RECEIVED
JAN - 4 1994

OSTI

Continuation Progress Report, Dec. 1993

DOE Award: DE-FG03-93ER61606

Title: New Approaches to Recognizing Functional Domains in Biological Sequences

Project Director: Gary D. Stormo

The *GeneParser* program, for identifying coding regions in genomic DNA sequences, has undergone substantial improvement during the first year of this project. The primary enhancements are the following:

1. Four-state classification. Now, in addition to identifying introns and internal exons the program also can find the 5'-terminal and 3'-terminal exons, thereby making predictions for the entire gene structure.
2. New classification statistics. A few new statistics are added to the original set. One of these is for initiation codons so that 5'-terminal exons can be identified. In addition periodicity statistics are used to help in classifying exons and introns. The most important new statistic is the use of BLAST searches to help identify exons. Regions which show high amounts of similarity to known coding regions are more likely to be coding regions themselves, so including such results in the predictions significantly improves the accuracy.
3. G+C affects on statistics. It is well known that the human genome has a "mosaic" structure in which regions, which may be millions of bases long, have a consistent G+C content but different regions may have very different G+C contents. Some of the types of statistics that are used are quite sensitive to the G+C content of the region, especially statistics like codon usage. Therefore we have compiled different statistical tables for different G+C contents, dividing regions into three classes of low, medium and high. *GeneParser* determines the G+C content of the entire genomic sequence under consideration and then makes the coding regions predictions based on the statistical tables for that particular content. This has proven to increase the reliability of the method significantly.
4. Sub-optimal solutions. One of the big advantages of the Dynamic Programming approach used in *GeneParser* is that both optimal and sub-optimal solutions are easily obtained. The current version lets the user display the optimal solution and then obtain, in order, displays of any number of suboptimal solutions by simply "clicking" the pointing device (mouse) of the workstation. In addition, it is possible to obtain a graphic display which shows the optimal solution containing any particular junction point. For example, suppose there is additional information suggesting that position X in the sequence is a donor splice site. One can simply "point to" position X and obtain the optimal solution with that position constrained to be a donor site. In fact, one can generate a graphical display in which the optimal score is shown for all possible positions for all possible junction types. This display serves to show which regions of the sequence can be reliably predicted, because only one type of prediction shows up there with high values, and which parts of the predictions are less reliable because alternative solutions have comparable scores.

We have also begun testing the robustness of the method to errors in the sequence data, and this

MASTER

rp

will continue into the second year. We have generated data sets with errors introduced them, both substitution and frameshift errors, at different rates. We have found that *GeneParser* is somewhat more robust to those errors than other methods, but we haven't yet tried to optimize the robustness. That will be done by using error-containing data during the training process where, presumably, the more error-tolerant statistics will become more heavily weighted and the overall method will increase its performance in the presence of errors.

Some of the results of this year's work on *GeneParser*, and a comparison of it to other methods, has been included in a review we have written for the second edition of "Nucleic Acid and Protein Sequence Analysis: A Practical Approach", being published by Oxford University Press. A more thorough description will be submitted next year as well as included in Eric Snyder's thesis. Progress on this method has also been reported at several meeting during the last year: I was an invited speaker at the Cold Spring Harbor meeting on "Genome Mapping and Sequencing" last May and talked about *GeneParser*. It was also the topic of a talk at last year's Aspen workshop on "Identifying Features in Biological Sequences" and was included in a talk I gave in August at a "Workshop on Molecular Sequence Analysis" at Stanford University. Eric Snyder was also invited to give a talk about *GeneParser* at the "3rd International Workshop on the Identification of Transcribed Sequences" in New Orleans last October.

Work has not been quite so productive on the use of Landscapes for classifying sequences. Dr. Ulyanov has used the suffix array, from which the Landscape is built, to develop a "Multi-alphabet Consensus" method for identifying common features in unaligned sequences, and that work is now being written up for submission. He has also shown that this method is of some value in helping to classify sequences into different categories. That is the direction we will continue working as we think this may be quite useful, especially for the task of identifying promoter regions and other ill defined regions of genomic sequences. In particular it looks like the Landscape can be the basis for a "minimum description length" compression of sequences, and the amount of compression obtained is related to the similarity between the sequence and a "source" database. For example, if a region of sequence serves as a promoter it can probably be compressed to a greater extent by referring to sequences in a database of promoters than by referring to databases of other types of sequences. If this supposition is true, and preliminary evidence suggests it is, then by collecting databases of various types of sequences (promoters, exons, introns, intergenic regions, ...) we can classify new sequences based on their compressibility with respect to each type of database. This work can, in turn, be graphically displayed in the context of the normal Landscape and should be a valuable tool for understanding the important features in genomic DNA sequences.

DATE

FILMED

6 / 2 / 94

END