

## Gene expression

Using mRNAs lengths to accurately predict the alternatively spliced gene products in *Caenorhabditis elegans*

Ritesh Agrawal and Gary D. Stormo\*

Department of Genetics, Washington University School of Medicine, 660 S. Euclid, Campus Box 8232, St. Louis, MO 63110, USA

Received on November 3, 2005; revised on February 18, 2006; accepted on February 27, 2006

Advance Access publication April 4, 2006

Associate Editor: Martin Bishop

## ABSTRACT

**Motivation:** Computational gene prediction methods are an important component of whole genome analyses. While *ab initio* gene finders have demonstrated major improvements in accuracy, the most reliable methods are evidence-based gene predictors. These algorithms can rely on several different sources of evidence including predictions from multiple *ab initio* gene finders, matches to known proteins, sequence conservation and partial cDNAs to predict the final product. Despite the success of these algorithms, prediction of complete gene structures, especially for alternatively spliced products, remains a difficult task.

**Results:** LOCUS (Length Optimized Characterization of Unknown Spliceforms) is a new evidence-based gene finding algorithm which integrates a length-constraint into a dynamic programming-based framework for prediction of gene products. On a *Caenorhabditis elegans* test set of alternatively spliced internal exons, its performance exceeds that of current *ab initio* gene finders and in most cases can accurately predict the correct form of all the alternative products. As the length information used by the algorithm can be obtained in a high-throughput fashion, we propose that integration of such information into a gene-prediction pipeline is feasible and doing so may improve our ability to fully characterize the complete set of mRNAs for a genome.

**Availability:** LOCUS is available from <http://ural.wustl.edu/software.html>

**Contact:** stormo@genetics.wustl.edu

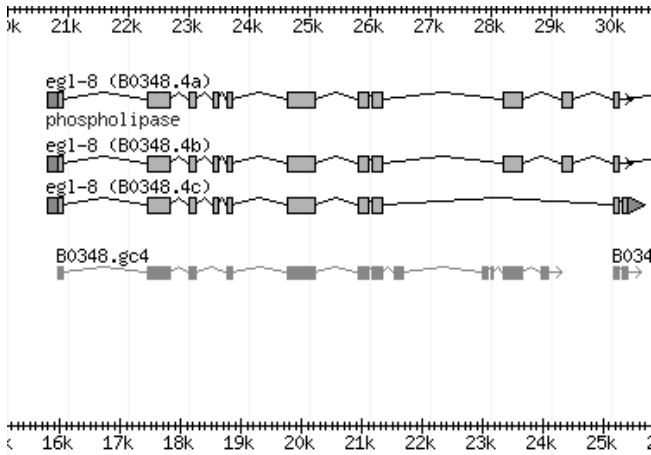
## INTRODUCTION

Determining the complete genomic structure for the entire set of an organism's transcripts remains far from solved even for well-studied organisms like *Caenorhabditis elegans*. Advances in *ab initio* gene structure prediction that make use of comparative genomics have produced reasonable gene-level prediction rates of up to 66% for constitutive spliceforms for compact genomes (Brent and Guigo, 2004) but need much improvement from the 15–20% accuracy on more complex genomes (Parra *et al.*, 2003; Flicek *et al.*, 2003). While evidence-based gene prediction (Hubbard *et al.*, 2002) provides nearly 80% accuracy on complex genomes at the exon level (Guigo and Reese, 2005), gene-level prediction is still quite low at ~40% (Guigo and Reese, 2005), perhaps owing to their reliance on EST evidence which rarely spans the full length of the gene (Suzuki *et al.*, 2002).

While clearly improvements are needed to reliably identify even a single product for each gene, obtaining the complete proteome requires knowledge of alternative splicing events as well. With experimental evidence for alternative splicing in over 80% of human genes (Kampa *et al.*, 2004), methods that predict all the products for each gene are critical. As most gene finders only provide single optimal solutions, most of the discovery of alternative splicing has relied on the use of ESTs (Brent and Guigo, 2004) which may underestimate alternative splicing because of their incomplete coverage and lack of information regarding combinations of exons that are utilized. While there is some suggestion that alternative splice products can be identified using sub-optimal solutions from general gene finders (Cawley and Pachter, 2003; Snyder and Stormo, 1995), the presence of multiple splice forms may confound the prediction of either spliceform (Fig. 1). Given the success of comparative genomics in identifying constitutive products, new approaches have begun to be developed to identify conserved alternative splices (Ohler *et al.*, 2005; Sorek *et al.*, 2004). The potential of these approaches will, however, be limited by the fact that alternative splicing is frequently species-specific (Pan *et al.*, 2005).

