A Computational Gene Prediction Pipeline Under Statistic Algorithms Adopted for Human

Xin LI, MCGD

INTRODUCTION

Automatic gene finding approaches are of practical interest in studying the human genome whose raw nucleotide sequences and transcripts (e.g. cDNAs, ESTs) are abundant but far from completely annotated, as well as cognitively meaningful in the sense that only the models being able to predict phenomena accurately are sound and functional. While computational gene predictions in prokaryotes have already achieved around 95% accuracy (Schiex et al., 2003), automatic gene identification in eukaryotes remains challenging (Guigó et al., 2000, Para et al., 2003) thanks to complicated genomic features, i.e. low gene density, exon-intron structure and alternative splicing. This problem promotes various automatic prediction methods in two categories (Rouze et al., 2002), *ab initio* (or intrinsic) and homology-based (or extrinsic) prediction programs.

Ab initio approaches rely on the interior composition features of gene structures, such as codon usage, G+C content; while homology-based ones refer to the similarity between nucleotide sequences and available transcripts (i.e. cDNAs, ESTs, proteins), or molecular evolutionary conservation among relevant but different species, human and mouse, for instance. GENSCAN (Burge et al., 1997) was one of the most commonly used and best *ab inito* programs in predicting high eukaryotic, especially human genes. It adopted a Hidden Markov Model of fifth order for exons and suggested the genome structure modeling mainly in coding regions. Recently another HHM-based method called AUGUSTUS (Stanke et al., 2003) was reported to outperform GENSCAN in dealing with long DNA sequences and gene structure prediction. Interestingly, AUGUSTUS employed a Hidden Markov Model with the order of four, albeit behaved more accurately than fifth-order GENSCAN. However, its advantages, especially on whole-gene structure predictions relating to exon, intron and splice sites censoring, might attribute to AUGUSTUS' introduction of a more detailed intron submodel. The potential for prediction power of AUGUSTUS remains impressive given the around-40% accuracy in human gene structure prediction.

On the other hand, homology-based prediction methods include those taking advantage of so-called spliced alignments and those based on comparative genomics among relevant species. The first class utilized local alignments to identify genes and solve human gene structures. As a matter of fact, both two largest human annotated gene databases, Ensembl and NCBI employed some versions of BLAST programs to interpret their data collection, BLAST and MegaBLAST, respectively (Durbin et al., 1997; Birney et al., 2002; Allen et al., 2004). Given the abundance and coverage of human ESTs, cDNAs and proteins, these methods often have higher specificity than *ab initio* approaches. Notably, human ESTs could provide important information to alternative splicing due to the database coverage (Bailey et al., 1998).

The other similarity-based category of prediction approaches appears as the completion of mouse genome sequencing. The estimates that 99% of mouse genes have human homologues legitimate these cross-species comparison efforts (Mouse Genome Sequencing Consortium, 2002). The basic assumption premising these methods is that the coding regions in genomes should be more conserved than non-coding regions. One of the earliest programs ROSETTA (Batzoglou et al., 2000) was directly derived from comparison of human and mouse ortholog. The predictor TWINSCAN comprised of GENSCAN module and BLASTN module between human and mouse genome; similarly, SGP2 was actually a combination of GENEID and TBLASTX (Parra et al., 2003; Flicek, 2003). Both of them outperformed any single predictor. Dewey et al.(2004) further made a three-species comparative prediction for novel human genes in human, mouse and rat, which used a pair-HMM based cross-species prediction program SLAM. The prediction accuracy they achieved was extremely high but at the cost of sensitivity.

With versatile approaches based on various models or algorithms, it is observed that most accurate results were produced if the most unrelated approaches were combined in usage (Dewey et al., 2004). The notion of complementing multiple predictors in a statistic manner rather than any direct overlapping comes from Allen et al. (2004). Allen et al. constructed several Combiner programs based on different algorithms in *Arabidopsis thaliana* and evaluated the sensitivity and specificity, with very promising improvements out of single predictors, i.e. GENSCAN, TWINSCAN.

Here I propose a combination program of human version, incorporating a series of single gene predictors i.e. AUGUSTUS, SGP2. The first-round outputs from these individual gene predictors are then combined in a statistical manner to give out scores in dynamic programming matrices, thus lead to optimal gene predictions.

METHODS

Input Gene Predictors

Multiple gene prediction programs, including both *ab initio*, homology-based ones and splice site predictors, are chosen as input predictors previous to pipeline assembly. They are selected as separately as possible to give out most accurate predictions. (Table 1).

Table 1, Input Gene Predictors

Predictors	Sources	Algorithms	Notes
AUGUSTUS	human genome	HMM (4th order)	
GENSCAN	human genome	HMM (5th order)	
RescueNet*	human genome	Self-organizing Map	
Protein match	human cDNA db	BLAST	Altenative Splicing***
EST match	human EST db	BLAST	
Splice Site Prediction**	human genome	EDA**	Intron Model
TWINSCAN	human, mouse	GENESCAN+BLASTN	
SGP2	human,mouse	GENEID+TBLASTX	
SLAM	human,mouse,rat	pair HMM	

* Mathony et al., 2004; ** Saeys et al., 2004&2003; EDA: Estimation of Distribution Algorithm

*** Foissac et al., 2004.

Pipeline Assembly Algorithms (Allen et al. , 2003)

1) Gene atomic sites and sequence states.

