Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

The PNarec method for detection of ancient recombinations through phylogenetic network analysis

Naruya Saitou[a,*], Takashi Kitano[b]

[a] Division of Population Genetics, National Institute of Genetics, Mishima 411-8540, Japan
[b] Department of Biomolecular Functional Engineering, College of Engineering, Ibaraki University, Hitachi 316-8511, Japan

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Recombination
Phylogenetic network

ABSTRACT

Recombinations are known to disrupt bifurcating tree structure of gene genealogies. Although recently occurred recombinations are easily detectable by using conventional methods, recombinations may have occurred at any time. We devised a new method for detecting ancient recombinations through phylogenetic network analysis, and detected five ancient recombinations in gibbon ABO blood group genes [Kitano et al., 2009. Mol. Phylogenet. Evol., 51, 465–471]. We present applications of this method, now named as “PNarec”, to various virus sequences as well as HLA genes.

1. Introduction

The good old days of constructing phylogenetic trees from relatively short sequences are over. Reticulated or “non-tree” structures are omnipresent in genome sequences, and the construction of phylogenetic networks is now the default for describing these complex realities. Recombinations (e.g., Kitano et al., 2009, 2012), gene conversions (e.g., Kitano and Saitou, 1999; Ezawa et al., 2010), and gene fusions (e.g., Tomiki and Saitou, 2004) are biological mechanisms to produce non-tree structures to gene phylogenies, while gene flow is well known factor for creating reticulations within population phylogenies.

The human ABO blood group system consists of three major allelic groups, A, B, and O. The A and B alleles are functional, and code for different glycosyltransferases, transferring N-acetylgalactosamine and galactose, respectively, to a common precursor glycosyl chain H (Yamamoto et al., 1990). O alleles are nonfunctional, and the most frequent allele has a point deletion in exon 6 (Yamamoto, 2000), which induces a frameshift, resulting in a truncated protein deprived of any glycosyltransferase activity (Yamamoto et al., 1990). Saitou and collaborators applied the phylogenetic network method (Bandelt, 1994; Huson and Bryant, 2006; Kitano, 2012) to nucleotide sequence data of ABO blood group genes of humans and non-human primates (e.g., Ogasawara et al., 1996; Saitou and Yamamoto, 1997; Kitano et al., 2000, 2009; Noda et al., 2000; Roubinet et al., 2004). Kitano et al. (2012) recently discovered that the current A allele was formed by a recombination between B101 and O01 around 200,000 years ago by conducting a simple sequence analysis involving a phylogenetic network construction method proposed by Kitano et al. (2009). We would like to show the general applicability of this method for detecting recent and ancient recombinations in this article.

2. Theory

Kitano et al. (2009) developed a new method to infer both the recent and ancient recombinations through construction of a phylogenetic network from extant sequences. Here we explain how to infer a recombination event from a phylogenetic network using a schematic model. Fig. 1 shows one recombination at some point between two parental alleles B and C, which diverged from the common ancestral sequence A. Two recombinant sequences (D and E) were produced, and later the three lineages (two parent sequence descendant lineages F and G and one recombinant lineage H derived from recombinant D) survived until the present time, while the lineage derived from recombinant E became extinct. Sequence I, which is outgroup to the rest of all sequences, is also depicted in this figure. Italic number denotes the number of nucleotide differences on each edge. The nucleotide configuration patterns of imaginary 30 nucleotide sites of these nine sequences (from A to I) are shown in Table 1. All the states of sequence A are set to be ‘–‘, and all mutational differences from that sequence are shown as ‘ + ’. For simplicity, no parallel changes was assumed, and if + status at some particular nucleotide site is shared with multiple sequences, this means they are identical by descent. One parental sequence B and its offspring sequence regions are underlined to indicate the recombination breakpoint, namely between sites 15 and 16.

Fig. 2 is a phylogenetic network for the five sequences (A–E) which do not exist at the present time. The numbers shown along each edge are nucleotide sites with changes, and those which are

* Corresponding author. Fax: +81 559 81 6789.
E-mail address: saitounr@lab.nig.ac.jp (N. Saitou).

1055-7903/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2012.09.015

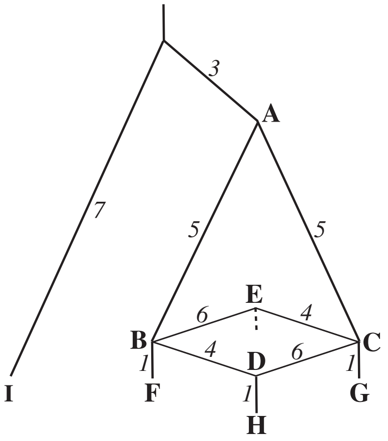


Fig. 1. A schematic view of recombination. A: ancestor, B: parent-1, C: parent-2, D: recombinant-1, E: recombinant-2, F: parent-1 lineage allele, G: parent-2 lineage allele, H: recombinant-1 lineage allele, I: outgroup. Italicized numbers designate numbers of nucleotide substitutions at corresponding branches. A broken line from sequence E shows that this lineage became extinct.

