Computational Identification of Evolutionarily Conserved Exons

Adam Siepel
Center for Biomolecular Science and Engr.
University of California
Santa Cruz, CA 95064, USA
acs@soe.ucsc.edu

David Haussler
Howard Hughes Medical Institute and
Center for Biomolecular Science and Engr.
University of California
Santa Cruz, CA 95064, USA
haussler@soe.ucsc.edu

ABSTRACT

Phylogenetic hidden Markov models (phylo-HMMs) have recently been proposed as a means for addressing a multi-species version of the ab initio gene prediction problem. These models allow sequence divergence, a phylogeny, patterns of substitution, and base composition all to be considered simultaneously, in a single unified probabilistic model. Here, we apply phylo-HMMs to a restricted version of the gene prediction problem in which individual exons are sought that are evolutionarily conserved across a diverse set of species. We discuss two new methods for improving prediction performance: (1) the use of context-dependent phylogenetic models, which capture phenomena such as a strong CpG effect in noncoding regions and a preference for synonymous rather than nonsynonymous substitutions in coding regions; and (2) a novel strategy for incorporating insertions and deletion (indels) into the state-transition structure of the model, which captures the different characteristic patterns of alignment gaps in coding and noncoding regions. We also discuss the technique, previously used in pairwise gene predictors, of explicitly modeling conserved noncoding sequence to help reduce false positive predictions. These methods have been incorporated into an exon prediction program called ExoniPhy, and tested with two large data sets. Experimental results indicate that all three methods produce significant improvements in prediction performance. In combination, they lead to prediction accuracy comparable to that of some of the best available gene predictors, despite several limitations of our current models.

General Terms

Algorithms, Experimentation, Performance

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Biology and genetics

Keywords

Gene prediction, phylogenetic hidden Markov model

1. INTRODUCTION

With three mammalian genomes now sequenced and assembled, more on the way, and the database of known genes steadily growing more accurate and more complete, the rules of the game are changing in mammalian gene prediction. In round numbers, some 70–90% of human protein-coding genes are probably now known—depending on how many genes exist and how many currently "known" are actually pseudogenes [40, 31]—and the situation is similar with other species. The challenge now is to reveal the intransigent genes laid bare by the genome sequencing projects, yet still hidden. Thus, it is no longer enough to do fairly well on average at predicting average genes, in newly sequenced regions of a genome. The next generation of gene-finding programs must be able to find new, possibly unusual, genes in well-studied regions, and with low enough false positive rates that predictions can be tested in the laboratory efficiently and economically (see [16]).

The missing genes most likely belong to several different classes, and a variety of computational approaches (some possibly quite specialized) will be required to identify them. One broad distinction that can be drawn is between ancient genes shared by most species, and newer, lineage-specific genes, e.g., resulting from recent gene duplications. In this paper, we are concerned with genes of the first type. Because they are present in most species, these "core" genes should be particularly amenable to comparative gene prediction, and they are a natural starting point for a multi-species method. We choose to approach these genes at the level of individual (coding) exons, which are more likely than complete genes to be conserved through evolutionary history. (If desired, predicted exons can be combined into complete or partial transcripts in post-processing.) A large percentage of exons appear to be well conserved across diverse sets of species—by our estimates, probably more than 80% for the placental mammals.

Our goal is to predict conserved exons using only the sequences of the genomes in question, so that genes can be found without cDNA evidence or homologous proteins. This problem is a multi-species version of the standard ab initio gene prediction problem addressed by programs such as Genie [21] and Genscan [6], and of the pairwise ab initio gene prediction problem addressed by programs such as Twin-

RECOMB'04, March 27–31, 2004, San Diego, California, USA.

scan [20], SGP2 [29], SLAM [1], and Doublescan [26]. In our case, we assume that a multiple alignment of orthologous sequences is available, and that the phylogeny of the species is known. Several new genomic-scale aligners help to make the first assumption possible [5, 4, 2] (but see Discussion), and an emerging consensus on mammalian phylogeny [28, 38] allows for the second assumption.

In addressing this problem, we make use of a type of probabilistic model called a phylogenetic hidden Markov model, or "phylo-HMM," which combines a hidden Markov model and a set of phylogenetic models [12, 43, 15, 30, 35, 25, 34]. Phylo-HMMs model molecular evolution as a Markov process in two dimensions: a substitution process over time at each site in a genome, which is guided by a phylogenetic tree, and a process by which the "mode" of evolution (as described by a phylogenetic model—see below) changes from one site to the next. Recently, phylo-HMMs have been applied to gene prediction with encouraging results [30, 25], but so far their performance has been evaluated only with simulated data or in small-scale experiments with real biological data. In one case [25], a phylo-HMM was used in the context of "phylogenetic shadowing" [3], which is based on a somewhat different set of goals from the ones we have stated (see Discussion).

We consider three ways of improving the performance of phylo-HMMs in exon prediction: the use of context-dependent phylogenetic models, explicit modeling of conserved noncoding sequences, and modeling of insertions and deletions (indels). These methods have been implemented in a computer program, called ExoniPhy[1], which predicts evolutionarily conserved exons from a multiple alignment, given a definition of a phylo-HMM. Using ExoniPhy, we have conducted the first large-scale experiments with real biological data in phylo-HMM-based exon/gene prediction. Comparisons of alternative versions of the program indicate that all three of the methods considered produce substantial improvements in prediction performance. When they are used in combination, ExoniPhy achieves a level of performance that is competitive with some of the best available gene predictors, despite the limitations of considering each exon separately and the fact that the program still lacks some of the basic features of current gene predictors—e.g, it does not allow for non-geometric length distributions of exons or use the best available methods for splice-site detection. Our results suggest that it may be possible, using a strategy based on phylo-HMMs, to predict conserved exons with very good sensitivity and near perfect specificity.

2. METHODS

2.1 Phylo-HMMs for exon prediction

A phylo-HMM is a hidden Markov model that has a phylogenetic model associated with each of its states (Figure 1A). It can be thought of as a probabilistic machine that generates a multiple alignment by randomly transitioning from one state to another, in discrete time steps, and at each step emitting an alignment column that is drawn from a distribution associated with the current state. These distributions of alignment columns are defined by probabilistic phylogenetic models, and reflect the topology and branch lengths of a phylogenetic tree, as well as a continuous-time Markov model of nucleotide substitution. A phylo-HMM can be used for prediction with a multiple alignment, much the way an ordinary HMM is used with a single sequence. Phylo-HMMs can be represented naturally as graphical models (Figure 1B) [25, 34].