The concept of gene atomic sites (Guigó et al. 1992) is a simplified gene structure model compared to a Hidden Markov Model (Burge et al. 1997; Stanke et al., 2003). According to Guigó et al. (1992) and Allen et al. (2003), four categories of atomic sites are considered: start codons, stop codons, acceptor splice sites (ending of an intron) and donor splice sites (beginning of an intron). Therefore, any base could be in one of the five states: start codon, stop codon, acceptor splice site, donor splice site and coding. In Allen’s model, they consider the possibility of one base to be in any of the five states a vector, while various dimensions in the vector represent outputs from different input predictors, named the evidence.

To paraphrase Allen et al. (2003), for DNA sequences defined linearly by atomic sites, the status of any interval sequences between two atomic sites could be determined by the states of both atomic sites. As every base in one strand of DNA could be “ Yes” or “No” for a specific atomic site state, a score of 1 (for yes) or 0 (for no) might be assigned to a state-decision matrix. A total of 10 biological meaningful states are get out of the 2^5=32 possible combinations, representing 10 possible states a position in DNA sequences could be. A complete DNA sequence label table was described in the report by Allen et al. (2003).

2) Probability computation and model construction

In brief, any input DNA sequence is corresponding to multiple gene structure models, while each gene model is represented by a probability, which is the product of all the probabilities of each base.

Assume l_j represents the 10 possible states for each position;

e_j is the evidence for sequence Ij to be a lj state, and comprises of five possibility vectors;

Then the probability is P (l_1, $l_2 \dots l_x | e_1$, $e_2 \dots e_x$)

Allen et al. (2003) then simplifies the computation of P (l_1, $l_2 \dots l_x | e_1$, $e_2 \dots e_x$) by assuming that the state l_j is only dependent of itself sequence Ij and the two adjacent ones I_{j-1} and I_{j+1}, thus P (l_1, $l_2 \dots l_x | e_1$, $e_2 \dots e_x$)=$\prod_{j=1}^{x} P(lj / e_{j-1}, ej, e_{j+1})$. Assume evidence vectors are Vs, Va, Vd, Vi, Ve; $P(lj / e_{j-1}, ej, e_{j+1})$ is the product of the five probabilities corresponding to evidence vectors. Meanwhile, several decision trees based on OC1 are built up to calculate each of the evidence vector probability. Finally, scores in dynamic programming matrices are calculated based on weighting of various resources.

Validation

1) RT-PCR sequencing would be taken to verify the accuracy of prediction.

Reverse-transcribed PCR includes: a) Primers design according to predicted exons; b) raw human RNA preparation; c) RT-PCR running for amplifying putative genes; d) PCR products sequencing compare and align with original pipeline results.

2) Alignment confirmation.

While RT-PCR sequencing could provide a direct confirmation to predicted exons, the predicted introns might be indirectly verified by successful alignments of PCR products and predicted genes containing introns. On the other hand, predicted introns might also be aligned through BLAST with human EST database, NCBI RefSeq and cDNA databases to confirm their existence.

DISCUSSION

One of the advantages of this pipeline prediction lays on the multiple data sources it base on. As is shown in Table1, various types of gene finding programs, either a homology-based approach, a *de novo* prediction model or some specific gene structure modeling for special signal sensors (i.e. splice sites) are incorporated to maximize the vitality of gene structure modeling and prediction. The RescuNet method (Mathony et al., 2004) based on relavie synonymous codon usage and Self-organizing Map neural network algorithm could be a complementary to HMM-based approaches AUGUSTUS and GENSCAN. Some previous results from multi-predictors combo show that the combination often outperforms than single ones. (Tech et al., 2003; Foissac et al. 2004)

Another merit of this strategy is the statistic assembly of multiple data resource inputs. Allen et al. 2003 has already demonstrated that in *Arabidopsis thaliana* genome, the “ statistic combier” gave out better results than those linear algorithms in which the weight for each data source was relatively subjective. Several decision trees based on OC1 will be constructed in order to compute the model probabilities. Every single probability is an average of multiple decision trees, which enhances the accuracy of the prediction.

Last but not the least, the problem of alternative splicing is very difficult in the scenario of *ab initio* gene finding programs. Often suboptimal gene models might be considered alternative splicing products (Brent et al., 2004). Meanwhile, this problem could be better tackled in similarity-based methods, given the comparison between cDNAs and nucleotide sequences or ESTs. Here in the computational prediction pipeline, the inputs from protein or EST matches and from dual/tri-genomic comparisons might be able to attack this problem.

REFERENCES

Roderic Guigó, S. K., Neil Drake and Temple Smith (1992). Prediction of Gene Structure. J Mol Biol *226*, 141-157.