Table 1
Nucleotide configuration patterns of 9 sequences shown in Fig. 1.

```
==================================
            111111111122222222223
   123456789012345678901234567890
==================================
A  ------------------------------
B  --+-+----+-------+--------+---
C  +----+------+-------+--+------
D  --+-+----+----------+--+------
E  +----+------+----+--------+---
F  --+-+----+----+--+--------+---
G  +----++-----+-------+--+------
H  --+-+----+----------+--+---+--
I  -+-----+--+--+-+--++-+--+---+-
==================================
```

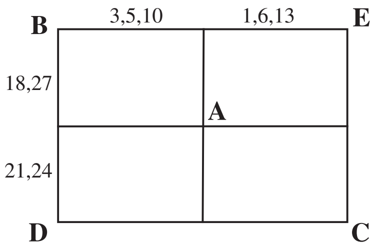


Fig. 2. The phylogenetic network for ancient sequences A–E of Fig. 1 and Table 1. See text for a detailed explanation.

identical with parallel edges are not shown. The ancestral sequence A is located within the rectangle formed by the two parental sequences (B and C) and the two recombinant sequences (D and E). It is clear that horizontal and vertical edges of the rectangle represent upstream and downstream nucleotide differences punctuated by the recombination, respectively. The recombination breakpoint is expected to be located after position 13 and before position 18 if we examine positions experienced substitutions. In fact, this estimation is compatible with the real breakpoint between positions 15 and 16 (see Table 1).

The phylogenetic network for the extant four sequences (F–I) is shown in Fig. 3. Numbers along each edge are nucleotide sites, as shown in Table 1. The phylogenetic network shown in Fig. 2 is overlayed with dotted lines. Although there is a rectangle in this figure, it is clearly smaller than that shown in Fig. 2. Because outgroup sequence I was included in this phylogenetic network, the location of the ancestral sequence A was determined, and it became one of four apexes of the rectangle. Two parent sequence descendant lineages (F and G) are located on opposing vertices of the rectangle with long edges, while the recombinant lineage sequence H has a short edge and is located on the vertex opposing to the outgroup sequence I. One nucleotide difference at site 28 on the edge D–H correctly reproduced one nucleotide substitution occured after the recombination event (see Fig. 1). In contrast, edges J–F and K–G contain nucleotide substitutions both before and after the recombination event. If we consider Fig. 1, nucleotide substitutions before the recombination are located at edges A–B and A–C, while those after the recombination are located at edges B–F and C–G, for edges J–F and K–G of Fig. 3, respectively. This is why the lengths of edges J–F and K–G are longer than that of D–H.

It should be noted that, unlike nodes A and D, nodes J and K are not real sequences which once existed in the evolutionary history, but were created when the phylogenetic network for four extant sequences F, G, H. and I was constructed. In fact, nodes J and K correspond to unnamed nodes between edges B–D and C–E in Fig. 2, respectively. If we examine the phylogenetic network in Fig. 1, it is clear that edges B–D, B–E, C–D, and C–E were created by a recombination, while the all other edges created by accumulating mutations. Although they are all edges mathematically, biological characteristics of these two groups of edges are different. Because of this nature, now the recombination breakpoint estimate becomes after position 10 and before position 21, through examination of nucleotide positions of substitutions on edges K–D (corresponding to the upstream region) and J–D (corresponding to the downstream region). This estimate is inevitably wider than that (after position 13 and before position 18 when we have the full recombination rectangle shown in Fig. 2, yet still the true recombination breakpoint is included in this zone.

The phylogenetic network shown in Fig. 3 suggested a new and simple method to detect a recombination, either ancient or recent (Kitano et al., 2009). Three nucleotide sequences (alleles) are sampled from one population, and with a clear outgroup sequence, ideally from the closest different species, a phylogenetic network of four sequences is constructed. If this quartet sequence network shares the characteristic shown in Fig. 3, a recombination event is estimated to happen. Characteristics that should exist in this case are as follows:

(1) One rectangle should be observed.
(2) Two kinds of edges of the rectangle (horizontal and vertical ones of Fig. 3) should be dominated by upstream and downstream nucleotide sites of the sequence in question. These two edges, or partitions IG–FH and IF–GH, correspond to two different types of phylogenetically informative sites. If there are some violations, they might be caused by parallel changes and gene conversions after recombinations.
(3) The node connected to the outgroup sequence should be the ancestral sequence to the two parental sequences, and the node opposing that should be the recombinant sequence.

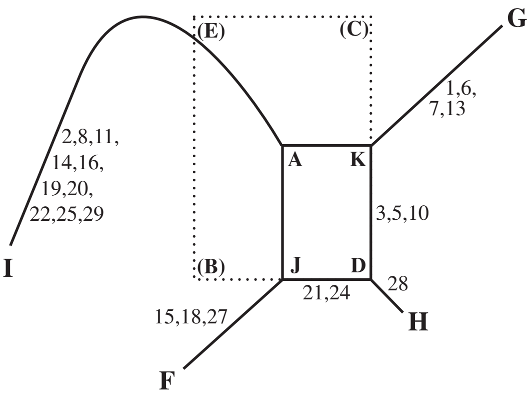


Fig. 3. The phylogenetic network for extant sequences F–I of Fig. 1 and Table 1. See text for a detailed explanation.

The remaining two nodes of the rectangle are parental sequences.

(4) Under the assumption of the approximate constancy of the evolutionary rate, the edge connecting the ancestral sequence and the outgroup sequence (edge A–I of Fig. 3) is expected to be the longest, while that connecting the recombinant and its descendant sequence (edge D–H of Fig. 3) should be the shortest. The two edges connecting parental sequences and their descendants (edges J–F and K–G of Fig. 3) should be longer than that involving the recombinant.

