Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction

Axel Bernal  
*University of Pennsylvania*

Koby Crammer  
*University of Pennsylvania*

Artemis Hatzigeorgiou  
*University of Pennsylvania*

Fernando C.N. Pereira  
*University of Pennsylvania*, pereira@cis.upenn.edu

Follow this and additional works at: [http://repository.upenn.edu/cis_papers](http://repository.upenn.edu/cis_papers)

Recommended Citation


© 2007 Bernal et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Reprinted from *PLOS Computational Biology*, Volume 3 Issue 3, e54.  
Publisher URL: [http://dx.doi.org/10.1371/journal.pcbi.0030054](http://dx.doi.org/10.1371/journal.pcbi.0030054)

This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/cis_papers/342](http://repository.upenn.edu/cis_papers/342)
For more information, please contact libraryrepository@pobox.upenn.edu.
Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction

Abstract
Most ab initio gene predictors use a probabilistic sequence model, typically a hidden Markov model, to combine separately trained models of genomic signals and content. By combining separate models of relevant genomic features, such gene predictors can exploit small training sets and incomplete annotations, and can be trained fairly efficiently. However, that type of piecewise training does not optimize prediction accuracy and has difficulty in accounting for statistical dependencies among different parts of the gene model. With genomic information being created at an ever-increasing rate, it is worth investigating alternative approaches in which many different types of genomic evidence, with complex statistical dependencies, can be integrated by discriminative learning to maximize annotation accuracy. Among discriminative learning methods, large-margin classifiers have become prominent because of the success of support vector machines (SVM) in many classification tasks. We describe CRAIG, a new program for ab initio gene prediction based on a conditional random field model with semi-Markov structure that is trained with an online large-margin algorithm related to multiclass SVMs. Our experiments on benchmark vertebrate datasets and on regions from the ENCODE project show significant improvements in prediction accuracy over published gene predictors that use intrinsic features only, particularly at the gene level and on genes with long introns.

Keywords
craig, crf-based ab initio genefinder, crf, conditional random fields, hmm, hidden markov model, mira, margin infused relaxed algorithm, pwm, position weight matrices, svm, support vector machines, tis, translation initiation site, wam, weight array model, wwam, windowed weight array model

Comments
© 2007 Bernal et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. Reprinted from PLOS Computational Biology, Volume 3 Issue 3, e54.
Publisher URL: http://dx.doi.org/10.1371/journal.pcbi.0030054

This journal article is available at ScholarlyCommons: http://repository.upenn.edu/cis_papers/342
Global Discriminative Learning for Higher-Accuracy Computational Gene Prediction

Axel Bernal¹*, Koby Crammer¹, Artemis Hatzigeorgiou², Fernando Pereira¹

1 Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America,
2 Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America

Most ab initio gene predictors use a probabilistic sequence model, typically a hidden Markov model, to combine separately trained models of genomic signals and content. By combining separate models of relevant genomic features, such gene predictors can exploit small training sets and incomplete annotations, and can be trained fairly efficiently. However, that type of piecewise training does not optimize prediction accuracy and has difficulty in accounting for statistical dependencies among different parts of the gene model. With genomic information being created at an ever-increasing rate, it is worth investigating alternative approaches in which many different types of genomic evidence, with complex statistical dependencies, can be integrated by discriminative learning to maximize annotation accuracy. Among discriminative learning methods, large-margin classifiers have become prominent because of the success of support vector machines (SVM) in many classification tasks. We describe CRAIG, a new program for ab initio gene prediction based on a conditional random field model with semi-Markov structure that is trained with an online large-margin algorithm related to multiclass SVMs. Our experiments on benchmark vertebrate datasets and on regions from the ENCODE project show significant improvements in prediction accuracy over published gene predictors that use intrinsic features only, particularly at the gene level and on genes with long introns.

Introduction

Most ab initio gene predictors use a probabilistic sequence model, typically a hidden Markov model, to combine separately trained models of genomic signals and content. By combining separate models of relevant genomic features, such gene predictors can exploit small training sets and incomplete annotations, and can be trained fairly efficiently. However, that type of piecewise training does not optimize prediction accuracy and has difficulty in accounting for statistical dependencies among different parts of the gene model. With genomic information being created at an ever-increasing rate, it is worth investigating alternative approaches in which many different types of genomic evidence, with complex statistical dependencies, can be integrated by discriminative learning to maximize annotation accuracy. Among discriminative learning methods, large-margin classifiers have become prominent because of the success of support vector machines (SVM) in many classification tasks. We describe CRAIG, a new program for ab initio gene prediction based on a conditional random field model with semi-Markov structure that is trained with an online large-margin algorithm related to multiclass SVMs. Our experiments on benchmark vertebrate datasets and on regions from the ENCODE project show significant improvements in prediction accuracy over published gene predictors that use intrinsic features only, particularly at the gene level and on genes with long introns.
We describe a new approach to statistical learning for sequence data that is broadly applicable to computational biology problems and that has experimentally demonstrated advantages over current hidden Markov model (HMM)-based methods for sequence analysis. The methods we describe in this paper, implemented in the CRAIG program, allow researchers to modularly specify and train sequence analysis models that combine a wide range of weakly informative features into globally optimal predictions. Our results for the gene prediction problem show significant improvements over existing ab initio gene predictors on a variety of tests, including the specially challenging ENCODE regions. Such improvements, particularly on initial and single exons, could benefit researchers who are seeking more accurate means of recognizing such important features as signal peptides and regulatory regions. More generally, we believe that our method, by combining the structure-describing capabilities of HMMs with the accuracy of margin-based classification methods, provides a general tool for statistical learning in biological sequences that will replace HMMs in any sequence modeling task for which there is annotated training data.

