

# SST versus EST in Gene Recognition

Andrey A. Mironov  
Laboratory of Mathematical Methods,  
National Center for Biotechnology NIIGENETIKA,  
Moscow, 113545, Russia.  
mir@vnigen.msk.su

Pavel A. Pevzner  
Departments of Mathematics and Computer Science,  
University of Southern California,  
Los Angeles, CA 90089-1113  
ppevzner@hto.usc.edu

## Abstract

*The EST data provide a powerful tool for identification of transcribed DNA sequences. However, since ESTs are relatively short, many exons are poorly covered by ESTs thus reducing the utility of EST data. Recently, SST (Signature Sequence Tags) fingerprints were proposed as an alternative to EST fingerprints. Given a fingerprint set of probes, SST of a clone is a subset of probes from the fingerprint set that hybridize with the clone. We demonstrate that besides being a powerful technique for screening cDNA libraries, SST technology provides for a very accurate gene predictions. Even with a small fingerprint set (600-800 probes) SST-based gene recognition outperforms many conventional and EST-based methods. The increase in the size of fingerprint set to 1500 probes provides almost perfect gene recognition. Even more importantly, SST-based gene predictions miss very few exons and therefore provide an opportunity to bypass cDNA sequencing step on the way from finished genomic sequence to mutation detection in gene hunting projects. Since SST data can be obtained in a highly parallel and inexpensive way, SST technology has a potential of substituting EST technology for gene hunting.*

## 1. Introduction

In the absence of accurate gene prediction programs gene identification and exon annotation in genomic DNA usually amounts to sequencing the corresponding

mRNA. This mRNA can be found by direct screening of cDNA libraries, northern blot analysis, or hybrid selection of cDNA ([1], [8]). A serious limitation of these techniques is the low signal-to-noise ratio in hybridization experiments and frequent failures of splicing-based exon amplification methods.

An alternative strategy is *in silico* gene prediction. Unfortunately, the error rate of gene recognition algorithms is still too high for accurate gene identification, particularly in the case of non-translated exons and exons with unusual codon usage and splicing sites [2]. To remedy this problem many recent studies ([14],[15],[16]) attempted to incorporate additional experimental information (and EST data, in particular) into gene prediction algorithms. However, since ESTs are rather short, EST data provided a relatively modest increase in accuracy of gene predictions.

Drmanac et al.,1993 and Meier-Evert et al.1993 proposed an alternative *SST approach* (*Signature Sequence Tags*) for screening cDNA libraries and identification of transcribed DNA. SST data can be obtained either with Hyseq, Inc. "clones on the chip" technology or with Affymetrix, Inc. "probes on the chip" technology. In the former case, the SSTs for the entire cDNA library can be obtained in just  $k$  parallel hybridization experiments where  $k$  is the size of the fingerprint set. It provides for a significant increase in the rate of producing fingerprints as compared to the intrinsically sequential EST approach. In December, 1997 Hyseq, Inc. announced that the company generates SSTs at a rate of 1 million bases per month and that their *HyGenomics<sup>TM</sup>* database already has 4 millions of SSTs. Given this rate of data accumulation

*HyGenomics*<sup>TM</sup> has a potential to become the largest proprietary database of genomic information in just a few months.

Drmanac et al., 1994 proposed to use SSTs for recognition of cDNA clones corresponding to already sequenced genes. Drmanac et al., 1996 and Milosavljevic, 1996a,b also showed how to identify potentially new genes in cDNA libraries with SST data (also called OSS - Oligonucleotide Sequence Signatures). However, the problem of gene prediction in *genomic* DNA using SSTs of cDNA libraries remained open. This problem is very different from the EST-based gene prediction since SST and EST approaches are, in some sense, complementary. While SSTs sample the entire cDNA clone in a rather sparse way, ESTs sample only a fraction of a full-length cDNA clone in a rather dense way. Below we show that *global and sparse* SST approach may be better for gene prediction than *local and dense* EST approach. It indicates that SST-based gene recognition may become an ultimate gene prediction tool as soon as SST data become widely available.

**SST-based gene recognition.** Let  $S$  be a fixed set of probes and  $C$  be a DNA sequence. An *SST* of  $C$  is a subset of probes from  $S$  that hybridize with  $C$ . Let  $G$  be a genomic sequence containing a gene represented by a cDNA clone  $C$ .

*SST-based gene recognition problem.* Given a genomic sequence  $G$  and the SST of the corresponding cDNA clone  $C$ , predict a gene in  $G$  (i.e. predict all exons in  $G$ ).

