Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction

Qian Liu1,*, Aaron J. Mackey2,3, David S. Roos2,3 and Fernando C. N. Pereira1

1Department of Computer and Information Science, 2Department of Biology and 3Penn Genomics Institute, University of Pennsylvania, Philadelphia PA 19104, USA

Received on October 15, 2007; revised on December 13, 2007; accepted on January 3, 2008

Advance Access publication January 10, 2008

ABSTRACT

Motivation: The increasing diversity and variable quality of evidence relevant to gene annotation argues for a probabilistic framework that automatically integrates such evidence to yield candidate gene models.

Results: Evigan is an automated gene annotation program for eukaryotic genomes, employing probabilistic inference to integrate multiple sources of gene evidence. The probabilistic model is a dynamic Bayes network whose parameters are adjusted to maximize the probability of observed evidence. Consensus gene predictions are then derived by maximum likelihood decoding, yielding n-best models (with probabilities for each). Evigan is capable of accommodating a variety of evidence types, including (but not limited to) gene models computed by diverse gene finders, BLAST hits, EST matches, and splice site predictions; learned parameters encode the relative quality of evidence sources. Since separate training data are not required (apart from the training sets used by individual gene finders), Evigan is particularly attractive for newly sequenced genomes where little or no reliable manually curated annotation is available. The ability to produce a ranked list of alternative gene models may facilitate identification of alternatively spliced transcripts. Experimental application to ENCODE regions of the human genome, and the genomes of Plasmodium vivax and Arabidopsis thaliana show that Evigan achieves better performance than any of the individual data sources used as evidence.

Availability: The source code is available at http://www.seas.upenn.edu/~strctlrn/evigan/evigan.html

Contact: qianliu@seas.upenn.edu

1 INTRODUCTION

As genome sequencing becomes more efficient, manual gene annotation has increasingly become a bottleneck in genome projects. Curated gene models are extremely valuable, but suffer from several inherent limitations. First, manual curation is slow and expensive, and thus lags increasingly behind as the acquisition of genomic sequence data accelerates. Second, successful manual curation depends heavily on biological evidence, which is often unavailable in the relatively understudied organisms that are becoming the target of sequencing efforts. Third, consistency and reproducibility are difficult to maintain, because individual curators use different annotation methods and evidence, as may even be the case for the same curator at different times. Finally, it can be difficult to track the sources of evidence that inform the manual curation process, complicating model revision as new data is acquired.

During manual curation, human annotators typically examine multiple sources of evidence, in order to produce a consensus view of the structure of a newly sequenced genome. We have developed an automated method that mimics this process of evidence integration and consensus building. Evigan (EVidence Integration for Genome Annotation using a Net-work) employs a dynamic Bayes net (DBN), a type of probabilistic graphical model that can accommodate multiple (possibly incomplete) gene predictions and other lines of evidence, yielding consensus gene models that maximize the probability of the evidence provided. The DBN model supports a wide variety of evidence types, including computational gene predictions, sequence homology search results, EST alignments and splice site predictions and is easily extensible to incorporate other evidence types, such as proteomics hits, predicted domain architecture, SAGE tags, or Affymetrix tiling array data. Evigan’s annotation process simulates an idealized human curator: different evidence sources are compared, those that tend to agree in particular contexts are assigned higher confidence and a consensus model is then created that reflects those confidence estimates. Evigan can produce a single consensus gene model or an ordered list of the n-best gene models, along with associated posterior probabilities for each. The resulting consensus may be edited by a human annotator to reflect biological knowledge and experience not available to the algorithm.

Previous algorithms for creating consensus gene models from diverse evidence can be classified into two main approaches: (i) voting and graph-based methods that assemble a consensus prediction using predefined rules to combine predictions from different sources (Murakami and Takagi, 1998; Rogic et al., 2002; Schiex et al., 2001; Howe et al., 2002; Brejova et al., 2005; Coghlan and Durbin, 2007), and (ii) machine-learning methods that learn rules for combining multiple sources by maximizing the prediction accuracy on a training set of manually annotated genes (Pavlovic et al., 2002; Allen et al., 2004, 2005). Evigan...
brings together the best features of those two approaches. Like approach (ii), Evigan adapts its combination rule to the actual data being integrated, and like approach (i), it does not require a manually annotated training set. Evigan achieves this by learning a dynamic Bayes network that maximizes the likelihood of the observed sources of evidence. The consensus prediction is then the most probable state sequence for the learned network (see below for further discussion).