Results

Datasets

All the experiments reported in this paper use a gene model trained on a nonredundant set of 3,038 single-gene sequences. We built this set by combining the Augustus training set [2], the GenScan training set, and 1,500 high-confidence CDSs from EnsMart Plus [22], which are part of the Genezilla training set (http://www.tigr.org/software/traindata.shtml). We then appended simulated intergenic material to both ends of each training sequence to make up for the lack of realistic intergenic regions in the training material, as described in more detail in Methods.

We compared CRAIG with GenScan, TwinScan 2.03 (without homology features, also known as GenScan++), Genezilla (formerly known as TigrScan), and Augustus on several benchmark test sets. We also ran predictions with HMMGene, the only other publicly available genefinder to use a discriminative structure training method; we present some prediction results with it in Methods. All programs we compare with are based on similar GHMM models with similar sequence features. Augustus uses two types of length distributions for introns: short intron lengths are modeled with an explicit distribution, but other introns use the default geometric distribution. This difference made Augustus run many times slower than the other programs in all our experiments.

We evaluated the programs on the following benchmark test sets.

**BGHM953.** This test set combines most of the available single-gene test sets in one single set. It includes the GeneParser I (27 genes) and II (34 genes) datasets [23], 570 vertebrate sequences from Burset and Guigo [24], 178 human sequences from Guigo et al. [25], and 195 human, rat, and mouse sequences from Rognic et al. [26]. Repeated entries were removed. We combined different sets to obtain more reliable evaluation statistics by smoothing out possible overfitting to particular sequence types.

**TIGR251.** This test set consists of 251 single-gene sequences, which are part of the TIGR human test dataset (http://www.tigr.org/software/traindata.shtml), and it is composed mostly of long-intron genes.

**ENCODE294.** This test set consists of 31 test regions from the ENCODE project [27,28], for a total of 21M bases, training to some extent. HMMGene uses a nongeneralized HMM model for gene structure, which does not include features associated with biological signals, but it is trained with the discriminative conditional maximum likelihood criterion [20]. However, conditional maximum likelihood is more difficult to optimize than our training criterion because it is required to respect conditional independence and normalization for the underlying HMM. GRAPE takes a hybrid approach for learning. It first trains parameters of a generalized HMM (GHMM) to maximize generative likelihood, and then it selects a small set of parameters that are trained to maximize the percentage of correctly predicted nucleotides, exons, and whole genes used as surrogates of the conditional likelihood. This approach is commonly used when training data is limited, and it usually provides superior results only in those cases [21]. However, the GRAPE learning method does not globally optimize the training criterion.
containing 294 carefully annotated alternatively spliced genes and 667 transcripts, after eliminating repeated entries and partial entries with coordinates outside the region’s bounds. This is the only test set that was masked using RepeatMasker (http://www.repeatmasker.org/) before performing gene prediction. Table 1 gives summary statistics for the training set and the three test sets.

Predictions on all tests and for all programs—including CRAIG—allow partial genes, multiple genes per region, and genes on both strands. Alternative splicing and genes embedded within genes were not evaluated in this work. Any other program parameters were left at their default values. For each program, we used the human/vertebrate gene models provided with the software distributions. In all tests, sequences with noncanonical splice sites were filtered out. Accuracy numbers were computed with the eval package [29], a standard and reliable way to compare different gene predictions.

Prediction in Single-Gene Sequences

Table 2 shows prediction results for all programs on BGHM957. CRAIG achieved better sensitivity and specificity than the other programs at all levels, except for somewhat lower base sensitivity but much higher base specificity than GenScan. The relative F-score improvement for initial and single exons over Genezilla, the second-best program overall for this set, was 14.6% and 5.8%, respectively. Single-exon genes were more difficult to predict for all programs, with specificity barely exceeding 50% for the best program, but CRAIG’s relative improvement in sensitivity was nearly 25% over runner-up Genezilla. Terminal exon predictions were also improved over the nearest competitors, but less markedly so. The improved gene-level accuracy follows from these gains at the exon level. GenScan++ and Augustus predicted internal exons with similar accuracy and their F-scores were only slightly worse than CRAIG, but the overall gene-level accuracy for GenScan++ looks much worse because it missed many terminal and single exons. GenScan also did well in this set, but overall performance was somewhat worse than the other programs.