Let probes from  $S$  appear in the genomic sequence  $G$   $k$  times. The *SST map* of  $G$  is a sequence  $x_1, \dots, x_k$  of length  $k$  such that  $x_i = 1$  if  $i$ -th occurrence of a probe from  $S$  in  $G$  corresponds to a probe from  $C$  and  $x_i = 0$  otherwise. A *1-position* (*0-position*) in genomic sequence is a position in  $G$  corresponding to  $x_i = 1$  (resp.  $x_i = 0$ ). In the absence of experimental errors sufficiently long exons are expected to be runs of ones in SST map (1s are shown as yellow bands in Figure 1). Of course, spurious runs that correspond to false exons and real exons that do not correspond to runs complicate the SST-based gene recognition. False negative experimental errors lead to shortening or breaking the candidate exons into smaller pieces whereas false positive errors lead to extending exons and creating additional false exons.

The algorithm *GeneSST* for SST-based gene prediction attempts to reconstruct broken/shortened/extended exons and eliminate false exons by analyzing runs in SST maps and matching them against the predicted splicing sites and exons in genomic DNA. A simple approach to SST-based gene recognition problem in the absence of

experimental errors finds all runs of sufficient length and attempts to flank them with sufficiently strong acceptor and donor sites located within a short distance from the endpoints of the run. For a given run let  $x$  ( $y$ ) be the rightmost (resp. leftmost) 0-position preceding (resp. following) the run. The algorithm searches for the leftmost candidate acceptor site and the rightmost candidate donor site *inside*  $[x, y]$  interval to find an exon corresponding the given run. This approach (implemented in *GeneSST* software) leads to predictions of *both* translated and non-translated exons.

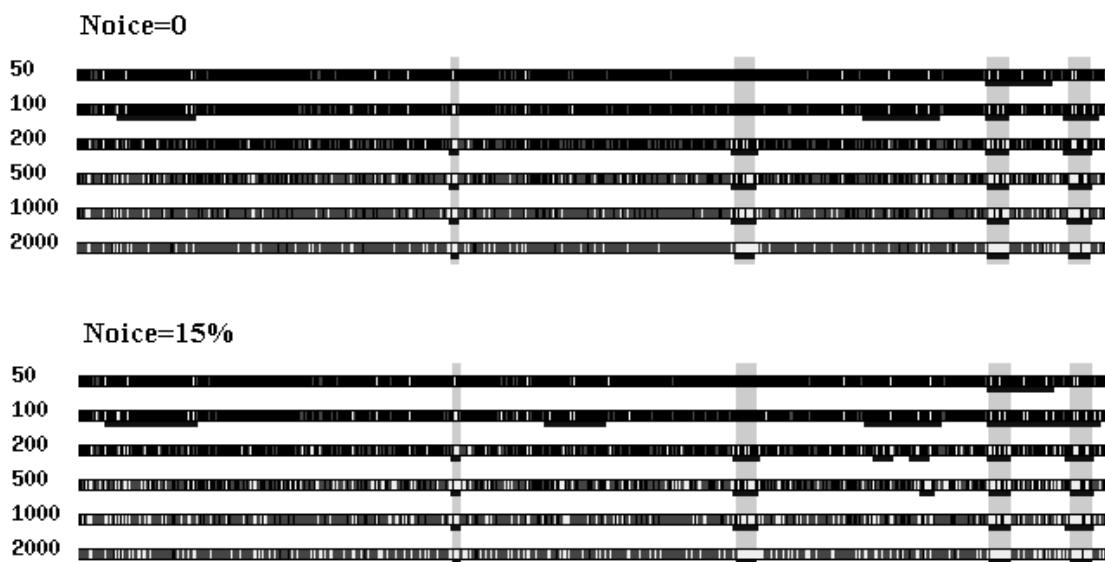
The shortcoming of the simple approach is that it searches for exons independently without trying to assemble the most likely exon structure. For recognition of translated exons, a special version of the *spliced alignment* algorithm [6] can be applied to solve SST-based gene recognition problem. Following [6] define a set of candidate exons which is obtained by a filtration of all segments flanked by the conventional acceptor and donor dinucleotides. A run of the *local spliced alignment* [13] between SST map (mimicking the genomic sequence in  $\{0, 1\}$  alphabet) and a long run of 1s leads to a solution of SST-based gene recognition problem.

Additional complications and constraints (minimal distance between exons, hybridization of probes at the exon junctions, analysis of very short exons, etc.) lead to further complications of the algorithm. Also, the current version of *GeneSST* ignores the hybridization intensities since very few details about the hybridization intensities distribution are currently available. As soon as such information becomes a public knowledge, *GeneSST* can be easily modified to work with hybridization intensities.

