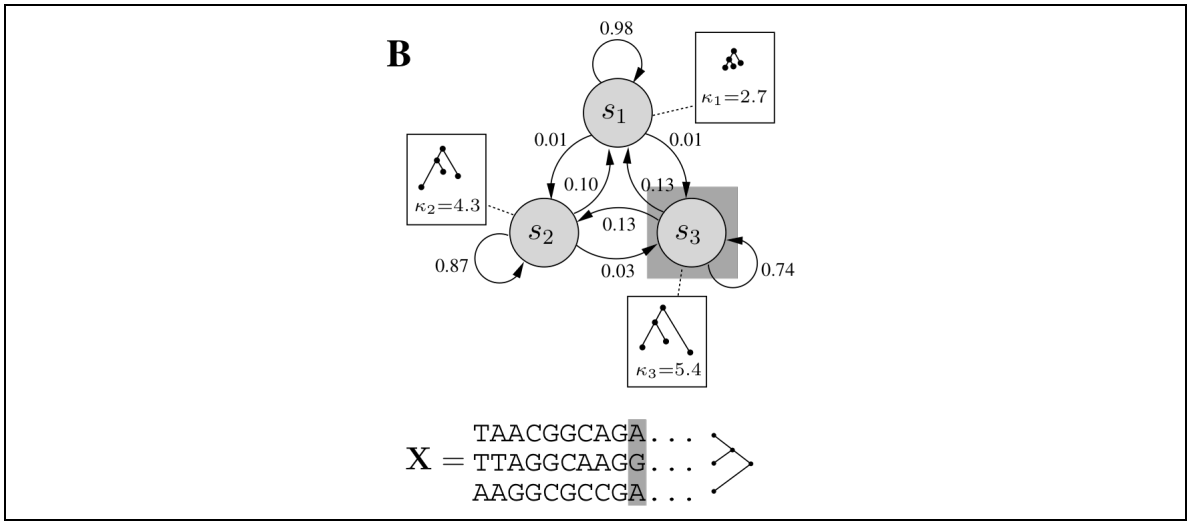

Figure 2: A phylo-HMM for representing a genomic sequence.

Each state in the HMM can be thought of emitting an entire phylogenetic tree, which is then run to produce the characters of a multiple alignment. While each tree has the same topology, each phylogenetic model can vary in its nucleotide substitution rates (see below). Each site in the resulting multiple alignment is associated with a phylogenetic model which operates independently from the other (a constraint that has proven to be unrealistic, see the Future Development section below).

A phylogenetic tree in this context works as follows: the background distribution of nucleotides is used to randomly generate a character, which is assigned to the root of the tree. The evolutionary substitutions from that starting character are then simulated, propagating a (possibly new) nucleotide to the next level of the tree and so on, until the leaves of the tree have nucleotides associated with them. These nucleotides at the leaves of the phylogenetic tree then represent a column in the multiple alignment [2].

The key parameter in a phylogenetic model within a phylo-HMM is the substitution rate matrix, which represents how often a nucleotide n_0 at time t will be substituted at $t + \Delta t$ by nucleotide n_1. These substitutions are the evolution events which transform a genome of one species into a genome of its descendent. The substitution rate matrix is used in conjunction with the branch lengths in its associated phylogenetic tree to calculate the probabilities that a particular nucleotide substitution will be made.

The selection of a substitution rate matrix is constrained by the need to limit the number of its parameters to ease the computational burden in optimizing it. Excluding the self-substitution diagonal columns, this matrix has 12 free parameters, which in practice is often too many to calculate efficiently. Most efforts to create a substitution matrix try to parameterize the entries in the matrix, albeit as little as possible to retain the model's flexibility [17, 20].

Below are descriptions of several of the more popular substitution rate matrix parameterizations. In each, a background distribution of nucleotides is represented by the 4 variables $\pi_A, \pi_T, \pi_C, \pi_G$.

