CABIOS

Vol. 12 no. 4 1996
Pages 291–295

A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences

Ingrid B.Jakobsen and Simon Easteal

Abstract

Reticulate evolution in molecular sequences is caused by recombination or gene conversion, and may interfere with the reconstruction of evolutionary history. This paper presents a program which calculates compatibility matrices for detecting reticulate evolution. In addition to visual inspection of matrices, they can be analysed statistically for clustering. The method is demonstrated using human and chimpanzee γ-globin sequences.

Introduction

Reticulate evolution of nucleotide and amino acid sequences involves the combining of sequence components from more than one source to form a new hybrid sequence, usually through recombination or gene conversion. When sequences have evolved in a reticulate fashion, different parts of the sequences have different phylogenetic relationships, and the pattern of evolution of the entire sequence cannot validly be described by a single phylogenetic tree. Standard phylogenetic reconstruction methods are not designed to identify reticulate evolution and their uncritical application to entire sequences can give quite misleading results when reticulate evolution has occurred.

There is, thus, a need for methods specifically aimed at identifying reticulate evolution, not only for its intrinsic interest, but also as a preliminary step in molecular phylogenetic analysis. Several approaches have been developed (Stephens, 1985; Sawyer, 1989; Fitch and Goodman, 1991; Hein, 1993), which vary in their applicability to different data sets. One of these involves the use of compatibility matrices (Le Quesne, 1969) to identify regions of aligned sequences within which there is relatively high phylogenetic compatibility of sites, but between which there is relatively low compatibility. This approach, which has been adopted for protein sequences by Sneath *et al.* (1975), is relatively fast and straightforward, and is particularly appropriate for exploratory analysis prior to phylogenetic reconstruction. However, no program with clear graphical output has been available to implement the method, and possibly for this reason, it has not been used.

Human Genetics Group, John Curtin School of Medical Research, Australian National University, Canberra, ACT 0200, Australia
E-mail: ingrid@jcsmr.anu.edu.au

Here we describe a program, called 'reticulate', for calculating and displaying compatibility matrices as an aid in identifying reticulate evolution in nucleotide sequences. The program includes a Monte Carlo approach to evaluating the significance of patterns in compatibility matrices, and we demonstrate its application by analysis of human and chimpanzee γ-globin sequences.

The program is primarily intended for the analysis of nucleotide sequences and will be described in those terms. However, it is equally possible to analyse amino acid sequences with this program.

Principle

Two sites in a set of aligned nucleotide sequence are defined to be compatible if there exists a possible evolutionary history of the sequences in which all nucleotide changes at both sites can be inferred to have occurred only once, i.e. the minimum number of possible changes (c) is one less than the number of distinct nucleotides (n) at each site: $c = n - 1$. Sites are incompatible when for all possible evolutionary histories of the sequences $c > n - 1$ for one or both sites. This means that one or both sites must have experienced repeated mutation or been involved in recombination or gene transfer.

Only parsimoniously informative sites need to be considered when determining compatibility. Sites are defined as parsimoniously informative if two or more nucleotides are present in two or more sequences each (Fitch, 1975). Other sites are either invariant or have only one or no nucleotides that occur in more than one sequence. Thus, at these sites, $c = n - 1$ for all possible trees and they cannot show incompatibility with any other site.

As an example, consider the alignment in Figure 1A. Three parsimoniously informative sites, X, Y and Z, are shown, and under each site are listed all trees for which $c = n - 1$ (Figure 1B). Sites X and Y share a tree for which $c = n - I$ and the two sites are thus compatible. Similarly, sites Y and Z are compatible. However, X and Z are incompatible as there is no tree for which $c = n - 1$ for both sites.

