Multiple-sequence functional annotation and the generalized hidden Markov phylogeny

Jon D. McAuliffe
Department of Statistics
University of California, Berkeley
jon@stat.berkeley.edu

Lior Pachter
Department of Mathematics
University of California, Berkeley
lpachter@math.berkeley.edu

Michael I. Jordan
Division of Computer Science
and Department of Statistics
University of California, Berkeley
jordan@stat.berkeley.edu

August 17, 2003

Technical Report 647

Abstract

Phylogenetic shadowing is a new comparative genomics principle which allows for the discovery of conserved regions in sequences from multiple closely-related organisms. We develop a formal probabilistic framework for combining phylogenetic shadowing with feature-based functional annotation methods. The resulting model, a *generalized hidden Markov phylogeny* (GHMP), applies to a variety of situations where functional regions are to be inferred from evolutionary constraints. In particular, we show how GHMPs can be used to predict complete shared gene structures in multiple primate sequences. We also describe SHADOWER, our implementation of such a prediction system. We find that SHADOWER outperforms previously reported *ab initio* gene finders, including comparative human-mouse approaches, on a small sample of diverse exonic regions. Finally, we report on an empirical analysis of SHADOWER's performance which reveals that as few as five well-chosen species may suffice to attain maximal sensitivity and specificity in exon demarcation.

1 Introduction

The prediction of functional regions in genomic sequences has traditionally been based on the identification of features associated with genes or regulatory regions (Zhang, 2002). Comparison of homologous genomic sequences facilitates such identification, because functional regions tend to be conserved in sequences which have evolved from a common ancestor, whereas non-functional regions are more likely to mutate. Information about the degree of conservation between pairs of sequences is known to help in the identification of genes (Alexandersson et al., 2003, Korf et al., 2001, Parra et al., 2003, Meyer and Durbin, 2002).

One drawback of pairwise comparative approaches to gene prediction is that non-functional regions are required to have diverged to a degree that enables statistical procedures to distinguish them from biologically active regions; typically, organisms such as human and mouse are used. These methods are therefore not applicable to discovering features present only at close evolutionary proximity, such as primate-specific genes. The *phylogenetic shadowing* principle of Boffelli et al. (2003) circumvents this problem by seeking to identify conserved regions among multiple closely-related organisms. This has numerous advantages: the alignment of the sequences is straightforward, the phylogenetic tree relating the sequences is easy to infer, and the identification of conserved regions among the sequences is possible using standard evolutionary models. The principle has been illustrated by Boffelli et al. (2003) in the identification of transcription factor binding sites in the primate-specific apo(a) gene.

To provide a systematic computational methodology for annotating genomic sequences based on the principle of phylogenetic shadowing, we have developed the *generalized hidden Markov phylogeny* (GHMP). The GHMP is a formal probabilistic model that combines conservation-based constraints deriving from multiple genomic sequences with algorithmic ideas that have proven useful in single-organism gene annotation systems. Our approach synthesizes generalized hidden Markov model gene finders, evolutionary models of nucleotide substitution, and phylogenetic trees. Similar ideas have been presented by Pedersen and Hein (2003) and Siepel and Haussler (2003); our extensions and contributions are described in Section 2. We have also implemented SHADOWER, a gene prediction system based on these ideas. We show that SHADOWER outperforms existing *ab initio* methods, including those taking comparative-genomics approaches, on a multiple-primate dataset of single exons from five separate gene regions. Furthermore, we present an empirical analysis of SHADOWER's performance on various subsets of our primates which reveals that just five species, selected according to a formal optimality criterion, suffice to deliver the best results SHADOWER can obtain for these data.

The remainder of the paper proceeds as follows. Section 2 presents theoretical and computational details of the GHMP, placing the GHMP within the general formalism of probabilistic graphical models. We report and discuss the data, parameter estimation procedure, and subset-selection optimization underlying our full empirical analysis in Section 3. Finally, Section 4 offers concluding remarks and outlook.

2 Methods and approach

2.1 The generalized hidden Markov phylogeny

Graphical models combine ideas from probability theory and graph theory to facilitate the use of sophisticated joint dependency structures in data analysis (Cowell et al., 1999, Jordan, 1999). The nodes of a graphical model correspond to random variables which relate to the problem and data at hand. The edges in the model encode marginal and conditional independencies among these random variables, according to a well-defined formal semantics. The generalized hidden Markov phylogeny (GHMP) is a *directed* graphical model—a model in which the underlying graph is directed and acyclic. In such models, there is a local conditional probability distribution associated with each node in the graph, conditional on its parents. The joint distribution over all random variables is defined to be the product of these local conditional distributions.

