Minireview

# Gene prediction: compare and CONTRAST

Paul Flicek

Address: European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Email: flicek@ebi.ac.uk

Published: 20 December 2007

*Genome* ***Biology*** 2007, **8:**233 (doi:10.1186/gb-2007-8-12-233)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2007/8/12/233



## Abstract

CONTRAST, a new gene-prediction algorithm that uses sophisticated machine-learning techniques, has pushed *de novo* prediction accuracy to new heights, and has significantly closed the gap between *de novo* and evidence-based methods for human genome annotation.

Gene prediction is one of the most important and alluring problems in computational biology. Its importance comes from the inherent value of the set of protein-coding genes for other analysis. Its allure is based on the apparently simple rules that the transcriptional machinery uses: strong, easily recognizable signals within the genome such as open reading frames, consensus splice sites and nearly universal start and stop codon sequences. These signals are highly conserved, are relatively easy to model, and have been the focus of a number of algorithms trying to locate all the protein-coding genes in a genome using only the sequence of one or more genomes. This technique, so-called *de novo* prediction, does not use information about expressed sequences such as proteins or mRNAs.

In this month's issue of *Genome Biology*, Gross and colleagues [1] describe the gene-prediction program CONTRAST, the latest significant advance in *de novo* gene prediction. The program exploits patterns inherent in multiple sequence alignments while making few assumptions about evolutionary processes. Its accuracy is considerably higher than any other *de novo* prediction program and has significantly closed the gap between *de novo* and evidence-based methods for human genome annotation.

There have been two previous significant breakthroughs in *de novo* human gene prediction. The first was the identification and optimization of algorithms to effectively model the problem. The second was the use of an evolutionarily related genome sequence to reliably increase both the sensitivity and specificity of the predictions. Both advances are briefly discussed below (for more on the history of gene finding see [2]).

Algorithms based on a generalized hidden Markov model (GHMM) framework have been particularly successful for gene prediction. A GHMM can be used to describe the relationship between the components of a protein-coding gene (such as exons and splice sites) and the sequence of genomic DNA in which the gene is found. The best-known example of this method is the program GENSCAN [3], which in 1997 was shown to be dramatically more accurate than the previous state-of-the-art prediction programs. GENSCAN was easy to use, very fast, and predicted genes in the long sequences of genomic DNA that would characterize the human genome project. Although subsequently shown to predict only 10–15% of genes correctly on realistic genome-wide datasets [4,5], GENSCAN remains a popular bioinformatics tool. GENSCAN predictions continue to be a standard feature for every genome released on both the University of California Santa Cruz (UCSC) [6] and Ensembl [7] genome browsers.

In 2002, with the publication of the mouse genome sequence [8], human gene prediction formally entered the era of comparative genomics (see Figure 1 for a comparison of the programs). A number of programs were developed to exploit this new data source. In both human-mouse comparisons and across the tree of life, the most successful of these dedicated algorithms was TWINSCAN [9], a gene-prediction program that exploited the signature of evolution using a reimplementation and extension of the GENSCAN GHMM model. TWINSCAN's improved accuracy featured a dramatic reduction in false-positive predictions, while managing to predict about 25% of human protein-coding genes completely accurately [5,10]. TWINSCAN itself was then extended with a more expressive model of evolutionary

