ParIS Genome Rearrangement server

I. Miklós[1], P. Ittzés[2] and J. Hein[3]

[1]Theoretical Biology and Ecology Modelling Group, Hungarian Academy of Sciences and Eötvös Loránd University, Pázmány Péter Sétány 1/c. H-1117 Budapest HUNGARY, [2]Collegium Budapest, Institute for Advanced Study, Szentháromság u. 2. H-1014 Budapest HUNGARY and [3]Genome Analysis and Bioinformatics Group, Oxford Centre for Gene Function, Department of Statistics, Oxford University, 1 South Parks Road, OX1 3TG Oxford, UK

ABSTRACT

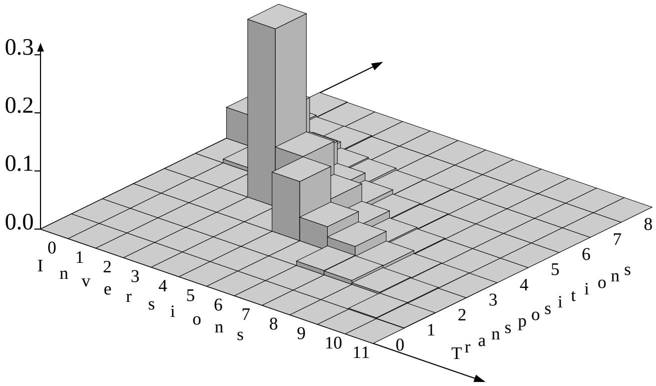
Summary: ParIS Genome Rearrangement is a web server for a Bayesian analysis of unichromosomal genome pairs. The underlying model allows inversions, transpositions and inverted transpositions. The server generates a Markov chain using a Partial Importance Sampler technique, and samples trajectories of mutations from this chain. The user can specify several marginalizations to the posterior: the posterior distribution of number of mutations needed to transform one genome to another, length distribution of mutations, number of mutations happened at a given site. Both text and graphical outputs are available. We provide a limited server, a downloadable unlimited server which can be installed locally on any linux/Unix operating system, and a database of mitochondrial gene orders.

Availability: http://www.colbud.hu:8765/paris.html and http://doob.stats.ox.ac.uk/˜miklos/paris.html

Contact: miklosi@ramet.elte.hu

Genome rearrangement events consists of not only inversions and (for multichromosomal genomes) translocations but transpositions and inverted transpositions, too. Unfortunately, parsimony algorithms which find the minimum number of mutations needed to transform one genome to another are available only for inversions and translocations (Hannenhalli , 1996; Hannenhalli & Pevzner, 1999; Tesler, G. , 2002). Algorithms considering all type of mutations are only approximations (Blanchette *et al.*, 1996; Gu *et al.*, 1999; Eriksen , 2001). We started a Bayesian approach for the problem of rearranging genomes by inversions, transpositions and inverted transpositions (Miklós , 2003), which combines the methods of Blanchette *et al.* (1996) and Larget *et al.* (2002). The former method considers all type of mutations, however, it is a greedy approximation, hence it is hard to statistically describe the distribution it gives. The later method is a Bayesian approach yielding a statistically well-defined posterior distribution, however, it considers only inversions. Our Bayesian approach is based on an evolutionary model allowing all type of mutations. We defined a Markov chain converging to the posterior distribution of trajectories of mutations transforming a genome to another, and empirically showed that the convergence was fast and the posterior distribution of number of mutations reasonably predicted the true number of mutations happened.

Fig. 1. The joint posterior distribution of number of inversions and transpositions rearranging fluke mitochondrial genomes of *Paragonimus westermani* and *Schistosoma japonicum*.



Several changes has been made since the first version of our MCMC sampler to make it faster and easier to use. The most remarkable change is that in the new version, we change only a part of the actual trajectory in the Markov chain. Although it provides slower mixing per accepted step in the Markov chain, the acceptance ratio is greater in this sampler, and thus, the overall performance became better. Minor algorithmic changes were also implemented, and hence a sampling step is performed about 10-15 times faster than in the first version.

The method has been integrated into a web server. We provide an on-line server, for which the size of the genomes and the length of the Markov chain is limited, and users are not allowed to submit a job until their previous job has finished. However, we offer a downloadable version, too, which can be installed

Fig. 2. The posterior length distribution of inversions and transpositions rearranging fluke mitochondrial genomes of *Paragonimus westermani* and *Schistosoma japonicum*.

