Improved Gene Prediction through Human/Mouse Ortholog Similarity Constraints

Stephanie Seneff[1], Chao Wang[1], and Chris Burge[2]

[1]MIT CSAIL
Massachusetts Institute of Technology
Cambridge, MA 02139

[2]Department of Biology
Massachusetts Institute of Technology
Cambridge, MA 02139

{seneff,wangc}@csail.mit.edu,cburge@mit.edu

Keywords

Gene prediction, genetic diversity, comparative genomics

Abstract

Given the availability of complete genome sequences from related organisms, it should be possible to exploit sequence conservation in predicting gene structure. In particular, one should be able to leverage knowledge about known genes in one species when trying to identify new genes in another. Such an approach is appealing in that high quality gene prediction can be achieved for newly-sequenced species, such as mouse and fugu fish, using the extensive knowledge that has been accumulated about human genes. In this research, we report a novel approach to predicting mouse genes by incorporating constraints from orthologous human genes using techniques that have previously been exploited in speech and natural language processing applications. Our approach uses a context-free grammar to parse a training corpus of annotated human genes. A statistical training procedure produces a weighted recursive transition network (RTN) intended to capture the general features of a mammalian gene. This RTN is expanded into a finite state transducer (FST) and composed with an FST capturing the specific features of the human ortholog. This model includes a trigram language model on the amino acid sequence as well as exon length constraints. A final stage uses the free software package, CLUSTALW to align the top N candidates in the search space. For a set of 98 orthologous human-mouse pairs, we achieved 96% sensitivity and 97% specificity at the exon level on the mouse genes, given only knowledge gleaned from the human genome.

1 Introduction

The computational biology community is experiencing an enormous growth in the number of sequenced genes that become available for research purposes every day. Looking to the future, it will become increasingly important to be able to leverage knowledge about one species to help in codifying the nucleotide sequences obtained for other species. At this time, the knowledge available for the human is much more precise and extensive than that for other vertebrates. However, with the recent milestone of the establishment of the complete mouse genome sequence [12], it becomes of paramount importance to accelerate

the pace at which new genomic sequences can be accurately decoded. It is now known that there is remarkably strong conservation of the nucleotide sequences within the coding exons for related species, on the order of 97% for humans compared with other primates, and about 85% for the much more distant pair of human-mouse orthologs. As discussed in [2, 12], there appears to be a remarkable conservation of individual exon length between the human and the mouse. This feature makes it feasible to exploit statistical methods that would otherwise fall apart because of an unwieldy search space.

1.1 Goals

Our goal in this work is to develop a statistical language model for gene finding by exploiting orthologous pairs, borrowing techniques previously applied to speech understanding. To begin our explorations, we conducted a preliminary experiment in which we used simple n-gram statistics to attempt to match up orthologous gene pairs. In particular, we trained an *amino acid trigram* language model for each human gene of a pair and selected the highest scoring mouse protein (among 102 candidates) as the proposed orthologous mouse gene. We found that the matching was nearly perfect. This, together with the knowledge that the lengths of individual exons are strongly conserved across different mammalian species [2], inspired us to design a gene-finding procedure that makes use of n-grams and exon length constraints as critical components. The other necessary ingredient to success would be a generic statistical model of a typical mammalian gene, that would map from the raw nucleotide sequence to the sequence of amino acids specifying the resulting protein.

1.2 Background

We have long exploited natural language techniques to aid in the process of understanding human speech. Our methods are based on parsing a corpus of orthographic transcriptions of users' utterances based on a context free grammar formalism [14] , then inducing a language model for the recognizer from an automatic analysis of the parse trees [15]. Our speech recognition framework [8] makes use of a finite state transducer (FST) formalism [13, 11] to define the search space. This formalism defines a space of interconnected "states," with a state transition matrix characterizing the connections among the states and supporting simultaneously a mapping from an input symbol to an output symbol, with an associated probability. For speech, we typically map in stages from phonetic (e.g., "flap") to phonemic ("/t/") realizations, subsequently grouping phoneme sequences into words ("gua*t*emala"), then optionally concatenating words into multi-word units ("guatemala_city") and finally word classes ("city_name"). A class n-gram language model provides critical constraint for the recognition task. A more sophisticated approach is to augment the FST with *recursive transition networks* (RTNs) [18], to support a hierarchical model where selected transitions on arcs are associated with an entire sub-network, identified by a unique name. This permits a direct encoding of a context free grammar into the recognizer's search space.

In our research on spoken dialogue systems, we have explored several options for integration between speech recognition and natural language understanding, where our goals were to deduce an effective sta-

tistical language model for the recognizer directly from the natural language grammar. We have recently been successful with two different techniques, both of which are based on parsing a large corpus of utterances and tabulating counts in the parse trees to determine the probability model. The distinction between the two approaches is in the complexity of the resulting recognizer language model. The simpler technique [15] induces a traditional class n-gram language model, whereas the more complex alternative [17] includes component categories that are represented by a recursive transition network (RTN) [18], allowing a structured encoding of layers above the preterminal layer in terms of a context-free grammar. We typically include bigram statistics on transitions within each layer of such an RTN, computed directly from the parse trees acquired for the training corpus. For speech applications, we have typically found that an RTN formulation is less successful. This is mainly because, for most applications, the RTN can not be expanded into a finite state network, and therefore suffers from performance loss in terms of computation required to evaluate the recursion on the fly.