Obtaining an organism's entire proteome may therefore require organism-specific search methods. Here we report a new algorithm, Length Optimized Characterization of Unknown Spliceforms (LOCUS), which tests the utility of prior knowledge of spliceform length, as can be obtained from high-throughput reverse transcription–polymerase chain reaction (RT–PCR) of regions of interest, on splice site prediction. Since many PCRs can be carried out in parallel and (when combined with a reverse transcriptase reaction) PCR can be used to amplify any transcribed region between a defined set of primers, this approach could provide a highly automated scheme towards characterizing the full proteome. The *C.elegans* ORFeome project has initiated such a study to identify all of the proteins encoded in that genome (Reboul *et al.*, 2003; Lamesch *et al.*, 2004). Using primers from the predicted first and last exons of each gene, they have amplified and cloned a large number of products whose lengths can be estimated from gels. End sequencing validated the correctness of the cloned products and at least one of the splice junctions for each product. But only a fraction of the products were completely sequenced to determine their entire set of alternatively spliced variants. The LOCUS program can greatly facilitate such a project because the length information can be obtained directly via gel electrophoresis which avoids the need for time-consuming cloning, transformation, selection and

\*To whom correspondence should be addressed.



**Fig. 1.** Splice product prediction difficulty in regions of alternative splicing. The *C.elegans* B0348.4 gene undergoes alternative splicing as a result of two exon skipping events. (Splicing differences between the first two isoforms occur downstream of the displayed region.) Actual gene structures (top three structures) and Genefinder predicted structure (bottom structure) are shown.

colony picking steps used in the ORFeome protocol. This length information leads to specific predictions of the gene structure which can be tested directly with spliceform-specific primers and PCR instead of requiring complete gene sequencing. Because a large majority of the predictions are correct (see Results), only a fraction of the genes will need to be sequenced to determine their exact structure. An outline of this strategy is given (Fig. 2).

## METHODS

### Training

We downloaded wormpep version 135 containing all curated worm cDNAs from wormbase (Chen *et al.*, 2005). Using BLASTN (Altschul *et al.*, 1997), the cDNAs were searched against the wormbase 135 build of the *C.elegans* genome. True GT splice donors and AG splice acceptors were identified as GT and AG sequences nearest to the splice boundaries as determined by BLAST. All other AG and GT dinucleotides within the sequence set made up the background set. Sequences occurring as a result of alternative splicing were identified by wormbase names containing multiple ending letters (e.g. B0280.12a and B0280.12b). Examples of alternative splicing containing an alternative splice flanked by shared sequence both upstream and downstream of the alternative splice were used for testing the program's predictions. AG and GT dinucleotides within these alternatively spliced regions were not included in the training.

From the training set, we obtained trinucleotide frequencies that were over-represented (as previously described in Solovyev *et al.*, 1994) in a 110 bp region around true acceptor AG sites (-80 to +30) and in an 80 bp region around true GT donor sites (-30 to +50) relative to their counterparts in the background set. The 'over-representation index' of each trinucleotide starting at each position in the window was calculated as the log probability of its frequency in the true set versus the background set using the following equation:

$$ORI_{t,i} = \ln \left( \frac{f_{t,i}}{p_{t,i}} \right),$$

where  $t$  represents each of the 64 possible trinucleotides,  $i$  represents all possible starting indices within the search windows and  $f$  and  $p$  represent the frequencies of the trinucleotides in the true and background sets,

respectively. The range of trinucleotide ORI at each position in these windows (Fig. 3) shows that some of the most over-represented trinucleotides begin at position -4 (TTT) and -3 (TTC) from the acceptor site and at positions +2 (AAG) and +3 (AGT) from the donor site, consistent with previous reports (Kent and Zahler, 2000).

