Comparative Analysis of Genomic Rearrangements in Microorganisms

Siv G. E. Andersson[1]
Siv.Andersson@ebc.uu.se
Daniel Dalevi[1]
Daniel.Dalevi@ebc.uu.se
Niklas Eriksen[2]
niklas@math.kth.se
Kimmo Eriksson[3]
Kimmo.Eriksson@mdh.se

[1] Department of Molecular Evolution, EBC, Uppsala University, Norbyvägen 18C, SE-752 36 Uppsala, Sweden
[2] Department of Mathematics, KTH, SE-100 44 Stockholm, Sweden
[3] IMa, Mälardalens högskola, Box 883, SE-721 23 Västerås, Sweden

Keywords: comparative analysis, genome rearrangements, GC content, *Chlamydia*, Derange II

1 Introduction

The theoretical genome rearrangement problem (sorting a permutation by transpositions and inversions) has attracted a good deal of attention from mathematicians and computer scientists, and comparative genome sequence data from closely related microbial strains and species are now accumulating at a rapid pace. Detailed inspection of this data will help us understand the evolutionary forces that determine genomic structures and stabilities in microbial systems. Important questions concern the extent to which microbial trees contain inconsistent branches caused by horizontal transfer events of individual genes. With the advent of completely sequenced data from closely related genomes, we expect many previously unknown properties of genome rearrangements to be uncovered, opening new questions for mathematicians and biologists alike.

The whole genome sequences of two chlamydian parasites, *Chlamydia trachomatis* and *Chlamydia pneumoniae*, have recently been published [1, 3, 4]. *Chlamydia* are obligate intra-cellular parasites, responsible for a variety of diseases in non-human mammals and birds, but also frequently transmitted to humans. For example, *C. trachomatis* is the causative agent of trachoma, a major cause of blindness in Asia and Africa and *C. pneumoniae* causes pneumonia and bronchitis. A comparative analysis of rearrangement events in the two genomes has revealed a large proportion of transpositions and inversions of segments of very small length, consisting of just one or two genes. This phenomenon calls for modified methods for estimating the number of different rearrangement events. In this project we have modified the Derange II computer program of Blanchette et al. [2]. We then use a simulation technique for determining appropriate parameters.

2 Method and Results

Calculations performed on the *Chlamydia* data by Derange II show that a significant part of the operations needed to restore the gene order are very local. In fact, about half of the inversions operate on a single gene, and about half of the transpositions operate on two neighbourly genes. There is also an overrepresentation of operations involving just a few genes. This behaviour was unexpected,

Niklas Eriksen was supported by NFR. Kimmo Eriksen was partially supported by NFR and SSF.

since no mechanism explaining such a bias for local rearrangements is known. To take this bias into account, we have modified Derange II so that short operations can be treated separately, with lower weights than longer operations. The problem of finding appropriate weights for the different types of events has been raised previously [2], but no definitive solutions have been proposed. We suggest the following simulation scheme to overcome this problem.

First we obtain a collection of simulated data sets of genome rearrangements of relevant size, by using both long and short operations at given proportions. In order to make data comparable with the real *Chlamydia* data, we use a string of 800 genes and make a sequence of events until there are 350 breakpoints. We vary the proportions of long and short inversions, transpositions and transversions to get a collection of simulated data of known properties. Then we run this through Derange II, using varying sets of weights. For some fixed set of weights, Derange II responds with the same proportions of operations as the ones used in the simulations. We take this as a good indication that this set of weights is the correct one to use. The model will be further refined so that there is a continuously decreasing rate of operations as a function of segment length.

In order to fully understand the dynamics of microbial genomes, we also need to understand the balance as well as the processes whereby new genes are being acquired and old genes are being removed. The finding that *C. pneumoniae* has 214 protein genes that are not present in *C. trachomatis* suggests that there is a significant rate of gene turnover in *Chlamydia* species. The unique genes may either have been introduced into one of the species rather recently or alternatively lost in the other species. One way of identifying recently introduced genes is by searching for genes with atypical base composition patterns. However, we show that the variation in GC content at third codon synonymous sites among genes in the two *Chlamydia* species is no greater than would be expected by chance, suggesting that there are no or few examples of recently introduced genes in the *Chlamydia* genomes. Accordingly, we suggest that a majority of the genes uniquely present in one genome have been lost from the other genome.

3 Discussion

There is a very strong bias towards extremely short rearrangement events in *Chlamydia* genomes. Thus, we may ask: What is the biological explanation for this bias? Are the rates of transposition and inversions related to segment lengths, and if so why? Is this a general phenomenon for microbial as well as for eukaryotic genomes, and if so then previous estimates of rearrangement frequencies based on fragments of restricted length may prove to be incorrect. Resolution of these issues can only be obtained by rigorous comparative analyses of closely related strains and species.

References

[1] Benson, D.A. *et al.*, Genbank, *Nucleic Acids Res.*, 28:15–18, 2000.

[2] Blanchette, M. Kunisawa, T., and Sankoff, D., Parametric genome rearrangements, *Gene*, 172:GC 11-17, 1996.

[3] Kalman, S. *et al.*, Comparative genomes of *Chlamydia pnemoniae* and *Chlamydia trachomatis*, *Nature Genetics*, 21:385–389, 1999.

[4] Stephens, R.S. *et al.*, Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*, *Science*, 23:754–759, 1998.