Our first thought was that techniques that worked best for speech would also be preferred for the genome parsing problem. In speech applications, words that form a natural set within a semantic class are grouped and replaced by their class label in the training sentences, with a within-class unigram statistic accounting for their internal distribution within the class. Word sequences must sometimes be concatenated into artificial compound words in order to simplify class membership to a list of items. Thus "salt lake city" becomes "salt_lake_city." The class then stands in for all of its member words (both singletons and compounds) in the sentence-level bigram statistics. A parallel in genomics would be to create compound words to account for all of the codings from nucleotides to amino acids. A properly constructed grammar could be used to tag exon-internal sequences according to their triple-code protein transformations, for example, producing "classes" like "<Leu>" containing "word sequences" like "t_t_a," "t_t_g," "c_t_t," "c_t_a," "c_t_c," "c_t_c," and "c_t_g." Nucleotides in introns could be tagged for the phase of the reading frame, in order to retain knowledge of the phase across the gap between the individual exons.

The alternative approach is to select a subset of the *nonterminal* categories in the NL grammar as classes in a class n-gram, and to expand those classes using a recursive transition network (RTN), coded directly from the rules in the grammar subsumed by the specially selected categories. While this approach is often impractical for speech applications, the complexity of grammars needed for genomics is considerably reduced, and it has the advantage that statistics can be shared across similar contexts. For example, it seems counterproductive to split the statistics on the introns into three distinct subgroups just because of the phase of the reading frame in flanking exons. An RTN can easily be configured such that the three intron classes can share a common nucleotide bigram language model, which can also be used for the nucleotide sequences flanking the outer edges of the gene.

It also becomes very straightforward to write rules to express positional bigram statistics in the 3' and 5' splice site motif patterns, which are then covered by a separate subnetwork within the RTN. We found that an RTN constructed for genomic sequences in this fashion could be automatically expanded into a finite state network, which could then be composed with FSTs representing other components of our model to produce an efficient search graph. This thus became our preferred strategy for representing the generic mammalian gene model.

1.3 Overview of the Approach

Our approach makes use of a generic mammalian gene model as well as specific constraints from the human orthologous gene when predicting the structure of a mouse gene. The generic gene model was obtained by parsing a training corpus of 400 annotated human genes using a context-free grammar[1]. A probabilistic training procedure produces a weighted recursive transition network (RTN) intended to account statistically for most of the distinct features of a typical gene (introns, exons, and 3' and 5' splice sites). This network, converted into a finite state transducer (FST), defines the basic search space used in predicting the structure of a mouse gene. The search space is further enhanced with exon length and amino acid n-gram model constraints obtained from the corresponding human ortholog. A search through the space, given an input mouse genomic sequence, produces an N-best list of alternative protein hypotheses, which can be resorted using standard sequence alignment tools, such as CLUSTALW [19]. Thus a formal alignment between the human and mouse orthologs is deferred until the post-processing stage.

All the models used in our approach make use of the finite state transducer representation, and the gene prediction procedure utilizes the FST toolkit developed in the Spoken Language Systems group at MIT, which is based on [13, 11].

2 Methodology

Central to our methodology is a statistical model for the genomic sequence, which is essentially a hidden Markov model (HMM). The hidden states in the model correspond to various basic functional constituents of a gene (e.g., exon, intron, splice sites, etc.), and the emission probability is defined as the likelihood of observing a particular nucleotide sequence conditioned on the state. Thus, the joint probability of an observed genomic sequence (x) and the corresponding state sequence (s) can be expressed as:

$$\mathrm{p}(x, s) = \mathrm{p}(s_0) \prod_{i=1}^{L-1} \mathrm{p}(x_i|s_i)\mathrm{p}(s_i|s_{i-1}) \tag{1}$$

in which x_i is the observed nucleotide sequence for state s_i, and L is the total number of states.

The problem of predicting the structure for a genomic sequence can then be solved by finding the state sequence that maximizes the probability:

$$s^* = \arg\max_s \mathrm{p}(x, s) \tag{2}$$

The state sequence encodes the proposed genetic structure of the input DNA sequence.

Although our gene model is equivalent to an HMM in the probability formulation, it was trained via an efficient parsing mechanism [14] and encoded as a weighted Recursive Transition Network (RTN). The top level of the RTN corresponds to the HMM model states (s_i). Some of the top level nodes are expanded

[1] Thus we are assuming that the human genome is representative of all mammals.

recursively, down to a sequence of terminal nucleotides (x_i), according to the rules of the grammar. The emission probability of observing that sequence can be computed by multiplying the probabilities on all the arcs visited by the expansion of the sub-level RTNs. For example, the 5' splice site is represented as a top level node that eventually expands into a sequence of 20 nucleotides, and the emission probability of this sequence, $\mathrm{p}(x_i|s_i = 5' \; splice \; site)$ is computed as a product of the RTN weights[2]. This generic gene model is enhanced with human ortholog-specific information, to provide effective constraints in processing the orthologous mouse gene.

