

1 Abstract

All living things are related. Understanding and describing the evolutionary relationship between organisms and ancestral relationship between individuals is at the heart of biology. This holds no less true in the field of bioinformatics. First, the new methods and genetic sequence data of bioinformatics have made significant contributions to this understanding. Secondly, bioinformatics has mostly been successful when comparative analysis can be brought to bear, e.g. protein folding by homology modelling, comparative gene finding, and disease mapping. Needless to say, an essentially correct understanding of the evolutionary relationships of a data set is a prerequisite for comparative methods to work. Most standard methods for inferring evolutionary relationships assumes a phylogenetic relationship globally conserved throughout the sequences, an assumption that breaks down in the context of recombinations. This proposal concerns novel methods capable of exact evolutionary analysis in the context of recombinations under well established models from population genetics. Preliminary work under a model suitable for the study of slow evolving organisms has shown the feasibility of the proposed methods. This project will further develop this software, especially in the areas of human and Internet interfaces. Furthermore, we will extend the techniques to a more complex model required for studies of fast evolving organisms like pathogens. We will work closely together with local pathogen researchers in applying the software to newly emerged strains of pathogens.

2 Background

The rise of sequence data in biology has been accompanied by a corresponding increase in our ability to infer the evolution of genomes and organisms during the last few decades. The revolution has been driven by both data and computation. Often analysis has been computationally limited: First investigating all phylogenies seemed a major obstacle, then combining phylogenetic analysis with alignment was a challenge, and finally using statistical modelling instead of combinatorial optimisation increased the computational burden 2-3 orders of magnitude. There are important areas of analysis that are still computationally limited, requiring further algorithmic advances. A key example is the analysis sequences undergoing recombination, where the presence of recombination undermines the methods based on finding a traditional phylogeny and necessitates more complicated descriptions of the ancestral relationship structure. Rather than being a feature of sexual reproduction with the occasional rare occurrence in other contexts, recombination seems to be an ubiquitous force driving much of the diversity observed in populations of organisms, including pathogens[2].

At our current stage of understanding, a single genetic sequence is usually of little use. We do not have the insight to identify all functional units of the sequence, let alone to predict their associated function, without detailed prior knowledge. So most information extracted from genetic sequences is obtained by comparative studies, analysing variation and conservation in a set of related genetic sequences. An essential prerequisite for comparative analyses is an understanding of the ancestral relationships among

the sequences.

Present day methods for analysing sets of homologous sequences usually assume the sequences to be phylogenetically related, i.e. the ancestral relationship can be viewed as a hierarchical clustering of the sequences. In cases with strong evidence of more complicated relationships, the sequences can be broken into segments adhering to the phylogenetic relationship assumption that are then analysed independently, but this is usually based on manual decisions. Automated methods for recombination analysis has only recently become feasible on standard equipment with advances in computational power, with the first exact inference of parsimonious recombination histories, or ancestral recombination graphs[3, 6] (ARGs), for the 1983 Kreitman data set[8] only being published in 2003[11]. Even this analysis stretched standard equipment to the limit and assumed a restricted model of substitutions only valid for relatively slowly evolving organisms.

With sequencing techniques reaching a level where small genomes can be sequenced as a matter of routine, it can be anticipated that the full genome sequences of novel pathogen strains or emergent diseases aggressively affecting humans, livestock, crops, or even pets will be available almost immediately after identification of the pathogen, due to their huge health and economic impacts. Automated assistance is crucial for the speedy analysis of newly sequenced genome sequences that can ensure a rapid response to new pathogen strains. Alignment to known close homologues will be straight forward, with mostly single nucleotide substitutions. The challenge will be in extracting information about the ancestral relationship between the new and known strains, and in particular about recombinations and recombination hot spots in the data.