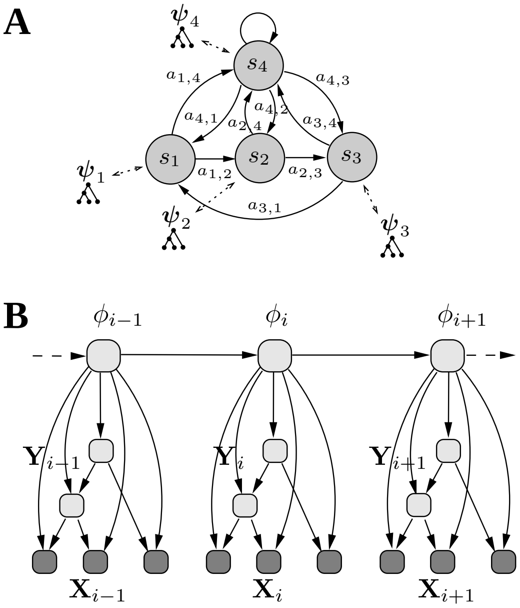

Figure 1: A state-transition diagram (A) and graphical model representation (B) of a simple phylo-HMM with three coding states (s_1, s_2, and s_3), corresponding to the three codon positions, and a noncoding state (s_4). In (B), the shaded nodes, collectively labeled $\mathbf{X}_{i-1}, \mathbf{X}_i$, and $\mathbf{X}_{i+1}$, represent observed random variables, defined by a given multiple alignment. The unshaded nodes represent latent variables for ancestral nodes in the tree ($\mathbf{Y}_{i-1}$, $\mathbf{Y}_i$, $\mathbf{Y}_{i+1}$) and states in the path (ϕ_{i-1}, ϕ_i, ϕ_{i+1}).

More formally, a phylo-HMM $\boldsymbol{\theta} = (S, \boldsymbol{\psi}, \mathbf{A}, \mathbf{b})$ is a four-tuple consisting of a set of M states, $S = \{s_1, \ldots, s_M\}$, a set of associated phylogenetic models, $\boldsymbol{\psi} = \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_M\}$, a matrix of state-transition probabilities, $\mathbf{A} = \{a_{j,k}\}$ $(1 \leq j, k \leq M)$, and a vector of initial-state probabilities, $\mathbf{b} = (b_1, \ldots, b_M)$. Model $\boldsymbol{\psi}_j$ is associated with state s_j $(1 \leq j \leq M)$, $a_{j,k}$ $(1 \leq j, k \leq M)$ is the probability of visiting state k at an alignment column i given that state j is visited at column $i - 1$, and b_j $(1 \leq j \leq M)$ is the probability that state j is visited first (thus, $\sum_k a_{j,k} = 1$ for all j, and $\sum_j b_j = 1$; note that the Markov chain for state transitions is assumed to be first-order and homogeneous). Let $\mathbf{X}$ be the given alignment, consisting of L columns (sites) and n rows (one for each species), with the ith column denoted $\mathbf{X}_i$ $(1 \leq i \leq L)$. The probability that a column $\mathbf{X}_i$ is emitted by state s_j is $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$, a quantity that can be computed with Felsenstein's "pruning" algorithm [11]. A "path" through the phylo-HMM is a sequence of states, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_L)$, such that $1 \leq \phi_i \leq M$ for $1 \leq i \leq L$. The joint probability of a path and an alignment is

$$P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta}) = b_{\phi_1} P(\mathbf{X}_1|\boldsymbol{\psi}_{\phi_1}) \prod_{i=2}^{L} a_{\phi_{i-1},\phi_i} P(\mathbf{X}_i|\boldsymbol{\psi}_{\phi_i}). \quad (1)$$

[1] Pronounced "ex-ON-if-I," like personify.

(For simplicity, transitions to an "end" state are omitted here.) The likelihood of a phylo-HMM is the sum over all paths, $P(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{\phi}} P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta})$, which can be computed with the forward algorithm, and the maximum-likelihood path is $\hat{\boldsymbol{\phi}} = \arg\max_{\boldsymbol{\phi}} P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta})$, which can computed with the Viterbi algorithm. Each phylogenetic model itself consists of several components, including a substitution rate matrix, a tree topology, a set of branch lengths, and a background distribution for nucleotides, which together describe the "mode" of evolution at a given site (the branching history of the species, the rate of substitution along each branch, the "pattern" of substitution, etc.). Details can be found in recent reviews of HMMs [10], phylogenetic models [22, 41], and phylo-HMMs [35, 34].

For a phylo-HMM to be applied to exon prediction, some strategy is required for associating its states with biological features of interest. In this paper, we use a very simple strategy, involving a one-to-one mapping between states and "labels" for individual sites. Let a *feature* be a biological entity that spans one or more sites in an alignment, such as an exon or splice site. (Features are assumed to be consistent across aligned sequences; see Discussion.) Two kinds of features are considered here: "variable-length" features (e.g., exons) and "signal" features (e.g., splice sites), which typically mark the boundaries of variable-length features. Each feature is associated with a set of *labels*, such that a labeling of sites defines a set of features, and a set of features defines a labeling of sites. Because labels and states are associated one-to-one, a phylo-HMM can easily be trained with a labeled alignment (with labels derived from sequence annotations), and a predicted state sequence (path) defines a predicted labeling, which in turn defines a set of predicted features. Our base model consists of variable-length features for exons and noncoding regions, and signal features for start codons, stop codons, $5'$ splice sites, and $3'$ splice sites (Figure 2). The simple strategy used here has obvious drawbacks (e.g., it erroneously implies a geometric distribution for exon lengths), but we consider it a reasonable place to start, given the many other sources of complexity in our models.

2.2 Context-dependent phylogenetic models

HMM-based gene finders often have states whose emission probabilities are conditioned on previous observations, so that differences can be considered in the relative frequencies of nucleotide tuples in regions of different biological function. Phylo-HMMs can be adapted to use such "high-order" states as well, in a way that allows not only the frequencies, but also the substitution patterns, of tuples of adjacent bases to be considered [35, 36].

High-order states can be allowed in a phylo-HMM through the use of phylogenetic models that are defined in terms of N-tuples of bases. We say that N is the *order* of such a model[2], and when $N > 1$, we call the model *context-dependent*. Context-dependent phylogenetic models can be treated much like ordinary phylogenetic models, but have larger numbers of free parameters, and are computationally more expensive to manipulate. Nevertheless, accurate parameter estimation is feasible for 2nd and 3rd order models, with even very general parameterizations of the substitution rate matrix, provided large enough quantities of data are available for training. Context-dependent models fit aligned biological sequences substantially better than ordinary, independent-site models, in both coding and noncoding regions, and even improve significantly on existing codon models in coding regions. These models permit amino acid substitution rates to be learned implicitly from nucleotide data, and at the same time capture phenomena such as the transition/transversion bias and the preference for synonymous substitutions over nonsynonymous substitutions [36].

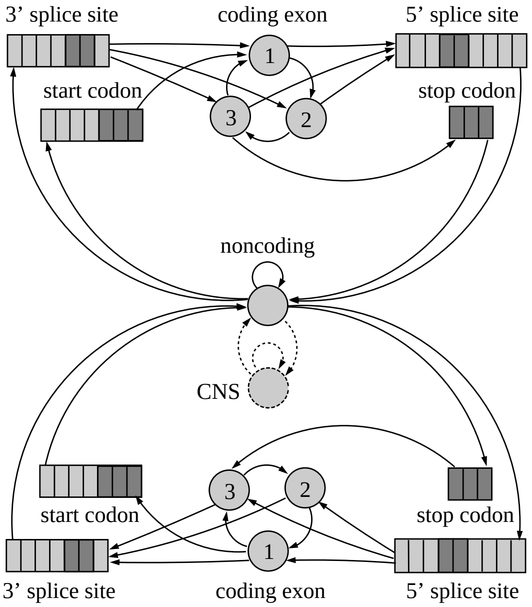

Figure 2: State-transition diagram for ExoniPhy. Each feature (indicated by a textual phrase in the diagram) is associated with a set of labels, each of which is identified with a state. States for variable-length features are represented by circles, and states for signal features by boxes. Signal features are generally defined as windows around positions of interest; the shaded boxes indicate the critical positions within each window (e.g., the start codon itself, the canonical "GT" in a $5'$ splice site). States for the positive strand are shown at top, and states for the negative strand at bottom. The CNS state is optional (see Section 2.3.)