### SITE SCORING

All potential acceptor and donor sites, which include all AG and GT dinucleotides, were scored as the sum of the log probabilities for each of their constituent trinucleotides at the positions of greatest information in the true set as follows:

$$\text{Raw Score}_k = \sum_i ORI_{t,i}$$

Here  $k$  represents the position of the potential splice site in the sequence, where we score trinucleotides ( $t$ ) at positions  $i = (-7, -6, -5, -4, -3, -2, +2)$  relative to putative AG acceptor sites and positions  $i = (-3, -2, +1, +2, +3, +4, +5)$  relative to putative GT donors. The raw score for a given acceptor or donor site is converted into a log-odds 'Score' for selecting each particular site. This is calculated as the log of the ratio for observing that score in the true acceptor or donor sites compared with the background sequences. To obtain a sufficient representation for each raw score, they were binned into discrete score values (i.e. all splice sites scoring between 0 and 1 are binned and therefore have the same log-odds score). Specifically, the Score for each potential splice site is

$$\text{Score}_k = \ln \left( \frac{f_{\text{Raw Score}_k}}{p_{\text{Raw Score}_k}} \right),$$

where  $f_{\text{Raw Score}_k}$  represents the number of times that score is seen in the positive set and  $p_{\text{Raw Score}_k}$  represents the number of times the score is seen in the background set.

### EXON SCORING

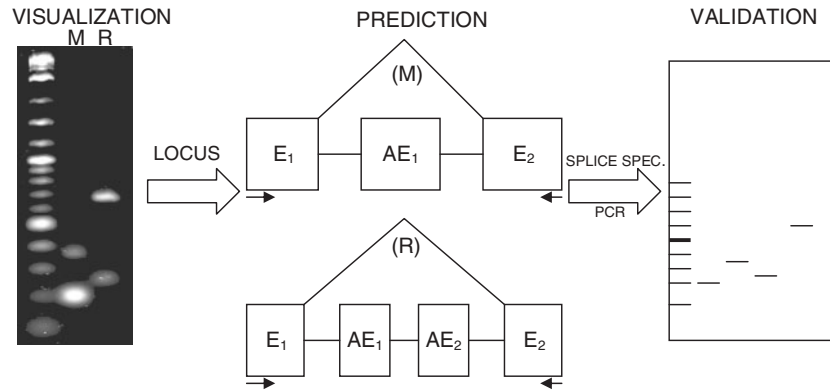
We use a very simple exon scoring scheme where each putative exon is scored based on its length and the sum of the acceptor and donor site scores. Only exons containing an open reading frame (ORF) are considered. An exon that goes from acceptor site  $j$  to donor site  $i$  would get

$$\text{Exon\_Score}_{j,i} = \text{Score}_j + \text{Score}_i + \ln \left( \frac{\text{Length}_{j,i}}{30} \right),$$

where  $\text{Length}_{j,i}$  is the length of the putative exon. The length correction value was chosen as 30 based on performance on the training set, although the results were fairly insensitive to its exact choice. Intuitively, 30 is about the length of an ORF expected by chance, so ones shorter than that should be penalized to prevent solutions with many very small exons. Introns do not contribute to the score directly, but must exceed a minimum length of 35 bases, a slightly conservative value given a *C.elegans* reported minimum intron length of 42 nt (Deutsch and Long, 1999).

### DYNAMIC PROGRAMMING ALGORITHM

Figure 4 outlines the general problem to be solved. The band on the gel indicates that there is a spliced product that includes the two primer sites,  $P1$  and  $P2$ , and is of length  $L$ . In the case of alternatively spliced genes there are two or more products of



**Fig. 2.** Computational approach to splice product characterization. The left lane of the gel contains size markers. The middle (M) and right (R) lanes are two different genes that have been amplified by RT-PCR and each shows two bands, indicating alternatively spliced products for those genes. The program LOCUS uses the genome sequence from the region and the size of each product (from the gel) to predict the overall genomic structure (under ‘prediction’) with alternative exons (AE) and regions removed through splicing (area under the caret) indicated. From the predicted product, splice-specific PCR primers can be designed to test if the prediction is correct. The bands in the ‘validation’ gel would demonstrate that each prediction is correct without actually sequencing any of the products.