3 Programme and Methodology

This project aims to develop new computational methods for parsimonious inference of ancestral recombination events in single nucleotide polymorphism (SNP) data. The methods will be applied to recently emerged bacterial strains. In more detail, the project entails the following subsidiary objectives

- To develop and refine our methods for ARG inference under the infinite sites model
- To extend the methods to ARG inference under the finite sites model
- To enable easy Internet access, both human and machine, to the software by developing web server access and embedding GRID functionality in the code
- To apply ARG inference under the finite sites model to pathogen analysis

In preliminary work for this proposal, we have recently been working on an initial version of a program for parsimonious analysis of SNP data under the infinite sites assumption[9]. The program uses a branch and bound method to determine the exact minimum number of recombinations required to explain a SNP data set under the infinite sites assumption. To our knowledge, only one other implementation exists for similar analysis of SNP data[11]. On a benchmark data set[8] consisting of 11 sequences with 44 segregating sites, the previous method required half an hour of

computation and 1.5GB of memory to complete the analysis. Our software analyses the same data set in less than a second using approximately 200KB of memory. We believe these preliminary studies exhibit significant potential for expanding the sizes of data sets for which exact recombination analysis is feasible. Examples of information derived from this study is illustrated in Figure 1.

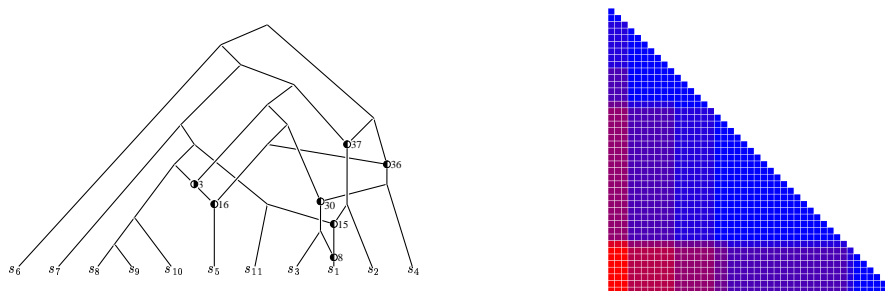


Figure 1: Two different types of inferences on the ancestral recombination history of the data set of [8]. The left hand illustration show one possible ARG with a minimum number of recombinations (7); recombinations are represented by nodes with two parents and one child, with \circ indicating that the prefix is the left parent, \bullet indicating that the prefix is the right parent, and the subscript indicating the position of the recombination. The right hand illustration shows local bounds on the minimum number of recombinations for all subregions of the data; blue corresponds to zero recombinations while red corresponds to 7 recombinations.

3.1 Models

Models for describing SNP data as a population of sequences evolving over time are well established in population genetics[5, Chapter 2]. In the context of recombination, usually three types of events are possible: i) substitution in a specific site for one of the sequences, ii) a copy of one of the sequences is added to the population, corresponding to two offsprings being present in the SNP data sample; based on a backwards-in-time perspective this event is usually called a coalescence, and iii) two sequences are replaced by a recombinant whose first part is a copy of one of the sequences and latter part is a copy of the other sequence.

3.1.1 Infinite Sites Model

The infinite sites model postulates that each site can be subject to at most one substitution. Hence, each segregating site – i.e. site containing two different nucleotides in the set of sampled sequences – is explained by a single substitution. This model is a reasonable approximation for slowly evolving sequences, e.g. of higher organisms. The data in [8] contains 2800 sites of which only 44 are segregating. Hence, we would expect only about one out of every 4,000 sites, or less than one site, to have been subject to more than one substitution. The infinite sites assumption simplifies analysis as exactly one substitution is known to have occurred in each segregating site, the total number of events being uniquely determined by the number of recombinations. Methods for finding lower bounds on the minimum number of recombinations

required[7, 10, 12] and stochastically sampling evolutionary histories[4, 1] have previously been studied for this model.

3.1.2 Finite Sites Model

For simpler organisms, e.g. like bacterial pathogens, where low fidelity replication, selection for dodging immune systems, and short generation times effects higher substitution rates, the infinite sites assumption is no longer valid. Finite sites models put no restriction on the number of substitutions that can occur in a particular site during the evolutionary history. This means that the total number of events of an evolutionary history is no longer uniquely determined by the number of recombinations in the evolutionary history. Indeed, any sample can be explained without postulating a single recombination. The problem corresponding to finding a history with a minimum number of recombinations under the infinite sites assumption thus becomes one of finding a minimum cost history, where the cost of an individual event should reflect its likelihood. Due to being combinatorially more complex lower bound techniques are a lot more sparse for finite sites models, but the methods of [10] do to some extent still apply.