The Jukes-Cantor model (JC69) [21]: The JC69 model has only one parameter, μ, which is the overall constant substitution rate for all nucleotides. This model assumes a uniform background distribution for nucleotides

$\left(\pi_A = \pi_T = \pi_C = \pi_G = 1/4\right)$. With only one parameter, this model is the easiest to calculate, but assuming that all substitutions are equally likely did not fit any known genomic data set [20].

The Kimura model (K80) [22]: The K80 model has two substitution rates: one for transitions (A⇔G or C⇔T) and one for transversions (A⇔C, A⇔T, G⇔C, G⇔T), reflecting the biochemical reality that converting one purine/pyrimidine into another is far simpler than going from a purine to a pyrimidine (or vice versa).

$$\mu_{xy} = \begin{cases} \alpha \text{ for transitions} \\ \beta \text{ for transversions} \end{cases}$$

This model also assumes a uniform background distribution for nucleotides $\left(\pi_A = \pi_T = \pi_C = \pi_G = 1/4\right)$, which also contradicted the existence of many DNA sequences with non-uniform distributions [20].

The HKY85 model [23]: The HKY85 model relaxes the uniform background distribution constraint of the K80 model, allowing $\left(\pi_A \neq \pi_T \neq \pi_C \neq \pi_G\right)$, and multiplies the transition and transversion rate by a background distribution computed from sequence data:

$$\mu_{xy} = \begin{cases} \alpha\pi_y \text{ for transitions} \\ \beta\pi_y \text{ for transversions} \end{cases}$$

This model has five parameters: one for each of the background distributions and one for the ratio of the transitions to the transversions (α/β) [2, 20].

The T92 model [24]: The T92 model modifies the HKY85 model by adding the constraint of strand symmetry. Because of the double-stranded nature of DNA, if one site is modified to become, say, a T, its matching site on the opposite strand must become an A (similarly for C and G). To reflect this restriction, this model modifies its calculated background distributions to set $\pi_C = \pi_G$ and $\pi_A = \pi_T$. These equalities can be derived from a calculation of π_{GC} as follows:

$$\pi_G = \pi_C = \frac{\pi_{GC}}{2}$$

$$\pi_A = \pi_T = \frac{(1 - \pi_{GC})}{2}$$

This model, then, has only two parameters: π_{GC} and the ratio of the transitions to the transversions (α/β) [20, 25].

The choice of which substitution matrix to be used in a phylo-HMM has thought to be a trade-off between simplicity and biological relevance. Indeed, the push to improve phylo-HMM models has generated more and more complex substitution matrices, while improving computing resources have allowed such complex models to become tractable. But, in a study which suggests a reversal to this trend, Fan et al. compared several different substitution matrices and found that the simpler models perform at least as well or better than the more complex models (for more on this topic, see below) [26].

Two Examples of Phylo-HMM Uses

In this section, two examples of how phylo-HMMs are used in comparative genomics studies will be described. For brevity and focus, only the actual phylo-HMM method and its immediate context will be examined, and most of the details regarding the background and results of the study will be elided.

Conserved sequences

Siepel et al. in 2005 incorporated a phylo-HMM in their search for conserved elements in multiple sequence alignments from across a wide variety of species, including MSAs from vertebrates, insects, worms, and yeast [13]. Their work is an example of using phylo-HMMs to segment genomic sequences into classes: conserved elements and non-conserved elements.

The architecture of their phylo-HMM can be seen below in Figure 3.

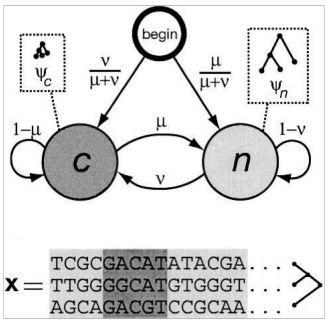

Figure 3: Phylo-HMM state transition diagram for identifying conserved/non-conserved elements.