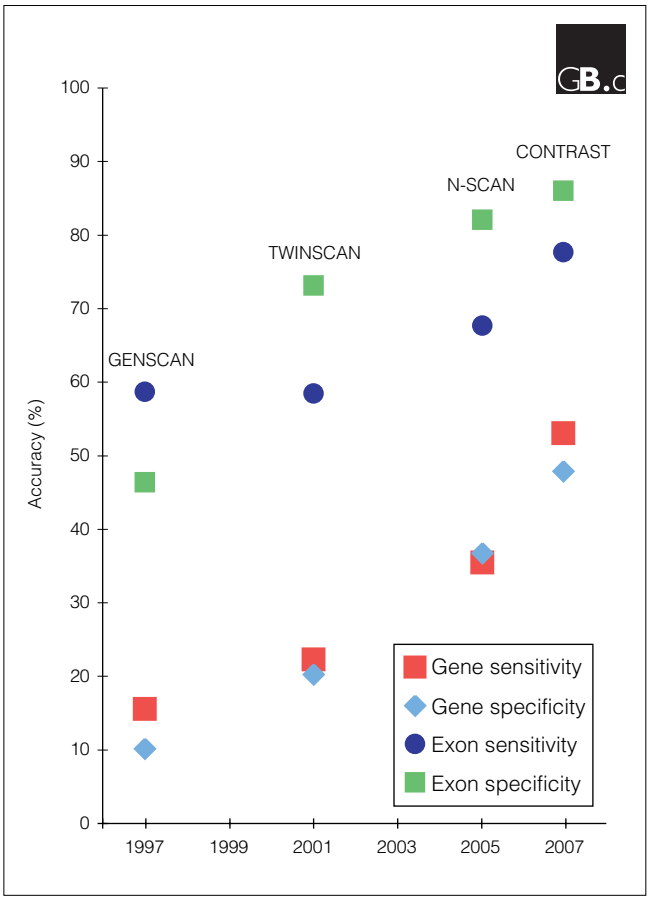


**Figure 1**
Increase in the accuracy of *de novo* gene prediction over time. The gene sensitivity and specificity and the exon sensitivity and specificity on the EGASP test set [5] are shown for several programs by year of initial publication. Included are GENSCAN (1997), TWINSCAN (2001), N-SCAN (2005) and CONTRAST (2007). Note the significant decrease in false positive predictions (as measured by the rise in TWINSCAN's exon specificity) with the inital use of evolutionarily related genome sequences. By comparison, the accuracy of the Ensembl evidence-based gene predictions used in the EGASP experiment at the gene level were 71.6% sensitivity and 67.3% specificity and 77.5% sensitivity and 82.7% specificity at the exon level.

conservation derived from a multiple sequence alignment of several complete genomes. This extension, known as N-SCAN, predicts approximately 35% of human genes correctly [5], but is no more accurate with a multiple sequence alignment than it is with the most informative pairwise genome alignment [10]. Thus, even though the N-SCAN model of evolutionary conservation is better than the one used by TWINSCAN, N-SCAN is not benefiting from the additional genome sequences used in the alignment.

At the same time as these advances in *de novo* gene prediction, evidence-based gene prediction was also progressing rapidly. The best evidence-based systems integrate data from sources such as mRNA and protein sequences to predict specific genes that are supported by a variety of expressed sequences [11,12]. These evidence-based gene sets are often used for other biological analyses such as [13].

CONTRAST is a dramatic advance on the previous state of the art [1]. Using the Consensus CDS (CCDS) [14] set as the gold standard, CONTRAST predicts nearly 60% of the genes correctly using only the human genome sequence and a multiple alignment with 11 so-called 'informant' genome sequences. This result is a stunning improvement on the previous state-of-the-art *de novo* gene-prediction algorithms both on the CCDS set and the gold standard manually annotated genes used for the ENCODE Genome Annotation Assessment Project (EGASP) [5] (Figure 1). Close examination of the EGASP results shows that CONTRAST compares very favorably with even the best evidence-based, expressed sequence prediction methods, especially for exon accuracy.

To achieve this, Gross *et al.* [1] did something unconventional in the gene-prediction field. They ignored what is known about evolutionary relationships and assumed that there must be additional information in the multiple sequence alignment even if they could not exactly say what sort of information was there. Doing this required a switch from generative models such as HMMs, which have been used by essentially all previous *de novo* prediction programs, to discriminative models such as support vector machines and conditional random fields. A support vector machine (SVM) is an example of the machine-learning technique called 'supervised learning', in which the algorithm is able to classify new items based on rules it has discovered from a correctly labeled training set. A conditional random field (CRF) can be used to classify sequential data and is applicable to many of the same problems as an HMM. CONTRAST uses both SVM and CRF techniques for different parts of the gene-prediction problem. The SVMs are used for coding region boundary detection (splice sites, start and stop codons), whereas a CRF is used to model the gene structure (that is, how all the pieces fit together). Readers interested in more information about these machine-learning techniques may like to start with a recent biology-based primer on SVMs [15].

There are limits for biological understanding with these new techniques. A process of evolution resulted in the extant sequences that we see, and understanding this process would be immensely valuable. Generative models such as HMMs attempt to explicitly describe the evolutionary process by generating the multiple sequence alignment of an evolutionarily conserved exon. For example, a phylogenetic HMM may use separate models of molecular evolution for the first, second and third positions of each codon [16]. Unlike phylogenetic HMMs, discriminative machine-learning techniques such as those used by Gross *et al.* [1] do not model the complexities of the evolutionary process, but they are able to find the subtle differences in the alignments associated with real genes from other, very similar alignments in the genome.

In the current implementation, training CONTRAST requires the target genome to be at least reasonably well annotated at the start. It is not yet clear how well it will perform when annotating a genome with no high-quality training data, although unpublished results from the CONTRAST team demonstrate substantial accuracy with only a few thousand training genes [17]. The situation where no training data is available could be simulated, at least for the case of human and mouse, by using one of the genomes as a well-annotated training set and the other to test predictive accuracy. As the most accurate *de novo* prediction program, CONTRAST will help complete the protein-coding gene set in well annotated genomes such as human and mouse, and may be vital for accurate annotation of complex genomes with informative sequence alignments to related species, but without significant expression data. Nevertheless, annotating a genome without the sequence of a closely related species is likely to remain a challenge.