(5) Positions of nucleotide differences at two parent descendant lineage sequences should be dominant with either upstream or downstream positions, depending on the location of the parental sequence in the recombinant sequence (see Table 1). In Fig. 3, nucleotide positions on edge J–F should be dominated by downstream sites like those for edge D–J (or K–A), while that on edge K–G should be dominated by upstream sites like those for edge D–K (or J–A).

If all the five characteristics are satisfied, we conclude that there was a recombination in the past in that population. We can also estimate the time of recombination based on the length of the edge between recombinant and its descendant (edge D–H of Fig. 3). Kitano et al. (2009) did not name this new method, and here we would like to name this method as the phylogenetic network-based recombination detection method, or PNarec.

The PNarec method proposed by Kitano et al. (2009) was still under development. For example, selection of quartet sequences (recombinant descendant lineage sequence, two parent descendant lineage sequences, and the outgroup) was rather ad hoc. We therefore would like to formulate the algorithm for the PNarec method as follows. First, we need a set of sequences to be analyzed with the prefixed outgroup sequence. We developed a PNarec program, written in perl (see Supplementary file 1), and the input file is an output of MISHIMA (Kryukov and Saitou, 2010). The following is the algorithm of the PNarec method.

- *Step A*: Count numbers of singleton sites for each sequence of the multiple alignment of all N sequences in question.
- *Step B*: Choose all possible three sequences out of N sequences, and after adding the outgroup sequence, count the three types of phylogenetically informative sites, corresponding to three kinds of splits, for all quartet sequences.
- *Step C*: Choose quartets in which one of splits had no phylogenetically informative site.
- *Step D*: Examine the distribution of the remaining two phylogenetically informative sites for each quartet chosen at the last step, and choose ones whose site distributions are compatible with each other. Here, "compatible" means that the two phylogenetically informative sites of the two splits do not overlap, and distribute to the upstream and downstream nucleotide positions.
- *Step E*: If there is any quartet remained, choose the quartet whose recombinant descendant lineage sequence candidate has the smallest number of singleton sites. If there are equally smallest cases, choose the quartet in which the two parent descendant lineage sequence candidates have the largest sum of the singleton sites. If there are quartets with equally largest sums, choose one whose sum of two phylogenetically informative sites is the largest. Eliminate the recombinant descendant lineage sequence included in the chosen quartet, then return to Step E. Otherwise, stop the operation.

3. Applications

3.1. PNarec method for preassigned quartet sequences

We applied the PNarec method to five kinds of nucleotide sequence data, as listed in Table 2. Four of them are viral sequences, and the last one is HLA. We first constructed a phylogenetic network by using the Neighbor-Net method (Bryant and Moulton, 2004; Huson and Bryant, 2006) from a distance matrix, then using that as guide, character-state type network was produced manually from multiple aligned sequence data.

The first example is HIV1 vpu protein coding sequences. It is known that HIV sequences experience frequent recombinations (e.g., Randaut et al., 2004). In particular, BF strains are known to be recombinants (Carr et al., 2001). De Candia et al. (2010) compared four HIV sequences including one BF strain sequence (see Table 2A), and predicted the recombination at nucleotide positions around 120–140 using the SimPlot v.3.5.1 program. Fig. 4A shows the 10 phylogenetically informative nucleotide sites for the four sequences, and Fig. 4B shows their phylogenetic network. Application of the PNarec method confirmed that sequence CRF12_BF is

Table 2
Nucleotide sequences used in this study.[a]

(A) HIV1 vpu protein coding sequences (from De Candia et al. (2010))
Subtype A: AF143901
Subtype B (NL4-3): AF324493
Subtype F1 (BR020.1): AF005494
CRF12_BF: AF385936

(B) Influenza A PA protein coding sequences (from He et al. (2009)):
A/Albany/6/58(H2N2): AY209992
A/duck/Guangxi/xa/2001(H5N1): DQ997517
A/chicken/Guangdong/4/00(H9N2): DQ064493
A/chicken/China/Guangxi 17/2000(H9N2): DQ485223

(C) Influenza A PB protein coding sequences (from He et al. (2009)):
A/New_York/233/2000(H 1N1): CY002647
A/chicken/Jiangsu/ 1/00(H9N2): DQ064561
A/duck/Shanghai/ 13/2001(H5 N 1): AY585519
A/chicken/ China/ Guangxi1/2000(H9N2): DQ485205

(D) HTLV (from Strimmer and Moulton (2000))
L26585: L26585
L76050: L76050
L76045: L76045
D13784: D13784
L76054: L76054

(E) HLA (from Gu and Nei (1999))
HLA_A: NM_002116
HLA_B: NM_005514
HLA_C: NM_002117
HLA_B7301: AJ31601

[a] DDBJ/EMBL/GenBank International nucleotide sequence database accession numbers are given after colons.