There are three states in the HMM: a begin state, and two states in the HMM representing conserved ('c') and non-conserved ('n') classes, respectively. The two class states have a self-transition for continuing a sequence as conserved or non-conserved, plus a transition to the other class state. There are two main transition probabilities: μ from conserved to non-conserved, and ν from non-conserved to conserved. The probabilities for the initial transitions from the begin state were calculated from these probabilities, and represent the overall probabilities at equilibrium that the running HMM will be in either of the two states.

Each state has associated with it a phylogenetic tree. Each tree is the same, except the branch lengths of the conserved tree have been scaled by a parameter ρ where $(0 \le \rho \le 1)$, which represents that the average rate of substitution in conserved regions is less than the average rate in non-conserved regions. By this difference in the two states, the sequences generated are segmented into two rate categories.

Like any phylo-HMM, it generates a multiple alignment probabilistically, and labels subsequences with 'conserved' or 'non-conserved' based on which HMM state emitted them. This phylo-HMM is embedded in a program called phastCons, which is available as a stand-alone application [27] or can be run as a track in the UCSC Genome Browser [28].

Siepel et al. note the following oversimplifications with this method of modelling conserved sequences:

- All sites evolve at one of only two evolutionary rates.
- These rates are uniform across the genome.
- Sites evolve independently, depending on whether they are in conserved or nonconserved subsequences.

- The phylogenetic models all have the same:
    - Branch-length proportions
    - Substitution patterns.

The researchers report that some attempts to improve their model in these areas resulted in a great slowdown in system execution without a matching improvement in the output [13].

Transcription factor binding sites

Wong and Nielsen in 2007 applied a phylo-HMM to the detection of transcription factor binding sites (TFBSs) [19]. Gene expression starts when a transcription factor binds to a matching TFBS, usually a sequence of 5-15 base pairs. Their work is an example of using phylo-HMMs to find functional regions in genome sequences.

The overall architecture for the phylo-HMM in this study is shown in Figure 4. Their model is divided into two types of states: silent states and emitting states. In Figure 4, the silent states are ovals/circles, and the emitting states are rectangles. The silent states do not emit nucleotides, and the emitting states output one or more columns when they are entered. The model has three silent states: Start, in which the traversal begins, End, in which it ends, and Intra, which links the emitting states. The model has *n* emitting states: one Background state, which emits a single column of the MSA from the background substitution distribution and the phylogenetic tree, and *n-1* TFBSs states, each of which emits multiple columns as follows. Each TFBS state has a number of sites associated with it, and each site within the TFBS state has its own substitution matrix. Entry into a TFBS state produces a symbol emitted for each associated substitution matrix, as dictated by the phylogenetic tree shared across all the states. As it is implemented in the HMM, each site within a TFBS "state" is a state within the HMM, and each of these substate transitions to the next substate with a probability of 1.

This phylo-HMM implementation is freely available in the program EvoPromoter [29], which is written in Java and uses BioJava [30] libraries.

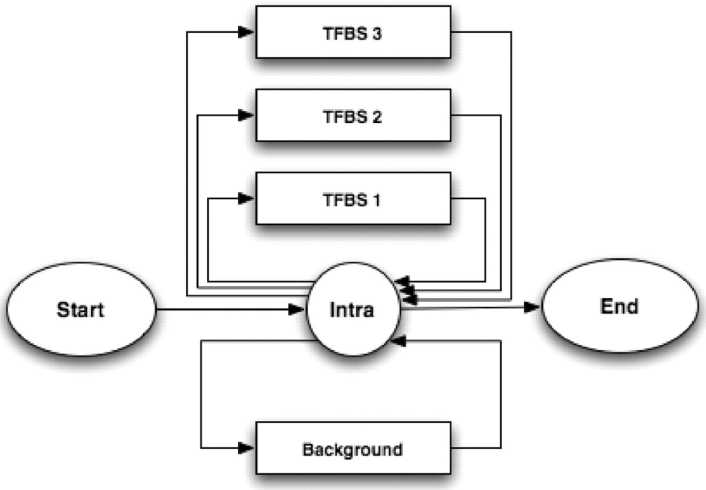

Figure 4: A Phylo-HMM for detecting transcription factor binding sites.