This section details the variables, independence structure, and local distributions peculiar to the GHMP. The graphical model perspective allows us to focus our attention on capturing, in the model definition, the essential ingredients of the multi-sequence functional annotation problem. Then, parameter estimation and probabilistic inference are handled using general-purpose graphical modeling algorithms (Jordan, 1999).

Many biologists are already acquainted with special cases of graphical model methods which preceded the recognition and elaboration of the general framework. Phylogenetic trees can be treated as graphical models, and the likelihood computation of Felsenstein (1981) is an instance of the general-purpose *junction tree algorithm* for graphical models (Cowell et al., 1999). Similarly, the forward, backward, and Viterbi algorithms for inference in hidden Markov models (HMMs) are also special cases of the junction tree algorithm. The GHMP synthesizes the ideas of HMMs and phylogenetic trees. The corresponding algorithms can be seen on the one hand as a synthesis of the HMM and tree inference algorithms, or on the other hand as simply another instantiation of the universal graphical model procedures.

Combinations of HMMs and evolutionary models have been previously described by Pedersen and Hein (2003) and Siepel and Haussler (2003), and our methods build on this earlier work. The GHMP and SHADOWER introduce:

- generalized hidden Markov dynamics (non-geometric exon length distributions)
- a dual-strand hidden state space
- GC isochore-specific parameters
- deterministic constraints on repeats, gaps, and in-frame stop codons inside aligned exons
- more complete splice site modeling
- an automated iterative procedure for alignment and tree building
- an analysis methodology for optimal species subset selection.

The principle behind our treatment of gaps also differs, as described below. The reader will recognize several ideas from the currently best-performing gene finders (Alexandersson et al., 2003, Korf et al., 2001, Burge and Karlin, 1997, Parra et al., 2003); indeed, our work represents an attempt to bring these ideas into a phylogenetic framework. Most important of all, we adhere to the phylogenetic

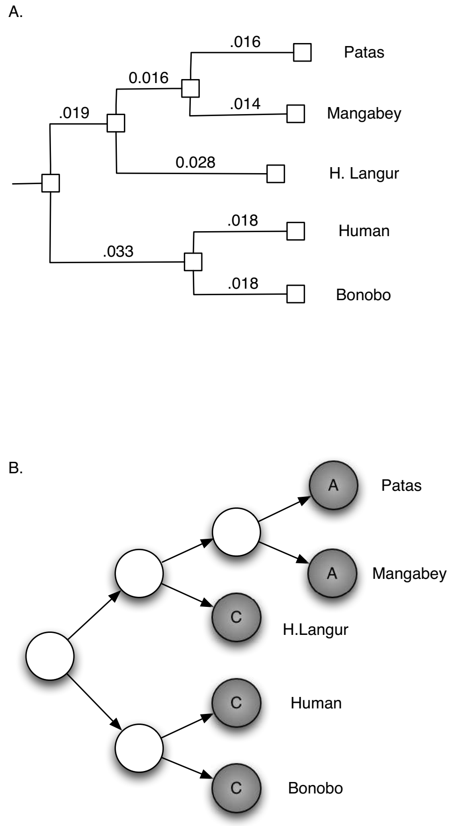

Figure 1: Two alternative representations of the same rooted phylogenetic tree. Panel A shows a diagram familiar to biologists, with annotated edge lengths. Panel B depicts the corresponding directed graphical model.

shadowing principle in our exclusive use of closely-related species. This motivates and justifies the seemingly stringent requirements we impose; for example, we rely on exactness of the multiple alignment and so consider only perfectly aligned splice signals and start/stop codons, with no gaps permitted. This approach should be contrasted with comparison-based gene finders using distantly-related organisms, which generally must search in the enormous space of possible alignments[1].

We begin by recasting familiar phylogenetic tree representations within the graphical model framework. Consider the rooted five-primate phylogenetic tree presented in Figure 1, Panel A, and the corresponding graphical model shown in Panel B. Both diagrams indicate the presence of a specific set of nucleotides at homologous sites in the five primates, and also indicate putative ancestral nucleotides. The distinction between observed nucleotides and ancestral nucleotides is captured in the graphical model by shading; in general, observed random variables are shaded, whereas unobserved (hidden) random variables are left unshaded. In Panel A, edge lengths are proportional to evolutionary distance. In the graphical model, on the other hand, edge lengths

[1]For example, there are approximately 10^{18} distinct alignments of five sequences, when each sequence is only five bases long.