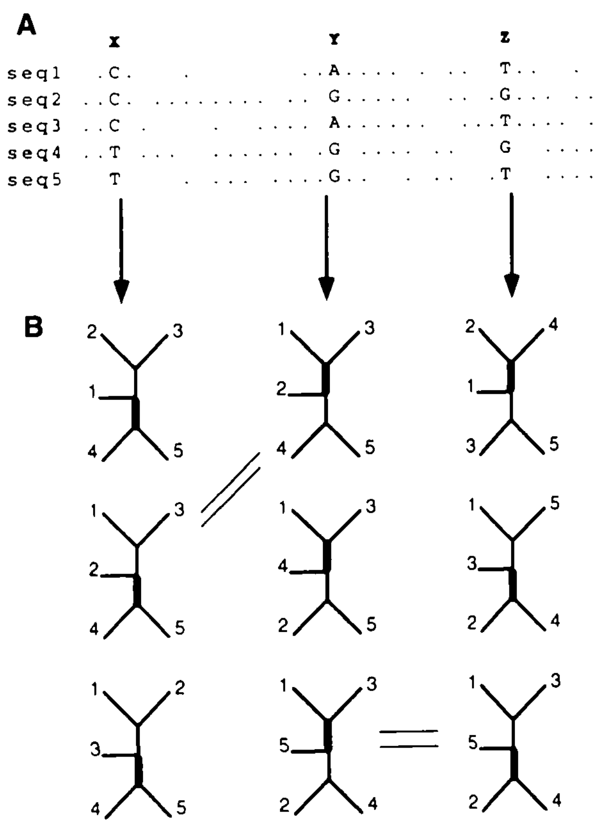

Fig. 1. Demonstration of compatibility determination. (A) Sequence alignment of five sequences, highlighting three sites to be used in compatibility determination. (B) The trees for which c = n − 1 (see the text) are shown below each site. The locations of nucleotide changes are marked by thick branches. Note that the second tree for site X and the first tree for site Y are identical; similarly, the third tree for site Y and the third tree for site Z

Implementation

The program for determining compatibility matrices was written in ANSI C on a Sun SparcStation 5 with SunOS 4.1.3. The program is command line based, and displays the matrix using the X11 graphics interface. It generates text and PostScript output. It has also been successfully compiled and run on a SiliconGraphics Indigo2 under IRIX V.4.

The program requires a set of aligned nucleotide or amino acid sequences. The data should be in FASTA/Pearson format. Once the sequences have been read by the program, all parsimoniously informative sites in the alignment are found. Pairwise compatibility of these sites is determined in the following way.

For a pair of sites, each sequence defines an ordered pair of nucleotides, with the first and second members of the ordered pair being the nucleotides present at the first and second sites, respectively. All distinct ordered pairs are found for the pair of sites. For example, for sites Y and Z from Figure 1, the distinct ordered pairs are AT, GG, and GT, while for X and Z they are CT, CG, TG and TT.

Any of the ordered pairs that have a unique nucleotide at one of the sites are eliminated, since it can be assumed that the change giving rise to that nucleotide has occurred only once. This process is repeated with the remaining pairs, eliminating pairs that progressively contain a unique nucleotide at either site. If all pairs can be eliminated in this way, the two sites are compatible. If any pairs remain, the sites are incompatible. This algorithm has been shown to determine compatibility in the sense of the existence or lack of a shared tree with $c = n - 1$ (Estabrook and McMorris, 1977).

Any sequences in which the nucleotide is unknown or not present due to a gap at a particular site are ignored for the purposes of determining compatibility between those sites and all others. An unknown nucleotide cannot contribute to compatibility or incompatibility. The interpretation of gaps is often difficult, hence the default is to treat them as unknown. The origin of a particular gap may be known and in such a case the event may be considered in the determination of compatibility. One of the sites making up such a gap can be edited in the datafile, replacing it with another character, e.g. ‘O’, in each sequence sharing the gap. The program will then treat it as equivalent to a nucleotide substitution for the purposes of determining compatibility.