In this report, we have applied Evigan to three large-scale data sets: the ENCODE regions of the human genome (ENCODE project consortium, 2004), and the genomes of *Plasmodium vivax* (TIGR, 2007) and *Arabidopsis thaliana* (Allen et al., 2004). These experiments demonstrate that for all three species, Evigan achieves better performance than any individual data source used as evidence.

## 2 EXPERIMENTS AND RESULTS

Evigan (Section 4 for algorithmic details) has been applied to ENCODE regions of the human genome, the genome of the protozoan parasite *P. vivax* (one of the causative agents of malaria), and the genome of the flowering plant *A. thaliana*. We assessed Evigan’s performance against well-curated gene models assumed to be the gold standard. We also compared Evigan with the following baseline gene models:

- Models used as individual evidence sources for Evigan;
- Majority voting among the evidence sources.
- For *A. thaliana*, the Combiner model (Allen et al., 2004).
- GLEAN (Elsik et al., 2007).

### 2.1 ENCODE

For this experiment, we run Evigan on 31 of the 44 fragments of the human genome selected in The Encyclopedia of DNA Elements (ENCODE) project (ENCODE project consortium, 2004), which was also used as the testing set in the ENCODE Genome Annotation Assessment Project (EGASP) (Guigo and Reese, 2005; Guigo et al., 2006) gene prediction evaluation. EGASP evaluated the performance of more than 20 gene finders on those 31 regions by comparing their results with gene models curated by the HAVANA group (Guigo et al., 2006). Participating gene finders were categorized into four groups based on the information that they were allowed to exploit in making predictions: the human genomic sequence only (*ab initio* gene prediction; EGASP category 2); genomic sequences from other species (EGASP category 4); EST, mRNA and protein sequence alignments (EGASP category 3); or any available information (EGASP category 1). The general trend found in EGASP was that gene finders in categories 1 and 3 exhibit better accuracy than those in categories 2 and 4 because they have access to richer features.

Source gene prediction sets used in our experiment include: (I) EGASP entries: AUGUSTUS-any (Stanke and Waack, 2003; Stanke et al., 2006), FgenesH++ (Solovyev et al., 2006), Jigsaw (Allen et al., 2005, 2006), Pairagon-any (Arunugam et al., 2006), GeneMark (Lukashin and Borodovsky, 1998), GeneZilla (Majoros et al., 2004), AceView (NCBI, http://www.ncbi.nlm.nih.gov/IEB/Research), Ensembl (Curwen et al., 2004), Exogean (Djebali et al., 2006), ExonHunter (Brejova et al., 2005), DogFish (Carter and Durbin, 2006), Mars (Flicek and Brent, 2006), and Saga (Chatterji and Pachter, 2005), (II) UCSC tracks: GeneID (Parra et al., 2000), Genscan (Burge and Karlin, 1997) and Twinscan (Korf et al., 2001; Flicek et al., 2001); (III) predictor published after EGASP: CRAIG (Bernal et al., 2007).

At present, Evigan can only use evidence about a single candidate transcript from each source (Section 3). While the gene finders in categories 2 and 4 predict single-transcript genes only, some gene finders in categories 1 and 3 predict multiple transcripts at a locus, based on evidence of alternatively spliced, overlapping EST or cDNA sequences. Where multiple transcripts are provided, Evigan takes just the longest available transcript, although all transcripts are retained in the annotation set used for performance evaluation. This filtering was applied to the Ensembl, Aceview, Exogean, FgenesH++ and Mars predictions. We also left out non-coding predictions, integrating and evaluating only predicted CDS between translation start and stop sites. Following the EGASP guidelines, an exon (or CDS) is counted correct if its boundaries and strand are exactly matched. A (coding) transcript is counted correct if all its CDS are exactly matched. A gene is counted as correct if one of its transcripts is predicted correctly. For incomplete transcripts in the annotation set, the EGASP analysis counts a predicted transcript as correct if its components are consistent with those of an annotated transcript; we have adopted a slightly more stringent measure wherein a predicted transcript is considered correct only if its components coincide precisely with those of an annotated transcript.

