GenePC and ASPIC integrate gene predictions with expressed sequence alignments to predict alternative transcripts

T. S. ALIOTO* and R. GUIGÓ

Center for Genomic Regulation
c/ Dr. Aiguader, 88
08003, Barcelona, Spain
**E-mail: tyler.alioto@crg.es*

E. PICARDI and G. PESOLE

Dipartimento di Biochimica e Biologia Molecolare
University of Bari
and Istituto Tecnologie Biomediche del C.N.R. (sede di Bari)
via Orabona 4
70126 Bari, Italy

We have developed a generic framework for combining introns from genomicly aligned expressed–sequence–tag clusters with a set of exon predictions to produce alternative transcript predictions. Our current implementation uses ASPIC to generate alternative transcripts from EST mappings. Introns from ASPIC and a set of gene predictions from many diverse gene prediction programs are given to the gene prediction combiner GenePC which then generates alternative consensus splice forms. We evaluated our method on the ENCODE regions of the human genome. In general we see a marked improvement in transcript-level sensitivity due to the fact that more than one transcript per gene may now be predicted. GenePC, which alone is highly specific at the transcript level, balances the lower specificity of ASPIC.

Keywords: Alternative Splicing; Gene Prediction; Combiner.

1. Introduction

1.1. *Gene Prediction Combiners*

The computational integration of multiple gene predictions into one unified or consensus prediction has increasingly become the first-pass annotation of choice for newly sequenced genomes. Examples of such programs are EuGène,[1] GAZE,[2] GLEAN, JIGSAW,[3] Genomix[4] and even the Ensembl pipeline.[5] These programs are popular because they perform a task that human annotators usually begin with — examining multiple predictions and looking for consensus. Of course, human annotators have the ability to look at all sources of information including, but not limited to, CpG islands, putative promoter elements, protein domain integrity, protein family conservation, genome integrity, and literature references. Neverthe-

less, the primary tools continue to be a combination of *ab initio*, homology-based, and expressed sequence tag (EST)/cDNA-based gene prediction. In this area, gene combiners are already demonstrating their utility. In the future perhaps combiners will be able to incorporate even more diverse sources of information such as those mentioned above.

1.2. *Diversity Combiners and Wise Crowds*

Because gene prediction combiners take advantage of the diversity of input, they can be considered a class of diversity combiner.[6] Diversity combiners used in the wireless telecommunication arena are effective at canceling out noise when receiving multiple independent inputs. The most effective of these are called maximal-ratio combiners; these combiners use a weighted sum of their inputs where the weight is based on their signal–to–noise ratio. We reasoned that a similar approach might help in integrating multiple gene predictions. Additionally, we tried to enforce the following four principles of collective wisdom:[7]

(1) *Diversity of opinion* — Input gene predictors should represent a diversity of algorithms.

(2) *Independence* — Input gene predictors should not be influenced by each others' predictions.

(3) *Decentralization* — Predictors should be able to specialize and draw upon local private knowledge.

(4) *Aggregation* — Some mechanism must exist to combine the predictions.

Our gene prediction combiner, GenePC, constructs a consensus prediction using the principle that exons co-predicted by diverse input prediction algorithms are more likely to be real. By explicitly correcting for correlated input, GenePC corrects for lack of diversity and interdependence. This procedure also takes the guesswork out of choosing which gene predictions to combine.

Decentralization is normally not an issue. However, when one organization is in charge of annotating a genome, it tends to run a set of gene prediction programs using the same set of aligned expressed sequence or proteins or, if using comparative genomics, they tend to use the same genomes for comparison. Likewise, when a gene annotation assessment project is undertaken (GASP,[8] EGASP,[9] nGASP), the allowed input and allowed training sets are usually fixed. Of course, this allows a fair assessment of individual gene predictors; however, in principle it should have a negative effect on the performance of gene prediction combiners.

The method we present here is an aggregation method that explicitly corrects for lack of diversity and independence among all pairs of inputs. Additionally, we correct for over-aggregation. A problem for consensus building is the special case in which more than one solution is correct. This is particularly relevant to gene prediction, where a gene can code for more than one alternative transcript. Our method

reintroduces high-quality transcript evidence during the last step of aggregation to guide the division of a single consensus into multiple likely consensus transcripts.

