

# Choosing Models for Similarity-Based Gene Prediction: Profiles versus Single Sequences

**William Reisdorf and Pankaj Agarwal**

Bioinformatics Research

SmithKline Beecham Pharmaceuticals R&D

King of Prussia, PA 19406 USA

{*William\_Reisdorf, Pankaj\_Agarwal*}@sbphrd.com

## Abstract

With the public release of the human genome draft sequence, one important challenge is to discriminate coding from non-coding regions. While no current computational methods, either alone or in combination, can locate exactly all exons and introns, progress continues in terms of the sensitivity and specificity of predictions. Homology-based gene prediction should improve as the number of database entries increase. However, it is not always clear which database sequence will provide the best model for a new gene. The top database match may only cover part of the sequence, and choosing a longer (even if lower-scoring) example may result in a superior gene prediction. Protein profiles often are a better representation of the variety inherent in a set of database matches. We present an evaluation of using profile-HMMs, generated by the HMMER package, to guide homology-based gene predictions from GeneWise.

We used a test set of 178 known human genes from Burset & Guigo (1), along with a recent set of WU-BLAST hits against NR. The BLAST hits were binned by P-value and for each bin, the results of using the best hit as a GeneWise model were compared against using a profile generated from all the hits in the bin. The comparisons were calculated using AcE (2). In general, for bins with excellent database matches, using the best hit was sufficient for a highly accurate gene prediction. However, as the database matches become poorer, profile-based predictions generally outperformed those using the best (single) hit. This came with an associated cost of increased execution time, and also an increased number of false positive predictions. But in cases where maximizing the number of true positives is important, a profile-based approach appears to have benefits.

## References

1. M Burset & R Guigo (1996) Evaluation of gene structure prediction programs. *Genomics* 34:353-367.
2. W Hayes (2000) AcE: A system for analyzing the accuracy of gene prediction programs.[poster submitted to ISMB-2000]