## 2. Results

The *GeneSST* program was tested on a set of 256 DNA sequences fragments containing non-homologous human complete genomic sequences [13]. A set of probes  $S$  was generated as a set of  $k$  random probes of size 7 for  $k$  ranging from 50 to 2000. Random experimental errors were introduced with rate varying from 10% to 30% (assuming equal false positive and false negative rates). The estimates of splicing sites strength and filtration of candidate exons was performed as in [13].

Figure 2 presents the results of *GeneSST* tests. In the error-free case the correlation coefficient for *GeneSST* exceeds 90% as soon as the size of the fingerprint set exceeds 600 probes. With 1000 probes the correlation coefficient becomes as high as 95%. The accuracy of SST-based gene recognition remains very



**Figure 1.** The SST map of the 4-exon human gene *HUMDZA2G* (14693 base pairs, exon sizes are 67, 261, 276 and 284 base pairs). The size of fingerprint set  $S$  varies from 50 to 2000, the error rate varies from 0 to 15 percent. 1-positions (yellow bands) represent the probes from  $S$  matching both genomic DNA and cDNA. 0-positions (red bands) represent the probes from  $S$  matching genomic DNA but not cDNA. Gray bars represent real exons, blue bars represent exons predicted by *ExonSST*.

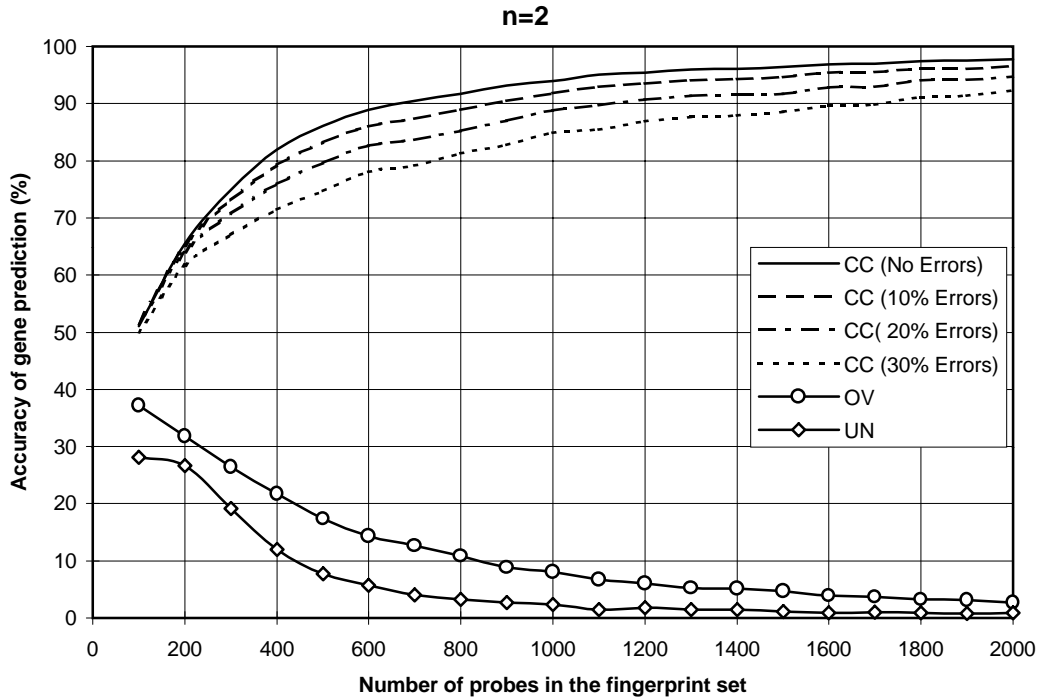
high even in the presence of significant experimental errors. The correlation coefficient decreases by as little as 1% – 3% for every 10% increase in the error rate. Moreover, Figure 2 illustrates that the prediction accuracy as high as 97% can be achieved by simply increasing the size of fingerprint set *without* improving experimental accuracy. Another important difference between SST-based and conventional algorithms is extremely low underprediction level in GeneSST (see Figure 2). This is very important for gene hunting projects during the transition stage from genomic sequencing to mutation detection.

### 3. Conclusions

In a recent paper Drmanac et al., 1998 [5] demonstrated the utility of massive SST data for resequencing and mutation detection. GeneSST demonstrates how to use SSTs for finding exons in genomic DNA. Since the information about the quality of SST data is not publicly available yet and since hybridization fingerprints have many artifacts, GeneSST uses a simplified model of binary SST fingerprints and does

not explore the specifics of *HyGenomics<sup>TM</sup>* information. From this perspective, GeneSST is rather a proof of concept than a software tuned for a particular type of data. At the same time, GeneSST performs well in the presence of 10% errors in SST data that are reported in Drmanac et al., 1998 [5].