A

```
==========================================
                          1 2 2 2 2 2
                    4 7 7 0 2 2 2 2 4 5
  Sequence name     4 0 3 5 2 3 4 6 1 2
------------------------------------------
 Subtype A          T A A T G T A G A G
 Subtype B (NL4-3)  C G T C A C A G G G
 Subtype F1 (BR020.1) T A A T G T T A A A
 CRF12_BF           C G T C A C T A G A
------------------------------------------
  Edge designation  X X X X X X Y Y X Y
==========================================
```

B

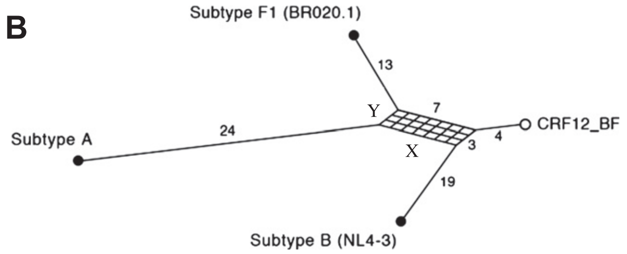


Fig. 4. Data analysis of four HIV sequences listed in Table 2A. (A) Nucleotide sequence data for 10 phylogenetically informative sites. (B) The phylogenetic network for the four HIV sequences. Numbers are edge lengths, and X and Y are edges formed by mutually incompatible sites.

descendant of the recombinant between ancestors of subtype B (NL4-3) and subtype F1 (BR020.1), for all the five criteria listed in the previous section were satisfied. However, sequence data shown in Fig. 4A suggest that the recombination breakpoint was most probably between nucleotide sites 223 and 224, in which the edge assignment changed from X to Y (see Fig. 4B for the two edges). Nucleotide site 241 was also assigned to edge X, but it can be considered as a result of a parallel substitution occurred after the recombination.

The next two examples are both influenza A virus genes. He et al. (2009) reported many avian flu strains which may be potential recombinants. They used a series of techniques, including the Neighbor-Net method. We chose two quartet sets of PA and PB protein coding sequences from He et al. (2009), as listed in Tables 2B and C, respectively. Their nucleotide configuration patterns are shown in Table 3. Because the numbers of phylogenetically informative sites are much larger than that shown in Fig. 4A, an abbreviated notation was used in this table. The majority of sites were one of two configurations, denoted as *X* or *Y*, and the minority configurations within these majority ones are numbered. These minority configurations were probably created with parallel changes after recombinations. Recombination breakpoints are estimated to be between sites 1333 and 1359 for the PA gene and between sites 1067 and 1085 for the PB gene. These locations are very similar to those (1332 and 1066) estimated by He et al. (2009). Fig. 5A and B are phylogenetic networks based on PA and PB nucleotide sequences (see Table 2), respectively. The five criteria for the PNarec method are again all satisfied in both networks, and it is clear that A/chicken/China/Guangxi17/2000(H9N2) and A/chicken/China/Guangxi1/2000(H9N2) are descendants of recombinants. It should be noted that the both phylogenetic networks are not two-dimensional but three-dimensional. However, the shortest edges are most probably results of parallel changes, and essentially the expected two-dimensional pattern shown in Fig. 3 was observed.

The next example is the five sequence set of human T-cell leukemia virus (HTLV) analyzed by Strimmer and Moulton (2000), as listed in Table 2D. A phylogenetic network based on their distance matrix produced by using the Neighbor-Net method is shown in Fig. 6A. The 11 phylogenetically informative sites are shown in Fig. 6B, and the phylogenetic network based on the multiply aligned nucleotide sequences, including phylogenetically non-informative sites, is shown in Fig. 6C. The topological relationship is identical between the two phylogenetic networks, however, proportions of edge lengths are slightly different.

If we ignore the sequence L76045, the two nucleotide sites 214 and 222 (designated as S and T at the last row) become singletons, and the remaining four sequences satisfy the five criteria of the PNarec method; sequence L76054 is the recombinant descendant lineage sequence, sequence L26585 is the outgroup, and the remaining two sequences (L76050 and D13784) are parent descendant lineage sequences. However, Strimmer and Moulton (2000) considered a node near the sequence L76050 as the most plausible root, and three other sequences as descendants derived from that root (see their Fig. 6 and result). Strimmer et al. (2001) reevaluated these sequences using the ancestral recombination graph method, and estimated that L76054 is the recombinant lineage sequence, and recombination occurred between ancestors of sequences L76050 and D13784, while locating L26585 as the outgroup. This new explanation is the same as our conclusion obtained by applying the PNarec method. It clearly shows that we do not need complicated methods to infer a recombination in the past. It should be added that the recombination breakpoint is probably located between site 466 and 573 according to the sequence pattern shown in Fig. 6B.

Table 3
Nucleotide configuration patterns of Influenza virus sequences.

Sequence name	X1	1	X2	2	X3	3	Y1	4	Y2
(A) Four Influenza A PA protein coding sequences[a]									
A/Albany/6/58(H2N2)	–	–	–	–	–	–	–	–	–
A/duck/Guangxi/xa/2001(H5N1)	+	–	+	+	+	+	–	+	–
A/chicken/Guangdong/4/00(H9N2)	–	+	–	+	–	+	+	–	+
A/chicken/China/Guangxi17/2000(H9N2)	+	+	+	–	+	–	+	+	+

Sequence name	X1	1	X2	Y1	2	Y2	3	Y3	4	Y4	5	Y5
(B) Four Influenza A PB protein coding sequences[b]												
A/New_York/233/2000(H1N1)	–	–	–	–	–	–	–	–	–	–	–	–
A/chicken/Jiangsu/1/00(H9N2)	+	+	+	–	+	–	–	+	+	–	+	–
A/duck/Shanghai/13/2001(H5N1)	–	+	–	+	–	+	+	–	–	+	–	+
A/chicken/China/Guangxi1/2000(H9N2)	+	–	+	+	+	+	+	+	+	+	+	+

[a] Notes: X1: 32 sites (from site 69 to site 630), 1: site 642, X2: 24 sites (from site 759 to site 1098), 2: site 1158, X3: 12 sites (from site 1170 to site 1332), 3: site 1335, Y1 (from site 1360 to site 1878): 31 sites, 4: site 1971, Y2: 8 sites (from site 1996 to site 2112).