Most of the genes in this set have short introns and the intergenic regions are truncated, so prediction was relatively easy and all programs did relatively well. The next section compares performance on datasets with long-intron genes and very long intergenic regions.

Prediction in Long DNA Sequences

As previously noted, TIGR251 has many genes with very long introns, so it is expected to be harder to predict accurately. This was confirmed by the results in Table 3. Performance was worse for all programs and levels when compared with the first set. However, CRAIG consistently outperformed the other programs with an even wider performance gap than in the first experiment. Here, base and internal-exon accuracies were also substantially improved. CRAIG’s relative F-score improvement for bases and internal exons over Genezilla, the second-best program in both categories, was 5.4% and 7.1%, respectively, compared with approximately 1% for BGHM953. Other types of exons also improved, as in the first experiment.
Because of these better base and exon-level predictions, the relative F-score improvement over runner-up Genezilla at the gene level was about 57%.

Our final set of experiments was on ENCODE294. The results are shown in Table 4. As previously mentioned, all sequences in this set were masked for low-complexity regions and repeated elements. Unlike previous sets, in which masking did not affect results significantly, prediction on unmasked sequences in this set was worse for all programs (unpublished data). In particular, exon and base specificity decreased an average of 8%.

We added a transcript-level prediction category to Table 4 to better evaluate predictions on alternatively spliced genes. We closely followed the evaluation guidelines and definitions by Guigo and Reese [28]. There, transcript and gene-level predictions that are consistent with annotated incomplete transcripts are counted correct, even in cases where the predictions include additional exons. We relaxed this policy to also mark as correct those predictions that contained incomplete transcripts whose first (last) exon did not begin (end) with an acceptor (donor). The reason for this change is that no program can exactly predict both ends of such transcripts. We developed our own programs to evaluate single-exon, transcript, and gene-level predictions for incomplete transcripts. Evaluations for other categories and for complete transcripts were handled directly with eval.

To ensure consistency in the evaluation, we obtained all of the programs except for Genezilla from their authors and we ran them on the test set in our lab. Genezilla predictions for this set were obtained directly from the supplementary material provided by [28] so that we could measure the potential differences between our evaluation method and that reported in [28], particularly at the transcript and gene level, for which we expected different results.

Overall, our results for all programs agree with those of Guigo and Reese [28]. Genezilla’s base and exon-level results using our evaluation program closely matched the published values. Transcript and gene-level results computed by our method were 1% better than the published numbers, which roughly match the percentage of incomplete annotated transcripts with no splice signals on either end. Computed predictions for GenScan and Augustus were also somewhat different, but not substantially so, from those reported by Guigo and Reese [28], presumably because of differences in program version and operating parameters.

Improvements in this set were similar to those obtained in our second experiment. The relative F-score improvements for individual bases and internal exons were 6% and 15.4% over GenScan++ and Augustus, the runner-ups in each respective category. Improvement in prediction accuracy on single, initial, and terminal exons is similar to that for the other test sets. Transcript and gene-level accuracies were, respectively, 30% and 30.6% better than Augustus, the second-best program overall. This means that our better accuracy results obtained in the first two single-gene sequence sets scale well to chromosomal regions with multiple, alternatively spliced genes.

### Table 1. Dataset Statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Genes</th>
<th>Single Exon Genes</th>
<th>Coding (Percent)</th>
<th>Average Number of Exons</th>
<th>Average Coding Length (bp)</th>
<th>Average Transcript Length (bp)</th>
<th>Average Intron Length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>17,875</td>
<td>3,038</td>
<td>721</td>
<td>6.8</td>
<td>6.27</td>
<td>1,213</td>
<td>17,683</td>
</tr>
<tr>
<td>BGHM953</td>
<td>4,544</td>
<td>953</td>
<td>84</td>
<td>24.9</td>
<td>4.77</td>
<td>860</td>
<td>3,453</td>
</tr>
<tr>
<td>TIGR251</td>
<td>1,496</td>
<td>251</td>
<td>43</td>
<td>5.8</td>
<td>5.96</td>
<td>1,044</td>
<td>17,857</td>
</tr>
<tr>
<td>ENCODE294</td>
<td>2,842</td>
<td>294</td>
<td>65</td>
<td>4.6</td>
<td>8.3</td>
<td>1,305</td>
<td>28,179</td>
</tr>
</tbody>
</table>