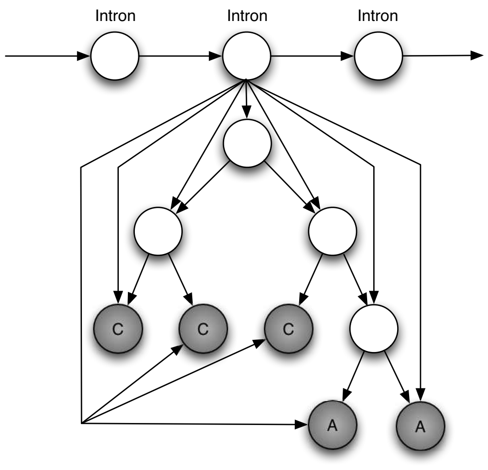

Figure 2: An excerpt of the GHMP graphical model corresponding to an aligned intronic column.

are uninformative. Instead, the pattern of edges formally encodes the following probabilistic assumption: given the nucleotide of an organism's immediate ancestor, that organism's nucleotide is conditionally independent of all other ancestral nucleotides.

To complete the specification of the graphical model, we require a local conditional distribution for each node v. At nodes other than the root, this distribution is simply given by an evolutionary nucleotide substitution model: for each possible initial parent nucleotide, the distribution specifies the probability that v evolved to a terminal nucleotide A, C, G, or T in time b at some evolutionary substitution rate. In the GHMP, the branch-length parameter b is specific to each node, while one substitution rate is shared by all non-root nodes. Finally, the distribution π of the root can be any probability distribution over the four bases, e.g., equilibrium base frequencies in a region.

In our implementation, we use the Felsenstein substitution model for the conditional distributions (Felsenstein and Churchill, 1996). This model requires both a transition-transversion ratio and an equilibrium base distribution; we take the latter to be the same as the root distribution π.

The GHMP uses this phylogenetic model to define a probability distribution on a single column of a multiple alignment. To define a probability distribution on a full alignment, the GHMP includes additional nodes that represent functional states. In the implementation of the GHMP that we consider in this paper, the functional states include intergenic regions, introns, coding exons, and coding exon boundaries (the last includes splice sites, start codons, and stop codons). State variables are unobserved (hidden) variables and thus are left unshaded. The nodes representing these variables are arranged as a chain in the graphical model, with the edges between these nodes representing the probability of transitioning between specific values of functional states; this is the graphical model representation of a Markov chain. The space of allowed functional states is structured so as to enforce frame and phase consistency across exons, with genes on either strand, as in Burge and Karlin (1997).

Let us refer to the collection of adjacent columns generated by a functional state as a *slice*. Consider first the simplest case, that of the intron state, in which a slice containing only a single column is associated with the state (see Figure 2). In the figure, the chain of functional state nodes

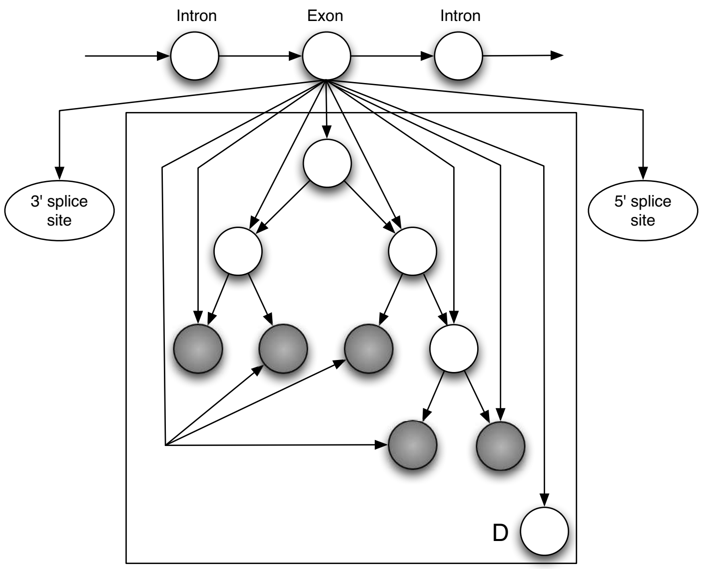

Figure 3: An excerpt of the GHMP graphical model corresponding to an aligned internal exon on the forward strand.

proceeds from left to right; for concreteness, the three nodes of the chain depicted here have all taken on the intron state. Below each hidden state node appears the phylogenetic tree of Figure 1, Panel B (for clarity, only the tree associated with the middle node is depicted). The same tree topology and parameters are used in every instance of the tree.