Statistical Power of Phylo-HMMs

The phylo-HMM technique has become prevalent enough to justify examining how varying its characteristics can affect the power of its predictions. Fan et al. [26] have run a series of controlled simulations of phylo-HMMs of the architecture used in [13] to determine which parameters change its predictions most and which aspects of the model are less important.

In their work, Fan et al. asked the following questions:

- Does the topology of the phylogenetic tree matter?
- Can we simplify the substitution model with minimal impact to accuracy?
- Should the phylogenetic tree include distantly related species?
- How many genomes do we need to detect a certain length of conserved sequences at a satisfactory statistical power?
- Is a constant conservation ratio across the genome biologically plausible?

Ideally, these questions could be answered from a theoretical analysis of the mathematical underpinnings of the phylo-HMM itself, but this type of model proved too complex for that type of analysis, so Fan et al. used simulations to assess the impact of parameter changes on model output. They first used sequence data from the promoter regions of four related species (dog, mouse, rat, human) to set the key

parameters of a phylo-HMM. This baseline model was then systematically modified to examine the influence of the key parameters identified.

The following subsections will describe the answers they found to the questions above.

Q: Does the topology of the phylogenetic tree matter?
A: Not much, if the branch lengths are approximately the same.

To address this question, Fan et al. ran their phylo-HMM model with three different types of tree topologies: star, balanced binary tree, and unbalanced binary tree. The binary tree topologies were further subdivided into depth-first and breadth-first trees, and the number of elements in the balanced trees was varied among 4, 6, and 8 units (the unbalanced tree was only evaluated at 8 units). The trees used in this study can be seen in Figure 5.

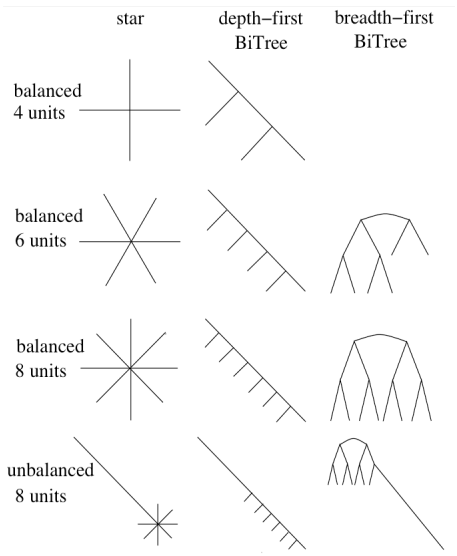

Figure 5: Tree topologies tested in simulations by [26].

Runs of these 11 trees (a balanced depth-first tree of 4 units is isomorphic with a balanced breadth-first one) were compared. Two significant results emerged from this analysis:

- There was little difference between the performance of star topologies and balanced tree topologies.

- Unbalanced tree topologies performed much worse than balanced tree topologies.

Their explanation of this phenomenon was that genomes that are clustered together, as in the concentrated parts of the unbalanced trees, lose their distinguishing power versus the remote species, reducing the overall power of the model. Their recommendation based on this result is to make the branch lengths roughly equal in whatever phylogenetic tree is used, by picking the appropriate species, and to then use a star topology because it is simpler than a binary tree, and just as effective.

Q: Can we simplify the substitution model with minimal impact to accuracy?
A: Yes.

The issue here is whether using a simpler substitution rate matrix is possible without sacrificing much in terms of model accuracy. Models with more free parameters, such as HKY85 and GTR/REV, would appear to capture more of the inherent variability in the evolution process, but creating models with those additional parameters requires many more resources, and they can actually overfit the data if the tree creation process is not done properly.