BGHM953 is a standard benchmark set of single-gene sequences with high protein-coding content (~23%), short average transcript length (~3,000 bp), and short average intron length (~700 bp). TIGR251 is also a single-gene sequence set but transcripts and introns are longer; genes in this set resemble our training set the most. BGHM953, and to some extent TIGR251, are not representative of the whole human genome, because of their high relative frequencies of single-exon genes and coding loci. In contrast, ENCODE294 is a highly curated dataset containing multiple-gene sequences with long intergenic regions and alternatively spliced genes with long introns. These characteristics more closely resemble real chromosomal DNA. Results on ENCODE294 may thus be better estimates of performance on biologically interesting genomic sequences.

doi:10.1371/journal.pcbi.0030054.t001

### Table 2. Accuracy Results for BGHM953

<table>
<thead>
<tr>
<th>Level</th>
<th>GenScan Sn</th>
<th>GenScan Sp</th>
<th>Genezilla Sn</th>
<th>Genezilla Sp</th>
<th>GenScan++ Sn</th>
<th>GenScan++ Sp</th>
<th>Augustus Sn</th>
<th>Augustus Sp</th>
<th>CRAIG Sn</th>
<th>CRAIG Sp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
<td>95.8</td>
<td>88.3</td>
<td>93.2</td>
<td>91.8</td>
<td>90.0</td>
<td>92.7</td>
<td>92.0</td>
<td>91.4</td>
<td>93.9</td>
<td>93.1</td>
</tr>
<tr>
<td>All</td>
<td>80.5</td>
<td>77.8</td>
<td>78.9</td>
<td>80.7</td>
<td>68.9</td>
<td>75.9</td>
<td>79.3</td>
<td>81.7</td>
<td>82.0</td>
<td>86.3</td>
</tr>
<tr>
<td>Initial</td>
<td>62.5</td>
<td>62.6</td>
<td>65.3</td>
<td>66.8</td>
<td>62.5</td>
<td>70.5</td>
<td>60.3</td>
<td>66.6</td>
<td>73.4</td>
<td>78.2</td>
</tr>
<tr>
<td>Internal</td>
<td>89.1</td>
<td>83.9</td>
<td>84.6</td>
<td>87.1</td>
<td>86.5</td>
<td>88.1</td>
<td>86.1</td>
<td>88.5</td>
<td>84.8</td>
<td>91.3</td>
</tr>
<tr>
<td>Terminal</td>
<td>73.9</td>
<td>78.6</td>
<td>76.1</td>
<td>78.1</td>
<td>23.1</td>
<td>52.9</td>
<td>78.1</td>
<td>80.7</td>
<td>81.6</td>
<td>86.0</td>
</tr>
<tr>
<td>Single</td>
<td>59.5</td>
<td>47.9</td>
<td>67.9</td>
<td>52.8</td>
<td>38.1</td>
<td>39.3</td>
<td>66.7</td>
<td>49.6</td>
<td>83.3</td>
<td>49.3</td>
</tr>
<tr>
<td>Gene</td>
<td>40.7</td>
<td>37.7</td>
<td>47.2</td>
<td>47.0</td>
<td>13.9</td>
<td>13.7</td>
<td>45.1</td>
<td>44.1</td>
<td>57.1</td>
<td>56.1</td>
</tr>
</tbody>
</table>

Sensitivity (Sn) and specificity (Sp) for each level and exon type.
doi:10.1371/journal.pcbi.0030054.t002
Significance Testing

In all tests and at all levels, CRAIG achieved greater improvements in specificity than in sensitivity. We investigated whether the improvements in exon sensitivity achieved by CRAIG could be explained by chance. Any exon belonging to a particular test set is associated with two dependent Bernoulli random variables for whether it was correctly predicted by CRAIG and by another program. We computed $p$-values with McNemar’s test for dependent, paired samples from CRAIG and each of the other programs over the three test sets, as shown in Table 5. The null hypothesis was that CRAIG’s advantage in exon predictions is due to chance. The $p$-values were $0.05$ for all entries, except for the TIGR251 experiments against Genezilla and the ENCODE294 experiments against Genezilla and GenScan; in general, these two genefinders proved to be very sensitive at the cost of predicting many more false positives. $p$-Values for the combined test sets were all below $0.001$, showing that CRAIG’s advantage was extremely unlikely to be a chance event.