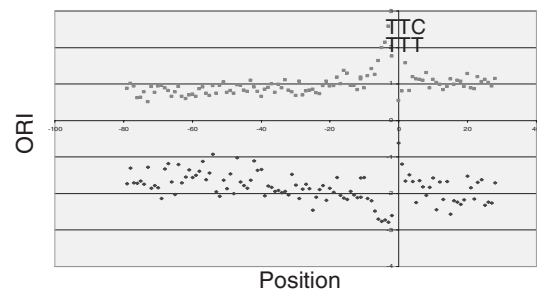
different lengths (Fig. 2), but each is treated separately. The sites labeled  $D_1$  to  $D_N$  are all of the potential donor sites between the primers (we consider all GT dinucleotides, although one might impose a score threshold). The sites labeled  $A_1$  to  $A_M$  are all of the potential acceptor sites (we consider all AG dinucleotides). The objective is to find the highest scoring combination of exons that have a combined length of  $L$ , and we allow a tolerance of  $\pm 15\%$  because the product length can only be estimated from the gel. We obtain solutions by employing a simple dynamic programming algorithm in both the forward and backward directions, and combining partial solutions that, in combination, have the appropriate length. Several possible solutions can be obtained and ranked by their Sum\_Score (see below).

From Figure 4, the only solution for donor site  $D_1$  is an exon from  $P1$  to  $D_1$  and its score is just  $\text{Score}_{D_1}$  plus the length contribution for that (partial) exon. The only solution for the acceptor site  $A_1$  is the exon ending at  $D_1$  followed by the intron from  $D_1$  to  $A_1$  (provided it is at least 35 bases downstream), and the score for that solution is the sum of the  $D_1$  score plus  $\text{Score}_{A_1}$ . For the donor site  $D_2$  there are two possible solutions, one of which is an exon from  $P1$  to  $D_2$  (provided there is an ORF between them), and its score would be calculated as for the  $D_1$  solution. The other possible solution is an exon from  $A_1$  to  $D_2$  combined with the previous solution for an intron ending at  $A_1$ . The solution with the highest score would be the best solution for the donor site  $D_2$ . For donors further downstream there are more potential solutions but we can obtain the highest scoring solution for each donor site by the following recursion:

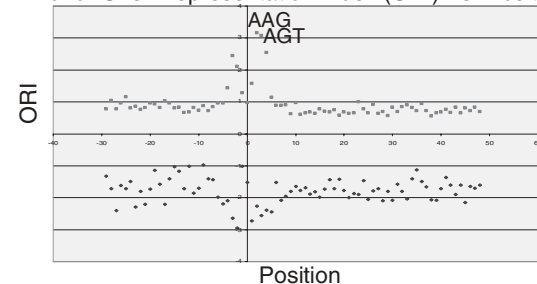
$$D_i = \text{Score}_{D_i} + \max_{j=0}^{\text{pos}_i < \text{pos}_j} \left[ A_j + \ln \left( \frac{\text{length}_{j,i}}{30} \right) \right], \quad (1)$$

with the additional constraint that there must be an ORF for the exon from  $A_j$  to  $D_i$  and that ORF must be in frame with the best solution for  $A_j$ . The special case of  $A_0$  corresponds to the primer site  $P1$  and  $A_0 = 0$ . This recursion ensures that the value  $D_i$  is the score of the best solution ending with donor site  $D_i$  that maintains an ORF for all of the exons between  $P1$  and  $D_i$ . The score is the value of  $D_i$  [Equation (1)], the Length is the length of that best solution

### A Acceptor Over-Representation Index (ORI) Vs. Position



### B Donor Over-Representation Index (ORI) Vs. Position



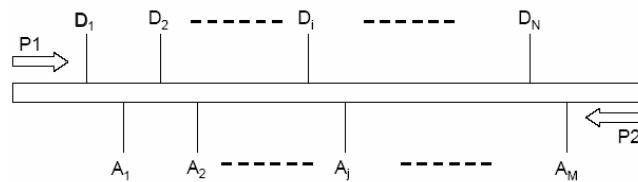
**Fig. 3.** Over-representation index values for trinucleotides surrounding acceptor sites (A) and donor sites (B). Vertical axis represents the over-representation index (ORI) for a particular trinucleotide. Horizontal axis represents the start position of the trinucleotide within a 110 bp (A) or 80 bp (B) window. The range [maximum (top) and minimum (bottom)] of ORI is shown at each position. The most over-represented trinucleotides for the acceptor and the donor sites are indicated.

