Gene Structure Prediction Using an Orthologous Gene of Known Exon-Intron Structure [1]

[*]Stephanie Seneff, [*]Chao Wang, and [+]Christopher B. Burge

Affiliation:

[*]Spoken Language Systems Group

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

[+]Department of Biology

Massachusetts Institute of Technology

[1]This article is published in Applied Bioinformatics 2004:3(2-3):81-90,

Send Correspondence to:

Stephanie Seneff

MIT Computer Science and Artificial Intelligence Laboratory

32 Vassar Street, 32-G438

Cambridge, Massachusetts 02139 USA

fax: +1 617 258 8642

phone: +1 617 253 0451

email: seneff@csail.mit.edu

or

Chao Wang

MIT Computer Science and Artificial Intelligence Laboratory

32 Vassar Street, 32-G362

Cambridge, Massachusetts 02139 USA

fax: +1 617 258 8642

phone: +1 617 253 7772

email: wangc@csail.mit.edu

or

Christopher B. Burge

MIT Department of Biology

77 Massachusetts Avenue, 68-223

Cambridge, Massachusetts 02139 USA

fax: +1 617 452 2936

phone: +1 617 258 5997

email: cburge@mit.edu

Number of pages: 30

Number of tables: 2

Number of figures: 10

Keywords: Gene prediction, comparative genomics, language models, exon length

Abstract

Given the availability of complete genome sequences from related organisms, sequence conservation can provide important clues for predicting gene structure. In particular, one should be able to leverage information about known genes in one species to help determine the structures of related 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 puffer fish, using the extensive knowledge that has been accumulated about human genes. Here, we report a novel approach to predicting the exon-intron structures of 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 annotated human genome.

1 Introduction

The biomedical research 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 leverage knowledge about one species to help in annotating the genome sequences obtained for other species. At this time, the knowledge available for the human genome is much more precise and extensive than that for other vertebrates. However, with the recent determination of the complete mouse genome sequence (Consortium 2003), it becomes of paramount importance to accelerate the pace at which new genomic sequences can be accurately decoded. It is well 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 pairs of human-mouse orthologs. As discussed in (Batzoglou et al. 2000; Consortium 2003), 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 be impractical because of an unwieldy search space.

1.1 Goals

Our goal in this work was 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 of human-mouse orthologs and selected the highest scoring mouse protein (among 102 can-

didates) 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 (Batzoglou et al. 2000), 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 (Seneff 1992) , then inducing a language model for the recognizer from an automatic analysis of the parse trees (Seneff et al. 2003). Our speech recognition framework (Glass et al. 1999) makes use of a finite state transducer (FST) formalism (Pereira and Riley 1997; Mohri 1997) 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) (Woods 1970), 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 statistical language model for the recognizer directly from the natural language (NL) 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 (Seneff et al. 2003) induces a traditional class n-gram language model, whereas the more complex alternative (Wang et al. 2000) includes component categories that are represented by a recursive transition network (RTN) (Woods 1970), 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 *non-terminal* 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 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[2]. 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 sorted using standard sequence alignment tools, such as CLUSTALW (Thompson et al. 1994). 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 (Pereira and Riley 1997; Mohri 1997).

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

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:

$$\mathrm{p}(x, s) = \mathrm{p}(s_0)\mathrm{p}(x_0|s_0) \prod_{i=1}^{M-1} \mathrm{p}(s_i|s_{i-1})\mathrm{p}(x_i|s_i) \tag{1}$$

in which x_i is the observed nucleotide sequence for state s_i, and M 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 this joint 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 (Seneff 1992) 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 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 $3'$ 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, $\mathrm{p}(x_i|s_i = 3'\ splice\ site)$ is computed as a product of the RTN weights[3]. 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