Binary sites, i.e. sites with only two different nucleotides, have the property that if a group of them are compatible under all pairwise comparisons, the group as a whole is compatible. However, it is possible for sites with more than two nucleotides to be pairwise compatible, but not compatible when considered as a group (Fitch, 1975). If the dataset contains many such sites, the pairwise comparisons used to determine the matrix may not accurately reflect compatibility in the dataset. To aid in the assessment of this, the user is presented with options for treating such sites. These are (i) to leave the sites with several nucleotides; (ii) to disregard the sites altogether; or (iii) to convert the sites into binary sites. In the case of nucleotide sequences, a site can be converted into a binary site by considering only transversion changes, with C and T (or U) equivalent, and A and G equivalent. Otherwise, an appropriate binary grouping is specified manually for each such site.

The program calculates the compatibility of all pairwise comparisons of informative sites. The results are displayed as a matrix with each side consisting of the informative sites in the order they are present along the sequence. Compatibility of any two sites is represented as a white square corresponding to the intersection of those two sites; incompatibility is represented by a black square. The matrix is triangular, but is presented as a square matrix symmetrical about the diagonal. The matrix is displayed using the X11 graphics interface and can be saved in two PostScript formats for printing. 'Plain' PostScript includes a text description of the dataset and numbering of the sites along the side of the matrix, while the encapsulated PostScript format contains just the matrix, suitable for importing into graphics packages. The sites used for the matrix, with site numbers, can also be saved as a text file.

Interpretation of matrices

The appearance of the matrix is determined partly by the nature of each informative site, and partly by the ordering of the sites along the sequence. If there has been no reticulate evolution, then the appearance is entirely based on the individual sites, while reticulate evolution causes clustering of sites showing compatibility and incompatibility. Two regions of the sequence that have experienced distinct histories can be identified by comparing the rectangle of inter-region comparisons to the squares of intra-region comparisons. The intra-region comparisons are more likely to be compatible, and so these regions appear proportionately whiter. A site experiencing repeated mutation is no more likely to be compatible with nearby sites than distant sites, so comparisons with that site create a dark line across the entire matrix.

Clustering of sites can be identified initially by visual inspection. To identify such clustering more easily, the matrix can be randomized using a Monte Carlo approach. The informative sites are shuffled, thus breaking up the structure of the matrix relating to the linear order of sites, while retaining the characteristics of each site. Visual inspection of a small number of such randomized matrices can allow the effect of reticulate evolution to be distinguished from the chance ordering of sites with varying levels of compatibility. It is important to note that the random matrices consist of the informative sites in random order, and are not randomizations of the individual pairwise comparisons of sites.

The clustering of compatible and incompatible comparisons in the matrix can be estimated by the 'neighbour similarity score', which is the fraction of **adjacent** squares in the matrix that are the same colour, i.e. both or neither are compatible. To determine whether the ordering of sites along the sequence has caused increased clustering and hence raised the neighbour similarity score above expectation, the neighbour similarity score is calculated for a user-defined number of random matrices generated as described. The probability of the observed clustering being random is the fraction of random matrices with at least as high a neighbour similarity score as the actual matrix.

Additionally, the alignment can be divided into regions and, for each region, the degree of compatibility calculated as the fraction of comparisons that are compatible within each region. Then the compatibility in the specified regions is calculated for random matrices generated as described above. The fraction of random matrices at least as compatible as the original matrix in each of the designated regions is determined and the average compatibility calculated for all random matrices.

Application to γ-globins

The γ-globin gene exists as two copies in most simian primates as a result of duplication of a 5 kb region including the gene. Gene conversion between the two copies has been extensively investigated (e.g. Fitch *et al.*, 1990). An alignment was prepared of the 5 kb repeat region in chimpanzees (*Pan troglodytes*) with both the human A and B haplotypes (accession numbers: M92294, M91036, M91037). The compatibility matrix of these six sequences showed that gene conversion had only occurred in the region of the gene itself, so a shorter alignment of 2320 nucleotides was prepared. This region includes the globin gene itself, the 5′ upstream regulatory regions, and extends 400 bp upstream of the 5′ CACCC element and 200 bp downstream of the stop codon.