Context-dependent phylogenetic models define joint distributions of N-tuples of alignment columns. To use them for high-order states in a phylo-HMM, joint probabilities must be converted to conditional probabilities. This conversion can be accomplished simply and efficiently with a two-pass dynamic programming algorithm, based on a missing data principle [35]. Once conditional probabilities are available, equation 1 can be replaced by (e.g., for $N = 3$):

$$P(\boldsymbol{\phi}, \mathbf{X}|\boldsymbol{\theta}) = b_{\phi_1} P(\mathbf{X}_1|\boldsymbol{\psi}_{\phi_1}) a_{\phi_1,\phi_2} P(\mathbf{X}_2|\mathbf{X}_1, \boldsymbol{\psi}_{\phi_2}) \times \prod_{i=3}^{L} a_{\phi_{i-1},\phi_i} P(\mathbf{X}_i|\mathbf{X}_{i-2}, \mathbf{X}_{i-1}, \boldsymbol{\psi}_{\phi_i}). \qquad (2)$$

In this paper, a general reversible 3rd order model (R3) is used for both coding and noncoding states, with a separate parameter describing the rate of substitution between every pair of nucleotide triples that differ by one base (multiple in-

[2] Somewhat confusingly, an Nth order substitution model is used for HMM states of order $N - 1$.

stantaneous substitutions are prohibited; the total number of rate-matrix parameters is 288). The use of richly parameterized context-dependent models allows the phylo-HMM to pick up on different context effects in coding and noncoding regions (e.g., strong CpG effect in noncoding regions, preference for synonymous substitutions in coding regions), in addition to differences in overall substitution rates. Training data tends to be sparse for the models associated with signal states, so the much simpler HKY model [17] is used for these states.

2.3 Distinguishing coding from conserved noncoding sequence

While comparative methods for gene prediction benefit enormously from differences in the average level and patterns of conservation in coding and noncoding regions, they can be mislead by islands of conservation in noncoding regions [1, 14]. This conserved noncoding sequence (CNS) is potentially a serious problem in gene prediction, because it may make up an even larger portion of mammalian genomes than does coding sequence (perhaps some 3% vs. 1.5% [8, 33]). Understanding what accounts for the surprising amount of mammalian CNS is an important topic of current research. At least until this area is better understood, however, it seems reasonable in comparative gene prediction to take the simple strategy of modeling all CNS uniformly, with a single, additional state in an HMM [1]. This is the strategy we take here.

In our case, the CNS state is associated with a context-dependent phylogenetic model, which can be trained on highly conserved sites that are believed not to code for proteins (see Section 2.5). Thus, not only the degree of conservation, but also characteristic patterns of context-dependent substitution (presumably distinct from those in coding regions) can be captured. While noncoding sites immediately adjacent to exon boundaries are often highly conserved, we currently exclude these sites, and only consider CNS that is isolated from exons[3]. Thus, the state transition diagram of Figure 2 is only slightly altered by the addition of a CNS state.

2.4 Modeling insertions and deletions

Although there has been progress in incorporating insertions and deletions (indels) into phylogenetic models [39, 18, 23], current solutions to this difficult problem remain complex and computationally intensive, and alignment gaps are still often ignored or addressed with simple heuristics [35, 25]. In multi-species exon/gene prediction, an appropriate way of modeling indels is particularly desirable, because patterns of alignment gaps provide one of the best indications of whether or not a segment of aligned DNA codes for proteins. The statistics on alignment gaps in coding regions are striking. We find, in multiple alignments of a large set of apparently conserved exons from human, mouse, and rat, that 89.5% of exons have no alignment gaps whatsoever and 9.5% have gaps with lengths that are perfect multiples of three, leaving only 1% with irregular gaps. In contrast, gaps appear frequently in noncoding regions, and with lengths that are approximately geometrically distributed. Notably, many conserved noncoding regions have a few short gaps with non-multiple-of-three lengths—an important clue that they do not code for proteins. To be useful in exon finding, a method for accommodating indels must be relatively simple and computationally efficient, and must allow alternative length distributions for alignment gaps to be defined. We describe such a method here.

A

species 1	b	–	b	b	b	–	–
species 2	b	b	–	b	–	b	–
species 3	b	b	b	–	–	–	b
gap pattern	0	1	2	3	4	5	6

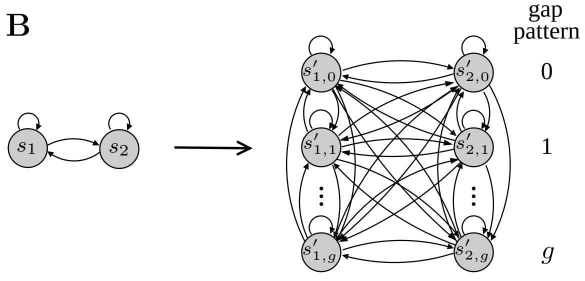

Figure 3: (A) A multiple alignment showing the possible gap patterns for three species ($g = 6$), with a different gap pattern appearing in each column. The symbol b is meant to indicate any base. (B) An example showing how a two-state HMM can be expanded in a cross-product type of construction, with a new copy of each state for each gap pattern.

Let the *gap pattern* of a column in an alignment be its unique pattern of gap and non-gap characters (Figure 3A). With n sequences, there are $2^n - 1$ such patterns, assuming columns that consist completely of gap characters are prohibited. (Below we show this can effectively be reduced to a number linear in n, by separately considering only the most essential gap patterns and treating the rest as a group.) Assume the gap patterns are numbered $0, \ldots, g$ ($g = 2^n - 2$), with 0 indicating the pattern with no gaps (the "null" gap pattern). Given a phylo-HMM $\boldsymbol{\theta} = (S, \boldsymbol{\psi}, \mathbf{A}, \mathbf{b})$, let a new "gapped" phylo-HMM $\boldsymbol{\theta}' = (S', \boldsymbol{\psi}', \mathbf{A}', \mathbf{b}')$ be defined with a set of "gapped states" S' consisting of $g+1$ copies of each state in S, $S' = \{s'_{1,0}, s'_{1,1}, \ldots, s'_{M,g-1}, s'_{M,g}\}$, and a set of "gapped models" $\boldsymbol{\psi}'$ consisting of $g+1$ copies of each model in $\boldsymbol{\psi}$, $\boldsymbol{\psi}' = \{\boldsymbol{\psi}'_{1,0}, \boldsymbol{\psi}'_{1,1}, \ldots, \boldsymbol{\psi}'_{M,g-1}, \boldsymbol{\psi}'_{M,g}\}$ (Figure 3B). If s_j represents the condition of an alignment column being assigned to some category of sites C_j, then $s'_{j,k}$ represents the condition of a column being assigned to category C_j *and* having gap pattern k. The emission probability of a column $\mathbf{X}_i$ for state $s'_{j,k}$ (with $1 \leq j \leq M$ and $0 \leq k \leq g$) is defined as

$$P(\mathbf{X}_i|\boldsymbol{\psi}'_{j,k}) = \begin{cases} P(\mathbf{X}_i|\boldsymbol{\psi}_j) & \text{if } \mathbf{X}_i \text{ has gap pattern } k \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

