Large Scale Sequencing Specific Neural Networks for Promoter and Splice Site Recognition*

Martin G. Reese, Nomi L. Harris and Frank H. Eeckman. Lawrence Berkeley National Laboratory, Genome Informatics Group, Berkeley, CA.

Recently several groups, including the LBNL Human Genome Center, are starting new projects in large scale sequencing of the human genome. One of the computational challenges in these projects is the automated detection of significant features in genomic DNA. We are especially interested in the recognition of promoter elements, coding regions, and splice site junctions. We believe that better detection of these features will lead to vastly improved gene finding. However, promoter elements and splice junctions have a complex structure, consisting of many individual elements, such as the TATA-box, transcription start signal for promoters and splice site consensus sequences. Furthermore, the relative positions of the individual elements are variable, and some elements may be absent altogether. Previous efforts in this area have been plagued by high rates of false positives. Given the nature of genomic sequencing in humans, where large introns are known to exist, we recognize the need for a very specific algorithm, with a small number of false positives.

We analyze the structure of the individual elements within promoters and splice sites using a novel technique that combines neural networks with weight pruning. A neural network is trained to recognize promoter or splice site elements until it reaches a local minimum. Then the pruning procedure deletes those weights in the network that add the lowest predictive value to the overall prediction. After pruning, the neural network is retrained until it is stuck again in a new minimum. This procedure is repeated until a defined error level is reached. Eventually, the pruned neural network gives clues about the importance of specific positions in the promoter element and splice junction by the distribution of the remaining weights.

To predict promoter sites, we use time-delay neural networks to combine the predictions that were made for each of the individual promoter elements. TDNNs are appropriate for recognizing promoter elements because they are able to combine multiple features, even those that appear at different relative positions in different sequences. Another advantage is the high selectivity of the TDNN, which is extremely important for promoter prediction systems, in order to avoid generating too many false positives.

Our TDNN predicts most of the annotated promoters in a set of human genes from Genbank (version 86.0). As an example, the TDNN finds the annotated promoter from a 13,865 basepair test gene, HUMTFPB, with a false positive score of 0.05%

---

* This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

(6 false positive predictions out of 13,865). On a test set containing 42 known human gene promoters and 84 random DNA sequences we were able to recognize 50% of the human gene promoters without false positive classification (correlation coefficient of 0.61).

We have applied this network and the splice site prediction networks to our most recently produced sequences and will present data at the conference.