The compatibility matrix of the shorter alignment is shown in Figure 2. The converted region stands out clearly against the unconverted (white) background. The 5′ and 3′ regions are likely to have the same evolutionary history since they are almost totally compatible with each other. The neighbour-joining tree of these regions is ((human A1, human B1), chimp 1, ((human A2, human B2), chimp 2)). Within the region of gene conversion, there are three smaller regions. These range from 70 nucleotides 5′ of the CACCC element to early in intron 2, from there to the end of the pyrimidine-rich element in intron 2, and finally to the end of the TG repeat in intron 2. The first and the third regions are mutually compatible.

To determine likely histories of the smaller regions, neighbour-joining trees were constructed. The first and third regions suggest conversion of γ-globin in both human haplotypes and in the chimpanzee (human A1, (human A2, (chimp 1, chimp 2)), (human B1, human B2)). The middle region suggests conversion only in the human A haplotype, (((human A1, human A2), human B1), chimp 1, (human B2, chimp 2)). This indicates that the region of conversion of the B haplotype is shorter than previously suggested (Fitch *et al.*, 1990). However, the sequences of both γ-globins from the B haplotype were not available for the earlier analysis.

Following the third region, there is a less distinct region, still within the second intron, showing compatibility both

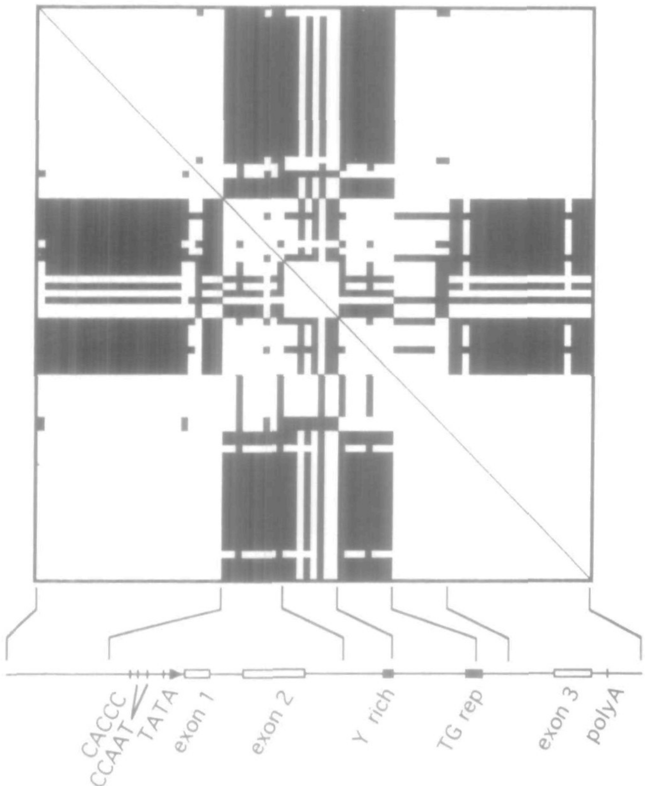

Fig. 2. Compatibility matrix of the γ-globin gene from chimpanzee and human haplotypes A and B. Below the matrix is a map of the gene where the corresponding regions with different histories are marked. The positions of the three exons are marked with open boxes and the cap site by a triangle. The locations of 5′ regulatory elements and the polyA addition site are as indicated. The black boxes mark the location of the intron 2 pyrimidine-rich element (Y rich) and the TG repeat element (TG rep). The locations of the informative sites making up the matrix are as follows: −641, −635, −628, −612, −607, −606, −601, −586, −584, −570, −567, −566, −563, −555, −550, −546, −543, −532, −531, −524, −450, −421, −408, −376, −369, −359, −323, −265, −65, 157, 158, 211, 303, 462, 479, 566, 609, 638, 656, 714, 716, 733, 741, 742, 795, 886, 890, 1024, 1064, 1066, 1068, 1070, 1077, 1090, 1091, 1113, 1125, 1155, 1188, 1189, 1193, 1194, 1221, 1453, 1490, 1491, 1492, 1493, 1504, 1565, 1580, 1583, 1591, 1604, 1605, 1640, 1649, 1654, 1666, 1668, 1673; where the start codon corresponds to positions 1–3.