It can be shown that, if gaps are handled as missing data when computing $P(\mathbf{X}_i|\boldsymbol{\psi}_j)$, then the models in $\boldsymbol{\psi}'$ remain valid probability models, so that $\boldsymbol{\theta}'$ is also a valid probability model (proof omitted). For our purposes, the transition

[3] So far, we find that it is better to force a choice between exons and CNS at the level of entire features, rather than to have such choices determine the placement of exon boundaries. Conservation patterns at exons boundaries are complex (see, e.g., [37]) and it is non-trivial to model them well.

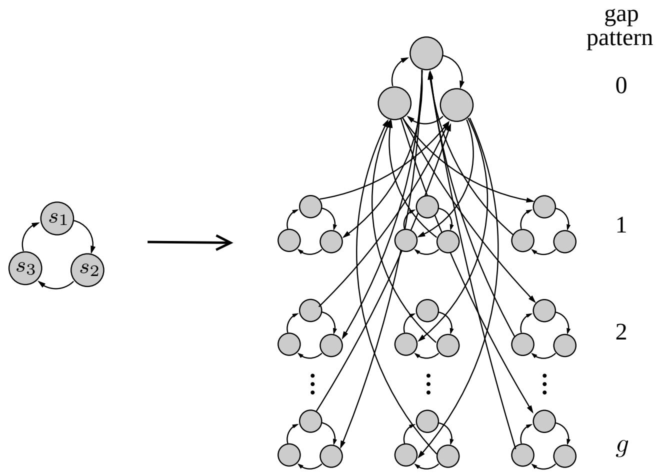

Figure 4: Illustration showing how a cycle of three states in an HMM, representing the three codon positions, can be expanded, with a new copy of the entire cycle for each codon position × non-null gap pattern. The arcs shown indicate the transitions that will have non-negligible probability if the expanded HMM is trained on data with predominately "clean" gaps. For simplicity, state labels have been omitted on the right.

A

spec. 1	b	–	b	b	b	b	b	b	b	–	–	–	–	–	–
spec. 2	b	b	–	–	b	b	b	b	–	b	b	–	–	–	–
spec. 3	b	b	b	–	–	b	b	b	–	–	b	b	–	–	–
spec. 4	b	b	b	b	b	–	–	b	–	–	–	–	b	b	–
spec. 5	b	b	b	b	b	b	–	–	–	–	–	–	–	b	b
gap pat.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

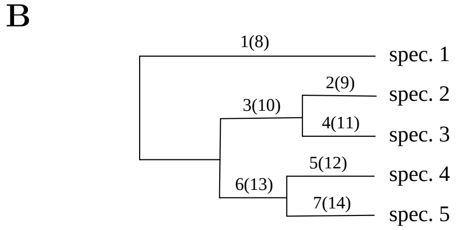

Figure 5: A multiple alignment (A) showing all phylogenetically simple gap patterns for a set of 5 species and the illustrated tree topology (B). Besides the null gap pattern, each simple pattern can be explained by a single deletion (insertion) on a branch of the tree (see numbers on branches). All patterns not shown are phylogenetically complex.

probabilities between gapped states ($\mathbf{A}'$) and a new set of initial state probabilities ($\mathbf{b}'$) can be estimated empirically, by simply considering the gap pattern as well as the label at each site in a set of labeled training data.

This gapped phylo-HMM is a multi-sequence generalization of a pair HMM [10], with several useful properties. (Note that, when $n = 2$, the gapped phylo-HMM reduces to a pair HMM.) As with a pair HMM, the model structure induces separate gap "open" and "extension" penalties (log probabilities of respectively entering and remaining in a state with a non-null gap pattern), consistent with a uniform distribution for the positions of indel events and a geometric distribution for the lengths of affected segments. The gapped phylo-HMM, however, has separate parameters corresponding to different kinds of indel events—for example, ones that have occurred on different branches of the phylogenetic tree (e.g., a deletion on the branch to the rodents or an insertion on the branch to the primates), or "nested" events involving multiple branches of the tree (e.g., an insertion on the branch to the rodents, followed by a deletion in mouse of part of the inserted segment). In this way, it can accommodate lineage-specific differences in insertion and deletion rates and expected gap lengths [27]. It is also capable of capturing asymmetries in the alignment procedure, such as when the alignment has been constructed with respect to a reference sequence, which is useful as long as the training and testing alignments are created in the same way. Moreover, the multi-species indel model is completely integrated with a phylogenetic, continuous-time Markov model of the substitution process (which is potentially context-dependent).

Furthermore, non-geometric length distributions for gaps can be modeled by creating additional copies of states. Here, we wish to represent the distinctive multiple-of-three-length gaps of coding regions. This can be accomplished as follows. Suppose the original set S contains one or more cycles of three states, corresponding to the three codon positions (Figure 4). When constructing the set S' of gapped states, these codon states are now duplicated as whole cycles of three, rather than individually, and $3g$ rather than g new copies are created, so that a set of g cycles is associated with each codon position. If the model is trained on a data set in which nearly all gaps in coding regions have multiple-of-three lengths and are non-overlapping, then the transitions estimated to have non-negligible probability will be as shown in Figure 4. Such a model will tend to assign high probability to candidate exons that have no gaps or "clean" gaps (non-overlapping with multiple-of-three lengths), and low probability to exons that have "dirty" gaps.

As mentioned above, the number of gap patterns for n species is $2^n - 1$, so the number of states in the gapped phylo-HMM quickly becomes prohibitively large. Let us define a "phylogenetically simple" gap pattern as a gap pattern that can be explained by at most one insertion or deletion on a branch of the phylogenetic tree (Figure 5). It can easily be shown that the number of phylogenetically simple gap patterns is $2(2n - 3) + 1$ for a (binary) phylogenetic tree with $n \geq 2$ species. (Whether or not a given gap pattern is simple with respect to a given tree can be determined with a slightly adapted version of the standard parsimony algorithm [13].) Let all other gap patterns (those that require multiple indel events to explain) be called "complex." If only phylogenetically simple gap patterns are distinguished in constructing a gapped phylo-HMM, and all "complex" patterns are lumped together, then (assuming a construction like the one shown in Figure 3) the number of states grows by a factor of $2(2n - 3) + 2$ rather than $2^n - 1$.