We also trained and tested an additional variant of CRAIG, in which we did not distinguish between short and long introns; this configuration corresponds closely to the state model representation used in most previous works. Following Stanke and Waack [2], we used the relative mean improvement:

$$r = \frac{\Delta S_{\text{ exon}} + \Delta S_{\text{ intron}} + \Delta S_{\text{ gene}} + \Delta S_{\text{ transcript}}}{4}$$

as the measure of differences in prediction accuracy between the CRAIG variant and CRAIG itself. The term $\Delta S_{\text{ exon}}$ denotes the mean increase in exon sensitivity and is defined as

$$\Delta S_{\text{ exon}} = \frac{\sum_{t \in T} n_t \times \Delta S_{\text{ exon}}^t}{\sum_{t \in T} n_t}$$

where $n_t$ is the number of annotated genes in dataset $t$, $T = \{ \text{BGHM953, TIGR251, ENCODE294} \}$, and $\Delta S_{\text{ exon}}^t$ is the difference in exon sensitivity between CRAIG and the CRAIG variant on dataset $t$. The other terms are defined similarly. The improvement obtained by CRAIG with respect to the variant was $r = 3.6$. This result was as expected: there was an improvement in accuracy from including the extra intron state in the gene model, but even the simpler variant was more than competitive with the best current genefinders.

Discussion

It is well-known that more gene prediction errors occur on regions with low GC content, which have higher intron and intergenic region density [15]. This behavior can also be observed on our combined results, as shown in Figure 2A. It also can be noticed that CRAIG had the best F-score for all intron lengths. Except for CRAIG and HMMGene, the F-scores for all other predictors were very close for all lengths. CRAIG’s advantage over its nearest competitors became more
apparent as introns increased in length. However, all gene-finders experience a significant drop in accuracy, at least 25% between 1,000 bp and 16,000 bp. For introns shorter than 1,000 bp, Augustus performs almost as well as CRAIG, in part because of its more complex, time-consuming model for short intron lengths.

Intron analysis of individual test sets, as shown in Figure 2B–2D, reveals that, except for ENCODE294, CRAIG consistently achieved an intron F-score above 75%, even for lengths more than 30,000 bp; in contrast, the F-scores of all other programs fell to lower than 65%, even for introns as short as 8,000 bp. The results show that CRAIG predicts genes with long introns much better than the other programs. This hypothesis was also confirmed with experiments on an edited version of ENCODE294 in which the original 31 regions were split into 271 contig sequences and all of the intergenic material was deleted except for 2,000 bp on both sides of each gene. This edited version was further subdivided into subsets with—ALT_ENCODE155—and without—NOALT_ENCODE139—alternative splicing. Figure 3 shows intron prediction results for this arrangement. It can be observed that intron prediction on NOALT_ENCODE139, a subset of 139 genes, has the same characteristics as either TIGR251 or BGHM953, that is, a rather flat F-score curve as intron length increases. The same cannot be said about complementary subset ALT_ENCODE155, whose significant drop in accuracy for long introns can be explained by the presence of alternative splicing.

We claimed in the Introduction that a key aspect of our model and training method is the ability to combine various genomic features and to find a global tradeoff among their

---

**Table 5. Significance Testing**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GenScan</th>
<th>Genezilla</th>
<th>GenScan++</th>
<th>Augustus</th>
</tr>
</thead>
<tbody>
<tr>
<td>BGHM953</td>
<td>0.03</td>
<td>1.66 × 10^-6</td>
<td>5.2 × 10^-6</td>
<td>1.3 × 10^-5</td>
</tr>
<tr>
<td>TIGR251</td>
<td>2.2 × 10^-7</td>
<td>0.22</td>
<td>1.4 × 10^-23</td>
<td>1.4 × 10^-25</td>
</tr>
<tr>
<td>ENCODE294</td>
<td>0.17</td>
<td>≥0.5</td>
<td>2.33 × 3^-22</td>
<td>4.7 × 10^-16</td>
</tr>
<tr>
<td>All</td>
<td>5.7 × 10^-6</td>
<td>6.2 × 10^-4</td>
<td>8.2 × 10^-105</td>
<td>3.45 × 10^-35</td>
</tr>
</tbody>
</table>

McNemar test results (p-value upper bounds) of paired exon sensitivity predictions between CRAIG and each of the other programs.
doi:10.1371/journal.pcbi.0030054.t005

---

**Figure 2. F-Score as a Function of Intron Length**

Results for all sets combined (A) and for individual test sets shown in subfigures (B–D). The boxed number appearing directly above each marker represents the total number of introns associated with the marker’s length. For example, there were 1,475 introns with lengths between 1,000 and 2,000 base pairs for all sets combined (A).
doi:10.1371/journal.pcbi.0030054.g002
contributions so that accuracy is maximized. Being able to identify introns longer than 30,000 bp with prediction accuracy comparable to that achieved on smaller introns is evidence that our program does a better job of combining features to recognize structure. Another way to see how well features have been integrated into the structure model is to examine signal predictions. It is well-known that translation initiation sites (TIS) are surrounded by relatively poorly conserved sequences and are harder to predict than the highly conserved splice signals. Also, stop signals present almost no sequence conservation at all and their prediction depends solely upon how well the last acceptor (in multi-exon genes) or the TIS (in single-exon genes) was predicted. Therefore, a simple splice site classifier can perform fairly well using only local sequence information. In contrast, TIS and stop signal classifiers are known to be much less accurate.

