Nucleic Acids Research, 2003, Vol. 31, No. 13 ***3507–3509***
DOI: 10.1093/nar/gkg583

SLAM web server for comparative gene finding and alignment

Simon Cawley*, Lior Pachter[1] and Marina Alexandersson[2]

Affymetrix Inc., 6550 Vallejo St, Suite 100, Emeryville, CA 94608, USA, [1]Department of Mathematics, UC Berkeley, USA and [2]Fraunhofer-Chalmers Centre, Gothenburg, Sweden

Received February 14, 2003; Revised and Accepted April 3, 2003

ABSTRACT

SLAM is a program that simultaneously aligns and annotates pairs of homologous sequences. The SLAM web server integrates SLAM with repeat masking tools and the AVID alignment program to allow for rapid alignment and gene prediction in user submitted sequences. Along with annotations and alignments for the submitted sequences, users obtain a list of predicted conserved non-coding sequences (and their associated alignments). The web site also links to whole genome annotations of the human, mouse and rat genomes produced with the SLAM program. The server can be accessed at http://bio.math.berkeley.edu/slam.

INTRODUCTION

The near-complete genome sequences of various mammalian organisms are now available (in decreasing order of completeness, currently human, mouse and rat) and many more are in various stages of sequencing and assembly. With this growing wealth of available sequence there is a corresponding increase in the need for comparative genomic analysis tools. SLAM (1) is an alignment and *de novo* gene finding program which aligns homologous sequences and annotates them with respect to coding and conserved non-coding regions. *De novo* programs work by identifying pattern differences between the various regions, such as coding potential, splice signals and exon length distributions. Since there is incomplete understanding of the biological mechanisms by which the cell identifies gene sequences, all current single organism *de novo* models are somewhat incomplete and as a result, most tend to have an appreciable rate of error, typically resulting in a high rate of false positives. By incorporating sequence conservation into the *de novo* predictions the false positive rate can be considerably reduced. There now exists a number of programs which essentially use a single-organism dynamic programming approach augmented with sequence conservation information to achieve such increased specificity (2–5). SLAM achieves a similar increase in specificity using a novel approach which is symmetric with respect to the organisms being annotated and performs sequence alignment at the same time. One unique feature of SLAM is that predictions are guaranteed to have the same gene structure in the homologous sequences, another is the *ab initio* prediction of conserved non-coding sequences (CNSs).

METHODS

SLAM is a cross-species gene finder that works by simultaneously aligning and identifying complete exon/intron structures in two evolutionary related but unannotated sequences of DNA. The probabilistic framework used is a generalized pair hidden Markov model (GPHMM) (6), a hybrid of generalized hidden Markov models which have been used previously for gene finding (7,8), and paired hidden Markov models which have applications to sequence alignment (9).

The SLAM program is integrated in a pipeline that takes as input two sequences and outputs a set of annotations for each sequence. The first step of the process is to identify repeats with RepeatMasker (http://repeatmasker.genome.washington.edu). Sequences with repeats already masked out can be supplied, but non-masked sequences are preferable since SLAM does not blindly mask repeats—only repeats known not to occur within coding exons are used as constraints in subsequent gene finding.

In the second step an *approximate alignment* is formed to reduce the search space of the algorithms. This is necessary since a naive implementation of a GPHMM has a running time which scales as the product of the input sequence lengths and is therefore impractical for typical sequence lengths encountered. An approximate alignment is a pre-determined subset of the alignment space, hopefully containing the true alignment between the two input sequences. The approximate alignment is created by taking a global alignment produced by AVID (10) then expanding the base-to-base alignment to intervals surrounding each base and expanding further around candidate aligned exon boundaries.

In the final step of the pipeline, the masked sequences and approximate alignment are used to search for the most likely alignment and annotation of the pair of sequences, using the Viterbi algorithm for the GPHMM. The score of each possible

*To whom correspondence should be addressed. Tel: +1 510 428 8534; Fax: +1 510 428 8585; Email: simon_cawley@affymetrix.com

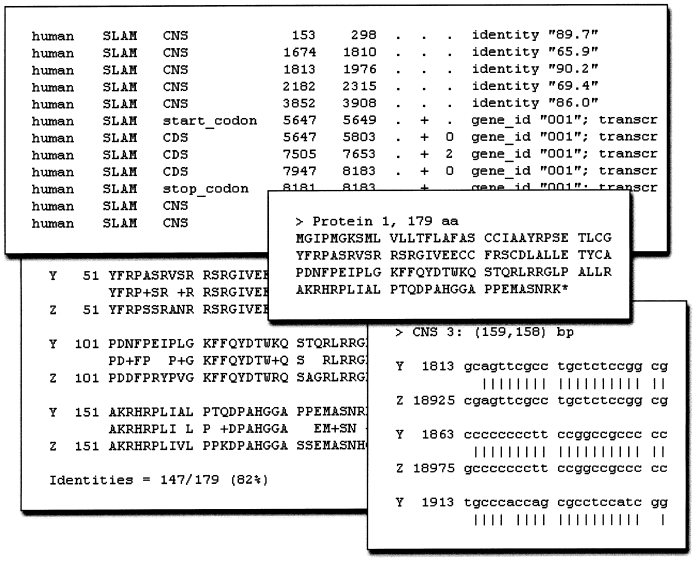

Figure 1. The SLAM output: predictions in GFF format, corresponding mRNA and peptide predictions and alignments of peptides and CNSs.