(the sum of the length of all the exons for that solution) and  $\text{Max}_j$  is the upstream acceptor site associated with that best solution.

An analogous recursion is used to determine the best solution for each choice of acceptor:

$$A_j = \text{Score}_{A_j} + \max_{i=1}^{\text{pos}_j + 35 \leq \text{pos}_i} [D_i] \quad (2)$$

## DYNAMIC PROGRAMMING ALGORITHM



**Fig. 4.** Problem outline. A cartoon depiction of a hypothetical genomic region of interest (middle line) flanked by a set of PCR primers (P1 and P2). Given all possible donors ( $D_1$  to  $D_N$ ) and acceptors ( $A_1$  to  $A_M$ ) within this region, LOCUS attempts to identify the set of splice sites that give rise to a product of the appropriate size and with the maximum score (see text).

which ensures that introns are at least 35 bases long because the position of the donor site  $D_i$  ( $pos_i$ ) must be at least 35 bases upstream of the acceptor site  $A_j$ . Together the recursions of Equations (1) and (2) allow for the determination of the highest scoring solution for each possible donor and acceptor site. But to apply the constraint that the complete solution must have length  $L \pm 15\%$  we apply the same recursions in the backward direction, from primer site P2 toward P1. These solutions are stored in analogous arrays  $\mathbf{D}^R$  and  $\mathbf{A}^R$ . Combined solutions are scored as follows:

$$\text{Sum\_Score}_{i,j} = \mathbf{D}(\text{Score}_i) + \mathbf{A}^R(\text{Score}_j)$$

subject to the constraints that the entire predicted product is an ORF, that  $pos_i + 35 \leq pos_j$  and that

$$0.85L \leq \mathbf{D}(\text{Length}_i) + \mathbf{A}^R(\text{Length}_j) \leq 1.15L$$

These constraints ensure that we only consider solutions in which the acceptor site is at least 35 bases downstream of the donor and that the total length of the predicted product is within 15% of the estimated length. It is important to note that we are only considering solutions that maintain an ORF between the primers, thereby ignoring those cases where alternative splicing leads to early translational stops. All such solutions are obtained and ranked by their Sum\_Score and the top 10 unique solutions are reported. Although there could be a large number of combinations to be evaluated, the imposition of those constraints reduces the allowed solutions to a reasonable number. Note that the recursions and the arrays have to be computed only once for any region, and then the Sum\_Scores can be determined for each separate product from the gel, each with their own length, using the same arrays.

## RESULTS

As a test of the accuracy of the predictions made by LOCUS we identified 151 regions of alternative splicing of internal exons in *C.elegans* consisting of skipped exons, 5' splice site changes, 3' splice site changes and combinations thereof. We chose primer sites within the constitutive exons on either side of the alternatively spliced region and provide the program with the genomic sequence of that region and the lengths of both products (as calculated from the reported gene structure), one at a time. LOCUS reports the top 10 ranked solutions for each length and we compare those with the known products. A predicted product is considered correct only if it matches the known product exactly, such that the predicted and

