Letter to the Editor

Recombination Analysis Using Directed Graphical Models

Korbinian Strimmer, Carsten Wiuf,† and Vincent Moulton‡*

*Department of Zoology and †Department of Statistics, University of Oxford, Oxford, England; and ‡FMI, Physics and Mathematics Department, Mid Sweden University, Sundsvall, Sweden

In Strimmer and Moulton (2000), we described a method for computing the likelihood of a set of sequences assuming a phylogenetic network as an evolutionary hypothesis. That approach relied on converting a given graph into a directed graphical model or stochastic network from which all desired probability distributions could be derived. In particular, we investigated how to compute likelihoods using split-graphs (Huson 1998). However, in the presence of recombination, split-graphs may not provide an appropriate choice of the underlying graph. In this letter, we propose basing the stochastic network on an ancestral recombination graph (ARG) (Hudson 1983; Griffiths and Marjoram 1996, 1997). We show that our approach using directed graphical models extends in a straightforward fashion to ARGs, and we outline the computation of their likelihoods. In particular, we provide an example of an ARG whose likelihood is greater than that of a competing nonnested tree, even though the ARG has a smaller number of free parameters.

Statistical phylogenetic analysis requires a model for the evolutionary relationships between the sequences in a given data set. For this purpose, directed acyclic graphs (DAGs) are suitable for describing the dependencies between the sequences. Several subclasses of DAGs can be easily derived from frequently used graphs such as trees, phylogenetic networks (Bandelt 1994), and split-graphs (Huson 1998). The latter two classes exhibit network-like structures for which the tree is a special case. Net-like models for sequence evolution can be particularly attractive when modeling statistical dependencies among sequences in the presence of recombination, where evolution is clearly non–tree-like. In order for a DAG-based phylogeny to provide a realistic model of recombination, we propose that the DAG should have at least the following properties:

1. The network contains tree nodes, i.e., nodes representing sequences that have exactly one direct parent sequence, to model tree-like evolution.
2. The network contains recombination nodes, i.e., nodes representing sequences that were created by the merging of two parent sequences.
3. Each node in the network except the root node is either a tree node or a recombination node.
4. Each node in the network, except those representing extant sequences, has two or more descendant sequences.
5. The network allows the inclusion of nodes representing all proposed recombination sequences.

Under these simple premises, it appears that split-graphs, even though they may give a good indication of when recombination is occurring, may not provide a suitable underlying graph for a stochastic network. For example, the number of recombination nodes in a DAG-based split-graph cannot be freely selected, and split-graphs are always generated as subgraphs of hypercubes, which can lead to somewhat restrictive constraints on any resulting DAG-based probabilistic model. However, split-graphs were not designed specifically with recombination in mind; they graphically portray incompatibilities in the data which may (or may not) be a consequence of recombination.

Taking the above considerations into account, another variant of net-like graphs, ARGs, may provide a more appropriate DAG-based phylogeny. These rooted graphs provide a way to represent linked collections of clock-like trees by a single network, and were originally developed in population genetics to describe stochastic processes generating hypothetical genealogies for a set of sequences subject to recombination (Hudson 1983; Griffiths and Marjoram 1996, 1997; Wiuf and Hein 1999). In addition to their use in coalescent simulations, they can also be employed as stand-alone models for sequence phylogeny. We suggest that ARGs can offer a useful basis for the statistical analysis of sequences whose evolution is net-like, and we demonstrate this by reanalyzing the HTLV data set we considered in Strimmer and Moulton (2000).

In figure 1, a hypothetical ARG on four taxa is presented that has five tree and two recombination nodes. If a genealogy contains no recombination nodes, then the ARG will degenerate to a rooted clock-like tree. Note that the ARG in figure 1 contains four embedded subtrees containing both the root and the tip nodes, which are pictured on the right of the figure. In general, if an ARG has r recombination nodes, it will contain 2^r such subtrees, and we will call these the canonical subtrees contained in the ARG. ARGs are parameterized by the heights of the tree nodes and the root node and by a breakpoint at each recombination node that specifies which part of the recombinant sequence represented by the node is derived from the parent sequences. Therefore, the number of free parameters for an ARG is precisely the number of internal nodes, so the ARG shown in figure 1 has seven parameters.

Computation of the likelihood of the data assuming some substitution model and a phylogeny based on an

Key words: ancestral recombination graph, likelihood-based sequence analysis, Bayesian network, phylogeny, split-graph.

Address for correspondence and reprints: Vincent Moulton, FMI, Physics and Mathematics Department, Mid Sweden University, S 851-70 Sundsvall, Sweden. E-mail: vince@dirac.fmi.mh.se.

