BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 11 2000
Pages 1040–1041

MaskerAid: a performance enhancement to RepeatMasker

Joseph A. Bedell*,‡, Ian Korf* and Warren Gish†

Genome Sequencing Center and Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA

Received on May 22, 2000; revised on July 18, 2000; accepted on July 21, 2000

Abstract

Summary: *Identifying and masking repetitive elements is usually the first step when analyzing vertebrate genomic sequence. Current repeat identification software is sensitive but slow, creating a costly bottleneck in large-scale analyses. We have developed* MaskerAid*, a software enhancement to RepeatMasker that increased the speed of masking more than 30-fold at the most sensitive setting.*

Availability: *On request from the authors (see http://sapiens.wustl.edu/MaskerAid).*

Contact: *maskeraid@watson.wustl.edu*

Main Text

Interspersed repetitive elements such as SINES and LINES represent a large fraction of vertebrate genomes. Current estimates suggest at least 40% of the human genome is repetitive (data not shown). The origin, evolution, and distribution of repetitive elements in the human genome have been a subject of intense study, both experimentally and computationally (Smit, 1996, for review). Identifying and masking of these repeats is often performed as a prelude to running gene prediction and database similarity search methods, to avoid false-positive results and to accelerate downstream computational steps. An effective program for identifying and masking repeats is RepeatMasker (A.Smit, unpublished; http://www.genome.washington.edu/uwgc/analysistools/repeatmask.htm), which searches through curated repeat databases using the alignment program CrossMatch (P.Green, unpublished; http://www.genome.washington.edu/uwgc/analysistools/swat.htm).

While the problem of repeat identification seems largely solved by RepeatMasker, it often consumes a large fraction of the total CPU time spent analyzing a sequence. At its most sensitive setting, RepeatMasker took an average of 55 min to mask typical human genomic clones on a 400 MHz processor (Table 1), with the vast majority of the time spent in CrossMatch. At this rate it would take more than 2 CPU-years to mask the entire human genome. Due to the expense involved in masking this volume of data, we developed an enhancement to RepeatMasker, called *MaskerAid*, that markedly increased the speed of masking while effectively maintaining sensitivity. MaskerAid acted as a software 'wrapper' around WU-BLAST (W.Gish, unpublished; http://blast.wustl.edu), to allow transparent replacement of CrossMatch by WU-BLAST. No changes to RepeatMasker itself were made.

To test the effectiveness of incorporating MaskerAid, we ran RepeatMasker with and without MaskerAid (respectively 'CrossMatch' and 'MaskerAid' in Table 1), using the four sensitivity modes of RepeatMasker ('slow', 'standard', 'quick', and 'qq' in Table 1). Execution times were averaged and bases masked were compared to RepeatMasker run in its slow, sensitive mode ('Relative Speed' and 'Comparison of Masked Sequences' in Table 1) on a Sun Microsystems E3500 (four 400 MHz UltraSPARC-II processors). Our test set of sequences consisted of 20 human genomic clones averaging 146 348 bp. Ten of the clones were finished sequences that form a contig on chromosome 7. The remaining ten were randomly chosen draft quality sequences, typical of what is expected to dominate vertebrate genomic sequence in the near future.

Overall, more than 98% of the bases were tagged identically using MaskerAid ('Same' in Table 1). At its most sensitive setting, MaskerAid helped RepeatMasker find 0.52% more repeat sequence ('Extra' in Table 1), but led it to miss 1.2% of the repeat sequence normally found ('Missed' in Table 1). These small discrepancies were apparently due to algorithmic differences between WU-BLAST and CrossMatch and generally involved repeats at the borderline of statistical significance. To avoid these differences entirely would require targeting future development of WU-BLAST specifically as a CrossMatch replacement.

