A Threading Approach for EST Data

Alexander Zien, Ralf Zimmer, Thomas Lengauer
Institute for Algorithms and Scientific Computing (SCAI)
GMD – German National Research Center for Information Technology
Schloß Birlinghoven, D-53754 Sankt Augustin, Germany
E-mail: {Zien, Zimmer, Lengauer}@gmd.de

One major use of expressed sequence tag (EST) data is the early discovery of new, possibly disease-related genes. Coding sequence (CDS) within ESTs can be found by the help of sequence alignment tools like `BLASTX` [3], `FASTX/FASTY` [8] and `SearchWise` [2]. However, these tools can only find genes of which homologues with detectable sequence similarity are already known and present in databases.

In this work, we investigate whether the number of identifiable genes can be increased by transfering the threading approach to nucleotide sequences. In order to do a comprehensive evaluation, we need a fast threading tool. Here, we utilize the `123D` program [1] from the `ToPLign` package developed at the GMD. `123D` is capable of calculating optimal alignments of protein sequences with respect to sequence similarity as well as several potentials reflecting structural preferences of amino acids. We extend this algorithm to work with nucleotide sequences as well by integrating it with an on-the-fly translation [2, 4, 9, 10]. We also allow for the compensation of frame shift errors, which are abundant in EST data [5].

A major part of this work is the generation of appropriate benchmark data sets. We build upon a non-redundant set of domains as defined by `SCOP` [7]. The EST sequences are real consensus sequences of assemblies picked from `STACK` [6].

References

[1] N. N. Alexandrov, R. Nussinov, and R. Zimmer. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. In *PSB'96*, pages 53–72, 1996.

[2] E. Birney, J. D. Thompson, and T. J. Gibson. PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *NAR*, 24(14):2730–2739, July 1996.

[3] W. Gish and D. J. States. Identification of protein coding regions by database similarity search. *Nature Genetics*, 3(3):266–272, March 1993.

[4] J. Hein. An Algorithm Combining DNA and Protein Alignment. *Journal of Theoretical Biology*, 167:169–174, 1994.

[5] L. Hillier et al. Generation and analysis of 280,000 human expressed sequence tags. *Genome Research*, 6:807–828, 1996.

[6] R. Miller, A. Christoffels, J. Burke, B. Karlak, A. Ptitsyn, T. Broveak, and W. A. Hide. A comprehensive approach to the determination of expressed human gene sequences: The Sequence Tag Alignment and Consensus Knowledgebase. In preparation.

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *JMB*, 247:536–540, 1995.

[8] W. R. Pearson, T. Wood, Z. Zhang, and W. Miller. Comparison of DNA Sequences with Protein Sequences. *Genomics*, 46(1):24–36, November 1997.

[9] D. J. States and D. Botstein. Molecular sequence accuracy and the analysis of protein coding regions. *PNAS*, 88:5518–5522, July 1991.

[10] Z. Zhang, W. R. Pearson, and W. Miller. Aligning a DNA Sequence with a Protein Sequence. In *RECOMB'97*, 1997.