Observe that every node in the tree has the functional state node as a parent. This allows the nucleotide substitution models to depend on the functional role of the slice being generated. In particular, the evolutionary substitution rate from ancestor to descendant varies with function. The version of the GHMP we implemented uses two substitution rates, *functional* (exons, exon boundaries) and *nonfunctional* (introns, intergenic regions). Since the substitution model we consider allows us to establish evolutionary rates only up to a positive scaling factor, we take the functional rate to be 1.0 with no loss of generality, leaving the nonfunctional rate as a free parameter $r > 1.0$.

While intronic states generate only a single-column slice, implying a geometric distribution for the lengths of aligned introns (Burge and Karlin, 1997), exonic states are associated with multiple-column slices. Consider the GHMP fragment shown in Figure 3, where the middle node has taken on the state of a shared exon. This hidden exon state is associated with a left exon boundary slice (here, a 3' splice site), an internal exonic slice, and a right exon boundary slice (here, a 5' splice site). The square containing the phylogenetic tree is a piece of graphical model notation called a *plate*. The plate indicates that the entire tree structure inside the plate is repeated D times, corresponding to an aligned exon spanning a D-column slice. Of course, different exons must be allowed different lengths, so D itself is a random variable, making the overall structure a *generalized plate* (whence *generalized* hidden Markov phylogeny). The conditional distribution of D given exon type is arbitrary, so that aligned exon length distributions appropriate to the species at hand may be modeled. Each tree in the figure, including those of the boundary slices we now describe, evolves

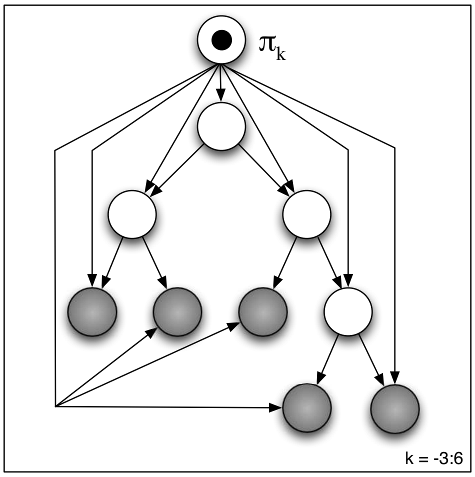

Figure 4: Detail of the 5' donor splice site submodel omitted from Figure 3.

at the functional rate 1.0.

Note that Figure 3 depicts a shared forward-strand internal exon. Other possibilities are an initial or terminal exon in a multi-exon gene, a single-exon gene, or any of these on the reverse strand. The splice-site ovals of Figure 3 are not part of the graphical model nomenclature; we utilize them only to simplify the diagram. We now consider the exon boundary models, i.e., splice site, start codon, and stop codon slices, that are substituted in the place of these ovals. In particular, Figure 4 shows the donor splice site model the reader should envision in place of the right oval in Figure 3. It is a plate denoting nine independent copies of the phylogenetic tree, numbered $k = -3$ to 6, with no zero index. Each copy generates one column of the full donor slice, that is, the window in the alignment surrounding the shared GT signal at columns 1 and 2. The columns are not identically distributed, but rather have position-dependent equilibrium base distributions π_k. These are explicitly depicted as the dotted node in the plate. (The dot indicates that the π_k are to be construed as fixed parameters, rather than in a Bayesian fashion as random variables.) The position-specific distributions allow us to exploit varying nucleotide usage in the splice signal's flanking region. This phenomenon has been studied in human genomic sequence by Zhang (1998) and others.

Treatment of the 5' acceptor site's window is exactly analogous to the case of the donor slice. Since the start codon ATG is non-stochastic, it requires no model. Finally, a stop codon slice is generated using a phylogenetic tree of stop codons (not pictured), as follows. A progenitor stop codon TAA, TAG, or TGA is chosen at the root of the tree according to a stop codon equilibrium distribution. This codon is evolved towards the leaves; they then constitute a three-column slice of the multiple alignment (each row of which is some valid stop codon). The stop-codon substitution model is defined by first independently evolving each of an ancestor's three nucleotides using a nucleotide substitution model. This evolution is then normalized by removing outcomes which are not stop codons and scaling the remaining outcomes by their total probability mass. In this manner, with probability one, a valid stop codon is produced at every node in the tree.

