Volume 12 Number 1 1984 Nucleic Acids Research

# Algorithms for restriction map comparisons

Michael S.Waterman, Temple F.Smith and Harold L.Katcher

Departments of Mathematics and Biological Sciences, University of Southern California, Los Angeles, CA 90089, USA

Received 12 August 1983

## ABSTRACT

An algorithm is presented which compares two restriction maps, yielding a measure of distance between the maps and relating the maps by an alignment. This new algorithm finds the minimum weighted sum of genetic events required to convert one map into the other, where the genetic events are the appearance/disappearance of restriction sites and changes in the number of bases between restriction sites. The algorithm is illustrated by comparison of the β-δ region of the globin gene cluster of four primate species. The results are in excellent agreement with known evolutionary relationships.

## INTRODUCTION

Restriction maps are usually constructed prior to sequencing of DNA and many mapped DNAs have never been sequenced. Such restriction maps contain information valuable for studies of sequence homology and similarity. It is natural, then, to consider computer methods for comparison of restriction maps. Existing sequence alignment algorithms are not applicable (Sellers, 1974, Waterman et al., 1976). This paper presents such an algorithm, which is distinct from usual sequence comparison algorithms and is designed for the specific nature of restriction maps. We have used the proposed algorithm to investigate the relationship among maps of homologous regions of the primate globin gene cluster. The results of these studies demonstrate the utility of this algorithm.

Other methods have been devised which infer phylogeny from restriction maps. Nei and Li (1974), Engels (1981) Ewens et al., (1981) and Hudson (1982) have, among others, estimated sequence variation from restriction site data. Templeton (1983) used convergent evolution considerations to investigate the ef-

fects of sequence evolution on restriction maps and proposed an algorithm for phylogenetic inference. The method proposed here differs from these. Rather than estimating sequence variation as in the above methods, we directly relate the maps in an alignment.

## METHODS

The distance we propose is the minimum weighted sum of the genetic events necessary to convert one map into another. These events are of two types: (i) the appearance or disappearance of restriction sites and (ii) a change in the number of bases between two restriction sites. Although it is possible to consider "mutation" of one restriction site into another, it is assumed here that such an event can be ignored because of its rarity.

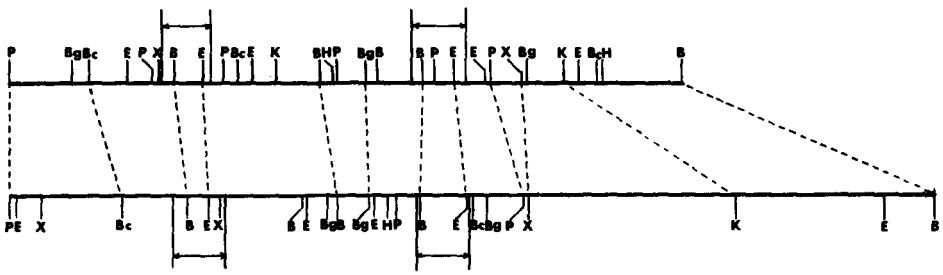
If a restriction site appears or disappears, that event receives weight $\lambda$. If the number of bases between two sites changes by x bases, that event receives weight $\mu(x)$.

Map A = $a_0a_1a_2\ldots a_na_{n+1}$ of n restriction sites is represented by an ordered set of pairs: $a_i = (r_i,p_i)$, where; $r_i$ = the restriction endonuclease producing site i and $p_i$ = the map position of site i. For example, if the fifth restriction site in map A is an EcoRI site that occurs 5.3 kilo-bases from the "origin", $a_5$ = (EcoRI, 5.3 kb). Also, $a_0 = (\alpha,1)$ and $a_{n+1}$ = $(\beta,N)$, where $\alpha$ and $\beta$ are simply to denote the ends of the sequence and need not be restriction sites themselves. A second map, B = $b_0b_1\ldots b_{m+1}$ is similarly represented with $b_j = (s_j,q_j)$.