The procedure for creating the gapped phylo-HMM remains as described above (in either the standard case or the special case of coding states) except that a copy of each original state (or cycle of states) is created only for each simple gap pattern, and a single copy is created for the class of all complex patterns. In practice, with coding states, copies are created for each simple gap pattern, but not for the class of complex patterns; thus, if a given column $\mathbf{X}_i$ has a complex gap pattern, then equation 3 evaluates to zero for every coding state. Copies of noncoding states corresponding to the class of complex gap patterns are given a fixed, small positive emission probability, which applies to all columns having complex gap patterns. As a result, alignment columns with complex gap patterns drop out of the maximization problem addressed by the Viterbi algorithm, except that they are required to occur in noncoding regions (they function as a constraint on the Viterbi path). The gapped phylo-HMM as a whole remains a valid probability model, up to a normalization constant.

2.5 Software and experimental design

A program called ExoniPhy was written to predict conserved exons in a multiple alignment, based on a phylo-HMM. ExoniPhy supports the indel model described above, as well as the use of context-dependent phylogenetic models. It assumes a one-to-one mapping between labels and states, and uses the Viterbi algorithm for prediction. ExoniPhy was tested with a 60-state phylo-HMM (including the CNS state), as shown in Figure 2; when the indel model is used, this number increases to 492 for 3 species or 2274 for 9 species (see data sets, below). The program and auxiliary software for estimating model parameters are available by request.

Using ExoniPhy, experiments were conducted to test the effect on prediction performance of the three methods in question, both individually and in combination. For each of two data sets (described below), predictions were produced using eight different versions of the program—all combinations with and without context-dependent phylogenetic models, the CNS state, and the indel model.

In addition, ExoniPhy was compared to a simpler program, called ExoniPhy--, that considers only the background distributions of (Nth order) phylogenetic models, and computes emission probabilities assuming independence of the sequences in the alignment (as with a profile HMM [10]). The simplified models associated with this program's HMM states are equivalent to phylogenetic models with "star" topologies and branch lengths approaching infinity. Thus, ExoniPhy-- is a non-phylogenetic version of ExoniPhy. It serves as a benchmark that indicates what level of performance is achievable with the same state-transition structure and state-specific background distributions as ExoniPhy, but ignoring the phylogeny, branch lengths, and substitution process. Like ExoniPhy, ExoniPhy-- can optionally use the indel model. Only the case of $N = 3$ (2nd order Markov models) was considered.

To assess the tradeoff between sensitivity and specificity, a tuning parameter called the "coding bias" was introduced, and predictions were made for several different values of this parameter. The coding bias is a constant added to the log probabilities of all transitions from noncoding states to coding states and signal states. (It can be thought of as the log of a multiplicative factor for these transition probabilities; after it is applied, the transition probabilities are renormalized.) As the coding bias is increased, more predictions are made, sensitivity tends to increase, and specificity tends to decrease. The coding bias acts as a kind of "bonus" (or "penalty," if negative) for predicting an exon, which can tip the balance between making and not making a prediction if it exceeds the difference in the (maximum) log probabilities of paths that go through, and do not go through, the exon. It is reasonable to leave this parameter free because the true density of exons is not reflected in the training data (and indeed, is unknown). Ten versions of each experiment were conducted, with coding bias values ranging from -20 to $+10$.

The first data set consisted of human-referenced multiple alignments of the complete human, mouse, and rat genomes (UCSC versions hg16, mm3, and rn3), produced with the program MULTIZ [2] (alignments available via the UCSC Genome Browser [19]). Models were trained on the alignments for human chromosomes 1–21, and tested on the alignments for chromosome 22. Noncoding sites were identified by subtracting sites corresponding to known genes, aligned mRNAs and ESTs, and predicted genes (according to any of several gene predictors; see http://genome.ucsc.edu) from all sites in regions where at least two species were aligned. Conserved noncoding sites were defined as the intersection of the set of noncoding sites and the 5% most conserved sites, as defined by the parsimony-based method of Margulies *et al.* [24]. The conserved noncoding sites were also required to be more than 100 bases away from any gene, mRNA, EST, or gene prediction. Sites corresponding to coding exons, start & stop codons, and splice sites were identified by mapping a set of (filtered) RefSeq annotations [32] onto the multiple alignments. Separate 1st order (REV) and 3rd order (R3) phylogenetic models were fitted by maximum likelihood to the sets of conserved noncoding, nonconserved noncoding, all noncoding, and coding (separate 1st, 2nd, and 3rd codon position) sites, using the discrete gamma model for rate variation ($k = 4$ rate categories) [42]. Similarly, separate 1st order (HKY) models were fitted to the sets of sites corresponding to each position of interest in the signal features. The RefSeq annotations for chromosomes 1–21 were also used to estimate state transition probabilities. Four different sets of transition probabilities were estimated, for the cases with and without the CNS state, and with and without the indel model. Sites coinciding with the 2nd data set (see below) were excluded from all training data.

The second data set was an alignment of about 1.8 Mb of the human genome, from the region on chromosome 7 of the gene mutated in cystic fibrosis (CFTR), and homologous sequence from 8 other eutherian mammals (chimpanzee, baboon, mouse, rat, cow, pig, dog, and cat) [38]. The alignment was constructed with the TBA program [2]. There were too few sites in this alignment to estimate all model parameters accurately, so a phylo-HMM was constructed based on estimates from the human/mouse/rat alignments, with some simple heuristics applied to adjust for the larger number of species and new phylogeny (details omitted).

Both data sets were partitioned into 50 kb segments (overlapping by 2 kb), and our programs were run separately on all segments, using the UCSC KiloKluster. Predictions were collected for all versions of ExoniPhy and ExoniPhy--, and all values of the coding bias parameter, then compared against sets of known exons. For the chromosome 22 validation set, we used a union of non-overlapping exons from the "RefSeq Genes," "Known Genes," and "Vega Genes" (see http://vega.sanger.ac.uk and [9]) tracks in the UCSC

Genome Browser (human July 2003 assembly), which had passed some simple filters designed to eliminate annotation errors (3888 exons, total). For the CFTR region, we used a non-overlapping set of exons from RefSeq, which contains all of coding regions described by Thomas *et al.* [38] (117 exons, total). Sensitivity and specificity of predictions with respect to these validation sets were measured at both the nucleotide and exon levels, using standard methods [7]. Because of the added difficulty of predicting exons boundaries in single exon prediction, we kept track of exons that were "nearly correct" (NC), as well as those that were exactly correct (CR). We define NC exons to be incorrect predictions both of whose boundaries are within 6 bases of being correct. We also recorded exon sensitivity levels with respect to the subset of exons that appear (superficially) to be conserved across species (i.e., with an ORF in all species and conserved splice sites or start & stop codons; 3263 of 3888 and 107 of 117 exons, respectively, met these criteria). For comparison, recent sets of predictions from the Twinscan [20], SGP2 [29], SLAM [1], and Genscan [6] programs[4] were compared to the same validation sets, using the same methods. We note that prediction specificity for all programs is almost certainly underestimated, at least for the chromosome 22 data set, because of as yet undiscovered genes.

3. RESULTS

We begin by discussing results for the human/mouse/rat alignments (chromosome 22), which are representative of what we observed with both data sets, then briefly discuss results for the CFTR alignment. In comparisons between versions of ExoniPhy and ExoniPhy--, we emphasize exon-level sensitivity and specificity based on correct or nearly correct (CR+NC) predictions with respect to known exons that are apparently conserved (the only ones currently detectable by our programs).

