BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 24 2006, pages 3096–3098
doi:10.1093/bioinformatics/btl474

Sequence analysis

GARD: a genetic algorithm for recombination detection

Sergei L. Kosakovsky Pond*, David Posada[1], Michael B. Gravenor[2], Christopher H. Woelk and Simon D.W. Frost

Department of Pathology, University of California San Diego, La Jolla, CA 92093, USA, [1]Departamento de Bioquímica, Genética e Inmunología, Facultad de Biología, Universidad de Vigo, Vigo 36310, Spain and [2]School of Medicine, University of Wales, Swansea, UK

Received on July 3, 2006; accepted on September 2, 2006
Advance Access publication November 16, 2006
Associate Editor: Christos Ouzounis

ABSTRACT

Motivation: Phylogenetic and evolutionary inference can be severely misled if recombination is not accounted for, hence screening for it should be an essential component of nearly every comparative study. The evolution of recombinant sequences can not be properly explained by a single phylogenetic tree, but several phylogenies may be used to correctly model the evolution of non-recombinant fragments.

Results: We developed a likelihood-based model selection procedure that uses a genetic algorithm to search multiple sequence alignments for evidence of recombination breakpoints and identify putative recombinant sequences. GARD is an extensible and intuitive method that can be run efficiently in parallel. Extensive simulation studies show that the method nearly always outperforms other available tools, both in terms of power and accuracy and that the use of GARD to screen sequences for recombination ensures good statistical properties for methods aimed at detecting positive selection.

Availability: Freely available http://www.datamonkey.org/GARD/

Contact: spond@ucsd.edu

1 INTRODUCTION

Recombination can have a profound impact on the evolutionary process and is of interest in its own right. In HIV-1, for instance, recombination rates can rival mutation rates (Zhuang *et al*., 2002). Recombination can adversely affect the power and accuracy of fundamentally important tools of molecular evolutionary analyses: phylogenetic reconstruction (Posada and Crandall, 2002), molecular clock inference (Schierup and Hein, 2000) and the detection of positively selected sites (Shriner *et al*., 2003). Consequently, reliable tools for discovering recombination are a critical part of any phylogenetic analysis. A diverse array of algorithms and software tools for detection of recombination have been published. However, when benchmarked on simulated (Posada and Crandall, 2001) and biological (Posada, 2002) data, the methods often gave contradictory results, and no definitive recommendation on which approach should be considered the 'gold standard' could be made. We developed a robust and extensible approach—Genetic Algorithm Recombination Detection (GARD)—to screen multiple sequence alignments for evidence of phylogenetic incongruence, identify the number and location of breakpoints and sequences involved in putative recombination events. Using simulated and biological datasets we have shown (Kosakovsky Pond *et al*., 2006) that GARD outperforms the best currently available tools in terms of power and accuracy in a wide-range of evolutionary scenarios.

*To whom correspondence should be addressed.

2 METHODS AND ALGORITHMS

We model recombinant sequences by allowing $S \geq 1$ non-recombinant alignment fragments, reconstructing a separate phylogenetic tree for each fragment and evaluating the goodness-of-fit for the model using the small sample Akaike's Information Criterion (Sugiura, 1978) computed with standard phylogenetic likelihood methods and point substitution models [see Kosakovsky Pond *et al*. (2006) for details]. The computationally challenging component of the model is the search for the locations of $S-1$ breakpoints—a problem of $O(L^S)$ complexity (L denotes the length of the alignment). When $S=2$, an exhaustive examination of all possible locations for the single breakpoint can be undertaken. This single breakpoint (SBP) method performs surprisingly well (Kosakovsky Pond *et al*., 2006) when a dichotomous classification of alignments into recombinant or non-recombinant is desired, and can be run quickly in a parallel computing environment.

When $S>2$, we utilize an aggressive population based hill-climber—the CHC genetic algorithm (Eshelman, 1991)—to search the space of breakpoint locations, encoded as a binary vector of sorted concatenated breakpoint positions. CHC always retains the most fit individual from the previous generation and performs two basic operations on individuals currently in the population:

(1) When two individuals, $\mathbf{b}_1$ and $\mathbf{b}_2$ are picked to mate, their offspring is equally likely to inherit bit b_i from either parent.

(2) If the diversity of the sample (measured by the range of AIC_c scores normalized by the score of the best individual) falls below a fixed threshold, then all individuals in the population, excluding the most fit one, have a proportion of randomly selected bits toggled.

For fixed S, the algorithm terminates if the best score remains unchanged over 100 consecutive generations. A typical GA run considers 10^3–10^4 possible models before converging. To infer S, we start with $S=1$ segments and increase S by 1 for subsequent GA runs, until the AIC_c score of the best model fails to improve further. GARD and SBP have been implemented as HyPhy (Kosakovsky Pond *et al*., 2005) language scripts enabled to run in an MPI environment. Presently, GARD is hosted on our 80-node cluster and can be accessed via a Web front-end. Standalone scripts or cluster installation instructions can be obtained from the authors upon request and will be made available online