[b] Notes: X1: 9 sites (from site 27 to site 387), 1: site 396, X2: 30 sites (from site 432 to site 1066), Y1: 20 sites (from site 1086 to site 1620), 2: site 1668, Y2: 4 sites (from site 1680 to site 1722), 3: site 1728, Y3: 5 sites (from site 1734 to site 1815), 4: site 1816, Y4: 6 sites (from site 1839 to site 2031), 5: site 2061, Y5: 4 sites (from site 2073 to site 2193).

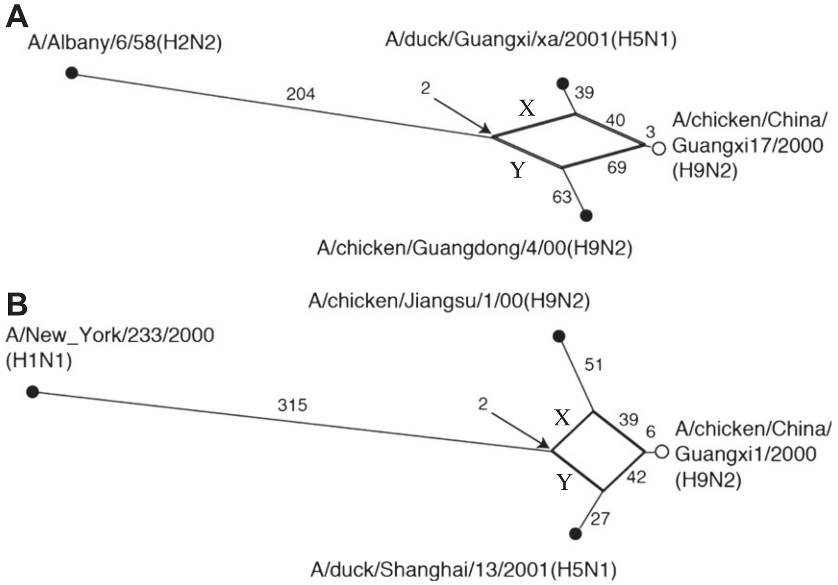
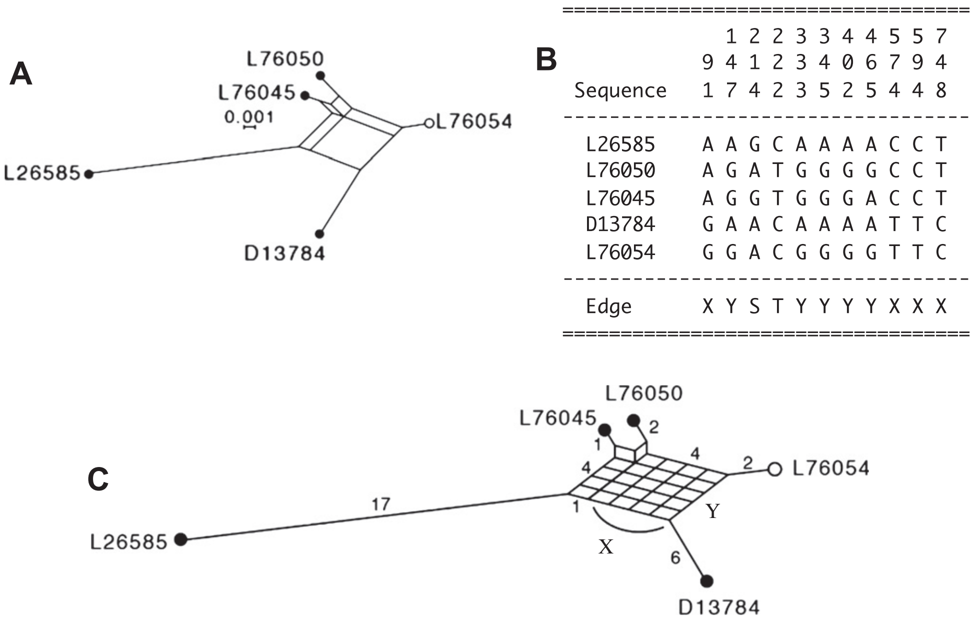


Fig. 5. Phylogenetic networks of influenza virus genes. (A) The phylogenetic network for the four influenza PA gene sequences listed in Table 2B. (B) The phylogenetic network for the four influenza PB gene sequences listed in Table 2C.



```
===================================
             1 2 2 3 3 4 4 5 5 7
             9 4 1 2 3 4 0 6 7 9 4
 Sequence    1 7 4 2 3 5 2 5 4 4 8
-----------------------------------
  L26585     A A G C A A A A C C T
  L76050     A G A T G G G G C C T
  L76045     A G G T G G G A C C T
  D13784     G A A C A A A A T T C
  L76054     G G A C G G G G T T C
-----------------------------------
  Edge       X Y S T Y Y Y Y X X X
===================================
```

Fig. 6. Data analysis of four HTLV sequences listed in Table 2D. (A) A distance matrix-based phylogenetic network. (B) Nucleotide sequence data for 11 phylogenetically informative sites. "Edge" at the bottom shows classification of each nucleotide site in terms of nucleotide configuration, corresponding to particular edge. (C) The phylogenetic network for the four HTLV sequences listed in Table 2D. X and Y correspond to edges X and Y in (B).