The effect on prediction performance of introducing context-dependent phylogenetic models is shown in Figure 6. The 3rd order models can be seen to produce a clear, substantial improvement over 1st order models, with both sensitivity and specificity rates increasing by more than 10% over all values of the tuning parameter. For this data set, the independent-site ($N = 1$) version of ExoniPhy shows fairly mediocre performance, while the context-dependent ($N = 3$) version shows respectable sensitivity (reaching more than 80%), and somewhat weaker but still improved specificity (near 50%, in the range of interest). Despite that only three species are available in this data set, the context-dependent models are able to perform reasonably well by exploiting the information encoded in the substitution patterns of tuples of adjacent bases.

A comparison with ExoniPhy-- (Figure 7A) indicates that consideration of the phylogeny and substitution process is indeed essential for ExoniPhy to perform as well as it does. ExoniPhy-- achieves a fairly high rate of sensitivity, but only because it makes a large number of predictions, most of which are false positives. At least with only three species, it appears not to be enough simply to find ORFs that are preserved across species and flanked by conserved splice sites and/or start & stop codons, with a nucleotide triple composition suggestive of coding exons. By making good use of the phylogeny and substitution process, ExoniPhy is able to perform dramatically better than ExoniPhy--.

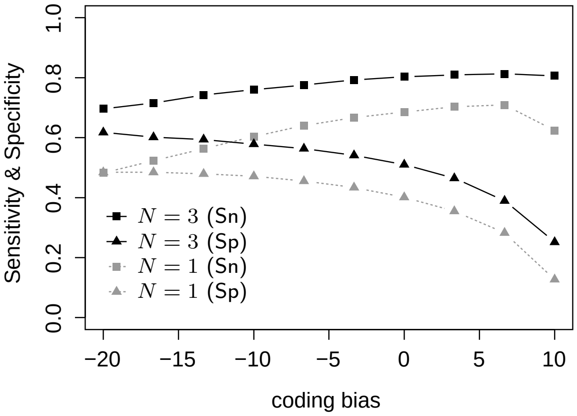

Figure 6: Exon-level sensitivity and specificity of ExoniPhy with context-dependent ($N = 3$) vs. independent-site ($N = 1$) phylogenetic models. Results are for correct or nearly correct (CR+NC) predictions with respect to 3263 apparently conserved exons on chromosome 22.

The addition of the CNS state results in an additional, significant improvement in performance. With the context-dependent model (Figure 7B), the CNS state improves specificity substantially (about 10%), with negligible cost in sensitivity (<1%, in the range of interest). A similar, though less pronounced, effect was seen with independent-site models (results not shown). Inspection of individual predictions indicates that, even with the CNS state, apparent CNS is sometimes still falsely identified as a coding exon. Nevertheless, a large number of false positives are eliminated with this simple strategy.

The indel model provides yet another substantial improvement in prediction accuracy, regardless of which version of the program is considered (Figure 8). The simpler models, in particular, benefit enormously from modeling of indels. When the indel model is used, the sensitivity of ExoniPhy-- approaches 80%, and with a strong negative coding bias, this otherwise very simple program achieves a respectable combination of sensitivity and specificity (panel A). The indel model brings the performance of the independent-site version of ExoniPhy nearly to the level of context-dependent versions lacking the indel model and CNS state (panel B). It improves the sensitivity of the context-dependent version without the CNS state (panel C) by a small but significant amount, but interestingly, does not change its specificity much; it may be that, while the indel model helps in some cases to distinguish CNS from coding sequence, it hurts in others (e.g., CNS that does not contain gaps). When added to the context-dependent version of ExoniPhy with the CNS state (panel D), the indel model produces a significant improvements in both sensitivity (2–3%)

[4] Predictions for the human July 2003 assembly for Twinscan and SGP2 were downloaded from http://genes.cs.wustl.edu/predictions and http://genome.imim.es/genepredictions, respectively. SLAM human/mouse predictions for the Nov. 2002 human and Feb. 2002 mouse assemblies previously obtained from the SLAM authors were mapped to the July 2003 assembly. Genscan is routinely run at UCSC on each new assembly.

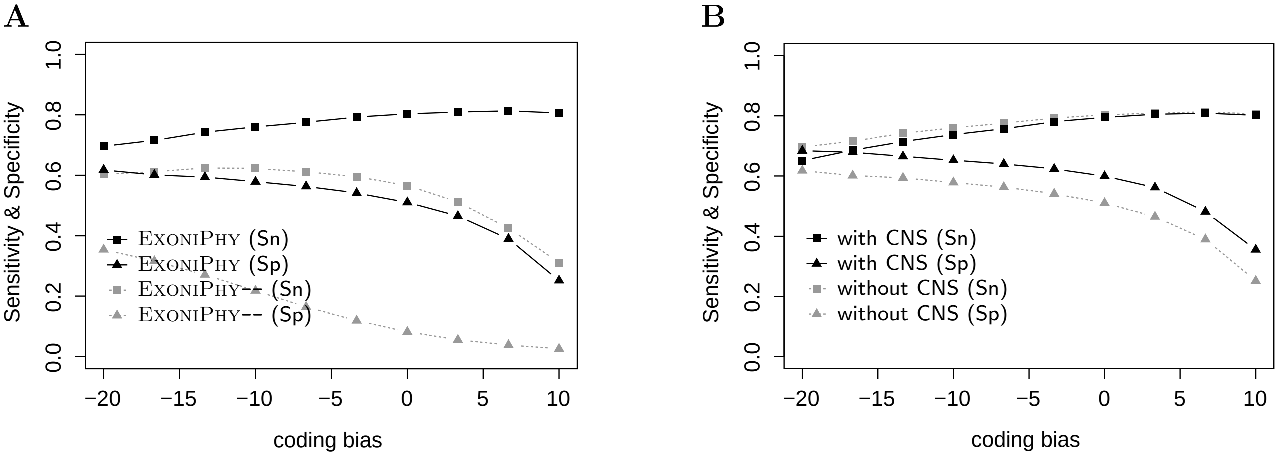

Figure 7: Sensitivity and specificity of (A) ExoniPhy versus ExoniPhy-- (both $N = 3$, no CNS state) and (B) ExoniPhy with and without the CNS state (both $N = 3$). Results are for the same set of exons as in Figure 6.

and specificity (4–5%, in the range of interest), for the best overall performance of all versions considered.

Next, we compared a version of ExoniPhy that incorporated all of these improvements with the other gene predictors. Before performing this comparison, however, we altered the program to use separate phylogenetic models trained on sites from five classes of G+C content (as determined in 50 kb windows): <40%, 40–45%, 45–50%, 50–55%, and >55% (G+C-specific models were used only for the coding, noncoding, and CNS states, not for models associated with signal states). The new version computes the G+C content of an input alignment (here, also 50 kb), then uses the corresponding set of phylogenetic models for prediction; it performs slightly better than the versions described in Figure 8. A performance comparison of this version of ExoniPhy and four other gene predictors, for the chromosome 22 data set, is shown in Table 1. There is clearly room for improvement, but ExoniPhy appears to be competitive with the other programs—somewhat remarkably, given the limitations inherent in single exon detection, the use of a simplified HMM structure, and relatively crude methods for splice-site detection. Notably, ExoniPhy was found to have the highest exon-level specificity of all programs tested (tied with SLAM in the CR case). Its sensitivity overall is somewhat weak, but increases substantially (by about 12%, for CR+NC predictions) when computed with respect to exons that are (apparently) conserved across species—our target class of exons.