Given these observations, we expected CRAIG to improve the most on TIS signal prediction accuracy, as all other programs examined in this work use individual classifiers for signal prediction, whereas CRAIG uses global training to compute each signal’s net contribution to the gene structure. Figure 4 shows the improvement in signal prediction accuracy for CRAIG when compared with the second-best program in each case. CRAIG shows improvement for all types of signals, but the improvement was most marked for TIS, especially in specificity. It can also be observed that the improvement on stop signals follows from the co-occurring improvement on both acceptor and TIS signals. The final outcome is that CRAIG makes fewer mistakes in deciding where to start translation and stop translation, which is one of the main reasons for its significant improvement at the gene level.

There is great potential for including additional informative features into the model without algorithm changes, for instance, features derived from comparative genomics. To facilitate such extensions, we designed CRAIG to allow model changes without recompiling the C++ training and test code. The finite-state model, the features, and their relationships to states and transitions are all specified in a configuration file that can be changed without recompiling the program. This flexibility could be useful for learning gene models on organisms that may require a different finite-state model or a different set of features.

Materials and Methods

Gene structures. In what follows, a gene structure consists of either a single exon or a succession of alternating exons and introns, trimmed from both ends at the TIS and stop signals. We distinguish two different types of introns: short—980 bp or less—and long—more than 980 bp. Figure 5 shows a gene finite-state model that implements these distinctions.

Linear structure models. In what follows, \( x = s_1, \ldots, s_P \) is a sequence and \( s = s_j, \ldots, s_q \) is a segmentation of \( x \), where each segment \( s_j \) starts at position \( pos(s_j) = p_j \) and has length \( \text{len}(s_j) = t_j \), and state label \( \text{lab}(s_j) \) is \( s_j \), with \( p_{j+1} = p_j + t_j \) and \( 1 \leq p_j \leq B \) for some empirically determined upper bound \( B \). The training data is \( \mathcal{F} = \{ (x^{(i)}, s^{(i)}) \}_{i=1}^T \), consists of pairs of a sequence and its preferred segmentation. For DNA sequences, \( x_t \in \Sigma_{DNA} = \{ A, T, G, C \} \), and each label \( \text{lab}(s_j) \) is one of the states of the model (Figure 5). A segment is also referred to as a genomic region; that is, an exon, an intron, or an intergenic region. A first-order Markovian linear structure model computes the score of a candidate segmentation \( s = s_j, \ldots, s_q \) of a given input sequence \( x \) as a linear combination of terms for individual features of a candidate segment, the label of its predecessor, and the input sequence. More precisely, each proposed segment \( s_j \) is represented by a feature vector \( f(s_j, x) \in \mathbb{R}^p \) computed from the segment, the label of the previous segment, and the input sequence around position \( pos(s_j) \). A weight vector \( w \in \mathbb{R}^p \) to be learned, represents the relative weights of the features. Then, the score of candidate segmentation \( s \) for sequence \( x \) is given by
values are precomputed and cached. For introns and intergenic regions the constant factor from these dot product computations, most like recurrence:

\[
M(i, y) = \left\{ \begin{array}{ll}
\max_{y \in [y_m, y_m]} & M(i-1, y^\prime) + w \cdot f((i-1, y), y, x) & \text{if } i > 0 \\
0 & & \text{if } i = 0
\end{array} \right.
\]

It is easy to see that \( M(i, y) = \max_{x \in \text{GEN}(x)} S_i(x, s) \) where \( \text{GEN}(x) \) is the set of all segmentations of \( x \) that end with label \( y \). Therefore, \( S_i(x) = M(P + 1, \text{END}) \), where \( \text{END} \) is a special synchronization state inserted at position \( P + 1 \). The actual segmentations are easily obtained by keeping back-pointers from each state-position pair \((y, i)\) to its optimal predecessor \((y, i-1)\). The complexity of this algorithm is \( O(Pm^2) \), where \( m \) is the number of distinct states and \( B \) is the upper bound on the segment length, because the runtime of \( w \cdot f \) is independent of \( P, B \), or \( m \). To reduce the constant factor from these dot product computations, most \( w \cdot f \) values are precomputed and cached. For introns and intergenic regions, the feature function \( f \) is a sum of per-nucleotide contributions, so the dynamic program in Equation 4 needs only to look at position \( i - 1 \) when \( y \) corresponds to such regions. Therefore, \( B \) needs to be only the upper bound for exon lengths, which was chosen following Stanke and Waack [2]. For long sequences, the complexity of the inference algorithm is therefore dominated by the sequence length \( P \).