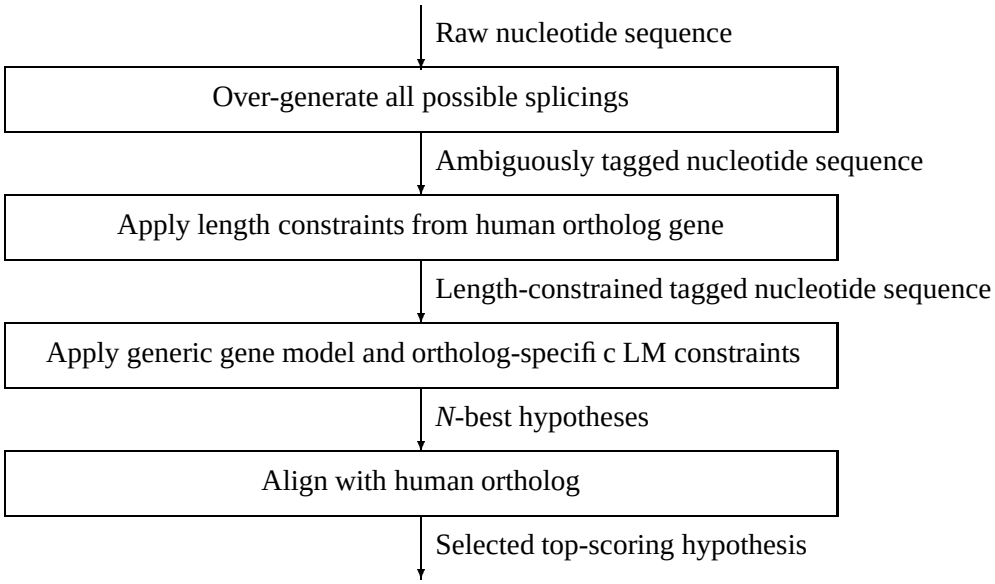

Figure 1: **Block diagram of procedure used to extract mouse gene structure by analogy with known human ortholog.**

[3]In practice, the weights on the RTN are negative log probabilities, so that a sum is used in computing the total probability.

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[4]. 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 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 (Thompson et al. 1994). 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 positions wherever they were possible according to strict rules for specific two- or three-nucleotide sequences at their boundaries, as illustrated in Figure 2. This results in a finite state transducer mapping raw DNA sequences to alternatively tagged sequences.

[4]Putative orthologs can be acquired using bidirectional BLAST search.

<**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.

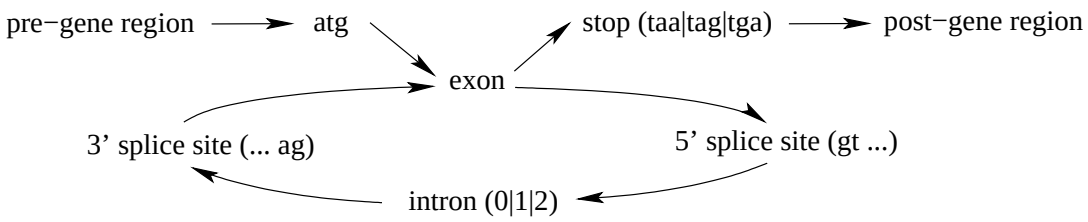

Figure 3: **Basic structure of the generic gene model.** Internal introns remember the reading frame to assure correct coding of the nucleotides into amino acids.

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 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, as illustrated in Figure 5. This model captures positional bigram statistics, which is equivalent to an inhomogeneous first-order Markov model (Burge

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."**

1998). 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 end of the exon (See Figure 2 in (Burge and Karlin 1997)). When the exon ends in phase 0 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 phase 1 and phase 2 exons. An example of the parse tree for an exon which ends in phase 2 and in a "not-g" configuration, is illustrated in Figure 6. Figure 7 shows that the distributions of the four nucleotides at the +5 position of the $5'$ splice site motif model are distinctly different for the "g"/"not-g" subsets, as reported previously.

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 parameters of the model. The training genes were truncated at 1000 nucleotides preceding the first coding exon and 1000 nucleotides subsequent to the end of the last

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."**

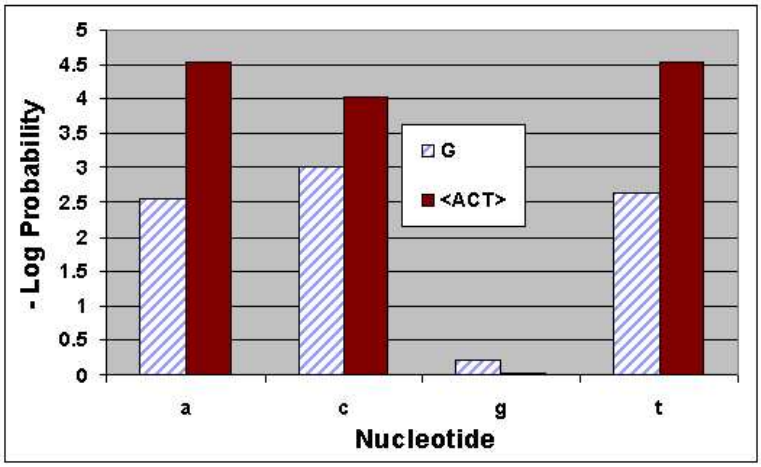

Figure 7: **Log probabilities obtained in the generic gene model for the four nucleotides in position $+5$ (labeled x) in the $5'$ splice site motif: "n n g|h </exon> g t n n x n", conditional on "g" or "h" (<act>) at position -1, the last base of the exon (labeled "g|h").**