In the remainder of this section, we first give an overview of our gene prediction procedure, followed by detailed descriptions of each component module.

2.1 Overview of gene prediction procedure

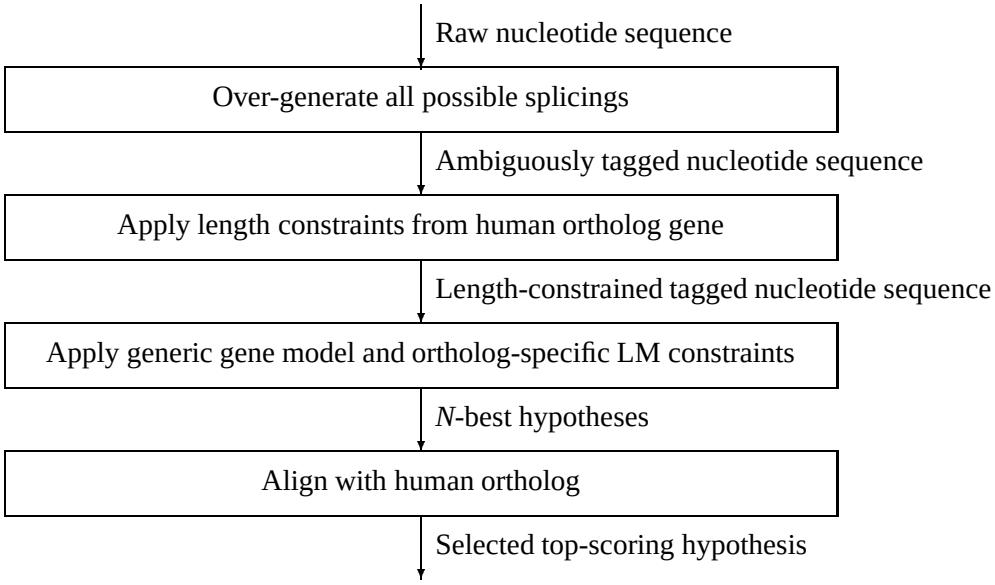

Figure 1: **Block diagram of procedure used to extract mouse gene structure via analogy with known human ortholog.**

The procedure to process a single mouse gene through our model requires several steps, as outlined in Figure 1. Each raw mouse sequence was pre-processed to over-generate all potential exons. This FST is then pruned by imposing exon length structure constraints, obtained from the annotated human orthologous gene[3]. The generic gene model is then applied to score alternative hypotheses available in the graph, as well as translating them into amino acid hypotheses. An amino acid trigram model, trained from the protein sequence of the human ortholog, is then applied. Finally, a hypothesized N-best list of the

[2]In practice, the weights on the RTN are negative log probabilities, so that a sum is used in computing the total probability.
[3]Hypothesized orthologs could in principle be acquired using blast analysis.

top-ranking candidates can be re-ranked by aligning each hypothesis with the human ortholog amino acid sequence, using a standard alignment tool such as CLUSTALW [19]. The final highest scoring alignment provides a hypothesized protein sequence for the mouse ortholog, segmented into a sequence of proposed exons.

2.2 Initial processing

Each raw mouse sequence was pre-processed to support hypothesized exon start and end loci wherever they were possible according to strict rules for specific two- or three-nucleotide sequences at their edges, as illustrated in Figure 2. This results in a finite state transducer mapping raw DNA sequences to alternatively tagged sequences.

<**exoni**>	before every	**atg**
<**exon**>	after every	**ag**
</**exon**>	before every	**gt**
</**exonf**>	after every	**STOP (taa | tag | tga)**

Figure 2: **Special tags inserted into raw genomic sequences in the initial processing phase.** <exoni> = beginning of initial coding exon; <exon> = beginning of internal exon; </exon> = end of internal exon; </exonf> = end of final coding exon.

2.3 Generic Gene Model

To train a generic gene model for the mammalian genome, we developed a context-free grammar that encodes critical aspects of the genomic structure, including accounting explicitly for substructure in the motif sequences at both the 3' and 5' splice sites of the intron, as outlined in Figure 3. The grammar also preserves reading frames between adjacent exons.

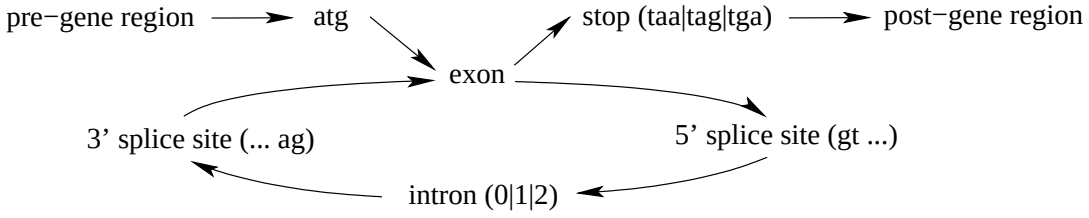

Figure 3: **Basic structure of our gene model.** Internal introns remember the reading frame to assure correct coding of the nucleotides into amino acids.