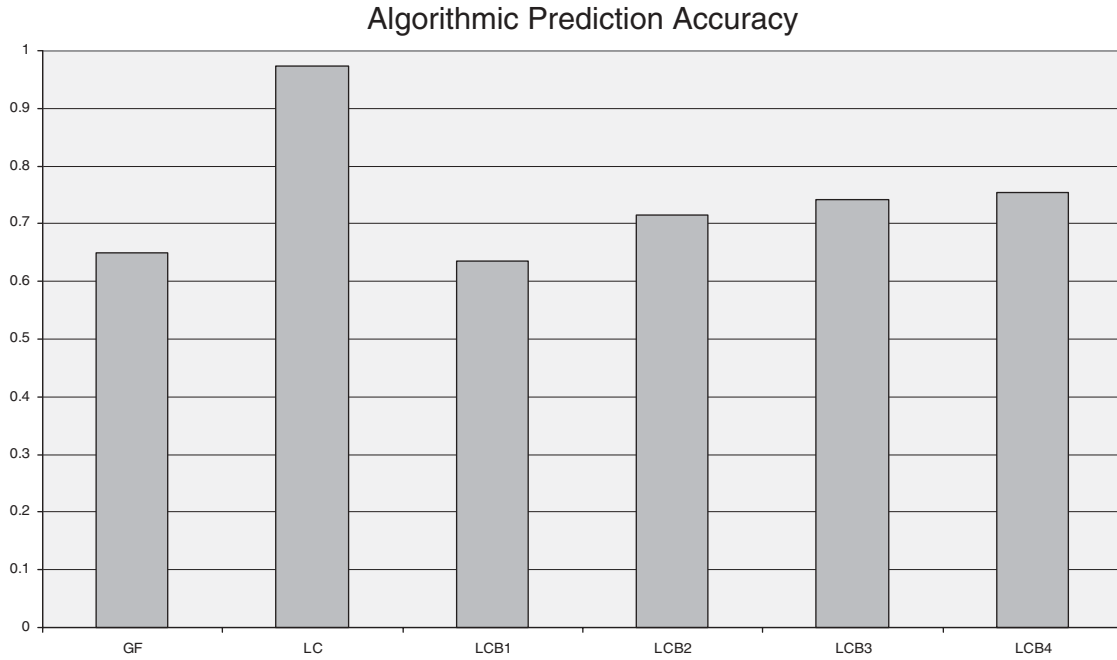
known protein products are identical. We compared our results with those of the well-established *ab initio* gene prediction program, Genefinder (Wilson, C., Hillier, L. and Green, P., unpublished data), available from the wormbase database (Chen *et al.*, 2005). Genefinder only attempts to predict a single product for each region, but we are unaware of other programs that attempt to predict all the alternative products.

For the 151 test regions, Genefinder correctly predicted one of the two correct products 65% (98/151) of the time. If we apply the same criteria for LOCUS, using only the highest ranked solution, it matches at least one of the correct products 97% (147/151) of the time. Using only the highest ranked solution for each length, LOCUS correctly predicts both of the alternative products 64% (96/151) of the time, similar to the accuracy of only one solution for Genefinder (see Fig. 5 for all comparisons). So, using only the highest ranked solutions for each length, LOCUS is able to obtain 81% (243/302) sensitivity on all of the gene products tested. This demonstrates that the added information of the length of the gene product can aid in its correct prediction. In those cases when the true solution is not returned as the optimal solution it is often found among one of the 10 suboptimal solutions. In fact, when considering the second through fourth best solutions, we find that our ability to correctly predict both spliceforms increases considerably to a range of 72–75% for these suboptimal solutions, respectively (LCB2-LCB4, Fig. 5). Among the entire set of 10 highest ranking solutions, both products are correctly predicted for 80% of the genes, and 89% of the complete list of gene products are accurately predicted. LOCUS works well for most *C.elegans* spliceforms as correct predictions were made from genomic regions up to 7 kb in length and regions where the true structure encompasses multiple alternative splicing events (Fig. 6). In fact, somewhat surprisingly, the prediction accuracy is nearly constant between short (<1 kb) medium (1–5 kb) and long (>5 kb) regions (data not shown).