Simulations were run with the following substitution rate matrices, listed in order of increasing complexity: JC69, F81, HKY85, and REV. Each matrix was substituted in for the matrix used in the initial baseline model. In each pairwise comparison, the simpler model performed as good or slightly better than the more complex model. The authors' explanation of this result is that it matched their "intuition that simpler models are easier to solve," implying that the simpler models, by virtue of their more limited feature space, could be better optimized than models with larger feature spaces.

Q: Should the phylogenetic tree include distantly related species?
A: No, it is actually counterproductive.

The issue of whether a phylogenetic tree should have one or more outgroups in it to optimize results was addressed in the subsection above regarding tree topologies, and the answer was that outgroup species actually reduce the power of the phylo-HMM model. This result is in line with two existing facts: 1) it is harder to create a good alignment with a species which is too distant from the others, negatively impacting the phylo-HMM downstream, and 2) some studies have shown that increasing the total branch length decreases the power in the model [26, 31, 32].

Fan et al. performed another test which would bear on this question: they varied the branch lengths of a given tree topology. They discovered that the modification which most greatly improved the model's accuracy was increasing the length of the shortest branch, making the branch lengths more uniform. Outgroups, then, make a phylo-HMM model *less* accurate.

Q: How many genomes do we need?
A: Between 6 and 10.

Deciding how many genomes to use to have a sufficient level of statistical power is a difficult problem in comparative genetics. Fan et al. examined this question by modifying their baseline model to equalize the branch lengths, reflecting their results from previous simulations. They then varied that common branch length and determined how many species their tree would require to produce a given sensitivity and specificity level.

Fixing the specificity at 0.95 and the common branch length to the one between mouse and rat, they determined that 6 genomes were needed to produce a sensitivity level of 0.90, and 10 genomes were needed to produce a sensitivity level of 0.95. In general, they found an inverse relationship between the number of genomes and the individual branch length, which is consistent with other studies [31].

Q: Is a constant conservation ratio across the genome biologically plausible?
A: It doesn't seem to be.

The conservation ratio of this phylo-HMM is the ratio of the average substitution rate of the conserved sites over that of the unconserved sites, and it is the crux of what is being modeled with this algorithm. When Fan et al. varied the conservation ratio in their baseline model and plotted sensitivity versus conservation rate (at a fixed specificity of 0.9), they found that these two values had an inverse sigmoid relationship, with the point of rapid descent in sensitivity occurring when the conservation ratio goes above 0.6 (the baseline model had a conservation ratio of 0.32).

This result seemed problematic to the authors. Noting that the multiple transcription factor binding sites that are associated with many promoter regions are each likely to have differing conservation ratios, they questioned the ability of a phylo-HMM with a single conservation ratio to be able to accurately model such

regions, given the great influence that a varying conservation rate was shown to have on model sensitivity.

Statistical Power: Further Study

Based on these results above and others in the paper, Fan et al. offered the following suggestions for the further development of phylo-HMMs:

- A goodness-of-fit test for phylo-HMM is needed on real data.
- The conservation state needs to change across species. Currently, this phylo-HMM decides on a conservation class for each site, which remains throughout the run. This parameter should be allowed to vary from species to species.
- The functional classes found for the genomes need to become more fine-grained beyond simply conserved/non-conserved, by more directly modeling functional elements like transcription factor binding sites.
- The ability on integrating the multiple alignment into the conservation analysis is desirable.

Future Development

While the phylo-HMM is over a decade old in its conception ([3, 4]), the greater availability of sequenced genomes in recent years has only now made the phylo-HMM a more prevalent tool in comparative genomics studies. With this increased interest has come a variety of proposals for improving the model. The subsections below outline some of them.

Varying the substitution matrix across the genomes.

Many of the criticisms of phylo-HMM models involve how rigidly that the critical substitution matrix is applied to the data. There are only as many substitution matrices as phylogenetic tree/HMM states, and each one of them enforces the same evolutionary rate along all the genomes. These restrictions are biologically implausible, since it seems apparent that there are a variety of different evolutionary rates across the genomes [13]. Also, a single substitution matrix is applied at a given site down the entire phylogenetic tree, implying that there are no changes to the evolutionary rate over time. This restriction, too, seems not to allow that different parts of the genome can be under selection at different times in

different species. Methods to relax these constraints on the substitution matrix are needed to create more biologically appropriate models for evolutionary behavior. [13, 26].