3.2 Methodology

Our approach is based on branch and bound techniques. This is a general heuristic technique for reducing computation time that does not sacrifice exactness. We can search all histories by tracing back the sequence sample population one event at a time. When a history with a minimum number of recombinations is encountered it is reported. To make this feasible, three tricks are applied: i) ancestral states, i.e. sequence populations that could be ancestral to our SNP sample, are stored so that we do not repeat a search back from an ancestral state already encountered, ii) using the lower bound techniques of [7, 10] we terminate the search of a particular branch as soon as it is evident that it cannot lead to a minimum recombination history; this is the application of the branch and bound technique, and iii) some events or series of events can a priori be recognised as not leading to a minimum number of recombinations that we would not otherwise find, and can thus be eliminated from the search.

During our preliminary work we have recognised the space used to store ancestral state to be the determining factor in limiting the size of data sets our method can feasibly be applied to. The first stage of the project will thus be to further reduce the amount of memory used. The main focus for this will be item iii) in the above outline of our approach. Concurrently with this we will start embedding GRID functionality in the software. From the outset, we will collaborate with Dr. Rosalind Harding and Dr. Rory Bowden on applying the software to the study of bacterial pathogen data. The main objective of this study will be to develop our understanding of the recombination structure of these data sets, but as a side effect it will help identify and develop relevant and useful interfaces to the software core. We expect this stage of the project to be at or near completion after the first six months of the project.

The second stage of the project will be to transplant the techniques to finite sites models. The key challenges in relation to the above outline of our approach will be item ii), as finite sites models have received little previous attention for development

of lower bound techniques, and item iii), as the unrestricted substitution process leads to a vast increase in the number of possible events. We will apply the software to data sets from both bacteria and higher organisms as it becomes available, and carefully analyse and compare the results to those obtained under the infinite sites model to assess the validity of the assumptions about the applicability of the two different models for recombination analysis. During the latter part of this stage, when the core software and interface are more or less completed, we will make the methods available through a web server for easy access.

Statement of Timeliness

Sequencing of the genome of new strains of pathogens has become a matter of routine. At the same time there has been an increasing appreciation of the importance of recombinations for pathogen diversification that may be instrumental for species jumping and development of multi-resistant strains. Unravelling the recombination structure of closely related sequences is an inherently hard problem that is only now coming within the means of relatively standard computational equipment. This proposal will make a significant contribution to making exact recombination inference broadly available for data sets that have hitherto been too large to allow such analysis.

Justification of Resources

We estimate that the satisfactory completion of this project will require the full time work of a post doctoral researcher for a period of one year. The intended researcher already has extensive experience from the preliminary work and will require a minimum of formalised supervision, but the principal investigator will be actively collaborating on the project with one or more weekly meetings to discuss the progress of the project and challenges encountered. The results of the research will be presented at conferences in the UK and abroad.

References

- [1] P. Fearnhead and P. Donnelly. Estimating recombination rates from population genetic data. *Genetics*, 159, 2001.
- [2] E. J. Feil and B. G. Spratt. Recombination and the population structures of bacterial pathogens. *Annual Reviews in Microbiology*, 55:561–90, 2001.
- [3] R. C. Griffiths. Neutral two-locus multiple allele models with recombination. *Theoretical Population Biology*, 19:169–186, 1981.
- [4] R. C. Griffiths and P. Marjoram. Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, 3, 1996.
- [5] J. Hein, M. H. Schierup, and C. Wiuf. *Gene Genealogies, Variation and Evolution*. Oxford University Press, 2005.

- [6] R. R. Hudson. Properties of the neutral model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.
- [7] R. R. Hudson and N. L. Kaplan. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics*, 111:147–164, 1985.
- [8] M. Kreitman. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature*, 304(5925):412–417, Aug. 1983.
- [9] R. B. Lyngsø. Beagle. Available from www.stats.ox.ac.uk/~lyngsøe/beagle/. Parsimonious recombination analysis under the infinite sites model by branch & bound.
- [10] S. R. Myers and R. C. Griffiths. Bounds on the minimum number of recombination events in a sample history. *Genetics*, 163:375–394, Jan. 2003.
- [11] Y. S. Song and J. Hein. Parsimonious reconstruction of sequence evolution and haplotype blocks. In *Proceedings of the 3rd Workshop on Algorithms in Bioinformatics (WABI)*, pages 287–302, 2003.
- [12] Y. S. Song and J. Hein. On the minimum number of recombination events in the evolutionary history of DNA sequences. *Journal of Mathematical Biology*, 48, 2004.