1.3. *Alternative Splicing Prediction*

The annotation of multiple alternative splice forms has been a stumbling block for both the standard class of gene predictors and for combiners. Programs that are successful at alternative transcript prediction are in general those that use expressed sequence tag (EST) or mRNA evidence (a notable exception is the latest version of Augustus,[10] which is able to output multiple transcripts per loci in the absence of expressed sequence evidence.) Many such programs, including ASPIC, rely only on transcript mapping and often output incomplete transcripts. EuGène is one of the first gene predictors/combiners to output full-length alternative transcripts using often incomplete transcript evidence. Our approach is similar in spirit to that of EuGène, but in implementation there are many differences. The most obvious of which is our use of ASPIC,[11] a software tool for Alternative Splicing PredICtion, to drive the construction of alternative gene models by GenePC.

2. Methods

Before we begin we must mention one caveat. Although more and more programs are predicting UTR segments of genes by using mRNA and EST evidence, we have decided to restrict ourselves to only the the coding segments (CDS) of genes as these are the most reliably predicted features of protein-coding genes. As such we will naturally not be able to annotate the variation in splicing of UTR sequence which actually represents a significant proportion of alternative splicing.

2.1. *Approach*

Our aggregation method combines ASPIC CDS predictions with GenePC, an extended version of GeneID[12] which combines gene predictions at the level of exact exons based on diversity and performance. Such a combination allows multiple consensus transcripts to be predicted per locus, depending on the number of mutually incompatible sets of introns inferred from the EST alignments.

Our method involves three steps. In the first step, we run GenePC using all available gene predictions on a genome or subset of the genome. GenePC scores every uniquely predicted exon as a function of the set of gene predictors that predicted it. We calculate an Exon Score S which is given by

$$S = \alpha \sum_{i=1}^{n} S_{i_{norm}} + \beta \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \left(1 - \frac{SNet_{ij} + SPet_{ij}}{2}\right) + \gamma \sum_{i=1}^{n} (x \cdot SNet_{ia} + y \cdot SPet_{ia}) \tag{1}$$

where $S_{i_{norm}}$ is the Z-score normalized exon score provided by program i and $SNet$ and $SPet$ are the exon-level sensitivity and specificity, respectively, of program i against program j or program i against the annotation a. The weighting parameters α, β and γ can be optimized for a particular combination of gene predictions and genomes. The coefficients x and y are adjusted to give more weight to sensitivity or specificity of a gene prediction program. We generally set $x = 0.25$ and $y = 0.75$. The parameters for First, Internal, Terminal and Single exon types are optimized independently using a training set of confirmed genes within the prediction region or on a separate set. If no annotation is available, performance values can either be estimated from the literature or γ may be set to zero. A graphical overview of this procedure is shown in Figure 1.

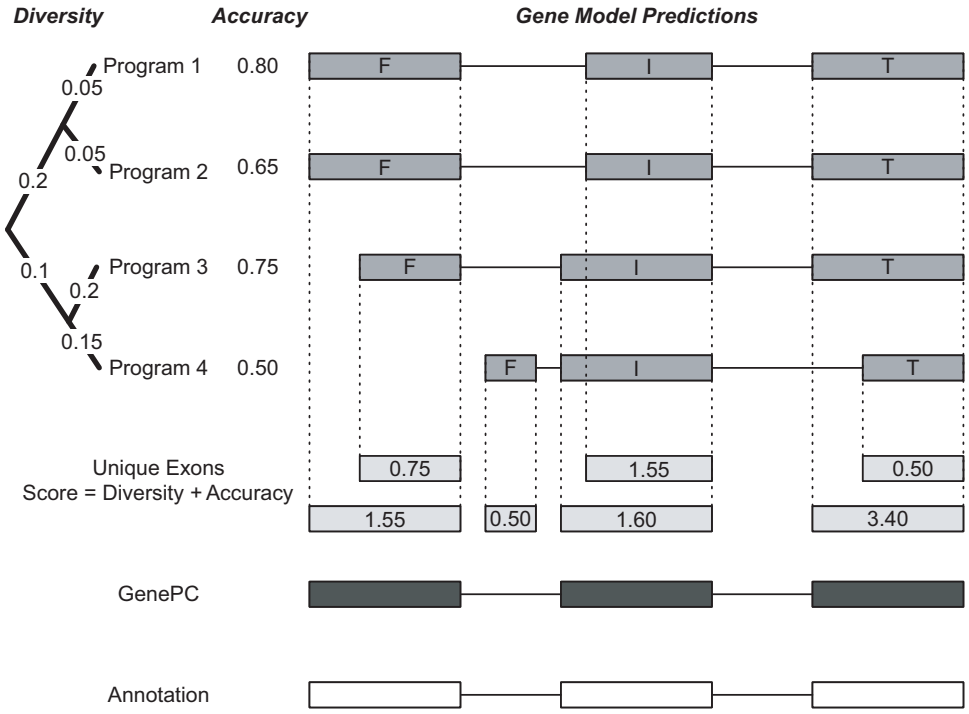

Fig. 1. Schematic of GenePC prediction aggregation using Diversity (distance between programs) and Performance (Accuracy against the annotation). Self-reported exon scores are not used ($\alpha = 0$.) In practice, we use a pairwise distance matrix rather than the phylogram shown here.