The last example is from human genes. Gu and Nei (1999) constructed phylogenetic trees for many human MHC genes, and proposed that allele HLA-B*7301 was generated by a interlocus recombination between HLA B and HLA C alleles. We applied the PNarec method to the four sequences shown in Table 2E; recombinant HLA_B7301 sequence, representative HLA B and C alleles as parental sequences, and one HLA A allele as the outgroup. Fig. 7A lists the phylogenetically informative sites. They are classified into three configurations corresponding to three edges. Fig. 7B is the phylogenetic network for these four sequences. Although the edge length (31) connected to one putative parental lineage (HLA_C) was longer than that (17) for the putative recombinant lineage (HLA_B7301) as expected, the edge length (13) for the another putative parental lineage (HLA_B) was shorter. However, if we consider the edge Z as produced by a parallel change occurred to the phylogenetically noninformative partition mapped to the edge connected to HLA_B, the short edge length becomes 15. In a similar manner, 3 partitions (or edge) X after the supposed recombination breakpoint (between site 748 and site 828) and 1 partition Y before that point can be produced through parallel substitutions to some phylogenetically noninformative partitions. If some of them happened on the edge connected to HLA_B, that edge length may become similar or even slightly longer than that for HLA_B7301.

Unlike many existing methods for detecting recombinations, we can estimate the time of recombination in the PNarec method. The three edge lengths connected to the extant sequences excluding

A

```
===========================================================
                                                          1
               1 1 1 2 2 2 3 3 5 5 7 7 7 8 8 8 8 9 9 9 9 0
           2 4 4 6 7 5 6 6 0 4 5 6 3 4 4 2 6 7 7 0 1 3 6 9
 Sequence  5 4 4 5 5 9 1 9 1 1 9 0 5 4 7 9 1 2 4 0 0 3 0 2
-----------------------------------------------------------
HLA_A      C G C G C G G A G C A C G G C G T C A G C C T T
HLA_B      G C A C A A C T A C G A C A T G T C A G T T T A
HLA_C      C G A G C G G A A A A C G G C C C A G A C C C T
HLA_B7301  G C C C A A C T G A G A C A T C C A G A T T C A
-----------------------------------------------------------
 Edge      X X Z X X X X X Z Y X X X X X Y Y Y Y Y X X Y X
===========================================================
```

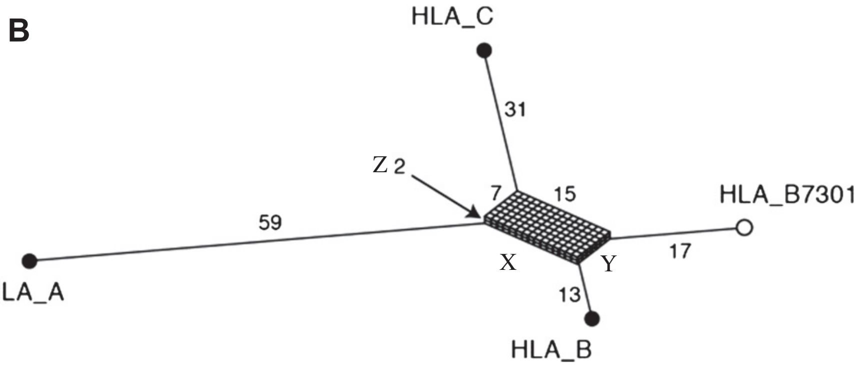


Fig. 7. Data analysis of four HLA gene sequences listed in Table 2E. (A) Nucleotide sequence data for 24 phylogenetically informative sites. "Edge" at the bottom shows classification of each nucleotide site in terms of nucleotide configuration, corresponding to particular edge. (B) The phylogenetic network for the four HL sequences. X–Z correspond to edges X–Z in (A).

the outgroup were simultaneously considered in their original method, we propose a simpler method; to use only the edge length going to the recombinant lineage. Under the assumption of the constancy of the evolutionary rate, λ, for the each situation, the time of recombination, *Tr*, can be estimated by:

$$Tr = d/\lambda, \quad (1)$$

where *d* is the evolutionary distance given as the number of nucleotide substitutions.

Table 4 shows the number (dS) of synonymous substitutions per site, the evolutionary rate (λ) for each situation, and estimated time of recombination with lower and upper limits, corresponding to 1 S.E. of dS values. Hanada et al.'s (2004) estimates on the synonymous rates were used for HIV, influenza virus, and HTLV, while the evolutionary rate estimated by Piontkivska and Nei (2003) was used for HLA. The recombination time for the vpu gene of HIV was estimated to be 7.7 years, but it was based on only 4 nucleotide differences, and the lower and upper limits are 0.0 years and 15.4 years, respectively. However, even this rough estimate may be able to give some biological significance. The times of recombinations for the two influenza virus protein genes were both around 1 year, consistent with the rapid change of strains of this virus. In contrast, the synonymous substitution rate of the HTLV was estimated to be ~1,000 times lower than influenza virus or HIV (Hanada et al., 2004), and accordingly, the time of recombination was estimated to be more than 1,000 years ago. The rate of synonymous substitutions for the human gene is much slower than that for HTLVs, and the time of the inter-locus recombination between an HLA_B and HLA_C allele was estimated to occur around 11 million years ago, before the divergence of humans, chimpanzees, and gorillas. This estimate is compatible with that of Abi-Rached et al. (2011) who estimated this interlocus recombination timing before the speciation of human, chimpanzee, and gorilla (see their Fig. 1B).