The results for the 9-species CFTR alignment were generally similar to those discussed above, but there were a few differences worth noting. The CNS state was found to be somewhat more important with the CFTR alignment (which includes a large number of highly conserved noncoding regions [24]), and the use of context-dependent models was found to be slightly less important (but still to result in a significant improvement). The indel model also generally produced a less dramatic improvement than with the first data set, and in the case of context-dependent phylogenetic models and the CNS state, it produced almost no net change in performance. (While it eliminated a few false positive predictions that had alignment gaps suggestive of noncoding regions, it also caused a few exons to be missed, because they included complex gap patterns or gaps with non-multiple-of-three lengths.) Table 1 also includes a comparison of ExoniPhy with other gene predictors for the CFTR data set. Most of the programs perform better in this relatively well-conserved region of chromosome 7 than on the whole of chromosome 22, making it difficult to gauge the effect on ExoniPhy of considering additional species. Our impression, however, is that the larger set of species improves performance where exon structure is well conserved, but also leads to more mispredictions due to violations of ExoniPhy's relatively rigid requirements for exon boundary conservation and indel patterns (see Discussion). As before, ExoniPhy lags behind the leading programs in terms of sensitivity, but has the highest specificity of all programs tested. Its nucleotide-level specificity is nearly perfect (98%), indicating that most false-positive exons are incorrect only in terms of the placement of their boundaries.

4. DISCUSSION

We have proposed three methods for improving the performance of phylo-HMMs in the prediction of conserved coding exons from aligned DNA sequence data: the use of context-dependent phylogenetic models, explicit modeling of conserved noncoding sequence (CNS), and a new way of modeling indels by expanding the state-space and altering the emission probabilities of the phylo-HMM. Experimental results based on two large data sets indicate that all three methods produce substantial improvements in prediction accuracy, and that they are useful in combination. These methods have been incorporated into a phylo-HMM-based exon predictor called ExoniPhy, which appears to be competitive with some of the best available gene predictors, despite that it is still quite primitive in certain respects. ExoniPhy's specificity is particularly good, and with some additional work, the program may become an important new tool for identifying candidate genes to be tested in the laboratory. (Interestingly, in our chromosome 22 experiments, 70% of "wrong" exons overlapped with mRNAs aligned to the genome; many may actually be coding exons.)

We are currently considering several ways of improving ExoniPhy. One possibility is to relax the constraint that exon boundaries must be consistent across all species, or

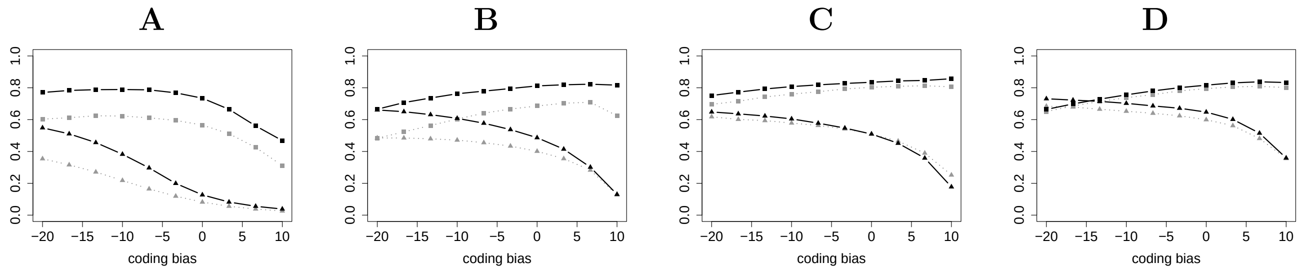

Figure 8: Sensitivity and specificity with and without the indel model for (A) ExoniPhy--, and for ExoniPhy with (B) independent-site models and no CNS state, (C) context-dependent models and no CNS state, and (D) context-dependent models and the CNS state. In all cases, the version with the indel model is represented by solid black lines and the version without by dotted gray lines. As in the previous plots, squares indicate sensitivity and triangles indicate specificity. Results are for the same set of exons as in Figure 6.

Chromosome 22

	Nucleotide		Exon (CR)		Exon (CR+NC)*		
Program	Sn	Sp	Sn	Sp	Sn	Sp	CSn†
Twinscan	0.834	0.711	0.768	0.641	0.785	0.671	0.849
SGP2	0.819	0.722	0.711	0.603	0.727	0.617	0.792
SLAM	0.602	0.895	0.534	0.664	0.558	0.693	0.633
Genscan	0.872	0.526	0.720	0.418	0.737	0.427	0.783
ExoniPhy	0.751	0.826	0.635	0.664	0.671	0.702	0.790

CFTR

	Nucleotide		Exon (CR)		Exon (CR+NC)*		
Program	Sn	Sp	Sn	Sp	Sn	Sp	CSn†
Twinscan	0.950	0.825	0.872	0.734	0.889	0.748	0.907
SGP2	0.973	0.801	0.863	0.682	0.897	0.710	0.925
SLAM	0.796	0.826	0.641	0.500	0.675	0.527	0.692
Genscan	0.847	0.577	0.778	0.526	0.778	0.526	0.785
ExoniPhy	0.854	0.980	0.701	0.752	0.769	0.826	0.841

*Allowing for "nearly correct" as well as correct exons (see text).

†Sensitivity with respect to apparently conserved exons.

Table 1: Comparison of ExoniPhy with four other gene predictors, based on a version with $N = 3$, the CNS state, and the indel model. The coding bias was chosen approximately to maximize exon-level Sn+Sp. Except for the last column, results are for complete versions of both data sets (3888 and 117 exons, respectively).

similarly, to allow whole exons to be absent in a subset of species. Generalizing the current model to allow for such differences in exon structure would contribute a certain amount of additional complexity, but appears to be feasible, and would likely result in a substantial improvement in prediction sensitivity. (Exon-level specificity would be improved as well, because exons with incorrect boundaries tend to be predicted where exon structure is not conserved.) This change may be essential as more species are considered, especially if more distant vertebrates (e.g., chicken) are included along with mammals (which may help in distinguishing between coding and conserved noncoding regions). In addition, certain features now de rigueur in gene prediction should be added, e.g., better methods for splice site detection and the ability to model non-geometric exon length distributions (perhaps via a generalized HMM framework [21, 25]). Other possibilities for improvement include using patterns of conservation more effectively in determining exon boundaries, and developing better models for CNS.

The new indel model also raises the interesting possibility of using a phylo-HMM for multiple alignment, similar to the way a pair HMM is used for pairwise alignment [10]. Multiple alignment might be addressed in conjunction with gene finding, as has been done for pairwise alignment in SLAM [1] and Doublescan [26]. Indeed, our experience with the CFTR data set, in which certain exons were missed because of "dirty" gaps, suggests that ExoniPhy might benefit from a certain amount of "on the fly" alignment. The best strategy may be to realign within a narrow band in alignment space, centered on a starting alignment that has been produced with an existing genomic-scale multiple aligner [30]. The state space and number of parameters are large in our current gapped phylo-HMM, and it may be necessary to find ways of restricting them further.