Integrating sequence alignment and evolutionary analysis.

The quality of a phylo-HMM is dependent on the quality of the multiple sequence alignment it starts with. Although it has been suggested that phylo-HMMs are reasonably robust with respect to its input MSA [26], there are still options for alignment generated by the MSA algorithm which are not then available to the phylo-HMM process because the MSA algorithm only produces one alignment. Some researchers have offered that this barrier can be removed by integrating the multiple sequence alignment and the phylo-HMM generation into a single process [13, 15]. This idea could be based on a body of work already using HMMs to do multiple sequence alignment (e.g., [18, 33-35]).

Allowing interaction between sites.

It is an artifact of the hidden Markov Model that the transition from any state is not dependent on any other information other than the transition probabilities associated with that state [1]. For a phylo-HMM, this artifact means that the evolution of any site is not influenced by the state of any other site, including its neighboring sites in the sequence. This independence is an oversimplification on its face – the sites within a codon are highly dependent. Intrasite dependency also seems to extend to noncoding regions as well [12].

To correct this problem, researchers have tried two approaches. The first is to model dinucleotides and trinucleotides explicitly, by simply extending the 4-character alphabet used in the original model to 16- and 64-characters, respectively [12, 36]. This technique allows the standard HMM algorithms to still be used, but makes their execution more computationally intensive. The second approach is to extend the representation of the HMM emissions to depend not only on the current state (via its emission probabilities), but one or two of the immediately previous states [2, 12, 36]. This technique requires some modifications to the HMM calculations for likelihood and parameter estimation, but it improves the goodness-of-fit of the phylo-HMM on coding regions [12].

Bibliography

1. Durbin, R., et al., *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. 1998: Cambridge Univ Pr.
2. Siepel, A. and D. Haussler, *Phylogenetic Hidden Markov Models*, in *Statistical Methods in Molecular Evolution*, R. Nielsen, Editor. 2005, Springer: New York, NY. p. 325-351.
3. Yang, Z., *A space-time process model for the evolution of DNA sequences.* Genetics, 1995. **139**(2): p. 993-1005.
4. Felsenstein, J. and G.A. Churchill, *A Hidden Markov Model approach to variation among sites in rate of evolution.* Mol Biol Evol, 1996. **13**(1): p. 93-104.
5. Goldman, N., J.L. Thorne, and D.T. Jones, *Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses.* J Mol Biol, 1996. **263**(2): p. 196-208.
6. Lio, P., et al., *PASSML: combining evolutionary inference and protein secondary structure prediction.* Bioinformatics, 1998. **14**(8): p. 726-33.
7. Thorne, J.L., N. Goldman, and D.T. Jones, *Combining protein evolution and secondary structure.* Mol Biol Evol, 1996. **13**(5): p. 666-73.
8. Husmeier, D., *Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models.* Bioinformatics, 2005. **21 Suppl 2**: p. ii166-72.
9. Husmeier, D. and F. Wright, *Detection of recombination in DNA multiple alignments with hidden Markov models.* J Comput Biol, 2001. **8**(4): p. 401-27.
10. Gibbs, R.A., et al., *Genome sequence of the Brown Norway rat yields insights into mammalian evolution.* Nature, 2004. **428**(6982): p. 493-521.
11. Thomas, J.W., et al., *Comparative analyses of multi-species sequences from targeted genomic regions.* Nature, 2003. **424**(6950): p. 788-93.
12. Siepel, A. and D. Haussler, *Phylogenetic estimation of context-dependent substitution rates by maximum likelihood.* Mol Biol Evol, 2004. **21**(3): p. 468-88.
13. Siepel, A., et al., *Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.* Genome Res, 2005. **15**(8): p. 1034-50.
14. Siepel, A. and D. Haussler. *Computational identification of evolutionarily conserved exons*. in *8th International Conference on Research in Computational Molecular Biology (RECOMB'04)*. 2004: ACM Press.
15. Pedersen, J.S. and J. Hein, *Gene finding with a hidden Markov model of genome structure and evolution.* Bioinformatics, 2003. **19**(2): p. 219-27.