Online large-margin training. Online learning is a simple, scalable, and flexible framework for training linear structured models. Online algorithms process one training example at a time, updating the model weights to improve the model's accuracy on that example.

Large-margin classifiers, such as the well-known SVMs, provide strong theoretical classification error bounds that hold well in practice for many learning tasks. MIRA [30] is an online method for training large-margin classifiers that is easily extended to structured problems [19]. Algorithm 1 shows the pseudocode for the MIRA-based training algorithm we used for our models. For each training sequence, \( x^i \), the algorithm seeks to establish a margin between the score of the correct segmentation and the score of the best segmentation according to the current weight vector that is proportional to the mismatch between the candidate segmentation and the correct one. MIRA keeps the current weight vector that is proportional to the mismatch between the correct segmentation and the incorrect one. The quadratic program in line 5 of Algorithm 1 formalizes that objective, and has a straightforward closed-form solution for this version of the algorithm. Line 11 of the algorithm computes \( w \) as an average of the weight vectors obtained at each iteration, which has been shown to reduce weight overfitting [31]. The training parameter \( N \) is determined empirically using an auxiliary development set.

Algorithm 1. Online Training Algorithm.

Training data \( \mathcal{F} = \{(x^0, y^0), (x^1, y^1), \ldots\} \) is some nonnegative real-valued function that measures the mismatch between segmentation \( \hat{s} \) and the correct segmentation \( s^0 \). The number of rounds \( N \) is determined using a small development set.

1: \( w^0 = 0 \); \( v = 0 \); \( i = 0 \)
2: for round = 1 to \( N \) do
3: \( i = i + 1 \)
4: \( s = \arg \max_{s \in \text{GEN}(s)} S_i(x, s) \)
5: \( w = \arg \min_{w} \| w - w^0 \|^2 \)
6: subject to \( s_L(x, y^0) - s_M(x, y_1) \geq L(s^0, s) \)
7: \( w^{(i + 1)} = w \)
8: \( v = v + w^{(i + 1)} \)
9: \( i = i + 1 \)
10: end for
11: \( w = v / (N^T) \)

Successful discriminative learning depends on having training data with statistics similar to the intended test data. However, this is not the case for gene training data. The main distribution mismatch is that reliable gene annotations available for training are for the most part for single-gene sequences with very small flanking intergenic regions.

To address this problem, we created long training sequences composed of actual genes separated by synthetic intergenic regions as follows. For each training sequence, we generated two extra intergenic regions and appended them to both sequence ends, making sure that the total length of both flanking intergenic regions followed geometric distributions with means 5,000, 10,000, 60,000, and 130,000 bp for each of four GC content classes, respectively [3,10]. The synthetic intergenic regions were generated by sampling from a GC-dependent, fourth-order interpolated Markov models (IMMs), with the same form as the models we used to score the intergenic state.

Algorithm 1 also requires a loss function, \( L \), and a small development set on which to estimate the number of rounds, \( N \). As loss function, we used the correlation coefficient at the base level [24], since it combines specificity and sensitivity into a single measure. The development set consisted of the 65 genes previously used in GenScan [1] to cross-validate splice signal detectors.

Features. The final ingredient of the CRAIG model is the feature function \( f \) used to score candidate segments based on properties of the input sequence. A typical feature relates a proposed segment to some property of the input around that segment, and possibly to the label of the previous segment.

Properties. We started by introducing basic sequence properties that features are based on. These properties are real-valued functions of the input sequence around a particular position. Some properties represent tests, taking the binary values \( 1 \) for true and \( 0 \) for false. For any test \( P_i \), \( ||P_i|| \) denotes the function with value \( 1 \) if the test is true, \( 0 \) otherwise.

The tests

\[
\text{sub}_{i}(x) = \| u = x[j : i + |u| - 1] \|
\]
check whether substring x occurs at position i ∈ X. For example, x = ATGCGCGGA would have sub1(x,1) = 1, sub1(x,2) = 0, and sub3(x,3) = 1.

The property score(x1) computes the score of a content model for state y at position i. This score is the probability that nucleotide i has label y according to a k-order interpolated Markov model [32], where k = 8 for coding states and k = 4 for noncoding states.

The property gcc(x1) calculates the GC composition for the region containing position i, averaged over a 10,000-bp window around position i.

Each feature associates a property to a particular model state or state transition.

**Binning.** Properties with multimodal or sparse distributions, such as segment length, cannot be used directly in a linear model, because their predictive effect is typically a nonlinear function of their value. To address this problem, we binned each property by splitting its range into disjoint intervals or bins, and converting the property into a set of tests that checked whether the value of the property belonged to the corresponding interval. The effect of this transformation was to pass the property through a stepwise constant nonlinearity, each step corresponding to a bin, where the height of each step was learned as the weight of a binary feature associated to the appropriate test.