We remark now on the simplicity of the exon model employed. Conditional on being in an exonic hidden state, the columns of the aligned exon interior are independent and identically distributed. The concepts of codon and peptide are not incorporated, nor is sequential dependence along the exon a part of the setup. This is in some sense the most naïve possible model of a shared exon: only a non-geometric length distribution and lower substitution rate, together with the boundary structures, distinguish exons from intronic and intergenic regions of the multiple alignment. Biologically, this exon representation does not stand up to the more sophisticated methods of Burge and Karlin (1997), Korf et al. (2001), Alexandersson et al. (2003) and others. On the other hand, our approach has scientific virtue: by comparing the performance of a simple model on multiple closely-related species to a complex model on a single organism, or distant paired organisms, we learn something about the relative advantages of these gene-finding strategies. The results of Section 3 speak for themselves in this regard.

Our discussion of the GHMP model closes with the issue of gaps in the multiple alignment. We have not attempted to include nucleotide insertion and deletion events in our model. Other authors (Pedersen and Hein, 2003, Siepel and Haussler, 2003) treat gaps as missing data, marginalizing gapped leaves out of aligned columns. This approach can be accommodated readily within the probabilistic inference mechanism of the GHMP, but it has practical drawbacks. A gap is not a nucleotide we failed to observe; instead, it is more like a nucleotide which evolved out of the phylogenetic tree at a given homologous position. As such, for purposes of functionally annotating the alignment, it evidences lack of conservation and should not just be integrated away during the probability computations. To incorporate this consideration into the model, we replace all gaps in an aligned column with that column's least-occurring base[2], as a heuristic penalization. However, before this is done, deterministic constraints involving gaps are enforced (see below). Note that, due to our use of closely-related species, the importance of any particular gap heuristic is greatly diminished: the aligned exons in our dataset were entirely gapless, so any approach which preferentially assigns gaps outside exons is likely to perform comparably.

2.2 Estimation and inference

We now discuss the parameter estimation methods used in our implementation. Starting with raw homologous sequence data from multiple organisms, we first obtain a multiple alignment and phylogenetic tree by repeated alternation between tree-based alignment and maximum-likelihood tree estimation over the aligned sequences. We use MAVID (Bray and Pachter, 2003) for the former and FASTDNAML (Olsen et al., 1994, Felsenstein, 1981) for the latter. The corresponding nucleotide substitution model is described in Felsenstein and Churchill (1996), with equilibrium base frequencies estimated by maximum likelihood from the raw sequence data. The transition-transversion ratio is fixed at 2.0, except where this is incompatible with the estimated equilibrium base distribution, in which case the smallest admissible value is utilized. Once the alignment and tree have been estimated, they are fixed during all subsequent inference on the GHMP, and the same tree topology and branch lengths are used for every column of the alignment.

The hidden Markov chain of functional states requires an initial probability distribution over the functional state space, as well as a matrix of transition probabilities. While these parameters can be estimated using expectation-maximization or other likelihood-based approaches, given appropriate data (Pedersen and Hein, 2003, Siepel and Haussler, 2003), the phylogenetic shadowing principle lets us finesse the issue. Since we work only with immediate primate neighbors of the human, a

[2]Ties are broken according to the equilibrium base distribution.

satisfactory approximation to the model's Markov chain parameters is obtained simply by using widely-available maximum-likelihood estimates from annotated human genomic sequence. Indeed, we transferred the reported GC isochore-specific H. sapiens parameters of Alexandersson et al. (2003) directly to the GHMP.

The same rationale applies to GC isochore-specific aligned length distributions for exons (by type), introns, and intergenic regions, as well as the equilibrium stop codon distribution: previously reported maximum-likelihood estimates on human genomic sequence are employed. However, since observed intron and intergene lengths in human sequence do not reflect the increased length in a multiple alignment due to gaps, the geometric distribution mean parameters are scaled up by a factor involving the fraction of gapped columns in the given alignment. This is not necessary for exonic lengths, because of the extreme rarity of gaps, as previously described. Furthermore, the position-dependent equilibrium nucleotide distributions of our donor and acceptor models are fixed at the human occurrence frequencies reported by Zhang (1998). This leaves only one parameter, the nonfunctional evolutionary substitution rate r. Its treatment as a model selection parameter is discussed in Section 3.