In the second step, we run the alternative splicing prediction program ASPIC, which is freely available at http://t.caspur.it/ASPIC/, with a set of ESTs aligned to the genome. We perform an initial alignment of ESTs to the genome with the program GMAP.[13] The ESTs are then clustered based on the sharing of at least one splice site (a fuzzy clustering methods that we developed allows for small shifts

in intron position due to spliced alignment error) and then each cluster is given to ASPIC. ASPIC then performs a multiple genome-EST alignment optimized for producing a set of transcripts with a minimal number of exons.[11,14] The longest open reading frame is then determined for each transcript output by ASPIC and overlapping CDS spans are assigned to different compatibility groups or bins. The number of bins is determined by the locus with the largest number of alternative transcripts. For loci with less than the maximum number of alternative transcripts, CDS spans are reassigned to empty bins so that we maximally cover the genome with the available evidence. Finally, for use in the next step, we convert the transcript data into sets of introns in GFF format for input into GenePC.

In the final step, we use the dynamic programming algorithm `genamic`[15] as implemented in GeneID to chain together the highest scoring set of exons from each intron compatibility set generated in the previous step. Only the introns from ASPIC are provided to GenePC so that exons not predicted by any gene predictor will not contribute to the final gene models. This feature is made possible in version 1.3 of GeneID, freely available at http://genome.imim.es/software/geneid/, which has been modified to use introns as evidence but will not output a gene model if no exons are present. In the last step, redundant transcripts are removed from the combined predictions of each run of GeneID.

3. Results

3.1. *Training and Test Sets*

We evaluated our method on the ENCODE[16] regions using 14 sets of gene predictions submitted to the EGASP competition.[9] The EGASP competition solicited gene predictions on the 44 ENCODE regions encompassing 1% of the human genome. Training was allowed on 13 of these regions. We trained GenePC on these same 13 regions, obtaining performance (prediction vs. annotation) and distance (prediction vs. prediction) values separately for First, Internal, Terminal and Single exons. Self-reported exon scores were not used since this information was not provided for all inputs. Input predictions included all non-combiner programs that predicted complete gene structures, thus excluding Spida and JIGSAW.

The set of ESTs used by the HAVANA team to annotate the ENCODE regions were downloaded from Genbank and aligned to the genome with GMAP. They were clustered based on shared splice junctions and given as input to ASPIC. The resulting transcripts (some of which are full-length but many of which are partial) were distributed into the minimal number of internally compatible sets of transcripts. In this case, the locus with the largest number of alternative incompatible transcripts was 12. Therefore, we were required to run GeneID twelve times on the ENCODE regions, each time using the same set of GenePC-scored non-redundant exons and a different set of ASPIC introns.

We evaluated ASPIC-GenePC predictions and all input predictions against the March 2007 Gencode[17] human ENCODE gene annotation release using the pro-

Table 1. Transcript-level performance.

Category	Program	SNt	SPt	$(SNt + SPt)/2$
Ab initio	GeneID	0.03	0.04	0.04
	Exonhunter	0.06	0.03	0.05
	Genemark	0.05	0.04	0.05
	AugustusAbinitio	0.13	0.16	0.15
Multi-genome	Sgp2	0.05	0.04	0.05
	Twinscan	0.10	0.08	0.09
	AugustusDual	0.15	0.17	0.16
	N-SCAN	0.21	0.35	0.28
EST/Protein and Pipelines	Ensembl	0.25	0.24	0.25
	AugustusAny	0.27	0.34	0.31
	AugustusEst	0.27	0.36	0.32
	Fgenesh	0.43	0.38	0.41
	Exogean	0.51	0.43	0.47
	Pairagon-N-SCAN	0.51	0.46	0.49
	ASPIC CDS	0.63	0.37	0.50
Combiners	JIGSAW	0.42	0.63	0.53
	GenePC	0.41	**0.65**	0.53
	ASPIC-GenePC	**0.65**	0.51	**0.58**

Note: All predictions except for ASPIC CDS, GenePC and ASPIC-GenePC predictions correspond to EGASP submissions downloaded from the UCSC genome browser.

gram `evaluation.pl` (Eduardo Eyras, personal communication) which takes into account alternative transcripts in both the annotation and predictions. Only full-length known and putative Gencode genes were used for evaluation.

