Chapter 26
Comparative Genomics

Xuhua Xia

Abstract Comparative genomics was previously misguided by the naïve dogma that what is true in *E. coli* is also true in the elephant. With the rejection of such a dogma, comparative genomics has been positioned in proper evolutionary context. Here I numerically illustrate the application of phylogeny-based comparative methods in comparative genomics involving both continuous and discrete characters to solve problems from characterizing functional association of genes to detection of horizontal gene transfer and viral genome recombination, together with a detailed explanation and numerical illustration of statistical significance tests based on the false discovery rate (FDR). FDR methods are essential for multiple comparisons associated with almost any large-scale comparative genomic studies. I discuss the strength and weakness of the methods and provide some guidelines on their proper applications.

26.1 Introduction

The development of comparative genomics predates the availability of genomic sequences. It has long been known that organisms are related, with many homologous genes sharing similar functions among diverse organisms. For example, the yeast *IRA2* gene is homologous to the human *NF1* gene, and the functional equivalence of the two genes was demonstrated by the yeast *IRA2* mutant being rescued by the human *NF1* gene [5]. This suggests the possibility that simple genomes can be used as a model to study complicated genomes. A multitude of such demonstrations of functional equivalence of homologous genes across diverse organisms has led to the dogmatic assertion that what is true in *E. coli* is also true in the elephant [attributed to Jacques Monod, [33], p. 290].

X. Xia
Department of Biology, University of Ottawa, Ottawa, Canada
e-mail: Xuhua.Xia@uottawa.ca

H. Horng-Shing Lu et al. (eds.), *Handbook of Statistical Bioinformatics*,
Springer Handbooks of Computational Statistics, DOI 10.1007/978-3-642-16345-6_26,

It is the realization that what is true in *E. coli* is often not true in the elephant that has brought comparative genomics into the proper evolutionary context. The impact of this realization on comparative genomics is best illustrated by a simple example. Suppose we compare a Cadillac Deville and a Dodge Caravan. The two are similar in functionality except that the Caddy warns the driver when it is backing towards an object behind the car. What is the structural basis of this warning function? Nearly all structural elements in the Caddy have their 'homologues' in the Dodge Caravan except for the four sensors on the rear bumper. This would lead us to quickly hypothesize that the four sensors are associated with the warning function, which turns out to be true. Now if we replace the Dodge Caravan with a baby stroller, then the comparison will be quite difficult because a stroller and a Caddy differ structurally in numerous ways and any structural difference could be responsible for the warning function. We may mistakenly hypothesize that the rear lights, the antenna or the rear window defroster in the Caddy, which are all missing in the stroller, may be responsible for the warning function. To test the hypotheses, we would destroy the rear lights, the antenna, the rear window defroster, etc., one by one, but will get nothing but negative results. What could be even worse is that, when destroying the rear lights, we accidentally destroy a part of the electric system in such a way that the warning function is lost, which would mislead us to conclude that the rear lights are indeed part of the structural basis responsible for the warning function-an 'experimentally substantiated' yet wrong conclusion. A claim that what is true in *E. coli* is also true in the elephant is equivalent to a claim that what is true in the stroller is also true in the Caddy. It will take comparative genomics out of its proper conceptual framework in evolutionary biology.

Evolutionary theory states that all genetic variation, including genomic variation, results from two sculptors of nature, i.e., mutation (including recombination) and selection. Thus, any genomic difference can be attributed to differences in differential mutation and selection pressure. This allows us not only to characterize evolutionary changes along different evolutionary lineages, but also to seek evolutionary processes underlying the character changes. In particular, evolutionary biology provides the proper comparative methods [7, 20, 28, 55, 71] for comparative genomics.

In what follows, I will numerically illustrate the comparative methods for analyzing genomic features that are either continuous or discrete. Large-scale comparative genomic studies almost always lead to multiple comparisons. So I will also illustrate the computation involved in controlling for false discovery rate which represents a key development in recent studies of statistical significance tests [8, 9]. One evolutionary process that has shaped bacterial genomes is the horizontal gene transfer, and the phylogenetic incongruence test used to detect such horizontal gene transfer events will be illustrated. The last section covers comparative genomic methods for detecting recombination events and mapping recombination points.

While molecular phylogenetics is often essential in comparative genomics, the subject has been treated fully elsewhere [22, 50, 66]. Simple overviews of the subject are also available [4, 66, 87]. A more egregious omission in this chapter is genome rearrangement, but interested readers may consult the publications of my

colleague at University of Ottawa, David Sankoff, who is a pioneer in the field and wrote excellent reviews on the subject [69, 70]. A large-scale empirical study of genome rearrangement in yeast species following a whole-genome duplication (WGD) event, featuring a meticulous reconstruction of gene order of the ancestral genome before WGD, has recently been published [26].