Having described the estimation of all free parameters in the GHMP, we turn now to the inference procedure. First we enforce a set of deterministic constraints: start codons, stop codons, and splice signals must be exactly aligned and gapless. Gaps are allowed only in codon-sized runs within exon slices. Additionally, in-frame stop codons are disallowed for every species inside an exon slice. Taken together, these constraints lead to the identification of all candidate aligned exons. These then underlie a generalized Viterbi algorithm, which computes the most probable trajectory through the hidden functional state chain, conditional on the observed alignment data. This version of the Viterbi algorithm supports non-geometric durations in exonic states, as well as the computation of phylogenetic-tree emission probabilities. We emphasize again that this algorithm, which involves conditioning on the alignment data and marginalizing out all ancestor branch nodes in the GHMP, is a special case of the general-purpose machinery for graphical model inference.

3 Results and discussion

3.1 All-species analysis

We have implemented SHADOWER, a system for automated functional annotation based on the ideas described in the previous section. Here we report on a re-examination of five exonic regions across a number of primates varying, by region, from 13 to 18. The datasets are described in Boffelli et al. (2003). Each region spans roughly 2 kb and contains a single exon from one of the five genes apolipoprotein(a), apolipoprotein(b), cholesteryl ester transfer protein (cetp), liver x-receptor α (lxr α), and plasminogen (plg). Human sequence was used in every region; beyond that, there is modest overlap among the sets of primates sequenced for each dataset.

In Table 1 we show the accuracy of SHADOWER's exon predictions as the nonfunctional rate r is varied from 1.0 (the functional evolutionary rate) to 2.5. For these datasets, predicted exon count increases monotonically with r. We estimate performance using cross validation, leaving out one dataset at a time. At each step, r is chosen on four of the datasets to maximize sensitivity, and accuracy is assessed on the remaining dataset. Total nucleotide-level sensitivity and partial-match exon-level sensitivity (which forgives inexact boundary demarcations) are both 100%—all coding

	Nonfunctional evolutionary rate r															
	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
apo(a)	×	×	→	→	→	→	1—	1—	1—	1—	1—	1—	1—	4—	4—	4—
apo(b)	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	2✓	2✓	4✓	4✓
cetp	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	2✓	2✓
lxr α	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
plg	—	—	✓	✓	✓	✓	✓	✓	1✓	1✓	1✓	1✓	1✓	1✓	1✓	3✓

Table 1: Each row shows the prediction results on a single-exon dataset as the nonfunctional rate parameter r varies from 1.0 to 2.5. ✓ means the exon is predicted exactly; × means the exon is completely missed; — means the exon is predicted but both boundaries are incorrect; → means the exon is predicted but only the downstream boundary is correct; the number n means there are additionally n false positive exons.

bases in all five exons are detected. Exact exon sensitivity is 80%, because of a single boundary failure—the upstream start codon boundary of the apo(a) exon is incorrectly localized at a nearby splice site. This is the unique initial exon in our data; we conjecture that this incorrect localization is partly due to the lacking observation of the remaining downstream apo(a) exons, whose presence would interact with the hidden Markov dynamics to create a stronger preference for an initial exon at this location.

Turning to false positive exon predictions using the cross-validated choice of r, we find a specificity of 89.6% at the nucleotide level, 83.3% at the partial-match exon level, and 66.7% at the exact exon level. The failure here is a single false positive terminal exon in the cetp region. It is interesting that this false positive appears at every value of r shown in Table 1, including even the functional rate 1.0. A look at the alignment reveals a highly-probable acceptor site slice and stop codon slice flanking an exon of typical length; taken together, these are the likely determinants of this prediction. It is less clear that additional upstream exons would ameliorate this problem, as they would in the case of apo(a). Instead, the situation calls for enhancement of our site-independent exon slice model. In Section 4 we expand on this point.

To contrast SHADOWER with state-of-the-art gene-finding methods, we ran GENSCAN, at default settings (Burge and Karlin, 1997), on the human sequence data from each of the five regions. Its nucleotide sensitivity came out at 44.7%, versus 100% using SHADOWER, and its nucleotide specificity was 34.0%, versus 89.6% with SHADOWER. GENSCAN entirely missed three of the five exons, partially matched one (lxr α), and demarcated one exactly (cetp), while producing one false positive exon in the apo(b) region.