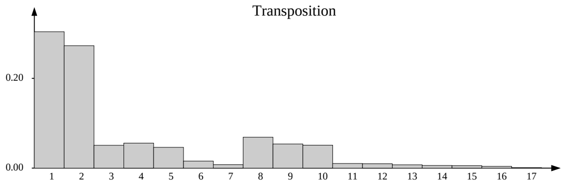
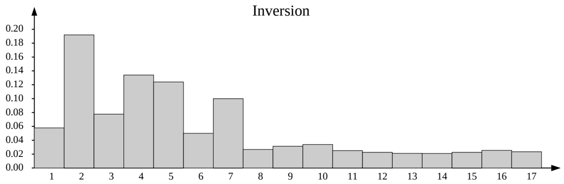
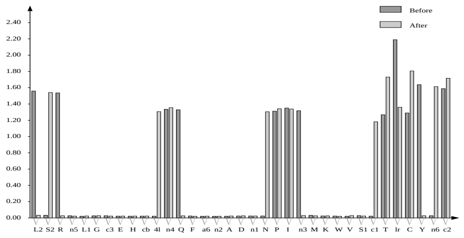



Fig. 3. The expected number of mutations happened before and after genes in mitochondrial genomes of *Paragonimus westermani* and *Schistosoma japonicum*.

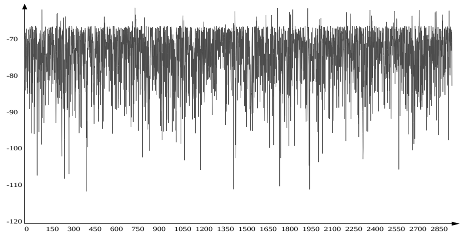


locally to the user's computer without such limitations. A database of gene orders in mitochondrial genomes is also downloadable.

The user does not need to transform genomes into signed permutations, the server automatically select the intersection of common genes in the two genomes and generates the corresponding signed permutation. The server provides the following statistics in text, .pdf and .ps files:

- The joint posterior distribution of number of inversions and transpositions happened. (See for example Fig 1).
- The posterior distribution of length of inversions and transpositions (See for example Fig 2).
- The posterior distribution of number of events happened at a site (See for example Fig 3).
- The log likelihood trace (See for example Fig 4).
- The trajectories sampled by the Markov chain. (See a sample trajectory in Table 1)

Fig. 4. The log-likelihood trace from the analysis of mitochondrial genomes of *Paragonimus westermani* and *Schistosoma japonicum*.



```
2 2 4 33|0 11 13 21|3 3 34 |0 16 17 27|2 4 7 35|2 0 2 34|
```

Table 1. A shortest trajectory transforming mitochondrial genome *Paragonimus westermani* to that of *Schistosoma japonicum*. Mutations are separated by '|'. The first number indicate the type of mutation, 0: transposition, 1: inverted transposition where the right block is inverted, 2: inverted transposition where the left block is inverted, 3: inversion. Following numbers indicate the positions of breakpoints of that mutation.

The MCMC sampler and postscript generating algorithms are written in C. The web interface is written in php 4.3. The php code generates a Perl script, which runs in background and invokes the compiled C codes. In the downloadable unlimited version, we provide details about how to use the compiled C codes in command line, which helps the user to write short scripts driving long runs of investigations.

ACKNOWLEDGEMENTS

This work was funded by EPSRC grant HAMJW and MRC grant HAMKA. I.M. was also supported by a Békésy György postdoctoral fellowship. We thank the Collegium Budapest for providing a webserver for this project. Special thanks to Susan Hutchinson and Irmtraud M. Meyer for their help.

REFERENCES

Blanchette,M., Kunisawa,T. and Sankoff,D. (1996) Parametric genome rearrangement. *Gene*, **172**, GC11–GC17.

Eriksen,N. (2001) (1+ε)-approximation of sorting by reversals and transpositions. *In: Proceedings of WABI2001, LNCS*, **2149**, 227–237.

Gu,Q-P., Peng,S. and Sudborough,I.H. (1999) A 2-Approximation Algorithm for Genome Rearrangements by Reversals and Transpositions. *Theor. Comp. Sci.*, **210(2)**, 327–339.

Hannenhalli,S. (1996) Polynomial algorithm for computing translocation distance between genomes. *In: Proceedings of CPM1996*, 168–185.

Hannenhalli,,S. and Pevzner,P.A. (1999) Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals. *Journal of ACM*, **46(1)**, 1–27.

Larget,B., Simon,D.L. and Kadane,J.B. (2002) Bayesian phylogenetic inference from animal mitochondrial genome arrangements. *J. Roy. Stat. Soc. B.*, **64(4)**, 681–695.

Miklós,I. (2003) MCMC Genome Rearrangement. *Bioinformatics*, **19**, ii130–ii137.

Tessler,G. (2002) GRIMM: genome gearrangements web server. *Bioinformatics*, **18**, 492–493.