Searls, D. S. and D. B. (1994). Gene structure prediction by linguistic methods. Genomics *23*, 540-551.

Birney, E. and Durbin, R. (1997). Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. ISMB *5*, 56–64.

Karlin, C. B. and S. (1997). Prediction of Complete Gene Structures in Human Genomic DNA. J Mol Biol *268*, 78-94.

Bailey, L. C., Searls,D.B. and Overton,G.C. (1998). Analysis of ESTdriven gene annotation in human genomic sequence. Genome Research *8*, 362-376.

Guigó, R., Agarwal, P., Abril, J.F., Burset, M., and Fickett, J.W. (2000). Gene prediction accuracy in large DNA sequences. Genome Research *10,* 1631-1642.

Serafim Batzoglou, L. P., Jill P. Mesirov, Bonnie Berger, and Eric S. Lander (2000). Human and Mouse Gene Structure: Comparative Analysis and Application to Exon Prediction. Genome

Research *10*, 950-958.

Birney, E., Clamp, M., and Hubbard, T. (2002). Databases and tools for browsing genomes. Annu Rev Genomics Hum Genet *3*, 293–310.

Catherine Mathe, M.-F. S., Thomas Schiex and Pierre Rouze (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research *30*, 4103-4117.

Consortium, M. G. S. (2002). Initial sequencing and comparative analysis of the mouse genome. Nature *420*, 520-562.

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

Kevin L. Howe, T. C., and Richard Durbin (2002). GAZE: A Generic Framework for the Integration of Gene-Prediction Data by Dynamic Programming. Genome Research *12*, 1418-1427.

Genis Parra, P. A., Josep F. Abril, Thomas Wiehe,James W. Fickett and Roderic Guigó (2003). Comparative Gene Prediction in Human and Mouse. Genome Research *13*, 108-117.

Jennifer L. Ashurst, J. E. C. (2003). GENE ANNOTATION: PREDICTION AND TESTING. Annu Rev Genomics Hum Genet *4*, 69-88.

Lavner, D. K. and Y. (2003). Gene Prediction by Spectral Rotation Measure: A New Method for Identifying Protein-Coding Regions. Genome Research *13*, 1930-1937.

Leila Taher , O. R., Saurabh Garg,Alexander Sczyrba, Michael Brudno, Serafim Batzoglou and Burkhard Morgenstern (2003). AGenDA: homology-based gene prediction. Bioinformatics *19*, 1575–1577.

Merkl, M. T. and R. (2003). YACOP: Enhanced gene prediction obtained by a combination of existing methods. In Silico Biology *3*.

Paul Flicek, E. K., Ping Hu,Ian Korf, and Michael R. Brent (2003). Leveraging the Mouse Genome for Gene Prediction in Human: From Whole-Genome Shotgun Reads to a Global Synteny Map. Genome Research *13*, 46-54.

Sohrab P. Shah, G. P. M., Alan K. Mackworth ,Sanja Rogic and B. F. Francis Ouellette (2003). GeneComber: combining outputs of gene prediction programs for improved results.

Bioinformatics *19*, 1296-1297.

Thomas Schiex, J. r. m. G., Annick Moisan and Yannick de Oliveira (2003). FrameD: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nucleic Acids Research *31*, 3738–3741.

Waack, M. S. and S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics *19*, ii215-ii225.

Yvan Saeys, S. D., Dirk Aeyels , Yves Van de Peer and Pierre Rouze (2003). Fast feature selection using a simple estimation of distribution algorithm: a case study on splice site prediction. Bioinformatics *19*, ii179–ii188.

Colin Dewey, J. Q. W., Simon Cawley, Marina Alexandersson,Richard Gibbs, and Lior Pachter (2004). Accurate Identification of Novel Human Genes Through Simultaneous Gene Prediction in Human, Mouse, and Rat. Genome Research *14*, 661–664.

Durbin, I. M. M. and R. (2004). Gene structure conservation aids similarity based gene prediction. Nucleic Acids Research *32*, 776-783.

Jonathan E. Allen, M. P., and Steven L. Salzberg (2004). Computational Gene Prediction Using Multiple Sources of Evidence. Genome Research *14*, 142-148.

Leila Taher, O. R., Saurabh Garg, Alexander Sczyrba and Burkhard Morgenstern (2004). AGenDA: gene prediction by cross-species sequence comparison. Nucleic Acids Research *32*, W305-W308.

Michael R Brent , R. G. (2004). Recent advances in gene structure prediction. Current Opinion in Structural Biology *14*, 264-272.

Schiex, S. F. and T. (2004). Integrating alternative splicing detection into gene prediction. BMC Bioinformatics *6*.

Shaun Mahony, J. O. M., Terry J Smith and Aaron Golden (2004). Gene prediction using the Self-Organizing Map: automatic generation of multiple gene models. BMC Bioinformatics *5*.

Yvan Saeys, S. D., Dirk Aeyels, Pierre Rouzé and Yves Van de Peer (2004). Feature selection for splice site prediction: A new method using EDA-based feature ranking. BMC Bioinformatics *5*, 64-75.