Longer repetitive regions found by native RepeatMasker tended to be reported as multiple, shorter segments when

*These authors contributed equally to this work.
† To whom correspondence should be addressed.
‡ Current address: Incyte Genomics, 4633 World Parkway Cir., St. Louis, MO 63134 USA, jbedell@incyte.com

Table 1. Comparison of RepeatMasker with CrossMatch vs MaskerAid

	Conditions		Overall performance			Comparison of masked sequences*		
	Setting	CPUs	Real time	Relative speed*	Fraction masked	Same†	Missed	New
CrossMatch	slow	1	3340	1×	41.5%	–	–	–
	standard	1	1224	2.7×	40.8%	99.2%	0.8%	0.02%
	quick	1	244	13.7×	38%	96.3%	3.6%	0.12%
	qq	1	64	52×	35.4%	93.6%	6.3%	0.13%
MaskerAid	slow	1	92	36×	40.9%	98.3%	1.2%	0.52%
	slow	2	73	46×	”	”	”	”
	slow	4	65	52×	”	”	”	”
	standard	1	78	43×	40.4%	98.1%	1.53%	0.37%
	quick	1	73	46×	39%	96.9%	2.8%	0.26%
	qq	1	77	43×	37%	95%	4.78%	0.2%

*Comparisons made to RepeatMasker run natively with CrossMatch at the slow setting.
†Total bases called identically (masked + unmasked).

MaskerAid was used (data not shown). While essentially maintaining sensitivity, MaskerAid improved the speed of RepeatMasker in its most sensitive, slow mode by a factor of 36. Additional speed was obtained on multiprocessor computers where one can exploit the multi-threaded design of WU-BLAST (CPUs in Table 1), however MaskerAid-specific operations, such as post-processing of alignments, continued to run single-threaded. When operating RepeatMasker in its fastest but least sensitive mode (‘qq’ in Table 1), the native program with CrossMatch actually ran faster (64 vs 77 s), although the MaskerAid replacement did tag more repetitive sequence (35 vs 37%). The fastest single-CPU setting using MaskerAid appeared to be ‘quick’. Similar speed-ups were observed in tests performed on a Compaq ES40 (dual 533 MHz Alpha 21264) and an Intel x86 clone (dual 450 MHz Pentium II). The increase in speed afforded by MaskerAid allows the entire human genome to be masked in less than 1 CPU-month, at the most sensitive setting of RepeatMaster.

Methods

Many compensable differences exist between the heuristic CrossMatch and WU-BLAST programs. Scoring matrices and databases are converted into BLAST format during MaskerAid installation, using bundled scripts. Subsequent invocations of MaskerAid then operate as follows: (1) CrossMatch arguments are translated into their WU-BLAST equivalents where possible; (2) WU-BLAST is executed and the report is parsed; (3) the results are post-processed to add alignment attributes present in CrossMatch but absent from WU-BLAST output; and (4) a report is generated similar to what CrossMatch would produce.

MaskerAid was developed under UNIX and is written in PERL. MaskerAid requires WU-BLAST 2.0 with its flexible parameter set; the version of WU-BLAST used here was dated 23-Apr-2000. An earlier version, dated 05-Feb-1998, worked satisfactorily in limited testing with MaskerAid, but may not be robust. This version is however freely available to download from http://blast.wustl.edu, whereas newer versions must be licensed (free for academic/nonprofit use). MaskerAid is unable to work with NCBI-BLAST (Altschul *et al.*, 1997) due to inherent limitations on word length settings, score thresholds, and scoring matrices. Replacement of CrossMatch with MaskerAid was tested using the 4-Apr-2000 version of RepeatMasker and the 31-Mar-2000 version of the Repbase Update repeat sequence database (copyright Genetic Information Research Institute).

The NCBI gi identifiers for the sequences used here were: 2842788, 3242759, 2588631, 2781383, 2275190, 3212967, 2337879, 2337878, 6358866, 6139277, 6289239, 6447142, 6139245, 6524328, 6289232, 6139227, 6447128, 6139182, 1669367, 1809226.

Acknowledgements

The authors thank Arian Smit for discussions and feedback on this work. This work was supported by NIH grant P50HG01458, with equipment support from Compaq Computer Corporation and Sun Microsystems, Inc.

References

Altschul,S.F., Madden,T.L., Schaffer,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.

Smit,A.F. (1996) The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.*, **6**, 743–748.