SHADOWER: A generalized hidden Markov phylogeny for multiple-sequence functional annotation

Jon D. McAuliffe, [1] **Lior Pachter,** [2] **Michael I. Jordan**[1, 3]

Keywords: Functional annotation, gene finding, phylogenetic shadowing, generalized HMM.

The prediction of functional regions in genomic sequences has traditionally been based on the identification of features associated with genes or regulatory regions. Comparison of homologous genomic sequences, e.g. from a pair of species, facilitates such identification [1, 5]. This is 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.

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. 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 [2] circumvents this problem by seeking to identify conserved regions among multiple closely-related organisms. This has numerous advantages: sequence alignment is straightforward, the relevant phylogenetic tree is easy to infer, and identification of conserved regions is possible using standard evolutionary models.

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 probabilistic graphical model [4] 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 [6] and [7]. Our extensions include generalized hidden Markov dynamics; a frame- and phase-consistent dual-strand hidden state space, supporting single-exon, multi-exon, and incomplete gene prediction; GC isochore-specific parameters; deterministic constraints on repeats, gaps, and in-frame stop codons; more complete splice site modeling; and an automated iterative procedure for alignment and tree building. The annotation is obtained as the most *a posteriori* probable trajectory through a hidden space of functional states; this trajectory is computed efficiently using algorithms for graphical model inference. Figure 1 shows a subcomponent of the GHMP graphical model corresponding to an aligned forward-strand internal exon.

To limit the number of sequenced organisms required for functional annotation, we have also developed a methodology for species subset selection. The method chooses subsets according to a maximin criterion on the weight of subtrees within the overall phylogenetic tree relating the species. Theory and efficient algorithms for this *maximal Steiner subtree* approach will be described at the conference.

We have implemented SHADOWER, a gene prediction system based on the GHMP. Table 1 shows that, by exploiting the additional constraints from multiple-species conservation, SHADOWER outperforms existing *ab initio* methods on a small dataset of single exons from five separate gene regions, across 13 primates. The data were originally reported by [2]. In addition, an analysis using species subsets of various sizes, each chosen by the maximal Steiner subtree criterion, revealed that SHADOWER needs only five of the available 13 primates to attain the performance reported in Table 1.

[1] Department of Statistics, University of California, 367 Evans Hall, Berkeley, CA 94720.
E-mail: {jon,jordan}@stat.berkeley.edu

[2] Department of Mathematics, University of California, 970 Evans Hall, Berkeley, CA 94720.
E-mail: lpachter@math.berkeley.edu

[3] Division of Computer Science, University of California, 387 Soda Hall, Berkeley, CA 94720.

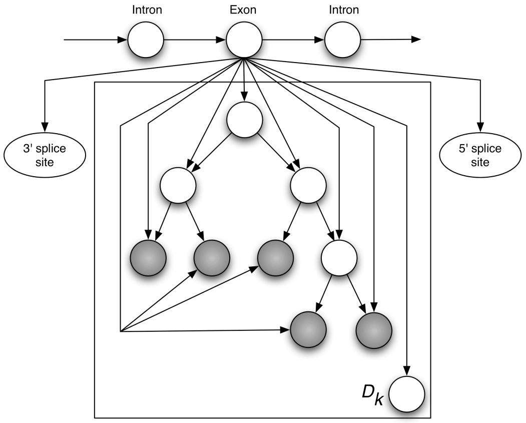

Figure 1: An excerpt of the GHMP graphical model corresponding to an aligned internal exon on the forward strand. The hidden chain of functional states runs along the top. Depicted underneath is a phylogenetic tree of nucleotides, both observed (shaded) and unobserved (unshaded). The bounding box (*plate*) around the phylogenetic tree denotes duplication, D_k times. Each copy of the tree corresponds to an alignment column, which populates the tree's leaves. D_k too is random, allowing the length of aligned exons to follow an arbitrary distribution (thus *generalized* hidden Markov phylogeny). The ovals labeled as splice sites are not part of the language of graphical models; they appear here to reduce visual clutter.

	Nucl.(%)		Exon Partial		Exon Exact	
	Sn	Sp	Sn	Sp	Sn	Sp
GENSCAN	44.7	34.0	2/5	2/3	1/5	1/3
MZEF	37.4	63.2	3/5	3/4	1/5	1/4
SHADOWER	**100.0**	**89.6**	**5/5**	**5/6**	**4/5**	**4/6**
SHADOWER^b	42.7	42.2	2/5	2/5	1/5	1/5
SLAM	80.2	100.0	3/5	3/3	3/5	3/3

Table 1: Sensitivity and specificity of various gene finders on the primate exon datasets. Results are shown at the nucleotide, partial exon (i.e. inexact boundaries), and exact exon level. GENSCAN [3] predicts complete or incomplete genes, using only the human sequence data. MZEF [8] predicts individual internal exons (without frame or phase consistency), using only the human sequence data. SHADOWER employs the GHMP to analyze multiple orthologous sequences. SHADOWER^b excludes exon boundary models, to exemplify a more limited approach based on multiple-species conservation. SLAM [1] uses human-mouse homology in a generalized pair HMM.

References

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

[2] 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* 299:1391–1394.

[3] Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* 268:78–94.

[4] Jordan, M. I., ed. 1999. *Learning in Graphical Models*. Cambridge, MA: MIT Press.

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

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

[7] 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 (RECOMB 03)*, New York: ACM. pp. 277–286.

[8] Zhang, M. Q. 1997. Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proceedings of the National Academy of Sciences USA* 94:565–568.