The portion of the grammar accounting for the amino acids, as illustrated in Figure 4, captures a statistical map from nucleotide sequences to amino acid sequences. A nucleotide bigram language model

exon_i_start				exon							stop_seq			exonf_end
				AA			...	AA						
				c	ca	Gln	...	c	cg	Arg	t1	a2	Stop	
<exoni>	a	t	g	c	a	g	...	c	g	a	t	a	a	</exonf>

Figure 4: **Schematic of our structural model for an exon, in the simple case of a very short single exon gene.** The preterminal symbol, "ca" stands for the specific situation of the nucleotide "a" following the nucleotide "c" in the second position of the triplet code. The third position in the model uniquely specifies the amino acid.

3'_motif										
Nt1	Nc2	Na3	....	Nt13	Nc14	Nc15	Nc16	Nc17	Nt18	ag
t	c	a	....	t	c	c	c	c	t	a g <exon>

Figure 5: **System's statistical model for the 3' splice site motif, consisting of the twenty nucleotide sequence up to and including the obligatory "ag."** This model captures positional bigram statistics, which is equivalent to an inhomogeneous first-order Markov model [3].

encodes the statistics of all introns. The model for the 3' splice site motif, which takes into account the 18 nucleotides preceding the "ag" signature of exon onset, is basically a positional bigram, as illustrated in Figure 5. The model for the 5' splice site motif is shorter, yet more intricate, as we wanted to account for the known distinction between situations where the nucleotide "g" is present or absent at the position just preceding the official end of the exon (See Figure 2 in [4]). When the exon ends in-phase with the reading frames, it seemed too difficult to encode this "G"/"not-G" distinction along with the protein coding process, so this distinction was only made for the out-of-phase exons. An example of the parse tree for an exon which ends in phase 2[4], and ends in a "not-G" configuration, is illustrated in Figure 6. Figure 7 shows that the distributions in the next-to-last position of the 4-sequence 5' splice site motif model are distinctly different for the "G"/"not-G" subsets, as predicted from the literature.

ab_end						
nuc1	h2	exon_end_h				
		exon_end	h1	h2	h3	h4
g	a	</exon> g t	g	a	g	t

Figure 6: **Model for the 5' splice site motif in the case where two nucleotides of the split codon have immediately preceded the exon boundary, and the last nucleotide before the boundary was not "G."**

The gene model is trained by parsing annotated human genes using this grammar. A corpus of about 400 human genes was used in estimating the probabilities of the model. The training genes were truncated

[4]Intron splits codon after second position.

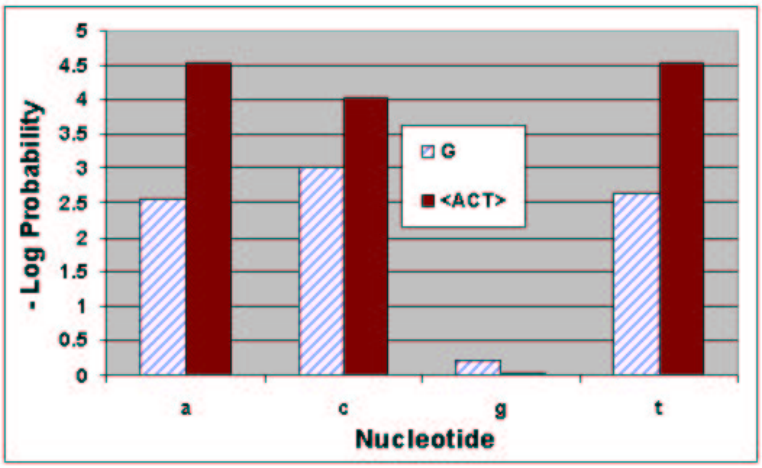

Figure 7: **Log probabilities obtained in our models for the four nucleotides in position X in the 5' splice site motif: "n n G|H </exon> g t n n X n", conditional on G or H (<ACT>) at position "G|H."**

at 1000 nucleotides preceding the first coding exon and 1000 nucleotides subsequent to the end of the last coding exon. Some characteristics on these genes are presented in Table 1. Statistics were tabulated from the parse trees for this corpus, and an RTN model was produced encoding the grammar, with negative log probabilities on transitions. This RTN was then expanded into a finite state transducer, such that it could be combined with additional constraints from the human ortholog.

2.4 Length constraints

As discussed in both [12] and [2], it appears that the lengths of corresponding exons of human and mouse orthologs are strongly conserved. Batzoglou *et al.* [2] found that 73% of exon lengths were identical, and the differences, when they occurred, were quite small and were nearly always a multiple of three. The introns, on the contrary, seemed to have considerably different lengths between the two species.

We used a finite state transducer to encode the intron/exon length constraints. In our FST length model, the introns are represented by a single state supporting all possible nucleotides in a self-loop, resulting in no length constraints for introns. The exons are represented as a cascade of one-nucleotide acceptors; the length of the cascade encodes the exon length explicitly. Given an annotated genomic sequence, we could derive a "strict" length model, essentially insisting that the length be conserved for all the exons in the gene. A more general solution would be to allow insertions and deletions of up to N codons (multiples of 3 nucleotides) in each exon, to support the most common types of variations.