Source gene prediction sets in the ENCODE dataset are of variable quality. To estimate the sources’ quality in the absence of reference gene annotations, we computed a quality score for each source based on the Evigan model parameters...
associated with the source, which represent an average level of agreement between the inferred consensus and the source, as detailed in Section 4.4. Evigan was first run with all 17 sources, producing the quality scores shown as blue diamonds in Figure 1. The results of this analysis show that gene finders from categories 1 and 3 tend to have better scores than the ones from categories 2 and 4, matching closely the EGASP results on reference annotation (see Supplementary Material), suggesting that Evigan quality scores are good indicators of the true quality of sources.

Although the ENCODE project provides a large collection of gene models for the human genome, far fewer data sources are likely to be available for other organisms. We therefore sought to determine how many and which data sources are sufficient for best Evigan performance. We ran Evigan on each subset of the top $k$ sources from the ranked list shown in Figure 1, for $k = 1, \ldots, 17$, yielding the F-score curves for the exon, gene and transcript levels shown in Figure 2. As more sources are added, performance improves until $k = 8$ for all three levels. For $k > 8$, performance starts decreasing as lower-quality sources are integrated. It is interesting to note that the average quality score for the top $k$ sources, computed from the parameters inferred by Evigan for those sources, tracks closely with the accuracy of the Evigan-produced consensus for those sources with respect to the EGASP reference annotation. This suggests that the Evigan individual source and average quality scores may be useful proxies for the accuracy of individual sources and of the Evigan consensus when no gold standard is available.

Following the above heuristic, we selected the eight sources with the highest quality scores for integration with Evigan. Figure 3 displays the sensitivity and specificity of the Evigan consensus of those eight sources, labeled Evigan-8g, in comparison with individual sources and with the Evigan combination of all sources, Evigan-17g (see Supplementary Material for detailed evaluation results). Overall, Evigan-8g outperforms any of the constituent data sources on which its analysis was based at all levels (exon, gene and transcript). Although the exclusion of multiple overlapping transcripts (such as those from differentially spliced cDNAs) from the source datasets supplied to Evigan decreases their accuracy, particularly with respect to sensitivity, a comparison of the performance statistics with the EGASP results indicates that this effect is small (see Supplementary Material). Using a larger collection of less accurate data sources (Evigan-17g) significantly reduces performance.

Evigan’s quality score estimate for a source depends on the set of sources being integrated. The quality score for a source inferred by running Evigan on all the 17 sources might be different than the one inferred by running Evigan on a subset of the sources. To evaluate the robustness of Evigan quality score estimates when different subsets of sources are integrated, 20 random subsets of eight sources were sampled and integrated by Evigan. Quality score estimates for each source from the experiments were collected and are shown as green dots in Figure 1. Overall quality scores of the sources inferred from the random subsets are robust and close to the ones inferred from all the 17 sources, although Genscan and Saga have outliers. Furthermore, Evigan’s performance using the top eight sources (Evigan-8g) is better than its performance using any of the random eight-source subsets, suggesting that Evigan-8g has near-optimal performance among any eight-source subsets.

Evigan’s quality score estimate for a source depends on the set of sources being integrated. The quality score for a source inferred by running Evigan on all the 17 sources might be different than the one inferred by running Evigan on a subset of the sources. To evaluate the robustness of Evigan quality score estimates when different subsets of sources are integrated, 20 random subsets of eight sources were sampled and integrated by Evigan. Quality score estimates for each source from the experiments were collected and are shown as green dots in Figure 1. Overall quality scores of the sources inferred from the random subsets are robust and close to the ones inferred from all the 17 sources, although Genscan and Saga have outliers. Furthermore, Evigan’s performance using the top eight sources (Evigan-8g) is better than its performance using any of the random eight-source subsets, suggesting that Evigan-8g has near-optimal performance among any eight-source subsets.