with the converted and unconverted regions. The neighbour-joining tree of this region suggests conversion between chimpanzee γ-globins. It does not show clearly as incompatibility with the rest of the sequence, as the pattern observed at these sites cannot be distinguished from rapid divergence of the human genes without reference to γ-globin sequences from other species.

The matrix was analysed by comparison with 10 000 random matrices. The program output for this analysis is presented in Figure 3. The observed neighbour similarity score was not reached by any random matrix. The matrix was divided into the five regions described earlier. The analysis (Figure 3) revealed the overall pattern of compatibility and incompatibility among sites to be non-random, as the degree of compatibility of the large, 5′ and

```
Summary statistics for non-randomness of compatibility matrix for
human and chimpanzee gamma globins

The matrix had 81 sites with overall compatibility 0.666358
The neighbour similarity score of the matrix was 0 876852

Statistics for 10000 random matrices·

Mean Neighbour similarity score for random matrices 0.629903
0 random matrices equalled or exceeded the observed

The P value is 0.000100

The individual regions analysed were
 1  1 to 27     27 sites from 5 to 323
 2: 28 to 36     9 sites from 381 to 1212
 3: 37 to 44     8 sites from 1255 to 1388
 4· 45 to 52     8 sites from 1441 to 1716
 5. 53 to 81    29 sites from 1723 to 2319

Compatibility of each region compared to random matrices
                 Shuffled   Fraction
   Observed      Average    exceeding
1  0.994302      0.666567   0 0000  +
2  0 944444      0 665267   0 0569  +
3  1.000000      0.663536   0.0624  +
4  0.964286      0.666307   0.0713  +
5  1.000000      0 667276   0.0000  +

'Fraction exceeding' is the fraction of random matrices with equal or
higher '+' or lower '-' values than observed in that region
Overall, 0 0000 exceeded observed for all regions at once

Observed compatibility of the inter-region comparisons.
Region 1 vs 2: 0 061728 for 243 squares
Region 1 vs 3  0 532407 for 216 squares
Region 1 vs 4: 0.064815 for 216 squares
Region 1 vs 5: 0.994891 for 783 squares
Region 2 vs 3. 0.472222 for 72 squares
Region 2 vs 4: 0.944444 for 72 squares
Region 2 vs 5: 0 260536 for 261 squares
Region 3 vs 4: 0 406250 for 64 squares
Region 3 vs 5: 0 568965 for 232 squares
Region 4 vs 5: 0 275862 for 232 squares
```

Fig. 3. Output from the matrix randomization routine, showing neighbour similarity scores; and compatibility of selected regions of the matrix, compared to randomized matrices

3′ regions of compatibility was not reached in any random matrix. However, the observed degree of compatibility of the three smaller regions internal to the gene conversion occurs at random 5–7% of the time, so each contains too few parsimoniously informative sites to establish reticulate evolution reliably by this analysis.

Discussion

Reticulate evolution may require different methods of detection in different data sets. We believe that the compatibility matrix is most suitable for detecting reticulate evolution in relatively large alignments with frequent events. Other methods may not perform as well under these conditions, due to the exponential growth of nucleotide patterns possible at each site.

The compatibility matrix is not a statistical or rigorous description of reticulate evolution. Rather, it can be considered a transformation of a multiple sequence alignment into a graphical form, taking advantage of the human ability for pattern recognition. Inspection of the matrix allows a more intuitive 'feel' for the data and may assist in the choice of phylogenetic method for further analysis of the data, or indicate regions that should be analysed separately or excluded.