3.2. *Evaluation*

We evaluated ASPIC-GenePC using a large number of measures including sensitivity and specificity at the level of nucleotides, exons, transcripts and genes (see Tbl. 1). JIGSAW and GenePC outperform all other predictors in nearly every category (results not shown). We believe the most significant result is the accuracy at the transcript level. GenePC and JIGSAW already perform quite well in predicting exact transcripts. Further gains in transcript-level sensitivity are realized when EST evidence is used to guide alternative best predictions. We observed an average 5% gain in accuracy. This is a combination of a 24% gain in sensitivity and 14% loss in specificity. ASPIC-GenePC predicts 466 of 719 Gencode-annotated transcripts in 381 of 397 genes. This is in contrast with the 295 transcripts predicted correctly by GenePC alone.

Table 2. Evaluation at gene, transcript, exon, intron and nucleotide levels.

	JIGSAW	GenePC	ASPIC CDS	ASPIC-GenePC
SNg	**0.96**	0.94	0.84	**0.96**
SPg	0.82	**0.83**	0.69	0.78
SNt	0.42	0.41	0.63	**0.65**
SPt	0.63	**0.65**	0.37	0.51
SNe	0.88	0.88	0.87	**0.93**
SPe	0.84	**0.86**	0.75	0.80
SNet	0.93	**0.96**	0.84	0.94
SPet	0.93	**0.97**	0.95	0.92
SNi	0.88	0.86	0.89	**0.95**
SPi	0.85	0.86	**0.87**	0.82
SNit	0.93	**0.95**	0.85	**0.95**
SPit	0.94	**0.96**	**0.96**	0.93
SNn	0.95	0.94	0.81	**0.96**
SPn	0.86	**0.87**	0.85	0.85
SNnt	0.95	**0.97**	0.86	0.96
SPnt	0.96	**0.98**	0.97	0.95

Note: SN indicates sensitivity. SP indicates specificity. Gene (g), transcript(t), exon(e), intron(i) and nucleotide(n) levels were assessed. For each exon, intron and nucleotide, we evaluated against the set of projected unique predicted features to the projected unique annotated features. We also compared the features of a predicted transcript to the features of the annotated transcript with the highest overlap (best hit), denoted by the suffix *t*. Gene sensitivity is defined as any exonic overlap.

For comparison, we also show the performance of JIGSAW, ASPIC, GenePC and ASPIC-GenePC at the exon and nucleotide levels in Table 2. It is interesting to note that with the addition of ASPIC introns, sensitivity increases and specificity decreases in almost every level compared with GenePC alone or with JIGSAW.

4. Discussion

The phenomenon of alternative splicing continues to represent a challenge to *ab initio* gene prediction. Two approaches to the problem have generally been taken in the past: (a) output of suboptimal predictions and (b) output of multiple transcripts supported by mutually incompatible evidence. We have taken the latter approach as it is straightforward and is backed by strong evidence of transcription. However, this does not preclude the use of a combiner to gather a consensus on how many alternative transcripts exist for a particular gene and output that number of transcripts, without relying on another source of evidence. We have not moved in this direction yet, considering most *ab initio* predictors have yet to produce more than one transcript per locus.

Nevertheless, using our straightforward EST-driven approach, we were struck by the 24% gain in transcript sensitivity over GenePC alone that is conferred by ASPIC-guided alternative transcript prediction. Part of this gain was offset by a

decrease in specificity. However, the number of additional genes predicted was not dramatic (28 new genes, 9 of which are annotated), meaning that the majority of the 171 new transcripts predicted are alternative transcripts of known loci. While all of them by definition have some level of EST support, many may actually be partial or non-coding transcripts predicted by ASPIC. Since the corresponding transcripts have been filtered out of the annotation, these predictions demonstrate themselves as false positives.

The success of our aggregation method is likely due to the high sensitivity of ASPIC in predicting reliable transcripts. ASPIC in fact has been mainly designed to perform a multiple alignment of ESTs to a genomic sequence based on the combined analysis of all available expressed sequence tags. Moreover, it refines the exon-intron boundaries by an appropriate dynamic programming module and generates the most likely transcripts using a new algorithm based on graph theory. However, where ASPIC fails to find exons or genes, GenePC can fill in the gaps. Furthermore, where ASPIC predicts transcripts with exons not present in the combined set of input gene predictions to GenePC, no transcripts are predicted.