Evigan for eukaryotic gene prediction
2.2 Arabidopsis thaliana

A more limited set of gene predictions is available for A. thaliana, but annotation of this plant genome has benefitted from Combiner (Allen et al., 2004), which uses a decision tree trained on a separate curated training set to adjudicate among the predictions of multiple gene finders. The complete set of predictions from all gene finders used by Combiner, gene predictions from Combiner itself, and a list of 2163 genes annotated based on full-length cDNA alignments and manual curation (Haas et al., 2002) were kindly provided by the authors. Evidence sources included predictions from Genemark HMM (Lukashin and Borodovsky, 1998), Genscan (Burge and Karlin, 1997), GlimmerA (Pertea and Salzberg, 2002), GlimmerM (Pertea and Salzberg, 2002) and Twinscan (Korf et al., 2001), splice site predictions (Pertea et al., 2001) and protein alignments. All gene finders were trained before the annotated gene set became available, and they only produce non-overlapping gene models. Protein alignments between (the translations) of candidate coding DNA sequences and elements of a non-redundant protein database were created with the dps and nap programs (Huang et al., 1997). Alignment data was filtered to remove proteins from A. thaliana, which might have biased Evigan’s predictions. Only HSPs with >50% amino acid sequence identity were used, and only the longest HSP was selected in cases of HSP overlap.

Following the original Combiner publication (Allen et al., 2004), the data set was split into a small training set (380 annotated genes) and a larger test set (1783 annotated genes), and Combiner parameters were estimated from the annotated training set. Evigan’s parameters were estimated on the whole data set (training and test) and GLEAN’s on the test set from which reference gene annotations had been removed. The performance of each of the individual gene finders, GLEAN, Evigan and Combiner was assessed by comparison with the 1783 gene test set. Slight differences from the results previously reported for Combiner are attributable to the use of a more recent version of Combiner in this study.

Two sets of experiments were carried out on the Arabidopsis genome, with the results summarized in Figure 4 (see the Supplementary Material for detailed performance tables). In the first set, analogous to our ENCODE study, five gene finders were used as data sources for both Evigan and Combiner (Combiner-G, GLEAN-G, Evigan-G). In the second set, these five gene finders were integrated with splice site and protein sequence alignment evidence (Combiner-A, Evigan-A) (GLEAN currently cannot incorporate protein alignment data). In both of these experiments, all of the data integration strategies (Combiner, GLEAN and Evigan) achieve higher sensitivity and specificity at both the gene- and exon levels than any of the gene finders used as data sources. In particular, both Combiner and Evigan improve upon Twinscan, which in turn far exceeds the results from GenemarkHMM, Genscan, GlimmerA, GlimmerM. That is, information from four gene finders that perform relatively poorly is actually helpful in correcting some of the mistakes of a much better gene finder. Integrating additional information (splice sites, protein sequence alignments) results in further improvements (compare -A with -G). The performance of Evigan and GLEAN are almost the same on the exon level and gene level. The performance of Combiner slightly exceeds that of Evigan and GLEAN, but Combiner depends on carefully curated training set, while Evigan and GLEAN infer their parameters solely from the data sources provided, none of which includes the training annotations used by Combiner.

2.3 Plasmodium vivax

Plasmodium vivax is one of four parasite species that cause human malaria. Although most deaths from malaria are attributable to P. falciparum, P. vivax infections may be even more prevalent (Mendis et al., 2001). In order to expedite research on this parasite, which cannot be cultivated in vitro, a reference P. vivax genome has recently been sequenced (TIGR, http://www.tigr.org/tdb/e2k1/pva1/pva1.shtml). The first-pass annotation of this genome exploited gene predictions from GeneZilla (Majoros et al., 2004) (6090 predicted genes), GlimmerHMM (Pertea and Salzberg, 2002) (12047), and Phat from GeneZilla (Majoros et al., 2004). The first-pass annotation of this genome exploited gene predictions from GeneZilla (Majoros et al., 2004) (6090 predicted genes), GlimmerHMM (Pertea and Salzberg, 2002) (12047), and Phat (Cawley, 2001) (4563), as well as ESTs and BLASTX hits. Manual curation steps yielded 5354 annotated genes. All of these datasets were downloaded from plasmodb.org (PlasmoDB, http://www.plasmodb.org/plasmo/home.jsp). All three gene finders predict non-overlapping gene models; in cases of overlapping ESTs, only the longest was used as input for Evigan; in cases where BLASTX hits overlapped, the one with the lowest E value was used.

