BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 9 2001
Pages 843–844

GeneMachine: gene prediction and sequence annotation

*Izabela Makalowska, Joseph F. Ryan and Andreas D. Baxevanis**

Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Building 49, Room 4A-22, Bethesda, MD 20892, USA

Received on February 12, 2001; revised on May 25, 2001; accepted on May 30, 2001

ABSTRACT

Motivation: A number of free-standing programs have been developed in order to help researchers find potential coding regions and deduce gene structure for long stretches of what is essentially 'anonymous DNA'. As these programs apply inherently different criteria to the question of what is and is not a coding region, multiple algorithms should be used in the course of positional cloning and positional candidate projects to assure that all potential coding regions within a previously-identified critical region are identified.

Results: We have developed a gene identification tool called GeneMachine which allows users to query multiple exon and gene prediction programs in an automated fashion. BLAST searches are also performed in order to see whether a previously-characterized coding region corresponds to a region in the query sequence. A suite of Perl programs and modules are used to run MZEF, GENSCAN, GRAIL 2, FGENES, RepeatMasker, Sputnik, and BLAST. The results of these runs are then parsed and written into ASN.1 format. Output files can be opened using NCBI Sequin, in essence using Sequin as both a workbench and as a graphical viewer. The main feature of GeneMachine is that the process is fully automated; the user is only required to launch GeneMachine and then open the resulting file with Sequin. Annotations can then be made to these results prior to submission to GenBank, thereby increasing the intrinsic value of these data.

Availability: GeneMachine is freely-available for download at http://genome.nhgri.nih.gov/genemachine. A public Web interface to the GeneMachine server for academic and not-for-profit users is available at http://genemachine.nhgri.nih.gov. The Web supplement to this paper may be found at http://genome.nhgri.nih.gov/genemachine/supplement/.

Contact: andy@nhgri.nih.gov

*To whom correspondence should be addressed.

SYSTEM ORGANIZATION

GeneMachine is based on a series of Perl modules, each of which runs one of the component gene-finding or homology search programs. These modules parse the resulting output and return the results in ASN.1 format, which is readable by Sequin. The GeneMachine executable processes run option and configuration files, passing the required information along to the individual program modules. The use of this type of modular system allows for the easy incorporation of new predictive or homology-based routines.

Homology searching

GeneMachine utilizes two of the most commonly-used tools for homology searching, namely BLAST (Altschul *et al.*, 1997) and RepeatMasker (Smit and Green, unpublished). After the input sequence is masked, a series of BLAST searches is performed on the query sequence. The EXPECT threshold for all BLAST searches is set to 0.001 to prevent the number of BLAST hits from becoming overwhelmingly large; this threshold can be changed by the user as appropriate. All BLAST searches can be fully customized, and can include searches against local sequence databases.

Gene prediction

Based on comparative studies assessing the predictive power of a number of algorithms, we chose four gene/exon prediction programs to be included in the initial implementation of GeneMachine. Two of these programs, GRAIL2exons (Mural *et al.*, 1992), and MZEF (Zhang, 1997) predict single exons. The other two programs, GENSCAN (Burge and Karlin, 1997) and FGENES (Solovyev *et al.*, 1995) predict complete gene structures. Each of these programs utilize a different algorithm and/or model for the prediction of transcriptional units.

In an effort to present the user with only the most significant results and in order to keep the final output from

becoming overwhelming, cutoffs have been implemented for some of the programs; these are described in the GeneMachine documentation and can be customized by the user.

Simple repeats

Searching for new polymorphic markers is one of the most important steps in a positional cloning project, since these markers can often be used to narrow down thecritical region defined by genetic mapping. We use Sputnik (Abajian, 1994) for the detection of simple repeats. The Sputnik output is pruned such that only microsatellites of length >30 bases are included in the annotated output file.

Data flow

The flow of data through GeneMachine is shown in Figure 1 of the Web supplement to this paper. A file containing one or more FASTA-formatted sequences is submitted to GeneMachine for processing, along with any desired options. On the predictive side, calls are made to GENSCAN, FGENES, MZEF, and GRAIL2exons, with the output returned in ASN.1 format. On the comparative side, an initial masking step is performed using RepeatMasker, with the output again returned in ASN.1 format. The now-masked sequence(s) are submitted to the component BLAST algorithms for homology searching. The BLAST output is automatically returned in ASN.1 format, removing the requirement for further parsing. Finally, Sputnik is used to identify simple repeats, as described above. Once all of these steps are complete, the user can open the resulting ASN.1 file using NCBI Sequin to view the results.

User interface

GeneMachine can be run from the command line on a UNIX-based machine, or can be run using a Web interface. The Web interface is designed for ease of use by biologists who may not be comfortable running UNIX-based programs, and minimizes the number of choices that have to be made prior to invoking the GeneMachine run. The public GeneMachine Web site is located at http://genemachine.nhgri.nih.gov. From this Web page, users can simply check-off which programs should be run, as well as specifying an organism (human, rodent, *Arabidopsis*, or *Drosophila*); specifying an organism will insure that the correct reference sets are used by some of the component gene prediction programs in making their predictions. Due to restrictions imposed by the authors of some of the individual gene prediction programs implemented within GeneMachine, this site has been restricted to academic and not-for-profit users to respect the wishes of those authors.

The user has greater flexibility when running from the UNIX command line. If the user has a commonly-used set of parameters, a configuration file can be created and automatically invoked; by doing this, custom requirements do not have to be specified each time that GeneMachine is run. The system accepts either a single FASTA-formatted file or a FASTA-formatted library as input. For example, contigs from the same, unfinished sequence can be placed in a single file and run simultaneously.

The Sequin graphical viewer

The Sequin graphical viewer is freely-available from the NCBI FTP site and runs on most platforms, meaning that GeneMachine results can be examined on most computers. Furthermore, Sequin is fully integrated with GenBank, a particularly important advantage when working with sequences that have been previously deposited into the public databases. The integration allows sequence data to be automatically updated, for easy retrieval of BLAST results, and for the import and merging of existing annotations with the results produced by GeneMachine.

The Sequin graphical viewer is also shown in the Web supplement (Figure 2). In the figure, most of the predicted exons/genes from each of the methods line up with one another, and they in turn line up with the hits obtained from the BLAST searches. The details of each BLAST hit can be observed by using the Edit Alignment feature of Sequin. The alignment pops up in a separate window for inspection (Web supplement, Figure 2, upper right). In addition, the coordinates of any annotated feature, gene, exon, or repeat can be viewed by either double-clicking on the feature itself in the main window, or by changing the display format to either GenBank or to Sequence. The feature editor allows users to add or modify relevant information, as well as to add additional annotations.

REFERENCES

Abajian,C. (1994) Sputnik. http://abaijan.net/sputnik/index.html

Altschul,S.F., Madden,T.L., Schffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Mural,R.J., Einstein,J.R., Guan,X., Mann,R.C. and Uberbacher,E.C. (1992) An artificial intelligence approach to DNA sequence feature recognition. *Trends Biotechnol.*, **10**, 67–69.

Smit,A.F.A. and Green,P. RepeatMasker. ftp://ftp.genome.washington.edu/RM/RepeatMasker.html

Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Ismb*, **3**, 367–375.

Zhang,M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA*, **94**, 565–568.