We also compared SHADOWER to SLAM (Alexandersson et al., 2003) using human-mouse homology. No homologous mouse sequence was found for the apo(a) and cetp exons; the remaining three exons were demarcated exactly. This gives SLAM a nucleotide sensitivity of 80.2% with 100% specificity on these exons. The important point of this comparison is that the evolutionary distance of human and mouse prevents SLAM from competing on functional annotation of some genomic regions under study.

3.2 Species-subset analysis and maximal Steiner subtrees

In the context of functional annotation using multi-species sequence datasets, the relationship between the set of species chosen and the consequent annotation quality has enormous practical significance. Given the expense and resources required for large-scale sequencing of an organism, we particularly need to determine how few species suffice to deliver adequate prediction of biological function. Of course, there are many available species sets of a given size, so one arrives naturally at a sequence of optimization problems: for every size, which collection of that size yields the highest-quality functional annotation?

To fix ideas, focus now on the 18 primates sequenced in the apo(a) region. There are 261,972 subsets of these primates having size at least two; we want to avoid producing a SHADOWER analysis for all of them. So instead we take the following approach. The results in the previous section show that the total evolutionary divergence among these primates is large enough to distinguish conserved from non-conserved regions but still small enough to enable exact alignments of exon boundaries. Now, we can measure divergence using the maximum-likelihood phylogenetic tree grown on the full apo(a) dataset, construing branch lengths as expected nucleotide substitution counts. In this tree, each available primate belongs to one leaf. From this viewpoint, the total divergence of all the apo(a) primates corresponds to the total weight of the phylogenetic tree, i.e., the sum of all branch lengths.

Similarly, the total divergence of any subset of the apo(a) primates corresponds to the weight of that subset's "family tree," that is, the lowest-weight subtree covering all the leaves in the subset (also known as the *Steiner subtree* for those leaves). This is the tree that a SHADOWER analysis restricted to the given subset would utilize. We take the Steiner subtree weight of an apo(a) primate subset as our surrogate for the annotation "quality" that subset would provide. Finding the subset of size k having maximal-weight Steiner subtree—the *k-mss problem*—is a well-defined optimization problem which admits a dynamic programming solution linear in tree size. A complete discussion of this topic is in preparation.

In Boffelli et al. (2003) it was shown that the percentage of total divergence attained by the k-mss primate subsets in these five regions increases rapidly for k up to five or six, then gradually for larger values of k. If our postulated connection between total divergence and annotation quality holds, we expect to see a similar relationship to exon sensitivity and specificity. To study this, we first solved the k-mss problem in all five regions, for each k from two to 13. Then, for each k, we ran SHADOWER on the k-mss primate subsets, region by region. The nonfunctional rate parameter r was chosen as described in the previous subsection. The resulting nucleotide-level exon sensitivity and specificity are shown as a function of k in Figure 5. As the figure shows, using just the five primates of each region's 5-mss allows SHADOWER to recover the same level of performance it obtained on the full primate collections. As expected, annotation quality improves rapidly at values of k up to five. The results are the same for exact and partial-match exon detection (not shown). The top panel of Figure 1 shows the five primates in the apo(a) 5-mss, situated in the phylogenetic tree grown on just their sequence data for the SHADOWER analysis.

4 Conclusions

We have developed the generalized hidden Markov phylogeny, a graphical model architecture that provides a rigorous probabilistic underpinning for the phylogenetic shadowing principle. Our results

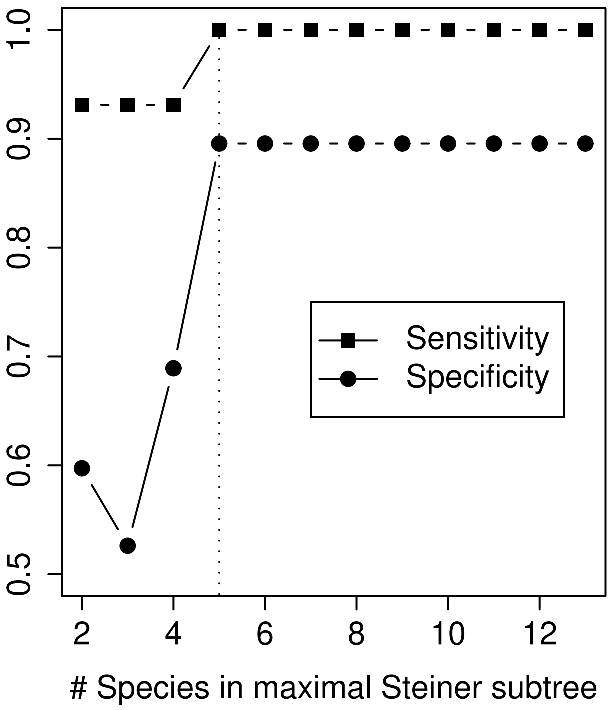

Figure 5: The performance of SHADOWER on various-sized primate collections corresponding to maximal Steiner subtrees.