As with A. thaliana, two sets of experiments were carried out on the P. vivax genome, integrating three gene finder models, with or without the inclusion of EST and BLASTX data, as shown in Figure 5 (see Supplementary Material for detailed evaluation results). Sensitivity and specificity were evaluated at the exon- and gene-level on a manually curated dataset. Once again, the performance of Evigan exceeds any individual gene finder. Evigan also outperforms majority voting (designated by ‘MV’ in Figure 5), because this analysis treats all evidence sources equally. Ascribing equal weight to the limited EST dataset (low sensitivity) and very noisy BLASTX evidence (low specificity) causes MV-A to perform worse than MV-G. Since Evigan learns to weigh individual data sources according to their
Evigan represents evidence from diverse sources (such as gene models, BLAST hits and EST matches) similarly to Combiner and Jigsaw, and uses a probabilistic graphical model (Jordan, 1999) like Pavlovic et al. do to represent the evidence integration task. Unlike those previous programs; however, Evigan does not rely on annotated training data. Instead, Evigan uses a carefully initialized expectation-maximization (EM) method to estimate model parameters directly from the observed evidence, making it suitable for systems where little or no annotated data is available.

The sources integrated by Evigan may have been developed with the help of annotated training sets, for instance ab initio gene finders. Evigan itself does not depend on what training data was used for the sources or on how they were trained. No training data is needed for Evigan to estimate its model parameters. In addition, our experiments with Evigan-derived source quality scores for EGASP sources suggest that scores based on Evigan’s comparison of sources are useful proxies for the expected accuracy of those sources, even when the sources are quite similar, and can thus be used to select sources likely to enhance the accuracy of the Evigan consensus. When a random subset of the sources is selected and quality scores inferred for the sources in the subset, the quality scores are consistent with the ones inferred from running Evigan on all the sources together.

Evigan is a refinement and generalization of the GLEAN program (Elsik et al., 2007). GLEAN represents sources of evidence through their support or non-support for proposed signals (coding starts and stops, splice donors and acceptors) while Evigan represents them as gene components at consecutive segments between signals. GLEAN learns a scoring function for each signal type and constructs the optimal gene parse with dynamic programming, whereas Evigan formulates evidence and gene parse in a single sequential probabilistic model so that learning and inference can be integrated. Like Evigan, GLEAN does not require a training set for parameter estimation. GLEAN’s signal-based evidence representation and scoring function, however, makes it difficult to incorporate segment-based evidence types, such as BLASTX (protein to genome sequence) alignments, SAGE tags, or tiling array data.

We made several important assumptions in choosing to learn the relative quality of evidence sources without resorting to a training set. First, we assume that there is a fair amount of agreement among sources because we actually learn their quality by checking how often they agree with each other via a proposed consensus. Second, we assume that majority voting is conditionally independent given the consensus gene model. This is a much weaker condition than asking for the sources to be unconditionally independent, since it only requires that there should not be strong correlations between source departures from the truth. Still, this is a matter of some concern, as many gene finders are trained on related datasets, use similar features, or are based on...
similar algorithms. Evigan could then fail if the errors of different sources were strongly correlated. For example, if one source provides good quality evidence whereas all the others are copies of a poor source, Evigan would wrongly learn to prefer the replicated poor sources.

Another limitation of Evigan is that the evidence provided by each source has to be extendible to a single gene model. In particular, gene models with alternative transcripts cannot be used, due to which we had to select just the longest transcript for each gene predictor that produces multiple transcripts. The reason for this limitation is that the probabilistic representation of each source (Section 4) is a single chain of random variables. We do not have yet a practical way of lifting this limitation, but this is an interesting area for further work.

In practice, however, Evigan performed quite well on several different datasets, as shown in Section 2. For all three genomes examined, Evigan outperformed all of the data sources used as evidence. Although these data sources are not always identical to the gene prediction sets generated by published algorithms, as some gene predictions were removed to provide the current version of Evigan with a non-overlapping set of gene models, the important point is that Evigan yields better gene predictions than any of the input data sources and also better even than the majority voting consensus (Fig. 5). Overall, the performance of Evigan is comparable to that of Combiner on *A. thaliana*, without Combiner’s need for a curated training set.