coding exon. Some characteristics of 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 (Consortium 2003) and (Batzoglou et al. 2000), it appears that the lengths of corresponding exons of human and mouse orthologs are strongly conserved. Batzoglou *et al.* (Batzoglou et al. 2000) 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, often 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 L 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 exons easily in our model, by providing a transition that by-passes the intron state. The inverse problem of splitting an exon into two is more difficult, due to the many possible sites at which splitting could occur. However, empirical studies have shown that the problem of "exon-splitting" is likely to be very rare when comparing mammalian genes. For example, in an analysis of 1,560 human-mouse orthologs and 360 mouse-rat orthologs, evidence was found for only about a half dozen intron loss events, and no intron gains (Roy et al.

2003). From a practical consideration, we could account for all variations of exon structure change with a more complex model, but at the expense of significantly increased ambiguity. We thus 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 L, the maximum number of codons we allow an exon to insert or delete.

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

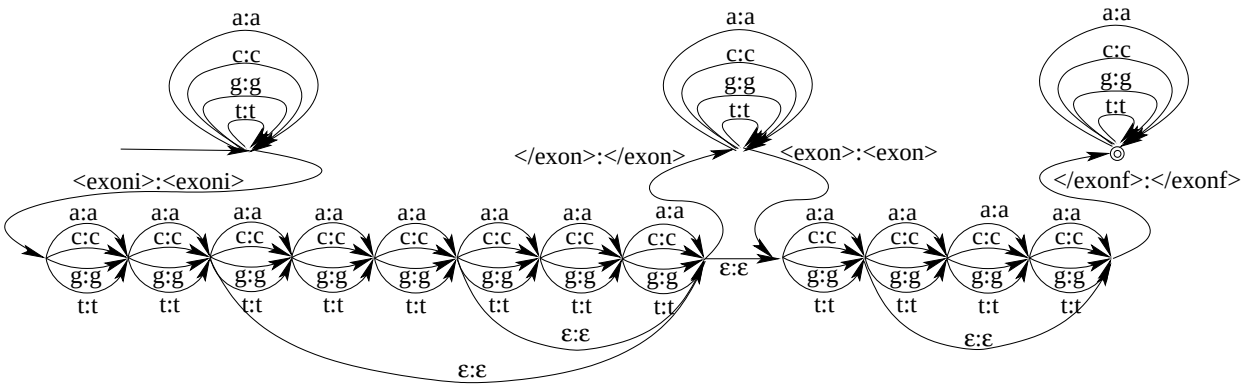

Figure 8: **An example length constraint FST for a hypothetical sequence "... <exoni> a t g t a </exon> g t ... a g <exon> a </exonf> ...".** In this example, we allow up to one codon insertion or deletion in each exon, as well as a merge of exons. In addition to the original exon length pattern "5 1", 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 (Bahl et al. 1991) was used for smoothing, with

probabilities estimated using a variation of the expectation maximization (EM) algorithm (Dempster et al. 1997). 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 (Thompson et al. 1994). 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 was fixed to be 100 in our experiments, although one could optimize this parameter if an independent set of development data were available.

3 Results and Discussion

We evaluated our approach using the same set of human-mouse ortholog pairs that had been used in (Batzoglou et al. 2000). 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 were 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). We could modify our algorithm to accommodate the "gc..ag" pattern. However, in our experience with related gene finding algorithms such as GENSCAN (Burge 1998) and GENOMESCAN (Yeh et al. 2001), allowing "gc" dinucleotides at the $5'$ splice site dramatically increases the search space without a significant improvement in accuracy. We thus evaluated our algorithm only on genes that have the "gt..ag" terminal dinucleotide pattern, leaving 102 ortholog pairs in our final test set. The human genes from the human-mouse orthologs in our test set 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 the 400 human genes selected for training the generic mammalian gene model, compared with distributions of the 102 human genes from the human-mouse ortholog test pairs.** There is no overlap in the two sets.