The *GeneSST* predictions indicate that SST data may have computational advantages over EST data. Since SST data are easier to obtain, we anticipate that SST technology will become a powerful gene hunting technique in the future. The current version of *GeneSST* intentionally ignores the codon potential to predict both coding and non-coding exons. Even with such a handicap and even in the presence of errors *GeneSST* generates very accurate predictions. Another important feature of GeneSST is that it can work with low quality first-pass DNA sequencing data. An interesting open problem is how to combine SST-based, EST-based and conventional gene prediction approaches for accurate gene predictions in the case of incomplete cDNA clones.



**Figure 2. Performance of GeneSST with different error rates as measured by correlation coefficient (CC) of gene predictions. UN (underprediction and OV (overprediction) values are shown for the error free case.**

### Acknowledgements

We are grateful to Mikhail Gelfand and Mike Waterman for many useful discussions. This work was supported by the U.S. Department of Energy under the grant DE-FG02-ER61919, Russian State Scientific Program "Human Genome", and Russian Fund of Basic Research.

### References

- [1] Brody, L.C., Abel, K.J., Castilla, L.H., Couch, F.J., McKinley, D.R., Yin, G., Ho, P.P., Merajver, S., Chandrasekharappa, S.C., Xu, J., Cole, J.L., Struewing, J.P., Valdes, J.M., Collins, F.S. and Weber, B.L. (1995) *Genomics* 25, 238-247.
- [2] Buset, M. and Guigo, R. (1996) *Genomics* 34, 353-375.
- [3] Drmanac R., Drmanac S., Labat I., Stavropoulos N. (1994) Requirements in Screening cDNA libraries for new genes and solutions offered by SBH technology. In U.Hochgeschwender and K.Gardiner, eds. *Identification of transcribed sequences*, Plenum Press.
- [4] Drmanac S, Stavropoulos NA, Labat I, Vonau J, Hauser B, Soares MB, Drmanac R *Genomics*, 37, 29-40.
- [5] Drmanac S, Kita D, Labat I, Hauser B, Schmidt C., Burczak J., Drmanac R. (1998) *Nature Biotechnology*, 16, 54-58.
- [6] Gelfand, M.S., Mironov, A.A. and Pevzner, P.A. (1996) *Proc. Natl. Acad. Sci. USA* 93, 9061-9066.

- [7] Gu, Y., Shen, Y., Gibbs, R.A. and Nelson, D.L. (1996) *Nature Genet.* 13, 109–113.
- [8] Hattier, T., Bell, R., Shaffer, D., Stone, S., Phelps, R.S., Tavtigian, S.V., Skolnik, M.H., Shattuck-Eidens, D. and Kamb, A. (1995) *Mammalian Genome* 6, 873–879.
- [9] Meier-Ewert S, Maier E, Ahmadi A, Curtis J, Lehrach H (1996) *Nature*, 361, 375–376.
- [10] Meindl, A., Dry, K., Herrmann, K., Manson, F., Ciccodicola, A., Edgar, A., Carvalho, M.R.S., Achatz, H., Hellebrand, H., Lennon, A., Migliacchio, C., Porter, K., Zrenner, E., Bird, A., Jay, M., Lorenz, B., Wittwer, B., D'Urso, M., Meitinger, T. and Wright, A. (1996) *Nature Genet.* 13, 35–42.
- [11] Milosavljevic A, Savkovic S, Crkvenjakov R, Salbego D, Serrato H, Kreuzer H, Gemmell A, Batus S, Grujic D, Carnahan S, Paunesku T, Tepavcevic J (1996a) *Genomics*, 37, 77-86.
- [12] Milosavljevic A, Zeremski M, Strezoska Z, Grujic D, Dyanov H, Batus S, Salbego D, Paunesku T, Soares MB, Crkvenjakov R (1996b) *Genome Res*, 6, 132-141.
- [13] Mironov, A.A., Roytberg, M.A., Pevzner, P.A. and Gelfand, M.S. (1998) *Genomics* (in press).
- [14] Sze S.H., Roytberg M.A., Gelfand M.S., Mironov A.A., Astakhova T.V., Pevzner P.A. (1998) *Bioinformatics* 14, 14-19.
- [15] Xu, Y, Uberbacher E.C. (1997) *Proc. of the First Annual International Conference on Computational Molecular Biology (RECOMB)*, Santa Fe, New Mexico, 330–336.
- [16] Xu, Y, Mural, R.J., Uberbacher E.C. (1997) *Proc. of the Fifth International Conference on Intelligent Systems for Molecular Biology (ISMB)*, Halkidiki, Greece, 344–353.