One further feature of Evigan is the ability to generate the n-best gene models, with associated posterior probabilities. Although the highest probability model is selected in the absence of further information, the larger list of candidates is potentially valuable, suggesting possible alternatively spliced transcripts, and as hypotheses to be integrated with additional computational and biological information. Automated annotation cannot completely replace manual annotation because human experts will use additional knowledge not present in the individual sources of evidence. However, a program like Evigan can serve as a fast first-pass annotation tool. Human curators can then quickly look through Evigan’s predictions to decide whether they want to make further changes. For example, if the posterior probability of the best consensus gene model is high, the curator may decide to keep it. If instead the posterior probability of the best model is below some threshold, the curator may examine the n-best gene models or go back to the individual sources of evidence to resolve the uncertainty.

Curation effort will thus be focused on just those cases in which the evidence is so conflicting that the automated method cannot come up with a high-probability consensus, rather than having to be applied uniformly everywhere. Although Evigan’s n-best gene models have the potential to suggest possible alternatively spliced transcripts, there is no guarantee that all the models proposed by Evigan are legitimate transcripts. To illustrate this, Figure 6 shows alternative models for a gene from the ENCODE set (EntrezGene ID: LAIR1, leukocyte-associated Ig-like receptor 1, on reverse strand). The curated transcripts and Evigan’s five best models are visualized on the top and bottom of the figure, respectively. The posterior probabilities of Evigan’s five best models are all similar. Evigan’s third and fourth best models Evigan_3 and Evigan_4 match exactly with the curated alternative transcripts AC0087461.1-001 and AC0087461.1-003, respectively. However, the other Evigan models differ from the curated transcripts. How to automatically recognize true alternative transcripts in Evigan’s n-best gene models is an open research question that we plan to investigate.

As shown in the experiments on *A. thaliana* and *P. vivax*, adding splice site prediction, protein and EST evidence improves Evigan’s performance over integrating gene prediction sets only. Aggregation of different types of evidence works well in our experiments, given they are complementary to each other and properly weighted according to their relative strength. It may seem contradictory that the best results for *A. thaliana* use all of the sources, whereas using all of the
sources leads to less good results in the ENCODE experiment. This apparent conflict can be explained by the differences between the sources in the two experiments. For *A. thaliana*, all of the sources are *ab initio* or *de novo* gene finders that exploit genomic sequence only, use similar features, have similar performance and often agree with each other. Therefore, each source seems mostly right from the point of view of the other sources, and Evigan’s task is the relatively simple one of adjudicating a relatively small number of disagreements. In contrast, the 17 sources available for the ENCODE experiment differ from each other drastically in terms of input features and performance. Their higher level of disagreement has the effect of diluting a minority of good sources with a majority of relatively bad sources, decreasing overall performance unless sources with low quality scores are excluded.

In addition to the experiments described in this study, we have also entered Evigan in the Nematode Genome Annotation Assessment Project (NGASP) evaluation for *C. elegans* gene prediction. Preliminary results of that evaluation show that Evigan performs competitively among combiner gene finders (NGASP, [http://www.wormbase.org/wiki/index.php/NGASP](http://www.wormbase.org/wiki/index.php/NGASP)). Furthermore, Evigan has recently been used to create gene models for a *P. falciparum* re-annotation workshop whose results will be soon available on PlasmoDB ([PlasmoDB](http://www.plasmodb.org/plasmo/home.jsp)).

As we already noted, the current version of Evigan cannot use evidence for alternative models from a single source, and is thus restricted to selecting a single hypothesis for each source. This might cause Evigan to underestimate the value of such sources. In addition, alternative model evidence could be very useful in helping Evigan select a set of *n*-best models that is more likely to include true alternative transcripts. This would involve representing multiple alternative states for each segment and associating them with appropriate state transitions at segment boundaries. These questions clearly deserve further research. Another area for further study is how to use partial sources of evidence to improve the consensus for specific gene model components. For example, EST and protein sequence alignments may be very informative for splice donor and acceptor prediction, even if they are less accurate than gene finders in predicting complete gene models. Preliminary results indicate that other noisy data types can also be very valuable. For example, SAGE tag results can improve the recognition of acceptor prediction, even if they are less accurate than gene finders in predicting complete gene models. Preliminary results indicate that other noisy data types can also be very valuable.