Mol. Biol. Evol. 18(1):97–99. 2001
© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

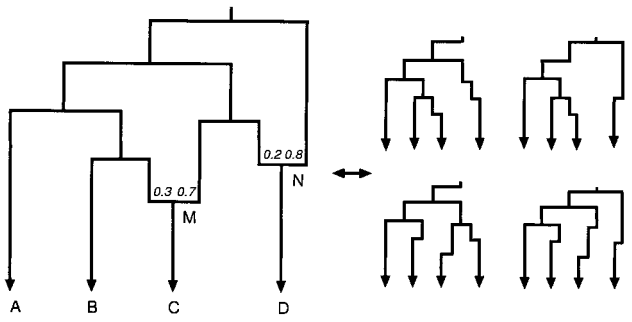

FIG. 1.—Ancestral recombination graph describing the genealogy of four sequences in which two recombinations have occurred, together with its four possible canonical subtrees. Note that two of the canonical subtrees share the same rooted topology but differ in branch lengths, and that node heights of all trees are linked. The ARG is formed by ''gluing together'' the canonical subtrees. *M* and *N* are recombination nodes, and the numerical values at each of these nodes indicate the proportion of each parent sequence that is to be found in the recombinant sequence represented by that node.

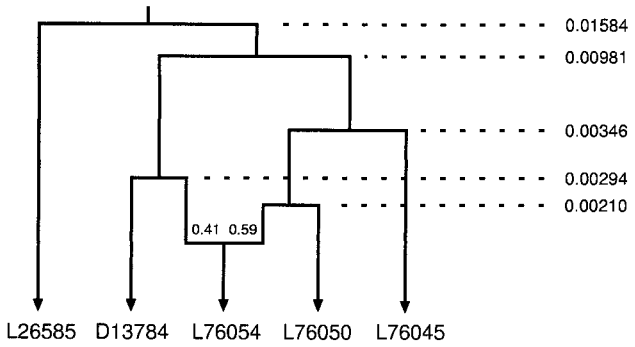

FIG. 2.—A possible ancestral recombination graph for the evolutionary history of the example data set presented in Strimmer and Moulton (2000). Optimized node heights are presented on the right (the node height at the recombination node cannot be easily estimated).

ARG is straightforward. The underlying network topology is interpreted as a Bayesian network, from which all desired probability distributions and likelihood values can be derived. Rather than giving a full description of this procedure, we emphasize some important points, referring the reader to Strimmer and Moulton (2000) for terminology:

1. Since any given ARG is rooted, no rooting procedure is required.
2. The directed graphical model requires the prescription of local node probabilities that specify the probability of observing some state at a selected node given the states of its direct parents. Various rules are conceivable for computing the probability $\Pr(x \mid y, z)$ of observing state x at a given site s of a recombinant sequence given the states y and z at site s of the parent sequences. Here, we consider two natural possibilities: the breakpoint model, which is implicit in Griffith and Marjoram (1996), and the mixture model, which was suggested for split-graphs in Strimmer and Moulton (2000). The breakpoint model requires knowledge of the precise tree-like history of each site, whereas the mixture model assumes knowledge only of the proportion of sites belonging to each parent sequence.

   In the breakpoint model, we assume that we are given a breakpoint p taking on some value in the set $\{0, 1/l, 2/l, \ldots, (l-1)/l, 1\}$, where l is the length of the recombinant sequence, and we define

$$\Pr(x \mid y, z) = \begin{cases} \Pr(x \mid y) & \text{for } \dfrac{s}{l} \leq p \\ \Pr(x \mid z) & \text{for } \dfrac{s}{l} > p. \end{cases} \quad (1)$$

   For the mixture model, we assume that the proportions of sites q and $(1-q)$ belonging to each parent sequence are known, and we define

$$\Pr(x \mid y, z) = q \Pr(x \mid y) + (1-q)\Pr(x \mid z). \quad (2)$$

3. The structure of an ARG permits the computation of the marginal likelihood at each site without having to refer to approximations from Monte Carlo simulations. For the breakpoint model, each site contributes the likelihood of one of the canonical subtrees to the overall likelihood (Kuhner, Yamato, and Felsenstein 2000), whereas for the mixture model, a linear combination of likelihoods for the canonical subtrees is obtained.
4. The node heights of the canonical subtrees are linked in both the breakpoint and the mixture models; in particular, various trees will share node heights. This is important to take into account when branch lengths are being optimized.
5. Likelihoods do not generally depend on the height of the recombination nodes (*M* and *N* in fig. 1). This reflects the fact that sequence evolution at each site is tree-like. However, if the sequence corresponding to a recombination node is known, then a corresponding node height can be inferred.
6. The location of the recombination breakpoint can, in theory, be estimated using both the breakpoint and the mixture models. In either case, this is, unfortunately, computationally quite intensive. For example, assuming the breakpoint model, essentially all possible mappings of sites to subtrees have to be considered and subsequently compared by computing the corresponding likelihoods. Since there are $l+1$ possible places for the recombination breakpoint for a recombinant sequence of length l with two parent sequences, there are $2(l+1)$ possibilities to assign the sites of the recombinant sequence to its parents. Therefore, if an ARG has r recombination nodes, then there are $(2l+2)^r$ possible mappings of sites to canonical subtrees.