Define $D_{ij}$ to be the minimum sum of weights of events required to convert map $a_0a_1\ldots a_i$ into $b_0b_1\ldots b_j$ where the site of $a_i$ is equal to the site of $b_j$ (that is they result from the same restriction endonuclease or $r_i = s_j$). If these sites are unequal, $D_{ij} = +\infty$. To initalize our matrix $D_{00} = 0$ and $D_{0,j} = D_{n+1,j} = +\infty = D_{i,0} = D_{i,m+1}$ if $0 < j < m+1$ or $0 < i < n+1$. For cases where $r_i = s_j$ the algorithm finds $D_{ij}$ by:

$$D_{ij} = \underset{\substack{0 < k \leq i \\ 0 < \ell \leq j}}{\text{minimum}} \{D_{i-k,j-\ell} + \lambda*(\ell+k-2) + \mu(p_i-p_{i-k}-q_j+q_{j-\ell})\}. \quad (1)$$

Here $\lambda$ is the weight associated with the appearence or disap-




a)

| $r_i$ | P | Bg | Bc | E | P | X | B | E | P | Bc | E | K | B | H | P | Bg | B | B | P | E | E | P | X | Bg | K | E | Bc | H | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p_i$ | 0.0 | 2.1 | 2.7 | 4.1 | 4.9 | 5.2 | 5.7 | 6.7 | 7.3 | 7.9 | 8.3 | 9.1 | 10.7 | 11.1 | 11.3 | 12.2 | 12.6 | 14.2 | 14.6 | 15.3 | 16.4 | 16.5 | 17.6 | 17.8 | 19.1 | 19.6 | 20.2 | 20.4 | 23.4 |

| $s_j$ | P | E | X | Bc | B | E | X | B | E | Bg | B | Bg | E | H | P | B | E | Bc | Bg | P | X | K | E | B |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $q_j$ | 0.0 | 0.2 | 1.1 | 3.9 | 6.2 | 6.9 | 7.3 | 10.1 | 10.2 | 10.9 | 11.3 | 12.4 | 12.5 | 12.9 | 13.2 | 14.2 | 15.7 | 15.9 | 16.4 | 17.7 | 17.8 | 25.1 | 30.3 | 32.3 |

b)

| | α | E | X | Bc | B | E | X | B | E | Bg | B | Bg | E | H | P | B | E | Bc | Bg | P | X | K | E | β |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| α | 0.0 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Bg | - | - | - | - | - | - | - | - | - | 12.9 | - | 16.2 | - | - | - | - | - | - | 21.7 | - | - | - | - | - |
| Bc | - | - | - | 4.1 | - | - | - | - | - | - | - | - | - | - | - | - | - | 21.1 | - | - | - | - | - | - |
| E | - | 4.5 | - | - | - | 6.4 | - | - | 9.6 | - | - | - | 13.2 | - | - | - | 17.3 | - | - | - | - | - | 29.1 | - |
| P | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 14.2 | - | - | - | - | 20.0 | - | - | - | - |
| X | - | - | 5.5 | - | - | - | 8.2 | - | - | - | - | - | - | - | - | - | - | - | - | - | 20.0 | - | - | - |
| B | - | - | - | 8.0 | - | - | - | 9.8 | - | - | 11.9 | - | - | - | - | 15.2 | - | - | - | - | - | - | - | - |
| E | - | 9.8 | - | - | 8.0 | - | - | 10.8 | - | - | - | 12.9 | - | - | - | - | 16.0 | - | - | - | - | - | 27.8 | - |
| P | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 13.9 | - | - | - | - | 19.2 | - | - | - | - |
| Bc | - | - | - | 8.5 | - | - | - | - | - | - | - | - | - | - | - | - | - | 17.5 | - | - | - | - | - | - |
| E | - | 13.6 | - | - | - | 11.3 | - | - | 13.2 | - | - | - | 16.6 | - | - | - | 17.2 | - | - | - | - | - | 28.5 | - |
| K | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 25.5 | - | - |
| B | - | - | - | - | 11.2 | - | - | 13.9 | - | - | 16.0 | - | - | - | - | 18.6 | - | - | - | - | - | - | - | - |
| H | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 19.1 | - | - | - | - | - | - | - | - | - |
| P | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 19.1 | - | - | - | - | 23.2 | - | - | - | - |
| Bg | - | - | - | - | - | - | - | - | - | 17.7 | - | 18.0 | - | - | - | - | - | - | 23.5 | - | - | - | - | - |
| B | - | - | - | - | 16.2 | - | - | 17.7 | - | - | 17.7 | - | - | - | - | 20.1 | - | - | - | - | - | - | - | - |
| B | - | - | - | - | 18.0 | - | - | 17.2 | - | - | 19.3 | - | - | - | - | 22.0 | - | - | - | - | - | - | - | - |
| P | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 21.7 | - | - | - | - | 2 | - | - | - | - |
| E | - | 25.8 | - | - | - | 19.0 | - | - | 19.2 | - | - | - | 21.3 | - | - | - | 23.0 | - | - | - | - | - | 33.5 | - |
| E | - | 27.6 | - | - | - | 21.2 | - | - | 20.8 | - | - | - | 23.2 | - | - | - | 24.6 | - | - | - | - | - | 34.2 | - |
| P | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 24.1 | - | - | - | - | 26.9 | - | - | - | - |
| X | - | - | 28.6 | - | - | - | 22.4 | - | - | - | - | - | - | - | - | - | - | - | - | - | 27.9 | - | - | - |
| Bg | - | - | - | - | - | - | - | - | - | 23.6 | - | 24.2 | - | - | - | - | - | - | 28.4 | - | - | - | - | - |
| K | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 32.3 | - | - |
| E | - | 34.2 | - | - | - | 27.8 | - | - | 26.4 | - | - | - | 26.6 | - | - | - | 28.9 | - | - | - | - | - | 35.1 | - |
| Bc | - | - | - | 31.6 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| H | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 27.6 | - | - | - | - | - | - | - | - | - |
| β | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 38.2 |