3.2. PNarec method for a set of sequences

Let us apply the PNarec algorithm to the sequence data determined and analyzed by Kitano et al. (2009). We developed a program written in perl based on the algorithm given in Theory section. Because it is better if the outgroup sequence is closely related to the population or species in question, we chose one sequence, SI-2, of siamang (*Symphalangus syndactylus*) from Fig. 1 of Kitano et al. (2009), and confined our search among sequences of the two more closely related gibbon species, agile gibbon (*Hylobates agilis*) and white-handed gibbon (*Hylobates lar*). We chose six

Table 4
Estimation of recombination times for five data.

Gene	dS (+SE)	Rate (Ref.)[a]	Time of recombination in year (a–b–c)[b]		
PA of Influenza A	0.00434+0.00307	6.84E−03 (1)	0.2	0.6	1.1
PB2 of Influenza A	0.00770+0.00386	6.84E−03 (1)	0.6	1.1	1.7
vpu of HIV1	0.01827+0.01832	2.38E−03 (1)	0.0	7.7	15.4
HTLV	0.00825+0.00584	5.20E−06 (1)	463	1587	2710
HLA	0.01770+0.00407	1.60E−09 (2)	8.5 MYA	11.1 MYA	13.6 MYA

[a] 1: Hanada et al. (2004), 2: Piontkivska and Nei (2003).

[b] a: Lower limit, b: estimated time, c: upper limit (in years, except for HLA in MYA: million years ago).

Table 5
Multiple alignment of six sequences.

Name	Sequence (variant sites only)	S
0: SI-2*	....	13
1: AG-10	G.....A.T.TTATA...GGCGTG.AACA...G.AGT.GC..A..TCC.ATG.CTGCGCGTAA	8
2: AG-1	....T...T.TTA.....GGC.TG...CAC.TG.AGT.GC.AA...CC...G.CTGCGCGTAA	1
3: WH-6	.TA.T..C..TTA...C.GGC.TG...CA...G.AGT.GC.AA..TCC.A.G.CTGCGCGTAA	2
4: WH-5	...TTG.C.CTTA..GCAGGC.TGA..CA...GAAGT.GC......CC.A.G.CTGCGCGTAA	1
5: WH-1	...TTG.C.CTTA..GCAGGC.TGA..CA.C.G.AGTTG.C..TA.CCT....C.GC......	6
6: AG-7	....T...T.TTA..GCAGGC.TG.....C..G.A.T.GC......CC....AC.G.......	1

Notes: Seven sequence names are from Kitano et al. (2009). When nucleotide of a sequence is identical with that of the outgroup sequence (SI-2), a dot is given. *S* values are number of singletons.

sequences which are representatives of 6 clusters defined in the phylogenetic network constructed by Kitano et al. (2009); (see their Fig. 2). The multiple alignment of these six sequences are shown in Table 5.

The result of step A of the new algorithm is shown as *S* values of Table 5. Because there are 6 sequences to be compared, there are 20 (=$_6C_3$) trios, and all of them are examined for number of three phylogenetically informative configurations for four sequences (the selected trio and the outgroup) at step B. We choose quartets in which one of splits had no phylogenetically informative site at step C, and 9 quartets were selected. The distribution of the remaining two phylogenetically informative sites were examined to be compatible or not at step D, and six quartets are left for three sequences as the recombinant descendant sequences; sequences 2, 3, and 4. These three sequences are sequentially chosen at step E in the order of sequence 4 (WH-5), sequence 2 (AG-1), and sequence 3 (WH-6) in a recursive way. The perl program, example input file, and output file are in Supplementary material.

The order of three recombinations from recent to ancient times is slightly different from that obtained by Kitano et al. (2009); (see their Fig. 5). Because we used a siamang sequence (SI-2) as the outgroup in this example, we could not detect the more ancient fourth and fifth recombinations discovered by Kitano et al. (2009), who used the human ABO A allele sequence as the outgroup.

4. Discussion

We definitely need more study to further strengthen the utility of the PNarec method. First of all, this method should be compared with the existing methods on recombination detection using computer simulations. Secondly, a statistical test should be introduced to the PNarec method. In this regard, Gauthier and Lapointe (2007) study on hybridization may be helpful. Although recombination and hybridization are biologically different, they have a similar mathematical framework, and they indeed used a quartet method with statistical test.

Another problem of the current PNarec algorithm is its simplistic nature. Quartets in which one of splits had no phylogenetically informative site are chosen at step C. However, we saw that all the three informative splits contained some sites for influenza A virus sequences (Fig. 5) and for HLA sequences (Fig. 7). Another too-stringent condition is step D in which consistency between two phylogenetically informative sites are required. We saw some inconsistent patterns in viral sequences. In fact, if we use the human ABO A allele sequence as the outgroup following Kitano et al. (2009), only two sequences (4 and 5) are selected (data not shown), because some parallel changes eliminated quartets including authentic recombinant descendant sequences. Therefore, it is clear that we have to devise a more sophisticated algorithm to handle such cases.