The general framework we have outlined — dividing transcript evidence into compatible sets and providing them as intron evidence for exon structure assembly — should be extensible to nearly any gene prediction method that directly takes EST/cDNA evidence or external constraints such as introns, as we have used here. While substantial gains in accuracy from the incorporation of expression evidence would be expected (and have been demonstrated) for *de novo* or *ab initio* gene predictors, the enhanced performance of a combiner was not anticipated. In theory the expressed sequence evidence is already used by at least several input methods, making its re-introduction redundant. However, combiners such as GenePC are afflicted with the same problem faced by traditional gene finders: they output only the optimal gene model. Thus, the enhancement we observe is acheived by essentially unflattening the consensus gene model determined by GenePC into multiple models that maximally represent the transcript evidence. Incompatibilities among ESTs are not ignored but are, if they are of sufficient quality, utilized to their full diagnostic potential.

In the future, we anticipate an increased focus on alternative transcript prediction. While we currently have to rely on the crutch of EST and cDNA sequences, advances in our understanding of the regulation of alternative splicing may allow the prediction of alternative transcripts based on knowledge of the critical factors present and active, for example, in particular cell-types or at particular developmental stages or in response to particular external stimuli. We suggest that our iterative approach in conjunction with an evidence combiner would be equally applicable in this context.

Acknowledgments

This work was supported by grants from the European Commission FP6 Programme (BioSapiens Network of Excellence) to RG and TA and from the Ministero Università e Ricerca, Italy (FIRB project "Laboratorio Italiano di Bioinformatica") to GP.

References

1. S. Foissac and T. Schiex, *BMC bioinformatics* **6**, p. 25 (2005), 10.1186/1471-2105-6-25.
2. K. Howe, T. Chothia and R. Durbin, *Genome research* **12**, 1418(Sep 2002), 10.1101/gr.149502.
3. J. Allen and S. Salzberg, *Bioinformatics (Oxford, England)* **21**, 3596(Sep 2005), 10.1093/bioinformatics/bti609.
4. A. Coghlan and R. Durbin, *Bioinformatics (Oxford, England)* **23**, 1468(Jun 2007), 10.1093/bioinformatics/btm133.
5. T. Hubbard, B. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal and E. Birney, *Nucleic acids research* **35**, 610(Jan 2007), 10.1093/nar/gkl996.
6. D. Brennan, *Proc. IRE* **47**(Jun 1959).
7. J. Surowiecki, *The Wisdom of Crowds* (Anchor, August 2005).
8. M. Reese, G. Hartzell, N. Harris, U. Ohler, J. Abril and S. Lewis, *Genome research* **10**, 483(Apr 2000).
9. R. Guigó, P. Flicek, J. Abril, A. Reymond, J. Lagarde, F. Denoeud, S. Antonarakis, M. Ashburner, V. Bajic, E. Birney, R. Castelo, E. Eyras, C. Ucla, T. Gingeras, J. Harrow, T. Hubbard, S. Lewis and M. Reese, *Genome biology* **7 Suppl 1**, 2 (2006), 10.1186/gb-2006-7-s1-s2.
10. M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack and B. Morgenstern, *Nucleic acids research* **34**, W435(Jul 2006), 10.1093/nar/gkl200.
11. P. Bonizzoni, R. Rizzi and G. Pesole, *BMC bioinformatics* **6**, p. 244 (2005), 10.1186/1471-2105-6-244.
12. G. Parra, E. Blanco and R. Guigó, *Genome research* **10**, 511(Apr 2000).
13. T. Wu and C. Watanabe, *Bioinformatics (Oxford, England)* **21**, 1859(May 2005), 10.1093/bioinformatics/bti310.
14. T. Castrignanò, R. Rizzi, I. Talamo, P. D'Onorio, A. Anselmo, P. Bonizzoni and G. Pesole, *Nucleic acids research* **34**, W440(Jul 2006), 10.1093/nar/gkl324.
15. R. Guigó, *Journal of computational biology : a journal of computational molecular cell biology* **5**, 681 (1998).
16. *Nature* **447**, 799(Jun 2007).
17. J. Harrow, F. Denoeud, A. Frankish, A. Reymond, C.-K. Chen, J. Chrast, J. Lagarde, J. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S. Antonarakis and R. Guigo, *Genome biology* **7 Suppl 1**, 4 (2006), 10.1186/gb-2006-7-s1-s4.