Finally, it is worth contrasting the "ancient exons" strategy presented here with the strategy of McAuliffe *et al.* [25], who used a phylo-HMM for gene prediction in the context of "phylogenetic shadowing" [3]. The idea of phylogenetic shadowing is to use aligned sequences from many closely related species (e.g., a dozen or more primates) so that features that have developed recently in evolutionary history (which will be undetectable by our methods) can be identified. On the other hand, the use of closely related species (even if large in number) may limit one's ability to detect more ancient genes, because not enough time has passed for a distinctive pattern of substitution to become evident. Thus, our approach may be more effective for distinguishing ancient, conserved exons from CNS. Our strategy is also motivated by the availability of whole-genome data; the current trend is to sequence a set of phylogenetically diverse vertebrate genomes rather than a large number of primate genomes. (We note that the question remains mostly open of what species—how many and with what phylogenetic relationships—will be optimal for multi-species gene finding, despite a small experimental study by McAuliffe *et al.* [25]; the similar question of the optimal distance between genomes for pairwise gene prediction has been studied carefully [44].) In any case, the "ancient exons" strategy of this paper and the phylogenetic shadowing strategy are likely to be complementary. Indeed, a program that allowed for changes in exon structure across species would potentially be able to combine the strengths of both approaches.

Acknowledgments

We thank Webb Miller and Jim Kent for their work on multiple alignment, Mathieu Blanchette for identification of conserved sites, Lior Pachter and Gill Bejerano for stimulating conversations about multi-species gene finding, and the anonymous reviewers of this paper for constructive comments. Access to the CFTR data set prior to its publication was generously provided by Eric Green and colleagues. Michael Brent, Paul Flicek, and Roderic Guigó provided helpful information about recently posted Twinscan and SGP2 predictions. This work was funded by the Howard Hughes Medical Institute (D.H.), NHGRI grant IP41HG02371 (A.S.), and an ARCS foundation scholarship (A.S.).

5. REFERENCES

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

[2] M. Blanchette, W. J. Kent, C. Riemer, et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 2004. In press.

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

[4] N. Bray and L. Pachter. MAVID multiple alignment server. *Nucleic Acids Res.*, 31:3525–3526, 2003.

[5] M. Brudno, C. Do, G. Cooper, M. F. Kim, E. Davydov, E. D. Green, A. Sidow, and S. Batzoglou. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.*, 13:721–731, 2003.

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

[7] M. Burset and R. Guigó. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[8] F. Chiaromonte, R. J. Weber, K. M. Roskin, M. Diekhans, W. J. Kent, and D. Haussler. The share of human genomic DNA under selection estimated from human-mouse genomic alignments. In *Cold Spring Harbor Symp. Quant. Biol.*, 2003. In press.

[9] J. E. Collins, M. E. Goward, C. G. Cole, et al. Reevaluating human gene annotations: a second-generation analysis of chromosome 22. *Genome Res.*, 13:27–36, 2003.

[10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[11] J. Felsenstein. Evolutionary trees from DNA sequences. *J. Mol. Evol.*, 17:368–376, 1981.

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

[13] W. M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.*, 20:406–416, 1971.

[14] P. Flicek, E. Keibler, P. Hu, I. Korf, and M. R. Brent. Leveraging the mouse genome for gene prediction in human: From whole-genome shotgun reads to a global synteny map. *Genome Res.*, 13:46–54, 2003.

[15] N. Goldman, J. L. Thorne, and D. T. Jones. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.*, 263:196–208, 1996.

[16] R. Guigó, E. T. Dermitzakis, P. Agarwal, et al. Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes. *Proc. Natl. Acad. Sci. USA*, 100:1140–1145, 2003.

[17] M. Hasegawa, H. Kishino, and T. Yano. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.

[18] I. Holmes and W. J. Bruno. Evolutionary HMMs: a Bayesian approach to multiple alignment. *Bioinformatics*, 17:803–820, 2001.

[19] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Res.*, 12:996–1006, 2002.

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

[21] D. Kulp, D. Haussler, M. G. Reese, and F. H. Eeckman. A generalized hidden Markov model for the recognition of human genes in DNA. In *Proc. 4th Int'l Conf. on Intelligent Systems for Molecular Biology*, pages 134–142, 1996.

[22] P. Liò and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res.*, 8:1233–1244, 1998.

[23] G. A. Lunter, I. Miklós, Y. S. Song, and J. Hein. An improved algorithm for multiple alignment on arbitrary phylogenetic trees. *J. Comp. Biol.*, 10:869–889, 2003.

[24] E. H. Margulies, M. Blanchette, NISC Comparative Sequencing Program, D. Haussler, and E. D. Green. Identification and characterization of multi-species conserved sequences. *Genome Res.*, 13:2507–2518, 2003.

[25] J. D. McAuliffe, L. Pachter, and M. I. Jordan. Multiple-sequence functional annotation and the generalized hidden Markov phylogeny. Technical Report 647, Department of Statistics, University of California, Berkeley, 2003.

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

[27] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[28] W. J. Murphy, E. Eizirik, S. J. O'Brien, et al. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science*, 294:2348–2351, 2001.

[29] G. Parra, P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett, and R. Guigó. Comparative gene prediction in human and mouse. *Genome Res.*, 13:108–117, 2003.

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

[31] E. Pennisi. Gene counters struggle to get the right answer. *Science*, 301:1040–1041, 2003.

[32] K. D. Pruitt and D. R. Maglott. RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, 29:137–140, 2001.

[33] K. M. Roskin, M. Diekhans, and D. Haussler. Scoring two-species local alignments to try to statistically separate neutrally evolving from selected DNA segments. In *Proc. 7th Annual Int'l Conf. on Research in Computational Molecular Biology (RECOMB'03)*, pages 257–266, 2003.

[34] A. Siepel and D. Haussler. Phylogenetic hidden Markov models. Submitted book chapter.

[35] A. Siepel and D. Haussler. Combining phylogenetic and hidden Markov models in biosequence analysis. In *Proc. 7th Annual Int'l Conf. on Research in Computational Molecular Biology (RECOMB'03)*, pages 277–286, 2003.

[36] A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.*, 2004. In press.

[37] C. W. Sugnet, W. J. Kent, M. Ares, and D. Haussler. Transcriptome and genome conservation of alternative splicing events in humans and mice. In *Proc. 9th Pacific Symp. on Biocomputing*, 2004.

[38] J. W. Thomas, J. W. Touchman, R. W. Blakesley, et al. Comparative analyses of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793, 2003.

[39] J. L. Thorne, H. Kishino, and J. Felsenstein. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33:114–124, 1991.

[40] N. Wade. Gene sweepstakes ends, but winner may well be wrong. New York Times, June 3, 2003.

[41] S. Whelan, P. Liò, and N. Goldman. Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends Genet.*, 17:262–272, 2001.

[42] Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.*, 39:306–314, 1994.

[43] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.

[44] L. Zhang, V. Pavlovic, C. R. Cantor, and S. Kasif. Human-mouse gene identification by comparative evidence integration and evolutionary analysis. *Genome Res.*, 13:1190–1202, 2003.