Acknowledgments

We appreciate Dr. Derek Wildman and other people who organized the symposium to memorize late Dr. Morris Goodman, to which one of us (N.S.) attended in October, 2011 with great pleasure. We also appreciate two anonymous reviewers for their valuable comments.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2012.09.015.

References

Abi-Rached, L. et al., 2011. The shaping of modern human immune systems by multiregional admixture with archaic humans. Science 334, 89–94.
Bandelt, H.J., 1994. Phylogenetic networks. Verh. Natwiss. Ver. Hamburg 34, 51–71.
Bryant, D., Moulton, V., 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol. Biol. Evol. 21, 255–265.
Carr, J.K. et al., 2001. Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America. AIDS 15, F41–F47.
De Candia, C., Esparda, C., Duette, G., Ghiglione, Y., Turk, G., Salomon, H., Carobene, M., 2010. Viral replication is enhanced by an HIV-1 intersubtype recombination-derived Vpu protein. Virol. J. 7, 259.
Ezawa, K., Ikeo, K., Gojobori, T., Saitou, N., 2010. Evolutionary pattern of gene homogenization between primate-specific paralogs after human and macaque speciation using the 4-2-4 method. Mol. Biol. Evol. 27, 2152–2171.
Gauthier, Lapointe, 2007. Hybrids and phylogenetics revisited: a statistical test of hybridization using quartets. Syst. Bot. 32, 8–15.
Gu, X., Nei, M., 1999. Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. Mol. Biol. Evol. 16, 147–156.
Hanada, K., Suzuki, Y., Gojobori, T., 2004. A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes. Mol. Biol. Evol. 21, 1074–1080.
He, C.-Q. et al., 2009. Homologous recombination as an evolutionary force in the avian influenza A virus. Mol. Biol. Evol. 26, 177–187.
Huson, D.H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Kitano, T., 2012. Application of phylogenetic network. In: Hirai, H., Imai, H., Go, Y. (Eds.), Post-genome biology of primates. Springer, Tokyo, pp. 181–190 (Chapter 12).

Kitano, T., Saitou, N., 1999. Evolution of Rh blood group genes have experienced gene conversions and positive selection. J. Mol. Evol. 49, 615–626.

Kitano, T., Noda, R., Sumiyama, K., Ferrell, R.E., Saitou, N., 2000. Gene diversity of chimpanzee ABO blood group genes elucidated from intron 6 sequences. J. Hered. 91, 211–214.

Kitano, T., Noda, R., Takenaka, O., Saitou, N., 2009. Relic of ancient recombinations in gibbon ABO blood group genes deciphered through phylogenetic network analysis. Mol. Phylogenet. Evol. 51, 465–471.

Kitano, T., Blancher, A., Saitou, N. 2012. The functional A allele was resurrected via recombination in the human ABO blood group gene. Mol. Biol. Evol. electrically published on March 8, 2012.

Kryukov, K., Saitou, N., 2010. MISHIMA - a new method for high speed multiple alignment of nucleotide sequences of bacterial genome scale data. BMC Bioinformatics 11, 142.

Noda, R., Kitano, T., Takenaka, O., Saitou, N., 2000. Evolution of the ABO blood group gene in Japanese macaque. Genes Genet. Syst. 75, 141–147.

Ogasawara, K., Yabe, R., Uchikawa, M., Saitou, N., Bannai, M., Nakata, K., Takenaka, M., Fujisawa, K., Juji, T., Tokunaga, K., 1996. Molecular genetic analysis of the variant phenotypes of the ABO blood group system. Blood 77, 2732–2737.

Piontkivska, H., Nei, M., 2003. Birth-and-death evolution in primate MHC Class I genes: divergence time estimates. Mol. Biol. Evol. 20, 601–609.

Randaut, A., Posada, D., Crandall, K.A., Holmes, E.C., 2004. The causes and consequences of HIV evolution. Nat. Rev. Genet. 5, 52–61.

Roubinet, F., Despiau, S., Calafell, F., Jin, F., Bertranpetit, J., Saitou, N., Blancher, A., 2004. Evolution of the O alleles of the human ABO blood group gene. Transfusion 44, 707–715.

Saitou, N., Yamamoto, F., 1997. Evolution of primate ABO blood group genes and their homologous genes. Mol. Biol. Evol. 14, 399–411.

Strimmer, K., Moulton, V., 2000. Likelihood analysis of phylogenetic networks using directed graphical models. Mol. Biol. Evol. 17, 875–881.

Strimmer, K., Wief, C., Moulton, V., 2001. Recombination analysis using directed graphical models. Mol. Biol. Evol. 18, 97–99.

Tomiki, T., Saitou, N., 2004. Phylogenetic analysis of proteins associated in the four major energy metabolism systems: photosynthesis, aerobic respiration, denitrification, and sulfur respiration. J. Mol. Evol. 59, 158–176.

Yamamoto, F., 2000. Molecular genetics of ABO. Vox Sang. 78, 91–103.

Yamamoto, F., Clausen, H., White, T., Marken, J., Hakomori, S., 1990. Molecular genetic basis of the histo-blood group ABO system. Nature 345, 229–233.