There are other types of exon length variations, including merging and splitting of exons, and lengths differing by other than a multiple of three. We can account for the merging of two human exons easily in our model, by providing a transition by-passing the intron state. The inverse problem of splitting an exon into two is more difficult, due to the many possible sites in which splitting could occur. In general,

one could account for all the variations with a more complex model, but at the expense of significantly increased ambiguity. We chose to ignore the less common variations (except merging) in our model, recognizing that our approach will not be able to recover those exons correctly. In Section 3, we will describe an experiment analyzing the trade-offs in selecting N, the maximum number of codons we allow an exon to insert or delete.

Figure 8 illustrates our model (for $N = 1$) with a simple example.

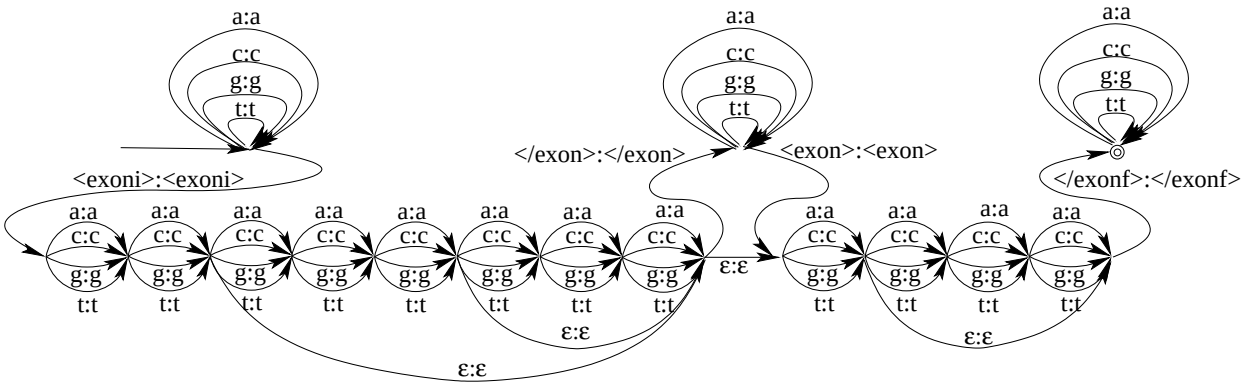

Figure 8: **An example length constraint FST for a made-up sequence "... <exoni> a t g t a </exon> g t ... a g <exon> a </exonf> ...".** In this example, we allow up to one codon insertion or deletion in each exon, as well as a merge of exons. In addition to the original exon length pattern "5 1", this FST also supports the following combinations: "2 1", "8 1", "2 4", "5 4", "8 4", "3", "6", "9", and "12".

2.5 Amino-acid language model

We applied an amino acid trigram model, also encoded as an FST, to adapt the generic gene model to the particular ortholog under consideration. The model is estimated from the amino acid counts in each human protein sequence. The Deleted Interpolation technique [1] is used for smoothing, with probabilities estimated using a variation of the expectation maximization (EM) algorithm [6]. This technique is identical to that used for our speech applications. The vocabulary of this language model is based on the 20 amino acids, but is enhanced with three phase markers at exon boundaries.

2.6 Post-processing via alignment

Global alignment between human and mouse orthologous protein sequences can in theory provide stronger constraints than n-gram models, which are simply based on frequencies of localized patterns. Thus, it is possible to further improve the system performance after the n-gram model is applied, by explicitly aligning the human ortholog with each of the N-best hypotheses produced by the system, in a re-ranking step. For this purpose, we used the publicly available general purpose multiple sequence alignment program

CLUSTALW. CLUSTALW can calculate the best match between multiple DNA or protein sequences, and produce a score associated with each match. We converted the N-best hypotheses into protein sequences and aligned each of them with the known protein sequence of the human ortholog. The one with the highest alignment score is then chosen to be the system output. We used the default settings of CLUSTALW, so that no special tuning was done to adapt the tool for aligning human-mouse orthologs. The N-best list size is fixed to be 100 in our experiments, although one could optimize this parameter if an independent set of development data is available.

3 Results and Discussion

We evaluated our approach using the same set of human-mouse ortholog pairs that had been used in [2]. The original data set contains a total of 117 pairs of orthologs. However, some of the genes contain alternatively spliced coding sequences based on the GenBank "CDS" annotation. We also found that there are about 3 mouse genes whose introns have the non-consensus terminal dinucleotides ("gc..ag"), a recognized variant, and 6 mouse genes with dinucleotides other than "gt..ag" or "gc..ag" (possibly due to sequencing or annotation errors). These data were excluded from our evaluation, leaving 102 ortholog pairs in our final test set. These genes are on average shorter than the ones we used for training our generic gene model, as shown in Table 1.