The criterion we used for evaluation is based on exactly matched coding exons. In particular, we use the exon-level *sensitivity* and *specificity* measures (Burset and Guigó 1996), 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 test sequences; specificity is defined as the ratio of the number of correctly identified exons over the total number of predicted exons.

3.1 Results

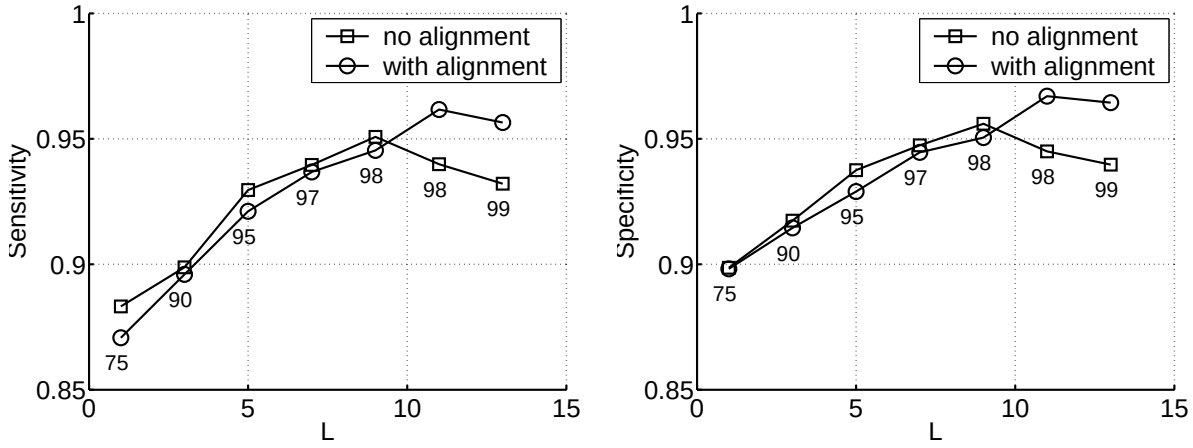

Figure 9: **Sensitivity and specificity on correctly identifying mouse exons as a function of L, the maximum number of codon insertion/deletions allowed in the length FST model.** L varies from 1 to 13 in the plots. The labels next to the data points indicate the total number of genes that our system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

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

The sensitivity and specificity measures in the plots were calculated on the subset of genes that our system can produce an answer for, for different values of L. As shown in the figure, there is clearly a trade-off in choosing L. Since we have no chance of correctly identifying those mouse exons that varied by more than L codons, a small L 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 L, we can generally produce outputs for more genes. However, with a large L, the performance could degrade due to increased ambiguity, as indicated by the downward trend in the figure beyond $L = 9$. The optimal performance was 96.2% sensitivity and 96.7% specificity for coding exons, which was achieved with L 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 L reaches 11. This seems to suggest that, when the exon length constraints are relatively strict, the trigram model is adequate for incorporating human protein 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 L 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.)

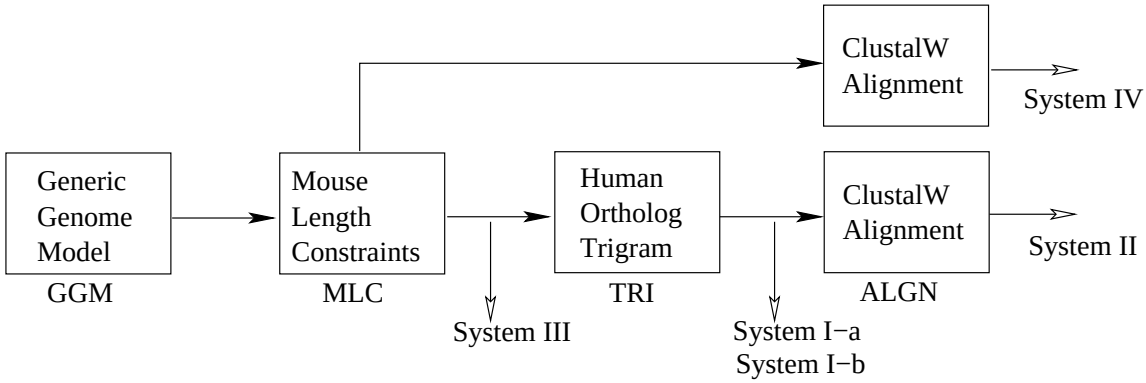

Figure 10: **Schematic of experiments on different system configurations for gene prediction of the mouse gene based on the human ortholog.**

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 are reduced 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 imper-

fect 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 was 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

Even though our algorithm was developed within the context of predicting mouse gene structure using information from the human orthologous gene, it can be easily extended to compare multiple species. We can simply run the procedure on all pairs of species for which the structure is known in one but not in the other and compare the results. This could provide additional confirmation in the typical cases when the predictions agree, or indicate uncertain exons or splice sites (or true changes in gene structure) in cases where the predictions disagree.