26.2 The comparative Method for Continuous Characters

26.2.1 Variation in Genomic GC% Among Bacterial Species

Studies of the variation in genomic GC% among bacterial species serve as the easiest entry point into comparative genomics. Wide variation in genomic GC% is observed in bacterial species. A popular selectionist hypothesis is that bacterial species living in high temperature should have high genomic GC% for two reasons. First, an increased GC usage, with more hydrogen bonds between the two DNA strands, would stabilize the physical structure of the genome [42, 64]. Second, high temperature would need more thermostable amino acids [3] which are typically coded by GC-rich codons. This implies that genomic GC% should increase with optimal grow temperature (OGT) in bacterial species. While this prediction is not supported, either based on results of sequence analysis [24] or by experimental studies [94], it has been found that GC% of rRNA genes is highly correlated with OGT [24, 30, 49, 79], [18, p. 535]. In particular, when the loop and stem regions of rRNA are studied separately, it was found that the hyperthermophilic bacterial species not only have higher proportion of GC in the stems but also longer stems [80]. In contrast, the GC% in the loop region correlates only weakly with OGT. Because stems function to stabilize the RNA secondary structure which is functionally important, these results are consistent with the hypothesized selection for RNA structural stability in high environmental temperatures.

When studying the relationship between two quantitative variables, such as OGT and stem GC%, a phylogeny-based comparison is crucially important to avoid violation of statistical assumptions. Figure 26.1 illustrates a case in which one may mistakenly conclude a positive relationship between X and Y when the 16 data points are taken as independent. A phylogenetic tree superimposed on the points allows us to see immediately that the data points are not independent. All eight points in the left share one common ancestor, so do the eight points in the right. So the superficial association between X and Y could be due to a single coincidental change in X and Y in one of the two common ancestors. One needs to use the phylogeny-based method, such as independent contrasts [20], [22, pp. 432–459] or the generalized least-squares method [46, 56, 57] when assessing the relationship between quantitative variables.

While the derivation and mathematical justification of the phylogeny-based comparative method is quite complicated, the most fundamental assumption is the

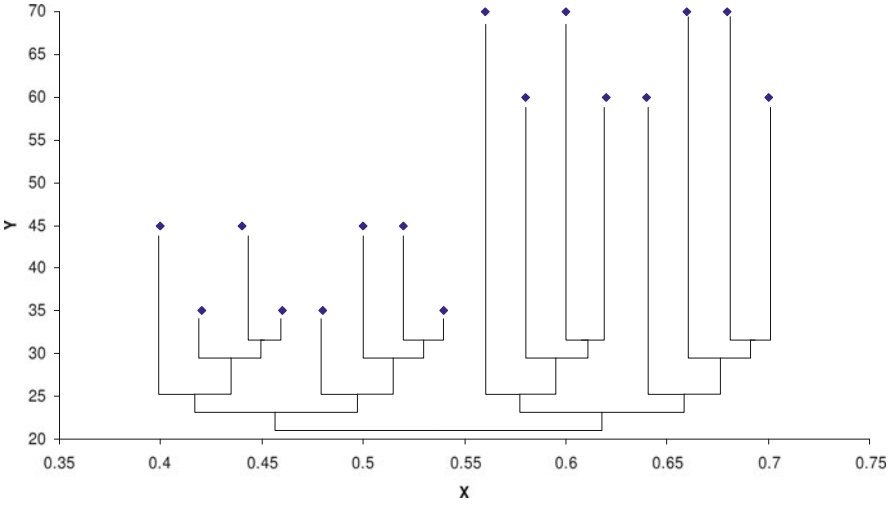

Fig. 26.1 Phylogeny-based comparison is important for evolutionary studies. The data points, when wrongly taken as independent, would result in a significant positive but spurious relationship between Y and X (which represent any two continuous variables, e.g., GC% and OGT)

Brownian motion model [22, pp. 391–414] which appears reasonable for neutrally evolving continuous characters. Here I illustrate the actual computation of independent contrasts with a numerical example to facilitate its application to comparative genomics, prompted by my personal belief that one generally cannot interpret the results properly if one does not know how the results are obtained.

Suppose a phylogeny of eight bacterial species whose OGT and GC% of rRNA genes have been measured, with the eight species referred to hereafter as s_1 to s_8 from left to right in Fig. 26.2. The computation is recursive, and is exactly the same for any quantitative variable. So we will only illustrate the computation involving OGT. One may repeat the computation involving GC% as an exercise.