	Training		**Testing**	
Property	**MIN**	**MAX**	**MIN**	**MAX**
total length (nucleotides)	1500	17,000	700	13,500
total length of coding sequences (nucleotides)	200	4000	200	2100
total number of exons	2	25	1	18

Table 1: **Distributions of 400 genes selected for training and testing our generic mammalian gene model.**

The criterion we used for evaluation is based on exactly matched coding exons. In particular, we use the *sensitivity* and *specificity* measures [5], which correspond to precision and recall used in information retrieval evaluations. Sensitivity is defined as the ratio of the number of correctly identified exons over the total number of exons in the reference; specificity is defined as the ratio of the number of correctly identified exons over the total number of hypothesized exons.

3.1 Results

The only significant parameter we chose to tune in our system is N, the maximum number of codons we allow to insert or delete in the exon length constraints. Figure 9 summarizes the impact of N on the system performance. We are not always able to find an orthologous mouse exon-intron structure for every human

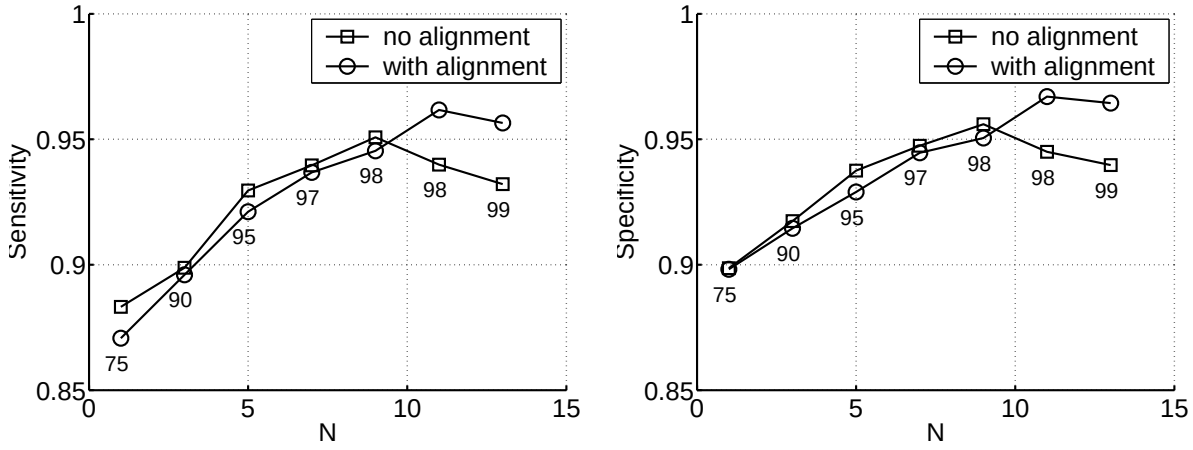

Figure 9: **Sensitivity and specificity on correctly identifying mouse exons as a function of N, the maximum number of codon insertion/deletions allowed in the length FST model.** N varies from 1 to 13 in the plots. The labels next to the data points indicate the total number of genes that our system is able to predict under each N.

gene. For example, we are able to predict gene structures for 98 mouse sequences (out of 102 in total) when we allow up to 9 codon insertions/deletions in each exon. This is due to the restrictions imposed by the length constraints; i.e., when the mouse exon length variation is beyond the coverage of the length constraints FST, the search could fail to find any gene in the mouse genomic sequence. We consider this a favorable feature of our algorithm - it is probably better to fail than to produce an erroneous result. For the failed cases, one can relax the length constraints, or adopt a different approach such as those based on alignments.

The sensitivity and specificity measures in the plots are calculated on the subset of genes that our system can produce an answer for, for different values of N. As shown in the figure, there is clearly a trade-off in choosing N. Since we have no chance of correctly identifying those mouse exons that varied by more than N codons, a small N will result in a significant number of errors due to those hard failures. It also results in more null outputs due to total search failures. As we increase N, we can generally produce outputs for more genes. However, with a large N, the performance could degrade due to increased ambiguity, as indicated by the downward trend in the figure beyond $N = 9$. The optimal performance was 96.2% sensitivity and 96.7% specificity for coding exons, which was achieved with N equal to 11 and with post-processing using the CLUSTALW alignment tool.

3.2 Discussion

It is interesting to observe that post-processing using CLUSTALW did not yield any further improvement over using the simple amino acid trigram model until N reaches 11. This seems to suggest that, when the exon length constraints are relatively strict, the trigram is an adequate model for incorporating human pro-

tein sequence constraints. However, the explicit alignment with human protein sequence via CLUSTALW provides stronger "language model" constraints than n-grams, and eventually out-performs the trigram model as N grows. (The trigram model, even though it was not able to predict the correct gene structure as the top candidate, was able to produce the correct answer in its N-best outputs.)

