DOE/ER/61606--T2

Final Report, Dec. 1997

DOE Award: DE-FG03-93ER61606

Title: New Approaches to Recognizing Functional Domains in Biological Sequences

Project Period: April 1, 1993 - March 31, 1997

Project Director: Gary D. Stormo

The purpose of this project is to develop new approaches and programs for determining the function of DNA domains. This will aid in the understanding of the sequence data obtained through the Human Genome Project. One of the great challenges of that project is to abstract important biological information from the raw sequences that emerge. Our efforts have focused on several areas: determining the protein coding regions (i.e. exon boundaries) in genomic DNA; recognizing patterns of DNA binding proteins, including nucleosomes, from the sequence using "multi-alphabet" analyses; better recognition methods for RNA genes and other patterns where structural considerations are important along with sequence; enhancing the "Sequence Landscape" approach to pattern recognition and applying it to various problems in domain classification.

The progress obtained in each of these areas is described below, along with the references of publications that have appeared, are in press or in preparation. All of these areas have resulted in new programs that are publically available via anonymous ftp from our site in the pub directory at beagle.colorado.edu.

GeneParser is the program we developed to identify optimal classification boundaries in genomic DNA. This was the first approach to combine several types of evidence into the classification and obtain optimal and suboptimal predictions by a Dynamic Programming algorithm. We also explored the use of neural networks to obtain the optimal weighting of the different types of evidence. This project was begun as a "proof of principle" endeavor, and it has been exceedingly successful in that regard. While the *GeneParser* program itself is not widely used, it has influenced essentially all of the programs now being used or in development. Our program was never optimized for speed, and further developments have allowed for progress beyond that which we achieved. Our own continuing work in this area is now in collaboration with David Haussler, UCSC, on developing Genie. That program is an example of a Generalized Hidden Markov Model, which is a direct descendant of the approach pioneered with *GeneParser*. The program is available via anonymous ftp. Publications resulting from this work are:

1. Snyder, E.E. and Stormo, G.D. (1995) Identification of Protein Coding Regions in Genomic DNA. *J. Mol. Biol.* 248:1-18.
2. Snyder, E.E. and Stormo, G.D. (1996) Identification of Protein Coding Regions in Genomic DNA Sequences. In: *Nucleic Acid and Protein Sequence Analysis, A Practical Approach, Second Edition* (M.J. Bishop and C.J. Rawlings, eds.) IRL Press, Oxford, pp. 209-224.
3. Snyder, E.E. and Stormo, G.D. (1996) Biology and Biochemistry: Neural Networks for identification of protein coding regions in genomic DNA sequences. In: *Handbook of Neural Computation*, Oxford Univ. Press.
4. Stormo, G.D. and Haussler, D. (1994) Optimally Parsing a Sequence into Different Classes Based on Multiple Types of Evidence. In: *Proceedings of the Second International Conference on Intelligent Systems in Molecular Biology*, pp. 369-375.

Multi-alphabet consensus analysis of word patterns in sequences was aimed at utilizing efficient word matching algorithms to discover important patterns, but allowing for the ambiguity usually

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

**Portions of this document may be illegible
electronic image products. Images are
produced from the best available original
document.**

RECEIVED

JUL 27 1998

observed with regulatory sites in DNA (as well as other protein-DNA interactions). This led to the development of a new program that is available via anonymous ftp, and to the demonstration of the utility of the approach on a couple of different data sets. Two publications resulted from this work:

1. Ulyanov, A. and Stormo, G.D. (1995) Multi-Alphabet Consensus Algorithm for Identification of Low Specificity Protein-DNA Interactions. *Nucl. Acids Res.* 23:1434-1440.
2. Ulyanov, A.V. and Stormo, G.D. (1996) Pattern-Recognition Methods for Classifying Domains of DNA Sequences. In: *Computational Medicine, Public Health and Biotechnology: Building a Man in the Machine* (M. Witten and D. Vincet, eds.) World Scientific Pub. Co., Singapore.

RNA folding using Graph Theory is an approach we developed to deal with the problem of identifying domains for RNAs, which are often composed of both sequence and structure constraints. In addition, there are often pseudoknot elements in the RNA that are not modeled by other RNA structure prediction methods, such as Dynamic Programming or SCFG methods. We developed two types of programs, one for predicting the structure (including pseudoknots) of an aligned set of RNAs, and the other to align a new sequence to an existing model of the RNA sequence and structure constraints. Together these programs can be used to automatically align, and develop a structural model for, previously unaligned RNAs via an EM-type algorithm. Both programs are available via anonymous ftp. Two papers have been published describing these approaches, and a third is currently in preparation:

1. Cary, R.B. and Stormo, G.D. (1995) Graph-Theoretic Approach to RNA Modeling Using Comparative Data. In: *Proceedings of the Third International Conference on Intelligent Systems in Molecular Biology*, pp. 75-80.
2. Tabaska, J. and Stormo, G.D. (1997) Automated Alignment of RNA Sequences to Pseudoknotted Structures. In: *Proceedings of the Fifth International Conference on Intelligent Systems in Molecular Biology*, pp. 311-318.
3. Tabaska, J., Cary, R.B., Gabow, H.N. and Stormo, G.D. (1997) An automated RNA modeling approach capable of identifying pseudoknots and base-triples. in preparation.

Sequence Landscapes is an approach that utilizes indexing methods to identify the frequency of all words in a database which also occur in a "target" sequence. This can be done very rapidly and efficiently by good indexing methods, and Gene Myers of the University of Arizona has collaborated with us on this. This approach can be utilized in a variety of ways on several different problems. For example, it can be used to identify repeated regions in sequences as well as under-represented words. We have mostly been interested in applying it to the problem of domain classification, which is done by building databases of specific domain types (i.e. exons, introns, promoters, etc). Then one can use the word frequencies in those databases to ascertain the likelihood that a new region of sequence belongs to one or the other, or to find the optimal boundaries between different class types. We are continuing to work on this project, through alternative funding and with a new proposal to DOE pending. We think this general method will be valuable for several purposes, including finding optimal oligos to be used in expression arrays. This purpose will utilize a fast approximate matching algorithm developed by Gene Myers that we have put into use with

the landscape program. Other developments, especially the use of this approach to help identify promoter regions in genomic DNA, are also being continued. Papers that have been published, or are in press, on this project are:

1. Levy, S. and Stormo, G.D. (1997) DNA Sequence Classification Using DAWGs. In: *Structures in Logic and Computer Science*, Lecture Notes in Computer Science, Vol. 1261 (J. Mycielski, G. Rozenberg, and A. Salomaa, eds.), Springer-Verlag, pp. 339-352.
2. Levy, S., Compagnoni, L., Myers, E.W. and Stormo, G.D. (1998) Xlandscape: a tool for the graphical display of word frequencies in sequences. *BioInformatics*, in press.