We now reanalyze the HTLV data set presented in Strimmer and Moulton (2000). In figure 2, an ARG representing a possible history for this data set is shown, where sequence L76054 is a putative recombinant. This ARG was obtained by ''gluing together'' two tree topologies obtained from a standard breakpoint analysis using the diversity plot (Robertson, Beaudoing, and Claverie 1999). The diversity plot also gave the estimate for the breakpoint given in figure 2. Optimizing the five node heights and employing the same substitution model

as in Strimmer and Moulton (2000), a log likelihood (log L) of $-1{,}496.46$ was obtained using the breakpoint model. Intriguingly, this likelihood is greater than that of the maximum-likelihood tree (log $L = -1{,}505.88$), which has seven free parameters (branch lengths), even though the ARG has only six parameters. Note that the unconstrained tree is not nested within the ARG. Also note that for each canonical subtree in the ARG, a statistical test could not reject a molecular clock for the respective parts of the sequences, which supports the use an ARG-based phylogeny in this example. It also confirms that incorporating recombination in a sequence analysis can reveal clock-like evolution that would otherwise have remained hidden (Schierup and Hein 2000).

In conclusion, we believe the computation of the likelihood of data related by evolutionary networks may be of use when assessing and comparing competing evolutionary hypotheses. If sequences are subject to recombination, then ARG-based phylogenies could provide a suitable way to model the statistical dependencies in the sequence data. Here, we have described how the likelihood of the data under a genealogy that is based on an ARG can be computed in the framework of stochastic networks. We have presented two variants: the breakpoint model, which requires a detailed mapping of sites to subtrees of the ARG, and the mixture model, which does not assume knowledge of the tree-like history at each site. If there is more than one recombination event, then it can be difficult to trace the ancestral history of each site so that the mixture model may then provide an appropriate approximation. Thus, the mixture model may be advantageous in the presence of strong mosaic-like evolution, as is observed, for example, in HIV (Robertson et al. 1995).

Reconstruction of the ancestral history of a set of sequences subject to recombination is difficult (see, e.g., Hein [1993], where a parsimony approach is taken). Thus, it is expected that inferring an ARG for a given data set will be just as difficult, although heuristic approaches, such as gluing together trees as we did in the example above, may deserve some attention. However, ARGs also impose some implicit constraints on the sequences (such as clock-likeness) that may not be valid for all data sets. Therefore, it seems that the ''best'' netlike model for sequence evolution under recombination will probably be some relaxed variant of the ARG.

Acknowledgments

We thank Arndt von Haeseler and Dirk Metzler for providing stimulating questions concerning Strimmer and Moulton (2000), and the referees and Michael Hendy for valuable comments. K.S. also wants to thank Mid Sweden University for its hospitality during a visit during which this letter was completed. This work was supported by an Emmy-Noether-Fellowship of the DFG to K.S., by BBSRC grant 43/MMI09788 and the Carlsberg Foundation, Denmark (C.W.), and by a grant from the Swedish Natural Science Research Council to V.M.

LITERATURE CITED

BANDELT, H.-J. 1994. Phylogenetic networks. Verh. Naturwiss. Vereins Hamburg **34**:51–71.

GRIFFITHS, R. C., and P. MARJORAM. 1996. Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. **3**:479–502.

———. 1997. An ancestral recombination graph. Pp. 257–270 *in* P. DONELLY and S. TAVARÉ, eds. IMA volumes in mathematics and its applications, Vol. 87. Progress in population genetics and human evolution. Springer Verlag, Berlin.

HEIN, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. J. Mol. Evol. **36**:396–406.

HUDSON, R. R. 1983. Properties of the neutral allele model with intergenic recombination. Theor. Popul. Biol. **23**:183–201.

HUSON, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**:68–73.

KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN. 2000. Maximum likelihood estimation of recombination rates from population data. Genetics **156**:1393–1401.

ROBERTSON, D. L., E. BEAUDOING, and J. M. CLAVERIE. 1999. HIV/SIV phylogenetic analysis page (http://igs-server.cnrs-mrs.fr/anrs/phylogenetics). Marseille, France.

ROBERTSON, D. L., P. M. SHARP, F. E. MCCUTCHAN, and B. H. HAHN. 1995. Recombination in HIV-1. Nature **374**:124–126.

SCHIERUP, M. H., and J. HEIN. 2000. Recombination and the molecular clock. Mol. Biol. Evol. **17**:1578–1579.

STRIMMER, K., and V. MOULTON. 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. **17**:875–881.

WIUF, C., and J. HEIN. 1999. The ancestry of a sample of sequences subject to recombination. Genetics **151**:1217–1228.

MIKE HENDY, reviewing editor

Accepted September 28, 2000