16. McAuliffe, J.D., L. Pachter, and M.I. Jordan, *Multiple-sequence functional annotation and the generalized hidden Markov phylogeny.* Bioinformatics, 2004. **20**(12): p. 1850-60.
17. Siepel, A. and D. Haussler, *Combining phylogenetic and hidden Markov models in biosequence analysis.* J Comput Biol, 2004. **11**(2-3): p. 413-28.
18. Satija, R., L. Pachter, and J. Hein, *Combining statistical alignment and phylogenetic footprinting to detect regulatory elements.* Bioinformatics, 2008. **24**(10): p. 1236-42.
19. Wong, W.S. and R. Nielsen, *Finding cis-regulatory modules in Drosophila using phylogenetic hidden Markov models.* Bioinformatics, 2007. **23**(16): p. 2031-7.
20. Galtier, N., O. Gascuel, and A. Jean-Marie, *Markov Models in Molecular Evolution*, in *Statistical Methods in Molecular Evolution*, R. Nielsen, Editor. 2005, Springer: New York, NY. p. 3-24.
21. Jukes, T.H. and C.R. Cantor, *Evolution of Protein Molecules*. 1969, New York: Academic Press.
22. Kimura, M., *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences.* J Mol Evol, 1980. **16**(2): p. 111-20.
23. Hasegawa, M., H. Kishino, and T. Yano, *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.* J Mol Evol, 1985. **22**(2): p. 160-74.
24. Tamura, K., *Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases.* Mol Biol Evol, 1992. **9**(4): p. 678-87.
25. Wikipedia contributors. *Models of DNA evolution*. 2009 23 October 2009 22:00 UTC [cited 2009 10 December 2009 07:44 UTC]; Available from: http://en.wikipedia.org/w/index.php?title=Models_of_DNA_evolution&oldid=321651045.
26. Fan, X., et al., *Statistical power of phylo-HMM for evolutionarily conserved element detection.* BMC Bioinformatics, 2007. **8**: p. 374.
27. Siepel, A. *PhastCons HOWTO*. 2005 [cited 2009 December 10, 2009]; Available from: http://compgen.bscb.cornell.edu/phast/phastCons-HOWTO.html.
28. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
29. Wong, W.S. and R. Nielsen. *EvoPromoter*. 2004 [cited 2009 December 11, 2009]; Available from: http://evopromoter.sourceforge.net/.
30. Holland, R.C., et al., *BioJava: an open-source framework for bioinformatics.* Bioinformatics, 2008. **24**(18): p. 2096-7.
31. Eddy, S.R., *A model of the statistical power of comparative genome sequence analysis.* PLoS Biol, 2005. **3**(1): p. e10.

32. McAuliffe, J.D., M.I. Jordan, and L. Pachter, *Subtree power analysis and species selection for comparative genomics.* Proc Natl Acad Sci U S A, 2005. **102**(22): p. 7900-5.
33. Mitchison, G.J., *A probabilistic treatment of phylogeny and sequence alignment.* J Mol Evol, 1999. **49**(1): p. 11-22.
34. Holmes, I., *Using guide trees to construct multiple-sequence evolutionary HMMs.* Bioinformatics, 2003. **19 Suppl 1**: p. i147-57.
35. Loytynoja, A. and M.C. Milinkovitch, *A hidden Markov model for progressive multiple alignment.* Bioinformatics, 2003. **19**(12): p. 1505-13.
36. Jojic, V., et al., *Efficient approximations for learning phylogenetic HMM models from data.* Bioinformatics, 2004. **20 Suppl 1**: p. i161-8.