**4 METHODS**

We formulate the gene evidence combination task as a sequence labeling problem, and then represent the sequence labeling problem with a dynamic Bayes network. This representation allows us to model different types of evidence in a uniform way. Standard algorithms are used to estimate the DBN parameters and to compute the consensus gene model. (See Supplementary Material for a flow chart of the system.)

**4.1 Sequence segmentation**

Figure 7 provides a schematic view of how evidence and consensus are represented in Evigan. In this toy example, track 1 displays a gene model, produced by gene finder 1, consisting of two exons and one intron; tracks 2 and 3 display slightly different gene models, generated by gene finders 2 and 3. The splice predictor showed in track 4 proposes possible donor and acceptor sites, while tracks 5 and 6 display HSPs representing transcript and protein alignments.

The cumulative set of potential signals proposed by any of these sources is then used to divide the genomic sequence into segments, with signal sites as boundaries, as shown on the top track in the figure (16 segments in this case). Alternatively, one could segment the sequence based on each occurrence of a conserved motif for each signal type, but this would produce a significantly larger number of segments, and therefore increased computational cost.

Next, sources of gene evidence are mapped onto these segments as possible states (exon, intron or intergenic). For example, the gene model predicted by gene finder 1 (track 1) maps to the 16 sequence segments to yield the following series: one intergenic state, five frame 1 exon states, two intron states, seven frame 1 exon states and one intergenic state. Track 0 represents the emerging consensus gene prediction, where the consecutive segments will be labeled with exon, intron, or intergenic states. (See Supplementary Material for details on the various state alphabets.)

**4.2 Dynamic Bayes networks, state alphabets and probability constraints**

Dynamic Bayes networks (DBNs) (Murphy, 2002) are probabilistic graphical models that encode dependencies among a set of random variables evolving along a single dimension, conventionally called “time”. DBNs generalize Hidden Markov Models (HMMs (Rabiner, 1989)) by allowing complex probabilistic dependencies among the variables at a given time, and between variables at consecutive times. In our DBN model, observed sequences of variables represent the sources of evidence, and another, unobserved sequence represents the hidden
computing the posterior probability of a (possibly partial) consensus gene structure. This again can be computed in time linear to the number of segments by dynamic programming using a straightforward variant of the forward–backward algorithm for HMMs.

The inference algorithms implemented in Evigan rule out impermissible consensus sequences. For example, it is required that the consensus state for each segment be supported by at least one evidence source for the segment or that it extends from the state for the previous segment; it is also required that only the state transitions consistent with a signal type be allowed: for example, only exon-to-intron transition is allowed at a proposed donor position.

4.4 Quality scores and evaluation

Evigan’s source quality score is calculated from the conditional probability table associated with the source, averaging the entries where the source variable and the consensus variable have consistent values. (See Supplementary Material for more details.) The quality score is a surrogate for source’s performance that does not need a curated training set.

Gene prediction accuracy was evaluated using the Eval software package (Keibler and Brent, 2003) by sensitivity and specificity on exon, transcript and gene levels. (See Supplementary Material for more details.)

ACKNOWLEDGEMENTS

We wish to thank Axel Bernal for providing CRAIG predictions on the ENCODE regions, and Jonathan Allen for evidence sources and Combiner predictions on A.thaliana. We would also like to thank Anat Caspi and Zhongqiang Chen for helpful suggestions for improving the manuscript. This work was supported in part through the Health Research Formula Fund from the Commonwealth of Pennsylvania, NSF ITR grant EIA 0205456 and grants from the NIH. Implementation was supported under the auspices of a Bioinformatics Resource Center contract from NIAID. DSR is an Ellison Medical Foundation Senior Scholar in Global Infectious Diseases.

Conflict of Interest: none declared.

REFERENCES


