Using GeneWise in the *Drosophila* Annotation Experiment

Ewan Birney[1] and Richard Durbin

Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

The GeneWise method for combining gene prediction and homology searches was applied to the 2.9-Mb region from *Drosophila melanogaster*. The results from the Genome Annotation Assessment Project (GASP) showed that GeneWise provided reasonably accurate gene predictions. Further investigation indicates that many of the incorrect gene predictions from GeneWise were due to transposons with valid protein-coding genes and the remaining cases are pseudogenes or possible annotation oversights.

The critical assessment of machine learning techniques is necessary to assess the effectiveness of computational methods. The critical assessment of protein structure prediction (CASP) has become a benchmark for protein structure assessment worldwide (Moult et al. 1999). We welcomed the opportunity offered by Reese and coworkers (2000) to independently assess the gene prediction methods available and provided one of the methods we developed, GeneWise, for this study.

The use of protein and EST similarity to help gene prediction is widespread, including methods such as Genie (Kulp et al. 1996) and GRAIL (Uberbacher et al. 1996). The GeneWise approach builds on the success of hidden Markov models (HMMs) for modeling both protein family information (Krogh et al. 1994; Eddy 1998) and gene predictions (Kulp et al. 1996; Burge and Karlin 1997; Krogh 1997). GeneWise is a HHM that is formed by the principled combination of two separate HMMs (E. Birney and R. Durbin, in prep.). GeneWise therefore can be thought of as considering every possible gene prediction in a genomic sequence and comparing each one to the protein profile–HMM. The best combined score of both the gene prediction and the protein profile–HMM is used to provide a simultaneous gene prediction and protein alignment.

To use GeneWise for gene prediction one needs a source of homology information. In this case, we used protein profile–HMMs from PFAM (Bateman et al. 2000). One of the major drawbacks to using GeneWise is the prohibitive computational cost of the method. This was solved in this case by using the halfwise methods, which prefilters the protein profile–HMM used in the comparison (see Methods). The results presented here were the completely automatic annotation from GeneWise without any manual intervention in the process.

RESULTS

A total of 165 gene predictions with 252 exons were made in the 2.9-Mb genomic segment. Of the 252 exons, 216 overlapped in some way with the std3 dataset of definite and possible predictions. This left 36 exons in 23 predictions outside of this set. A number of these (16) were profile HMMs of transposons or retroviral transposons. The remaining 20 exons were potential mispredictions or annotation mistakes. By manual examination of these cases we found four potential mispredictions by GeneWise, in each case a trailing exon in an otherwise correct gene prediction. Of the remaining 16 exons, 10 were clear annotation oversights, leaving 6 that were less clear cut, for example, pseudogenes might explain the presence of these hits. There were no predictions by GeneWise of completely wrong genes, in line with our expectation, as GeneWise only predicts genes by virtue of their homology to other genes. We would place our base pair accuracy as far higher (in the 90% range) and the wrong gene predictions to be at 0.

DISCUSSION

The GASP assessment was a valuable exercise in providing independent evaluation of gene prediction effectiveness. Providing clear-cut assessment of gene predictions is a difficult task and was not helped by the time pressures of both the contributing groups and the assessing group to provide this study. It is clear that the rules for what predictions will be considered as real need to be detailed in the future, and possibly the ability to assess such things as pseudogene predictions, will be important. Ideally there should be experiments by the assessing group after the gene predictions have been made, so that it is clearer that people have at least attempted to verify a gene prediction experimentally.

Present address: European Molecular Biology Laboratory–European Bioinformatics Institute (EMBL–EBI) , Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.
[1]**Corresponding author.**
E-MAIL Birney@ebi.ac.uk; FAX 44-1-2223-494468.

10:547–548 ©2000 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/00 $5.00; www.genome.org

The predictions made by GeneWise were very much in line with the predictions made using the BLOCKS method (Henikoff et al. 2000). The BLOCKS method considers smaller, ungapped and unspliced motifs drawn from a broader database than PFAM. The result is that there are differences due to the different database source and due to the method—in particular GeneWise tends to predict more coding sequence than BLOCKS for a particular family.

The effectiveness of GeneWise in this study was reported at below the levels we believe to be correct. It is our belief that the specificity numbers for all methods are not well assessed in this study, and that people should not quote them without considerable discussion of the shortcomings of this assessment, that is, the calling of transposon genes as errors and annotation oversights. Even so, this exercise is valuable to raise awareness of the problems in both prediction and assessment. We look forward to participating in future studies.

METHODS

The method used in this study, halfwise, is part of the Wise2 package available from http://www.sanger.ac.uk/Software/Wise2. halfwise is a PERL script that uses BLASTX to compare the DNA sequence against a protein database designed to represent the protein space covered by PFAM database. The BLASTX search selects a number of potential PFAM models to be used in the more computationally expensive GeneWise method.

The DNA sequence was split up into 100-kb chunks with no overlaps, and each chunk was run through the halfwise method. The resulting GFF output was then processed to assemble the complete GFF file. The total time to perform the analysis was a weekend of off-peak computer resources at the Sanger Centre.

ACKNOWLEDGMENTS

This work was supported by the Wellcome Trust. E.B. is a Wellcome Trust Prize Student.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Bateman, A., E. Birney, R. Durbin, S.R. Eddy, K.L. Howe, and E.L.L. Sonnhammer. 2000. The pfam protein families database. *Nucleic Acids Res.* **28:** 263–266.

Burge, C. and S. Karlin. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.

Henikoff, J.G. and S. Henikoff. 2000. Genomic sequence annotation based on translated searching of the BLOCKS+ database. *Genome Res.* (this issue).

Krogh, J. 1997. Two methods for improving performance of a HMM and their application for gene finding. *Intell. Syst. Mol. Biol.* **5:** 179–186.

Krogh, A., M. Brown, I.S. Milan, K. Sjolander, and D. Haussler. 1994. Hidden Markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.* **235:** 1501–1531.

Kulp, D., D. Haussler, M.G. Reese, and F.H. Eeckman. 1996. A generalized hidden Markov model for the recognition of human genes in DNA. *Intell. Syst. Mol. Biol.* **4:** 134–142.

Moult, J., T. Hubbard, K. Fidelis, and J.T. Pedersen. 1999. Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins Suppl.* **3:** 2–6.

Reese, M., G. Hartzell, N.L. Harris, U. Ohler, and S.E. Lewis. 2000. Genome annotation assessment in *Drosophila melanogaster. Genome Res.* (this issue).

Uberbacher, E.C., Y. Xu, and R.J. Mural. 1996. Discovering and understanding genes in human DNA sequence using GRAIL. *Methods Enzymol.* **266:** 259–281.