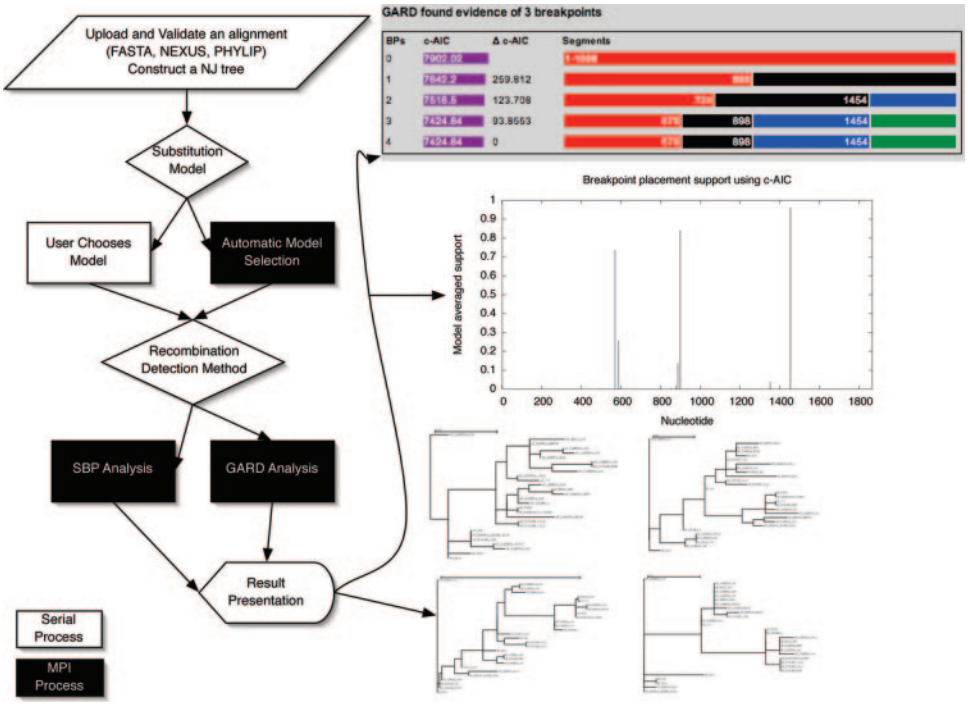

Fig. 1. GARD and SBP server schematic flowchart and sample output.

if there is sufficient interest. The current implementation, shown schematically in Figure 1, allows the user to:

(1) Upload an alignment of sequences to screen. At present up to 50 aligned DNA/RNA sequences with up to 10 000 nt will be accepted. Both numbers will be increased periodically.

(2) Select an appropriate model of nucleotide evolution (Kosakovsky Pond and Frost, 2005) and specify the distribution used to model site-to-site variation in substitution rates.

(3) Run SBP or GARD screens for recombination.

(4) Visualize and download the results of recombination screens, including: (i) the number and best location of inferred breakpoints, and the improvement in AIC_c score achieved by the multiple breakpoint model (if any); (ii) model averaged supportfor the location of breakpoints, useful for assessing the degree of confidence; (iii) phylogenetic trees inferred from each non-recombinant breakpoint; (iv) a NEXUS file containing the alignment, inferred partitions and trees.

(5) Result files and HyPhy scripts needed for additional processing and inference (e.g. for further tests of phylogenetic incongruence) can be downloaded and run locally.

We intend to add new features and analysis options (e.g. protein sequence analysis) with time.

3 DISCUSSION

In practice many widely-used molecular analyses may be confounded by the presence or absence of recombination. Hence, screening for recombination should be an integral part of phylogenetic analyses. We have developed an intuitive and powerful method for detecting evidence of recombination in alignments of DNA sequences. It is able to provide estimates for the number and location of breakpoints, and infer segment-specific phylogenetic trees. GARD does not require a non-recombinant reference alignment and recombination between ancestral sequences is also accommodated. Arbitrarily complex models of point substitution (e.g. those allowing site-to-site variation in substitution rates, or codon models) can be easily incorporated. GARD outperforms other methods and can be run in parallel on a cluster of computers, and so is well suited to screen for recombination in large datasets.

ACKNOWLEDGEMENTS

This research was supported in part by the National Institutes of Health (AI43638, AI47745, and AI57167, R01-GM66276), the University of California Universitywide AIDS Research Program (IS 02-SD-701), and by a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to SDWF and SLKP (AI36214). D.P. was also supported by grant BFU2004-02700 of the Spanish Ministry of Education and Science and by the 'Ramón y Cajal' program of the Spanish government.

Conflict of Interest statement. None declared.

REFERENCES

Eshelman,L.J. (1991) The CHC adaptive search algorithm: How to do safe search when engaging in nontraditional genetic recombination. In Spatz, B.M. (ed.), *Foundations of Genetic Algorithms,* Morgan Kaufmann, San Mateo, CA, pp. 265–283.

Kosakovsky Pond,S.L. and Frost,S.D.W. (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics*, **21**, 2531–2533.

Kosakovsky Pond,S.L. (2005) HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, **21**, 676–679.

Kosakovsky Pond,S.L. *et al.* (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Mol. Biol. Evol.*, **23**, 1891–1901.

Posada,D. and Crandall,K.A. (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc. Natl. Acad. Sci. USA*, **98**, 13757–13762.

Posada,D. and Crandall,K.A. (2002) The effect of recombination on the accuracy of phylogeny estimation. *J. Mol. Evol.*, **54**, 396–402.

Posada,D. (2002) Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol. Biol. Evol.*, **19**, 708–717.

Schierup,M. and Hein,J. (2000) Recombination and the molecular clock. *Mol. Biol. Evol.*, **17**, 1578–1579.

Shriner,D. *et al.* (2003) Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet. Res.*, **81**, 115–121.

Sugiura,N. (1978) Further analysis of the data by Akaike's information criterion and the finite corrections. *Comm. Stat. Theory Meth.*, **A7**, 13–26.

Zhuang,J. *et al.* (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J. Virol.*, **76**, 11273–11282.