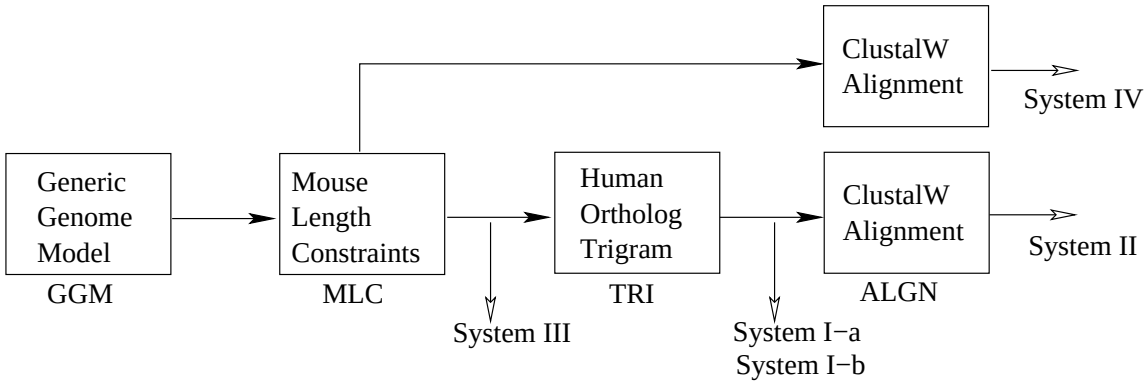

Figure 10: **Schematic of experiments on different system configurations for gene prediction of the mouse gene based on the human ortholog.**

To help us analyze the relative contributions of the various components of our system, we experimented with different system configurations, as outlined in Figure 10. All of these experiments were conducted with length constraints derived exclusively from the annotated mouse gene. Results are provided in Table 2. By replacing the length constraint with an exact length specification from the target mouse gene, we can determine an upper bound on how well the rest of the system is performing. In fact, this configuration (System I-a in the table) yielded 100% sensitivity and specificity, even without any CLUSTALW alignment post-processing.

However, if we add even a small amount of perturbation from perfection in the mouse length constraints (System I-b), by allowing deviations of ± 3 codons from the exact lengths on all exons, both sensitivity and specificity degrade to 98.4%. This reflects the tremendous ambiguity in allowable gene structures for the genomic sequences. It also seems to suggest that the loss of performance due to imperfect knowledge of mouse exon lengths (as deduced from human orthologs) is relatively small. We reach this conclusion since, with sufficient relaxation of length constraints from the human ortholog predictor, we are able to achieve results that are only slightly worse than the results for System I-b. As for most of our real experiments, addition of N-best selection from CLUSTALW alignment (System II) resulted in a slight degradation in performance.

The other question we were interested in addressing is the degree to which the trigram language model based on the human ortholog improves the quality of the N-best list. If the trigram is omitted from the above configuration, performance degrades significantly, down to only 87% sensitivity and specificity (System III). However, it is interesting that the correct hypotheses are often available within the 100-best list, since, in this case (System IV), CLUSTALW plays a much more critical role to bring the performance to the same level that is achieved by its analog, System II.

System	Configuration	Exon-level Sensitivity (%)	Exon-level Specificity (%)
I-a	MLC(exact) + GGM + TRI	100	100
I-b	MLC($\pm$3) + GGM + TRI	98.4	98.4
II	MLC($\pm$3) + GGM + TRI + ALGN	97.9	98.1
III	MLC($\pm$3) + GGM	87.2	87.2
IV	MLC($\pm$3) + GGM + ALGN	98.1	98.1

Table 2: **Results for various experiments discussed in text.** See Figure 10 for definitions of terms.

4 Relevance of Approaches and Results

We believe that the techniques developed here will be powerful for future tasks of gene discovery for novel species. For example, if a predicted exon structure for a mouse gene homologous to a known human disease genes can be obtained with high accuracy, then this information could be of value in designing knockout or transgenic mice experiments to help in understanding the underlying disease process. Another exciting possibility is to use these techniques to improve genetic modeling in species such as the zebra fish [7], which hold promise for inferring developmental processes through retroviral-induced mutations.

The algorithm described in this paper can also be applied to harvest genomic data for research on alternatively spliced genes, or isoforms. It is an interesting question as to what determines if an exon can be alternatively spliced. A promising approach to addressing this problem is to study alternatively spliced orthologous genes: if an exon exhibits similar behavior in the orthologs, the factors would likely be conserved across the two species [10]. However, ortholog information is generally available only on the gene level. Our technique could contribute by matching orthologs on the isoform level with high accuracy.

We are not aware of any reported research on the topic of gene prediction by analogy with a known orthologous exon-intron structure, although a related but harder problem of gene prediction for both human and mouse based on a joint genomic sequence model has been addressed by several researchers. For example, Meyer and Durbin [9] take the approach of a "probabilistic pair HMM" to jointly model the two sequences. I.e., during the "exon" state, the HMM can use known human-mouse confusion statistics to score the joint hypotheses for amino acids read off from the paired human/mouse genes in two parallel coding sequences. Their best results were 80% sensitivity and 79% specificity on the exon level, realized after a post-processing step to remove implausible hypotheses. Batzoglou et al. [2] align the human mouse orthologs through a novel iterative procedure that relies on exact matches of k-mers, with the value of k decreasing with each iteration. They made use of standard dynamic programming methods to completely score the final alignments that emerged from the iterative process. Statistical methods are used to score the quality of the candidate splice sites, as in our work, but they also make use of human-mouse confusions for the codons, as do Meyer and Durbin. Their length constraints were similar to ours except that they penalized lengths that did not match exactly. Their results for internal exons are nearly perfect, but performance degrades substantially on initial and final exons, where only one of the splice site motif patterns is

available. Here they get only 71% prediction accuracy.