c)

Figure 1. Aligned restriction maps a) fragments including the β and δ globin genes derived from the lowland Gorilla (top) and the Owl Monkey (bottom) genomic DNA; b) Detailed maps of lowland Gorilla (top) and Owl Monkey (bottom) restriction sites giving the restriction enzyme, (BamHI (B), BclI (Bc), BglII (Bg), EcoRI (E), HindIII (H), KpnI (K), PstI (P) and XbaI (X)) and the distance in kilobases from the (arbitrary) origin. c) The matrix (D) used to generate the alignment shown in a. The parameters used in this example are, $\lambda = 1$, $w_o = 0.5$, $w_1 = 0.5$ and $\Delta = 0.5$ kbp

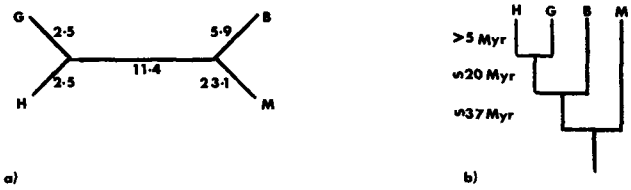



Figure 2. a) Dendrogram obtained from Table 1d. b) Probable evolutionary relationship given by Barrie and Jeffreys (1981)

pearance of sites and μ is the weight associated with the difference in length between aligned sites. If map A has n sites and map B has m sites, this algorithm runs in time $O(m^2n^2)$, (a constant multiple of $n^2m^2$). Fortunately the matrix is quite

Table 1

Map distances for different weights

| Weights* | | | | | Distance matrices | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | λ | $w_0$ | $w_1$ | Δ(kbp) | | H | G | B | M |
| a) | 1.0 | 1.0 | 0.0 | 0.5 | H | 0 | 5 | 19 | 29 |
| | | | | | G | | 0 | 19 | 32 |
| | | | | | B | | | 0 | 26 |
| | | | | | M | | | | 0 |
| b) | 1.0 | 1.0 | 0.0 | 0.2 | | H | G | B | M |
| | | | | | H | 0 | 11.0 | 25 | 32 |
| | | | | | G | | 0 | 25 | 33 |
| | | | | | B | | | 0 | 30 |
| | | | | | M | | | | 0 |
| c) | 1.0 | 0.5 | 0.0 | 0.2 | | H | G | B | M |
| | | | | | H | 0 | 8 | 18.0 | 26.0 |
| | | | | | G | | 0 | 18.5 | 27.0 |
| | | | | | B | | | 0 | 24.5 |
| | | | | | M | | | | 0 |
| d) | 1.0 | 0.5 | 0.5 | 0.5 | | H | G | B | M |
| | | | | | H | 0 | 5.0 | 19.8 | 35.7 |
| | | | | | G | | 0 | 19.9 | 38.2 |
| | | | | | B | | | 0 | 29.0 |
| | | | | | M | | | | 0 |

*The weighting function, μ, for gaps between sites was assumed to be a linear function $w_0 + w_1.x$; where x is the difference in length between the two aligned sites, if x exceeds the parameter Δ, otherwise μ is taken as zero. H represents human, G the lowland gorilla, B the yellow baboon an old world monkey and M owl monkey, a new world monkey.

sparse, having finite entries only where $r_i = s_j$.