For example, following GenScan [1], we mapped the GC content property gcc to four bins: -43, 43–51, 51–57, and >57. For other properties, we used regular bins with a property-specific bin width. For instance, exon length was mapped to 90 bp-wide bins.

**Test and feature combinations.** We used Boolean combinations of tests and binary features to model complex dependencies on the input. Conjunctions can model nucleotide correlations, for example donors of the form U3, that is, donors with either an A or a C at position 3.

In general, for two binary functions f and g, we denoted their conjunction by f ⋀ g and their disjunction by f ⋁ g.

**State features.** State features encode the content properties of the genomic regions associated to states: exons, introns, and intergenic regions. State features do not depend on the previous state, so we omitted the previous state argument in these feature definitions.

**Codingnoncoding potential.** This feature corresponds to the log of the probability assigned to the region by the content scoring model:

\[
pot(x) = \frac{1}{\text{len}(s)} \sum_{k=\text{pos}(s)}^{\text{pos}(s)+\text{len}(s)} \log \text{score}_{\text{gcc}}(x, k) - \text{bias}_y(x)
\]

where \( \mu \) is the arithmetic mean of the distribution of log score, on the training data. For coding regions, the sum is computed over codon scores instead of base scores. Other features related to log score, also included in \( \text{pot} \) are the coding differential and the score log-ratios between intronic and intergenic regions.

**Phase biases.** Biases in intron and exon phase distributions have been found and analyzed by Fedorov et al. [33]. We represented possible biases with the straightforward functions.
where $p$ is the offset, $w$ is the window width, and $a$ is the motif. This feature counts the number of occurrences of $w$ within $p \leq w/2$ bases of the start of segments.

In principle, all sequence positions are potential signal occurrences, but in practice one might filter out unlikely sites, using a sensitivity threshold proportional to level of signal conservation, thus decreasing decoding time.

Burge and Karlin [1] model positional biases within signals with combinations of position weight matrices (PWMs) and their generalizations, weight array models (WAMs) and windowed weight array models (WWAMs), with very good results. It is straightforward to define these models as sets of features based on our motif feature, as shown here in the WWAM case:

$$\text{WWAM}_{\sigma,w} (s, y', x) = \left\{ \text{motif}_{p,a,w} (s, y', x) : y \in \sum_{\text{DNA}}^{n} \right\}$$

PWMs and WAMs are special cases of WWAMs and can thus be defined by $\text{PWM}_{w} = \text{WWAM}_{\sigma,w} |_{\sigma = 1}$ and $\text{WAM}_{w} = \text{WWAM}_{\sigma,w} |_{\sigma = 0}$. This means that we can use all of these techniques to model biological signals in CRAIG with the added advantage of having all signal model parameters trained as part of the gene structure.

Correlations between two positions within a signal are captured by combinations of motif features. For example, feature conjunction $\text{motif}_{\sigma,w} (156, 20, 1), E_i \times \text{motif}_{\tau,w} (156, 20, 1), E_i \times x$ would be 1 whenever there is an $A$ in position --3 and a $T$ in position 2 relative to a donor signal occurring at position 156 in $x$, and 0 otherwise.

We can also extend the feature conjunction operator to sets of features: if $A$ and $B$ are sets of features, such as the WWAM defined in Equation 5, we can define the set of features $A \times B = \{ f \land g : f \in A, g \in B \}$.

In Equation 5, if we use the amino acid alphabet $\sum_{\text{DNA}}$, and work with codons instead of single nucleotides, we can model signal peptide regions. If we sum over disjunctions of motif features, we can easily model consensus sequences.

Table 7 shows the motif feature sets for each biological signal. The parameters required by each feature type were either taken from the literature [1] or by search on the development set.

We included an additional feature set, motivated by previous work [2], to learn splice site information from sequences that only contain intron annotations. For any donor (acceptor), we first counted the number of similar donors (acceptors) in a given list of introns. A signal was considered to be similar to another if the Hamming distance between them was at most 1. The features were induced by a logarithmic binning function applied over the total number of similarity counts for Hamming distances 0 and 1.

Acknowledgments

We thank Aaron Mackey for advice on evaluation methods, datasets, and software.

Author contributions. AB, AH, and FP conceived and designed the experiments. AB performed the experiments and analyzed the data. AB and FP wrote the paper. AB, KC, and FP contributed ideas to the model and algorithms and refined and implemented the algorithms. FP proposed the initial idea.

Funding. This material is based on work funded by the US National Science Foundation under ITR grants EIA 0205456 and IIS 0428193 and Career grant 0238295.

Competing interests. The authors have declared that no competing interests exist.