Since the method relies on parsimoniously informative sites, certain kinds of events cannot be detected using a compatibility matrix, most notably gene conversions leading to a change in relative distances among sequences rather than a change in phylogeny, and in some instances conversion by a sequence outside the considered data set. These kinds of events may be detected using another method, such as that of Stephens (1985). It could be argued that such differences have occurred because selection pressure has changed. If the conversion is from an omitted sequence, the case for reticulate evolution would be established more strongly if the template sequence could be found and included, in which case the present method should confirm the event. Also, as a consequence of the compatibility algorithm, reticulate evolution cannot be investigated for less than four sequences, while in general three sequences with the appropriate relationships are considered sufficient to establish recombination (Andersson *et al.*, 1991; Hughes, 1991). However, this is not a limitation for most datasets.

The compatibility matrix approach is a useful tool for identifying sequence regions with distinct phylogenetic histories. Because it is fast and results are easily interpreted, it can be used as a routine step in screening for reticulate evolution prior to phylogenetic reconstruction of sequences. The Monte Carlo randomizations provide a basis for statistical evaluation of matrices. More sophisticated graphical approaches to detecting reticulate evolution are possible, and are currently being investigated.

Availability

The C program described in this paper for calculating compatibility matrices is freely available at http:/jcsmr.anu.edu.au/dmm/humgen.html, along with the alignment of γ-globin genes.

Acknowledgements

The authors would like to thank Hugh Fisher for developing the X11 graphics module for the program, Sue Wilson and Elizabeth Thompson for discussions of the statistical analysis of matrices and the neighbour similarity score, and Gareth Chelvanayagam and Lars Jermiin for comments on the manuscript.

References

Andersson,L., Gustafsson,K., Jonsson,A. and Rask,L. (1991) Concerted evolution in a segment of the first domain exon of polymorphic MHC class II β loci. *Immunogenetics*, **33**, 235–242.

Estabrook,G.F. and McMorris,F.R. (1977) When are two qualitative taxonomic characters compatible? *J. Math. Biol.*, **4**, 195–200.

Fitch,D.H.A. and Goodman,M. (1991) Phylogenetic scanning: a computer-assisted algorithm for mapping gene conversions and other recombinational events. *Comput. Appl. Biosci.*, **7**, 207–215.

Fitch,D.H.A., Mainone,C., Goodman,M. and Slightom,J.L. (1990) Molecular history of gene conversions in the primate fetal γ-globin genes. *J. Biol. Chem.*, **265**, 781–793.

Fitch,W.M. (1975) Toward finding the tree of maximum parsimony. In Estabrook,G.F. (ed.), *Proceedings of the Eighth International Conference on Numerical Taxonomy*. W.H.Freeman & Co., San Francisco, CA., pp. 189–230.

Hein,J. (1993) A heuristic method to reconstruct the history of sequences subject to recombination. *J. Mol. Evol.*, **36**, 396–405.

Hughes,A.L. (1991) Testing for interlocus genetic exchange in the MHC: a reply to Andersson and co-workers. *Immunogenetics*, **33**, 243–246.

Le Quesne,W.J. (1969) A method of selection of characters in numerical taxonomy. *Syst. Zool.*, **18**, 201–205.

Sawyer,S. (1989) Statistical tests for detecting gene conversion. *Mol. Biol. Evol.*, **6**, 526–536.

Sneath,P.H.A., Sackin,M.J. and Ambler,R.P. (1975) Detecting evolutionary incompatibilities from protein sequences. *Syst. Zool.*, **24**, 311–332.

Stephens,J.C. (1985) Statistical methods of DNA sequence analysis: detection of intragenic recombination or gene conversion. *Mol. Biol. Evol.*, **2**, 539–556.

Received on January 9, 1996, accepted on May 30, 1996