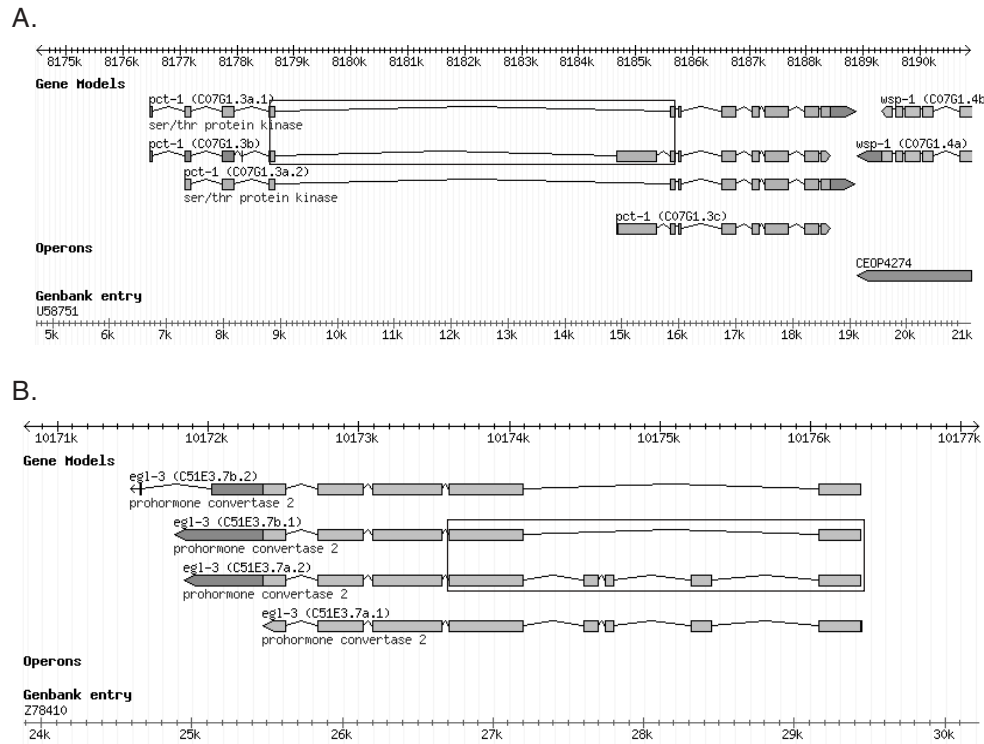
## DISCUSSION

Recent years have seen an explosion in the amount of DNA sequence data and in research efforts to translate that raw data into its functional transcriptional counterpart. A major focus of that research has been directed towards prediction of organismal proteomes through identification of gene structures. Recent analyses of the current predictive power suggest a great deal of progress in the ability to identify a gene's 'parts list', the set of exons of which it is composed (Brent and Guigo, 2004; Guigo and Reese, 2005). A full understanding of the proteome will require an enhanced ability to connect these parts into their respective strings and to handle regions of alternative splicing where parts may be interchanged (Brent and Guigo, 2004; Guigo and Reese, 2005; Lee and Wang, 2005).

We have described a new algorithm which, given the scalability of PCR and new techniques for automated band mapping (Zerr and Henikoff, 2005), can be added to a high-throughput gene prediction pipeline and aide in the resolution of the complete proteome. One can imagine using 3' and 5' EST libraries to identify regions of the genome that correspond to all of the ends of the gene products. Using primers based on those end sequences, RT-PCR could be used to amplify the complete set of gene products between them, using gels to get length estimates for each product. LOCUS could



**Fig. 5.** Prediction accuracy. Fraction of total dataset for which a correct prediction was made for a single spliceform by Genefinder (GF) or LOCUS (LC). Fraction of total dataset for which correct predictions were made for both spliceforms and returned as optimal (LCB1) or 2–4 ranking suboptimal solutions (LCB2–LCB4).



**Fig. 6.** Examples of alternative splicing correctly predicted from the test set. (A) Example of alternative splicing within 7 kb of genomic sequence and (B) example of alternative splicing which requires several different alternative splice sites. Regions under consideration are boxed.

then be used to predict, with high accuracy, the exact gene products for each observed mRNA, including those with alternative splicing. Even without EST libraries, one can imagine using the most confidently predicted exons, using information such as BLAST hits to protein databases, to design primers for a high-throughput RT-PCR detection of alternatively spliced genes, whose lengths can be used by LOCUS to predict the complete set of gene products.