support the general premise of phylogenetic shadowing, showing that a relatively straightforward implementation of a GHMP can yield state-of-the-art performance in the identification of primate-specific elements in the human genome.

We have described a relatively simple implementation of the GHMP, in which the elementary component models (splice sites, exons, introns, intergenic regions) do not attempt to capture probabilistic dependence among the aligned columns. Our success on the available datasets makes a strong case for the viability of such a simplified model. This success is of course predicated on the strength of the signal in the data. For basic *ab initio* annotation of exons, introns and splice sites, our analysis has suggested this signal is sufficiently strong for accurate annotation when the data consist of sequences from as few as five primates.

The graphical model framework underlying the GHMP readily accommodates architectural variations and extensions, and several are of immediate interest. First, the GHMP can be extended to allow for the identification of regulatory elements and binding sites. The known regulatory similarities of closely-related organisms suggest that such sites may be conserved in position and number; we already have empirical evidence for this from the apo(a) gene (Boffelli et al., 2003). Second, the GHMP model described here does not account for gaps. A model incorporating gapped slices would be of general interest, and in particular would be useful in the context of the regulatory element modeling problem, where for instance varying-sized boxes of short repetitive elements are known to be homologous across species. Finally, a more powerful exon model would not only help in the reduction of false positives (as, for example, in the cetp gene), but could also be used to incorporate functional annotation methods for proteins (e.g., Simon et al., 2002) into genomic sequence annotation.

SHADOWER software is available at `www.stat.berkeley.edu/users/jon/shadower`.

5 Acknowledgments

We thank Dario Boffelli and Eddy Rubin for the sequence data that we have analyzed as well as many helpful discussions about the concepts of phylogenetic shadowing. Lior Pachter was supported in part by a grant from the NIH (R01-HG02362-02). Michael Jordan was supported by a grant from the NSF (IIS-9988642).

References

Alexandersson, M., Cawley, S., and Pachter, L. (2003). SLAM—cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.*, 13:496–502.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., and Rubin, E. M. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science*, 299(5611):1391–1394.

Bray, N. and Pachter, L. (2003). Mavid multiple alignment server. *Nucl. Acids Res.*, 31(13):3525–3526.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268:78–94.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York, NY.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum-likelihood approach. *J. Mol. Evol.*, 17:368–376.

Felsenstein, J. and Churchill, G. A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.*, 13(1):93–104.

Jordan, M. I., editor (1999). *Learning in Graphical Models*. MIT Press, Cambridge, MA.

Korf, I., Flicek, P., Duan, D., and Brent, M. R. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17:S140–S148.

Meyer, I. M. and Durbin, R. (2002). Comparative *ab initio* prediction of gene structures using pair HMMs. *Bioinformatics*, 18(10):1309–1318.

Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, 10:41–48.

Parra, G., Agarwal, P., Abril, J. F., Wiehe, T., Fickett, J. W., and Guigo, R. (2003). Comparative gene prediction in human and mouse. *Genome Res.*, 1:108–117.

Pedersen, J. S. and Hein, J. (2003). Gene finding with a hidden Markov model of genome structure and evolution. *Bioinformatics*, 19:219–227.

Siepel, A. and Haussler, D. (2003). Combining phylogenetic and hidden Markov models in biosequence analysis. In *Proceedings of the Seventh Annual International Conference on Computational Biology*, pages 277–286.

Simon, A. L., Stone, E. A., and Sidow, A. (2002). Inference of functional regions in proteins by quantification of evolutionary constraints. *Proc. Natl. Acad. Sci. USA*, 99:2912–2917.

Zhang, M. Q. (1998). Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, 7(5):919–932.

Zhang, M. Q. (2002). Computational prediction of eukaryotic protein coding genes. *Nat. Rev. Genet.*, 3:698–709.