alignment/annotation combination is a probability which depends on a number of factors including statistical sequence properties (such as splice site signals and coding content) as well as amount and type (coding/non-coding) of conservation. In addition to determining the exon/intron structure, a novel feature in SLAM is the prediction of CNSs. The use of CNSs in the model reduces the false positive rate and often identifies UTRs, putative binding site regions and possibly other biologically important non-coding features. This enables the distinction between sequences that are highly conserved at the protein level and sequences conserved merely on the DNA level. The output of SLAM, illustrated in Figure 1, is a summary of the annotations along with a number of output files containing the predictions for both sequences in GFF-format, the corresponding mRNA and peptide predictions and the alignment of the peptides and the CNSs in a BLAST-like format.

The online version of SLAM currently runs in a preset default mode, the downloadable version comes with a number of additional options:

- The ability to run on a finished sequence versus a draft sequence.
- The ability to sample sub-optimal parses [which is useful to assign confidence to predictions and to predict possible alternative splicing (Cawley,S. and Pachter,L. manuscript in preparation)].
- The ability to use annotation constraints in one sequence to produce the most likely homologous prediction in the other sequence.

The running time of SLAM depends on the similarity of the input sequences and their length. Both these factors affect the size of the approximate alignment. Typically, running on a pair of BAC-sized human and mouse sequences the server responds in a few minutes. There is no restriction on input lengths, but long sequences (>100 Kb on the web server) will be cut up into smaller pieces.

RESULTS

With the draft sequences for both human, mouse and rat available, it is now possible to perform both two way and three way whole genome comparisons. SLAM was one of the programs used in the Mouse Sequencing Consortium analysis of the mouse draft sequence (11) and the results of our whole genome runs (human–mouse, human–rat) can be browsed and downloaded from the SLAM web site. The UCSC genome browser provides an excellent environment for comparing various sources of information (12) and it is used as a basis for browsing the SLAM whole genome results.

DISCUSSION

The symmetry of the SLAM method is based on the assumption that, in addition to having good alignment, exons must conform to a conserved gene structure, have consistent ORFs, matching splice sites and similar exon lengths. This results in reliable predictions with major improvements in boundary detection and false positive rates for cases where the assumptions hold true. On the other hand, these assumptions can be a disadvantage in cases where they do not hold—problematic situations include the insertion of introns or exons in one organism and differences in the order of genes between the sequences. The former issue can be addressed by changes in the underlying Markov model, the latter by more involved pre-processing of the input sequences, both of these issues remain as areas for future work. The issue of conserved gene order has not been a problem in the analysis of human, mouse and rat—for instance, human and mouse tend to have gene order preserved in chunks of up to 8 Mb on average (13).

The model's ability to distinguish between conserved coding and conserved non-coding sequence has led to further improvements in accuracy, moreover, the CNS feature holds great promise of identifying biologically important non-coding

features and will be an important part of future comparative studies.

The SLAM code is freely available for academic and non-profit purposes and can be downloaded from our web site.

ACKNOWLEDGEMENTS

We thank Nicolas Bray for helping to set up the web server site. Colin Dewey has helped in performing and analyzing the SLAM whole genome runs. L.P. and S.C. are partially supported by a grant from the NIH (ROI-HG02362-01). M.A. is supported by the Swedish Foundation for Strategic Research.

REFERENCES

1. Alexandersson,M., Cawley,S. and Pachter,L. (2003) SLAM—cross-species gene finding and alignment with a generalized pair hidden Markov model. *Genome Res.*, **13**, 496–502.
2. Batzoglou,S., Pachter,L., Mesirov,J.P., Berger,B. and Lander,E.S. (2000) Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, **10**, 950–958.
3. Bafna,V. and Huson,D.H. (2000) The conserved exon method for gene finding. *ISMB'00: Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (2000)*, 3–12.
4. Korf,I., Flicek,P., Duan,D. and Brent,M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics*, **1**, S1–S9.
5. Parra,G., Agarwal,P., Abril,J.F., Wiehe,T., Fickett,J.W. and Guigó,R, (2003) Comparative gene prediction in human and mouse. *Genome Res.*, **13**, 108–117.
6. Pachter,L., Alexandersson,M. and Cawley,S. (2002) Applications of generalized pair hidden Markov models to alignment and gene finding problems. *J. Comp. Biol.*, **9**, 389–399.
7. Burge,C. and Karlin,T. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
8. Reese,M.G., Kulp,D., Tammana,H. and Haussler,D. (2000) Genie—Gene finding in *Drosophila melanogaster. Genome Res.*, **10**, 529–538.
9. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis.* Cambridge University Press, Cambridge, UK.
10. Bray,N., Dubchak,I. and Pachter,L. (2003) AVID: a global alignment program. *Genome Res.*, **13**, 97–102.
11. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
12. Karolchik,D., Baertsch,R., Diekhans,M., Frey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J., *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
13. Mural,R., Adams,M.D., Myers,E.W., Smith,H.O., Miklos,G.L., Wides,R., Halpern,A., Li,P.W., Sutton,G.G., Nadeau,J. *et al.* (2002) A comparison of whole-genome-shotgun-derived mouse chromosome 16 and the human genome. *Science*, **296**, 1667–1671.