These results can not be directly compared with our results, because the problem is formulated as a joint prediction of two related genes rather than a prediction of one gene based on its similarity to a known ortholog. One would expect a substantially better performance for our system, which is confirmed by our results. It is interesting to note, however, that we have not yet utilized any known confusion statistics between human and mouse orthologous genes/proteins, except as they might be embedded in the CLUSTALW alignment algorithm. We could conceivably obtain improvements by building explicit models for these confusions into our final alignment stage. For this we could make use of another set of speech-based tools we have developed to account for a probabilistic mapping between the idealized phonemes of a word and their phonetic realizations in casual speech [16].

5 Acknowledgments

This research is supported by the NSF under subaward number 1120330-133982, administered through the Carnegie Mellon University.

We are indebted to Professor Bonnie Berger, a member of our laboratory and a faculty member of the MIT Math department, who first suggested that we attempt to find mouse genes based on human orthologs, and who provided us with the same set of human-mouse ortholog pairs that had been used in [2]. Gene Yeo and Dr. Dirk Holste provided the annotated human genes that were used to train the generic mammalian gene model. Mike Rolish helped with some of the data and script preparation. Dr. Lee Hetherington implemented the FST toolkit that was used in this work.

References

[1] L. Bahl, P. Brown, P. de Souza, R. Mercer and D. Nahamoo (1991) "A fast algorithm for deleted interpolation," *Proc. EUROSPEECH-91*, pp. 1209–1212, Genova, Italy.

[2] S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger, and E. S. Lander (2000) "Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction," *Genome Research*, Vol. 10, pp. 950–958.

[3] C. Burge (1998) "Modeling dependencies in pre-mRNA splicing signals," in *Computational Methods in Molecular Biology* (S. Salzberg, D. Searls, and S. Kasif, eds.), Amsterdam, Elsevier Science, pp. 127–163,

[4] C. Burge and S. Karlin (1997) "Prediction of Complete Gene Structures in Human Genomic DNA," *Journal of Molecular Biology*, Vol. 268, No. 1, pp. 78–94.

[5] M. Burset and R. Guigó (1996), "Evaluation of gene structure prediction programs," *Genomics*, Vol. 34, pp. 354–357.

[6] A. Dempster, N. Laird and D. Rubin (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, Vol. 39, pp. 1–38.

[7] N. Gaiano, M. Allende, A. Amsterdam, K. Kawakami, and N. Hopkins (1996) "Highly Efficient Germ-line Transmission of Proviral Insertions in Zebrafish," *Genetics*, Vol. 93, No. 15, pp. 7777–7782.

[8] J. Glass, T. J. Hazen, and I. L. Hetherington (1999) "Real-time telephone-based speech recognition in the JUPITER domain," *Proc. ICASSP*, Phoenix.

[9] I. M. Meyer and R. Durbin (2002) "Comparative Ab Initio Prediction of Gene Structures using Pair HMMs," *Bioinformatics*, Vol. 18, No. 10, pp. 1309–1318.

[10] B. Modrek and CJ. Lee (2003), "Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss," *Nature Genetics*, Vol 34, No. 2, pp. 177-180.

[11] M. Mohri (1997) "Finite-State Transducers in Language and Speech Processing," *Computational Linguistics*, Vol. 23, No. 3, pp. 269–311.

[12] Mouse Genome Sequencing Consortium (2003) "Initial Sequencing and Comparative Analysis of the Mouse Genome," *Nature*, Vol. 420, pp. 520 - 562.

[13] F. Pereira and M. Riley (1997) "Speech recognition by composition of weighted finite automata," in *Finite-State Language Processing* (E. Roche and Y. Schabes, eds.), Cambridge, MA, MIT Press, pp. 431–453.

[14] S. Seneff (1992), "TINA: A Natural Language System for Spoken Language Applications," *Computational Linguistics*, Vol. 18, No. 1, pp. 61 - 86.

[15] S. Seneff, C. Wang, and T.J. Hazen (2003) "Automatic Induction of N-Gram Language Models from a Natural Language Grammar," *Proc. EUROSPEEECH-03*, pp. 641–644, Geneva, Switzerland, September.

[16] (To Appear) S. Seneff and C. Wang, "Statistical Modeling of Phonological Rules through Linguistic Hierarchies," *Speech Communication*.

[17] C. Wang, D. S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue, "MUXING: A Telephone-access Mandarin Conversational System," *Proc. ICSLP '00*, Vol. II, pp. 715–718, Beijing, China, Oct. 2000.

[18] W. Woods (1970) "Transition Network Grammars for Natural Language Analysis," *Commun. of the ACM*, Vol. 13, pp. 591–606.

[19] http://www.ebi.ac.uk/clustalw/.