Given the length information for each spliced product in a region, LOCUS is nearly always (97% of the time) able to predict the correct splicing pattern of at least one of the products. But it gets both products correct only ~64% of the time (using only the highest ranked predictions), leaving ample room for improvement. Tests using known examples of alternative splicing from *Drosophila melanogaster* and *Homo sapiens* show a decreased prediction accuracy relative to that reported for *C.elegans* likely due to known species-specific differences in RNA splicing signals (Salzberg, 1997). Adaptation for use in these organisms will therefore likely require re-training of algorithm parameters. More complex scoring schemes for exons and introns, such as those used in generalized hidden Markov model methods (Mathe *et al.*, 2002; Brent and Guigo, 2004) would probably increase the accuracy. In addition, there is often other evidence available to help locate the exonic regions between the primer sites. Many species now have extensive EST datasets that may cover most of the exons, and combined with product length information could lead to very accurate predictions of which exons are included in each gene product. Similarities to protein databases and conservation between species are also types of information that should help identify the complete set of exons for a genomic region. The use of splicing enhancer (Cartegni *et al.*, 2003) and suppressor (Wang *et al.*, 2004) sequences could also help to identify alternative exon sequences. The results we have presented demonstrate that the product length information used by LOCUS, which greatly reduces the set of possible products that a genomic region might produce, significantly increases the accuracy of gene product prediction. Each additional type of information about probable exon sequences should make the rankings of the true products even more reliable.

*Conflict of Interest:* none declared.

## REFERENCES

- Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Brent,M.R. and Guigo,R. (2004) Recent advances in gene structure prediction. *Curr. Opin. Struct. Biol.*, **14**, 264–272.
- Cartegni,L. *et al.* (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
- Cawley,S.L. and Pachter,L. (2003) HMM sampling and applications to gene finding and alternative splicing. *Bioinformatics*, **19**, ii36–ii41.
- Chen,N. *et al.* (2005) Wormbase: a comprehensive data resource for *Caenorhabditis* biology and genomics. *Nucleic Acids Res.*, **33**, D383–D389.
- Deutsch,M. and Long,M. (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.
- Flicek,P. *et al.* (2003) Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map. *Genome Res.*, **13**, 46–54.
- Guigo,R. and Reese,M.G. (2005) EGASP: collaboration through competition to find human genes. *Nat. Methods*, **2**, 575–577.
- Hubbard,T. *et al.* (2002) The Ensembl Genome Database Project. *Nucleic Acids Res.*, **30**, 38–41.
- Itoh,H. *et al.* (2004) Computational comparative analysis of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA*, **10**, 1005–1018.
- Kampa,D. *et al.* (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, **14**, 331–342.
- Kent,J.W. and Zahler,A.M. (2000) Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*–*C. elegans* genomic alignment. *Genome Res.*, **10**, 1115–1125.
- Lamesch,P. *et al.* (2004) *C. elegans* ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.*, **14**, 2064–2069.
- Lee,C. and Wang,Q. (2005) Bioinformatics analysis of alternative splicing. *Brief. Bioinformatics*, **6**, 23–33.
- Mathe,C. *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.
- Ohler,U. *et al.* (2005) Recognition of unknown conserved alternatively splice exons. *Plos Comput. Biol.*, **1**, 113–122.
- Pan,Q. *et al.* (2005) Alternative splicing of conserved exons if frequently species-specific in human and mouse. *Trends Genet.*, **21**, 73–77.
- Parra,G. *et al.* (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.
- Reboul,J. *et al.* (2003) *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.*, **34**, 35–41.
- Salzberg,S.L. (1997) A method for identifying splice sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.
- Snyder,E.E. and Stormo,G.D. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, **248**, 1–18.
- Solovyev,V.V. *et al.* (1994) Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. *Nucleic Acids Res.*, **22**, 5156–5163.
- Sorek,R. *et al.* (2004) A non-EST-based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.
- Suzuki,Y. *et al.* (2002) DBTSS: database of human transcriptional start sites and full-length cDNAs. *Nucleic Acids Res.*, **30**, 328–331.
- Wang,Z. *et al.* (2004) Systematic identification and analysis of exonic splicing suppressors. *Cell*, **119**, 831–845.
- Zerr,T. and Henikoff,S. (2005) Automated band mapping in electrophoretic gel images using background information. *Nucleic Acids Res.*, **33**, 2806–2812.