Algorithms for the best alignment of a shorter map within a longer one can be obtained by modifications similar to those of Sellers (1980). Finding the best matching regions or segments of two maps can be accomplished by modification similar to those of Smith & Waterman (1980).

## RESULTS

To illustrate the algorithm of equation (1), we compare the restriction maps based on the action of eight restriction endonucleases on the ca. 30 kilobase pair long fragments of primate globin gene clusters containing the δ and β globin genes. These maps all start at the first PstI site 5' from the δ globin gene and end at the first BamI site 3' from the β globin coding sequence. The data for the gorilla, yellow baboon and the owl monkey were calculated from the restriction maps and restriction fragment lengths given by Barrie and Jeffrey (1981). The human restriction map was reconstructed from other data (Flavell et al., 1978; Bernards et al., 1979; Little et al., 1979). Reconstructed maps are available upon request.

A weight of 0.5 for the appearance/disappearance of a site ($\lambda$=1) and a weight of 0.5 for any difference in length between paired sites greater than 0.5 kilobases was initially employed. Figure 1 shows the map alignment and the matrix ($D_{i,j}$) generated by calculation of the alignment of the owl monkey and the gorilla restriction maps. Even for these naive weights, the implied phenetic relationships among these four primates (figure 2) is in good agreement with those obtained from more detailed information.

It should be noted that the restriction sites located within highly conserved coding regions of the β and δ globin genes are matched in the map alignments as expected. Con-served restriction sites occuring outside coding sequences may indicate regions important for control or structural purposes. Table 1 shows the effect on the inter taxa distances obtained for alternate choices of the weights. These data suggest that the implied evolutionary relationships are not an overly sensitive function of the weight, an important property. Secondly

some choices of weights, such as only counting intra aligned site deletions/insertions when they exceed 0.5k bases (in these data), results in distances nearly proportional to likely times of divergence (Barrie and Jeffreys, 1981).

**ACKNOWLEDGEMENT**

This work was supported by a grant from System Development Foundation.

**REFERENCES**

1. Barrie, P.A. and Jeffreys, A.J. (1981). J. Mol. Biol. 149 319-336.
2. Bernards, R., Little, P.F.R., Annison, G., Williamson, R. and Engels, W.R. (1981) Proc. Natl. Acad. Sci., U.S.A., 78. 6329-6333.
3. Ewens, W.J., Spielman, R.S., Harris, H. (1981). Proc. Natl. Acad. Sci., U.S.A., 78. 3748-3750.
4. Flavell, R.A. (1979). Proc. Nat. Acad. Sci., U.S.A. 76. 4827-4831.
5. Flavell, R.A., Kooter, J.M., De Boer, E., Little, P.F.R. and Williamson, R. (1978) Cell 15, 25-41.
6. Hudson, R.R. (1982). Genetics 100, 711-719.
7. Little, P.F.R., Flavell, R.A., Kooter, J.M., Annison, G. and Martin, S.L., K.A. Vincent, and A.C. Willson (1983) J. Mol. Biol. 164, 513-528.
8. Nei, M., Tajima, F. (1981) Genetics 97, 145-163.
9. Sellers, P.H. (1974). J. Appl. Math (SIAM), 26, 787-793.
10. Sellers, P.H. (1980). J. Algorithms, 1, 359-373.
11. Smith, T.F. & Waterman, M.S. (1981). J. Mol. Biol, 147, 195-197.
12. Templeton, A.R. (1983) Evolut. 37, 221-244.
13. Waterman, M.S., Smith T.F., & Beyer, W.A. (1976) Advan. Math.
14. Williamson, R. (1978). Cell 15, 25-41.
15. Williamson, R. (1979). Nature 278, 227-231.