Combinatorial Methods for Gene Recognition
DE-FG02-94ER61919
(Final report: October, 1994 - January, 1997)

Principal Investigator Pavel A. Pevzner

Department of Computer Science
The Pennsylvania State University
University Park, PA 16802
pevzner@cse.psu.edu

October 29, 1997

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, make 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 in electronic image products. Images are produced from the best available original document.

Combinatorial Methods for Gene Recognition (grant DE-FG02-94ER61919)

Specific Aims. Develop algorithms and software for gene prediction.

Results. The major result of the project is the development of a new approach to gene recognition called *spliced alignment* algorithm. We have developed an algorithm and implemented a software tool (for both IBM PC and UNIX platforms) which explores all possible exon assemblies in polynomial time and finds the multi-exon structure with the best fit to a related protein. Unlike other existing methods, the algorithm successfully performs exons assemblies even in the case of short exons or exons with unusual codon usage; we also report correct assemblies for the genes with more than 10 exons provided a homologous protein is already known. On a test sample of human genes with known mammalian relatives the average overlap between the predicted and the actual genes was 99%, which is remarkably well as compared to other existing methods. At that, the algorithm absolutely correctly reconstructed 87% of genes. The rare discrepancies between the predicted and real exon-intron structures were restricted either to extremely short initial or terminal exons or proved to be results of alternative splicing. Moreover, the algorithm performs reasonably well with non-vertebrate and even prokaryote targets. The spliced alignment software *PROCRUSTES* has been in extensive use by the academic community since its announcement in August, 1996 via the WWW server (www-hto.usc.edu/software/procrustes) and by biotech companies via the in-house UNIX version.

An important ingredient of this project is a collaboration with Russian visiting scientists M.S. Gelfand, A.A. Mironov and M.A. Roytberg and graduate student S.H.Sze (Computer Science) and S.Rahmanov (Biology). The results of this collaboration are reflected in the thirteen papers citing the grant DE-FG02-94ER61919 (see below) Seminars and demonstrations of *PROCRUSTES* were given at Moscow State University, June, 1995, DIMACS Gene Recognition Workshop Philadelphia, October, 1995, Princeton University, October, 1995, Sequana, Inc., November, 1995, National Center for Biotechnology Information, November, 1995, Department of Biology, University of Southern California, February, 1996, Department of Molecular Biotechnology, University of Washington, February, 1996, 2nd Sandia Workshop on Computational Molecular Biology Albuquerque, March, 1996, Amgen, Inc., April, 1996, SmithKline Beecham, Inc., May and November, 1996, Millenium, Inc, May, November and December, 1996, Israeli Computational Molecular Biology Workshop, June, 2, 1996, Genome Therapeutics, November, 1996, Pennsylvania State University, November, 1996, RWC, Tokyo, Japan, December, 1996, 1st Annual Computational Molecular Biology Conference (RECOMB 97), Santa Fe, January, 1997.

Significance. Recent developments in gene recognition and large-scale DNA sequencing indicate that combinatorial methods in this area are becoming increasingly important. From this perspective the spliced alignment approach for gene recognition may become a method of choice in the future.

The following papers describe different aspects of the project and acknowledge the support of DOE grant DE-FG02-94ER61919

1. M.S.Gelfand (1995) FANS-REF: a bibliography on statistics and functional analysis of nucleotide sequences. *Comput. Appl. Biosci.*, 11, 541. (included in EMBL data library).
2. Gelfand M.S. (1995) Prediction of function in DNA sequences *J. of Computational Biology*, 2, 87-115
3. Hannenhalli S., Pevzner P.A. (1995) Towards Computational Theory of Genome Rearrangements. In J. van Leeuwen, (ed.) *Computer Science Today: Recent Trends and Developments*. Lecture Notes in Computer

Science, 1000, 184-202 (invited paper in a jubilee volume)

4. Hannenhalli S., Pevzner P.A. (1995) Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Proc. of the 37th Annual Symposium on the Theory of Computing*, Las Vegas, Nevada, 178-189
5. Mironov A.A., Alexandrov N.N., Bogodanova N.Y., Grigoriev A., Lebedev V., Lunovskaya-Gurova, Grigoriev A., Lebedev V., Truchan M., Pevzner P.A. (1995) DNASUN: a software package for the biotechnology laboratory, *Computer Applications in Biological Sciences*, 11, 331-335
6. Bafna V., Pevzner P.A. (1995) Sorting by transpositions. *Proceedings of the Sixth Annual Symposium on Discrete Algorithms*, San Francisco, CA, 614-623
7. M.S.Gelfand, L.I.Podolsky, T.V.Astakhova, M.A.Roytberg. Prediction of the exon-intron structure and multicriterial optimization. In: *Bioinformatics and Genome Research* (H.A.Lim, C.R.Cantor, eds.), 173-180 (World Scientific, 1995)
8. Gelfand M.S., Podolsky L.I., Astakhova T.V., Roytberg M.A. (1996) Recognition of genes in human DNA sequences. *J. Computat. Biol.*, 3, 223-234.
9. M.A.Roytberg, T.V.Astakhova, M.S.Gelfand (1996) Algorithm for highly specific recognition of protein-coding regions. *Proc. of the Seventh Workshop on Genome Informatics*, December, 1996, Tokyo, Japan, 82-87
10. M.Gelfand, A. Mironov, P.Pevzner (1996) Spliced alignment: A New Approach to Gene Recognition. D.Hirschberg and E.Myers (eds.) *Combinatorial Pattern Matching 1996, Lecture Notes in Computer Science*, 1075, 141-159
11. Hannenhalli S., W. Feldman, H. Lewis, S.Skiens, Pevzner P.A. (1996) Positional Sequencing by Hybridization. *Computer Appl. in Biol. Sciences*, 12, 19-24
12. Hannenhalli S., Pevzner P.A. (1996) To cut... or not to cut: applications of comparative physical maps in molecular evolution. *Proc. of the Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, Atlanta, Georgia, 304-313
13. M.Gelfand, A. Mironov, P.Pevzner. (1996) Gene Recognition via Spliced Alignment *Proc. National Academy of Sciences USA*, 93, 9061-9066