We expect that the techniques developed here will be useful for future tasks of gene annotation for newly sequenced genomes. For example, if a predicted exon structure for a mouse gene homologous to a known human disease gene can be obtained with high accuracy, then this information could be of value in designing knockout or transgenic mouse 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 (Gaiano et al. 1996), which hold promise for genetic dissection of 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 whether 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 (Modrek and Lee 2003). 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 groups. For example, Meyer and Durbin (Meyer and Durbin 2002) took the approach of a "probabilistic pair HMM" to jointly model the two sequences. In the "exon" state, the HMM used known human-mouse confusion statistics to score the joint hypotheses for amino acids deduced 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.* (Batzoglou et al. 2000) aligned 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 were used to score the quality of the candidate splice sites, as in our work, but they also made use of human-mouse confusion statistics for the codons, as did 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 were nearly perfect, but performance degraded substantially on initial and final exons, where only one of the splice site motif patterns is available. Here they obtained 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 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 (Seneff and Wang).

5 Acknowledgments

This research was 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 the MIT Computer Science and Artificial Intelligence Laboratory, for helpful suggestions and for providing us with the set of human-mouse ortholog pairs that had been used in (Batzoglou et al. 2000). Gene Yeo and Dr. Dirk Holste provided the annotated human genes that were used to train the generic mammalian gene model. Michael Rolish helped with some of the data and script preparation. Dr. Lee Hetherington implemented the FST toolkit that was used in this work.

References

Bahl, L., P. Brown, P. de Souza, R. Mercer, and D. Nahamoo (1991). A fast algorithm for deleted interpolation. In *Proceedings of the 2nd European Conference on Speech Communication and Technology*, Genova, Italy, pp. 1209–1212.

Batzoglou, S., 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 10*, 950–958.

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

Burge, C. and S. Karlin (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology 268*(1), 78–94.

Burset, M. and R. Guigó (1996). Evaluation of gene structure prediction programs. *Genomics 34*, 354–357.

Consortium, M. G. S. (2003). Initial sequencing and comparative analysis of the mouse genome. *Nature 420*, 520–562.

Dempster, A., N. Laird, and D. Rubin (1997). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1–38.

Gaiano, N., M. Allende, A. Amsterdam, K. Kawakami, and N. Hopkins (1996). Highly efficient germline transmission of proviral insertions in zebrafish. *Genetics 93*(15), 7777–7782.

Glass, J., T. J. Hazen, and I. L. Hetherington (1999). Real-time telephone-based speech recognition in

the JUPITER domain. In *Proceedings of the 1999 International Conference on Acoustics, Speech and Signal Processing*, Phoenix.

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

Modrek, B. and C. 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 34*(2), 177–180.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics 23*(3), 269–311.

Pereira, F. and M. Riley (1997). Speech recognition by composition of weighted finite automata. In E. Roche and Y. Schabes (Eds.), *Finite-State Language Processing*, pp. 431–453. Cambridge, MA: MIT Presss.

Roy, S. W., A. Fedorov, and W. Gilbert (2003). Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proceedings of the National Academy of Sciences 100*, 7158–62.

Seneff, S. (1992). TINA: A natural language system for spoken language applications. *Computational Linguistics 18*(1), 61–86.

Seneff, S. and C. Wang. Statistical modeling of phonological rules through linguistic hierarchies. *Speech Communication*. To appear.

Seneff, S., C. Wang, and T. Hazen (2003). Automatic induction of n-gram language models from a natural language grammar. In *Proceedings of the 8th European Conference on Speech Communication*

and Technology, Geneva, Switzerland, pp. 641–644.

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research 22*, 4673–4680.

Wang, C., D. S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. Zue (2000). MUXING: A telephone-access mandarin conversational system. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Volume II, Beijing, China, pp. 715–718.

Woods, W. (1970). Transition network grammars for natural language analysis. *Communication of the ACM 13*, 591–606.

Yeh, R.-F., L. P. Lim, and C. B. Burge (2001). Computational inference of homologous gene structures in the human genome. *Genome Research 11*, 803–816.