CONTRAST is not the first program to apply these types of machine-learning technique to the problem of computational gene prediction. Bernal *et al*. [18] recently introduced a gene-prediction program called CRAIG, which does not use any sequence alignments, but does use a semi-Markov CRF. CRAIG shows notable improvement over a large selection of other non-alignment-based programs. However, it performed less well than HMM-based multi-genome prediction programs such as N-SCAN [5,18]. DeCaprio *et al*. [19] developed Conrad, a comparative gene-prediction program that also uses semi-Markov CRFs. Conrad shows striking improvements on fungal genomes compared with other leading prediction programs, but its current implementation makes its application to large mammalian genomes computationally prohibitive [19].

It is still the case that the best full-length gene predictions are done by mapping expressed sequences to the genome assembly. CONTRAST finds the initial and terminal exons of a gene relatively difficult to predict and this somewhat limits exact gene prediction. However, Gross *et al*. [1] show convincingly that there is complex information in the multiple sequence alignment of mammalian genomes and that this information can be exploited to create far more accurate gene predictions than those produced by the best HMM-based algorithms. The performance of CONTRAST suggests that the dominance of HMM-based programs in gene-prediction might be waning. Without doubt, further advances in machine-learning methods for large-scale biological analysis will help us integrate and understand complex biological data. A challenge for computational biologists is to transform the language of SVMs and discriminative learning techniques into biological models that will help us understand the complex processes of evolution that have created the extant species that we are now so busily sequencing.

The development of CONTRAST is a welcome result to those of us who believed that there must be additional information that could be used for gene prediction in multiple sequence alignments. Brent [2] recently suggested a number of possible reasons why multiple sequence alignments had failed to increase the accuracy of comparative gene prediction. These included sequence quality, alignment methods, and lack of splice site and exon conservation in the mammalian lineage. It looks as though his final reason - that designers of *de novo* gene prediction algorithms had not yet been clever enough to come up with a solution - might well have been the right one.

## Acknowledgements

I would like to thank Sean Eddy and Steve Searle for helpful discussions.

## References

1. Gross SS, Do CB, Sirota M, Batzoglou S: **CONTRAST: A discriminative, phylogeny-free approach to multiple informant *de novo* gene prediction.** *Genome Biol* 2007, **8:**R269.
2. Brent MR: **Genome annotation past, present, and future: how to define an ORF at each locus.** *Genome Res* 2005, **15:**1777-1786.
3. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA.** *J Mol Biol* 1997, **268:**78-94.
4. Flicek P, Keibler E, Hu P, Korf I, Brent MR: **Leveraging the mouse genome for gene prediction in human: from whole-genome shotgun reads to a global synteny map.** *Genome Res* 2003, **13:** 46-54.
5. Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, *et al.*: **EGASP: the human ENCODE Genome Annotation Assessment Project.** *Genome Biol* 2006, **7 Suppl 1:**S2.
6. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, *et al.*: **The UCSC genome browser database: update 2007.** *Nucleic Acids Res* 2007, **35:**D668-D673.
7. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al.*: **Ensembl 2008.** *Nucleic Acids Res* 2007, doi:10.1093/nar/gkm988.
8. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:** 520-562.
9. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction.** *Bioinformatics* 2001, **17 Suppl 1:**S140-S148.
10. Gross SS, Brent MR: **Using multiple alignments to improve gene prediction.** *J Comput Biol* 2006, **13**:379-393.
11. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system.** *Genome Res* 2004, **14:**942-950.
12. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC known genes.** *Bioinformatics* 2006, **22:**1036-1046.
13. Goodstadt L, Ponting CP: **Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human.** *PLos Comput Biol* 2006, **2:**e133.
14. **Consensus CDS** [http://www.ncbi.nlm.nih.gov/CCDS]
15. Noble WS: **What is a support vector machine?** *Nat Biotechnol* 2006, **24:**1565-1567.
16. Siepel A, Haussler D: **Combining phylogenetic and hidden Markov models in biosequence analysis.** *J Comput Biol* 2004, **11:** 413-428.
17. **CONTRAST** [http://contra.stanford.edu/contrast]
18. Bernal A, Crammer K, Hatzigeorgiou A, Pereira F: **Global discriminative learning for higher-accuracy computational gene prediction.** *PLoS Comput Biol* 2007, **3:**e54.
19. Decaprio D, Vinson JP, Pearson MD, Montgomery P, Doherty M, Galagan JE: **Conrad: Gene prediction using conditional random fields.** *Genome Res* 2007, **17:**1389-1398.