The computation is of three steps. First, we recursively compute the ancestral values for internal (ancestral) nodes x_1 to x_6. We treat these ancestors as if they were new taxa and compute the branch lengths leading to these ancestral nodes. We may start with the two sister species s_1 and s_2. The OGT of their ancestor (x_1) is a weighted average of the OGT values for s_1 and s_2 (weighted by the branch lengths):

$$OGT_{x_1} = \frac{v_2}{v_1 + v_2} OGT_{s_1} + \frac{v_1}{v_1 + v_2} OGT_{s_2} = \frac{3 \times 70}{4} + \frac{1 \times 74}{4} = 71 \quad (26.1)$$

One may note that the weighting scheme in (26.1) is such that the ancestral state is more similar to the state of the descendent node with a shorter branch than the other with a longer branch. This makes intuitive sense as a descendent node diverged much from the ancestor should be less reliable for inferring the ancestral state than a descendent node diverged little from the ancestor.

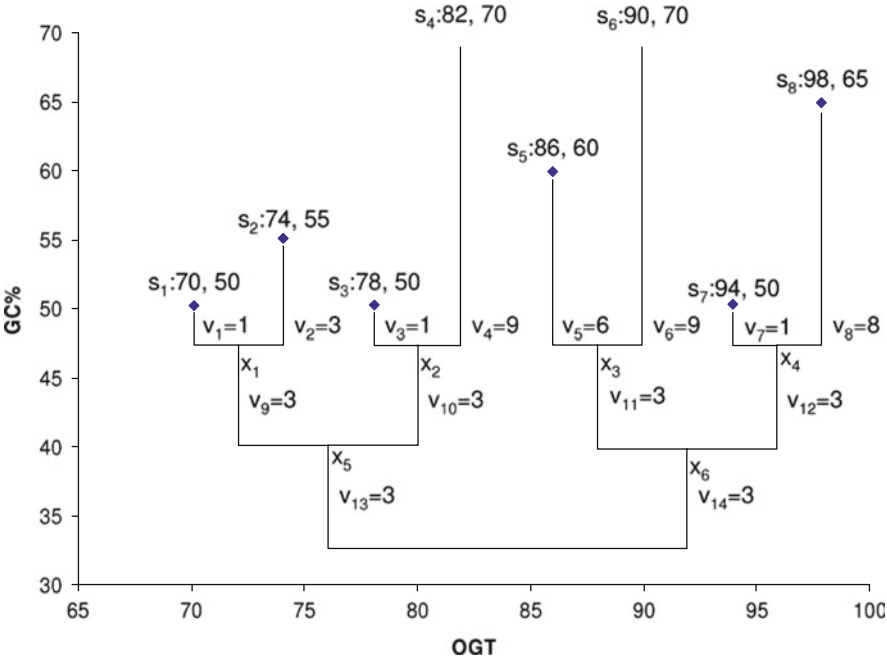

Fig. 26.2 A phylogeny of eight bacterial species (s_1–s_8) each labeled with optimal growth temperature (OGT) and GC% of the stem region of rRNA genes in the format of 'OGT, GC%'. The branch lengths ($v_1 - v_{14}$) are next to the branches. Ancestral nodes are designated by x_1 to x_6

We now treat x_1 as if it is a new taxon and compute the branch lengths leading to it from its ancestor (x_5) as

$$v_{x_1} = \frac{v_1 v_2}{v_1 + v_2} + v_9 = \frac{1 \times 3}{1 + 3} + 3 = 3.75 \tag{26.2}$$

We do the same for x_2 to x_4, and the associated OGT_{xi} and v_{xi} values are listed in Table 26.1. The computation of the ancestral states for x_5 and x_6 is similar to that in (26.1), e.g.,

$$OGT_{x_5} = \frac{v_{x_2} OGT_{x_1}}{v_{x_1} + v_{x_2}} + \frac{v_{x_1} OGT_{x_2}}{v_{x_1} + v_{x_2}} = \frac{3.9 \times 71}{7.65} + \frac{3.75 \times 78.4}{7.65} = 74.63 \tag{26.3}$$

Now we can take the second step to compute the unweighted contrasts (designated by C) as well as the sum of branch lengths linking the two contrasted taxa. With eight species, we have seven ($= n - 1$, where n is the number of species) contrasts (first column in Table 26.2). These unweighted contrasts, as well as the sum of branch lengths (SumV) associated with the contrasts, are illustrated for those between s_1 and s_2 and between x_1 and x_2 for OGT in (26.4). All the computed unweighted contrasts for both OGT and GC%, as well as the associated SumV

Table 26.1 Computed ancestral states (OGT_{xi} and GC_{xi}) and the branch lengths (v_{xi}) for the six ancestral nodes

x_i	OGT_{x_i}	v_{x_i}	GC_{x_i}
x_1	71.0000	3.7500	51.2500
x_2	78.4000	3.9000	52.0000
x_3	87.6000	6.6000	64.0000
x_4	94.4444	3.8889	51.6667
x_5	74.6275	4.9118	51.6176
x_6	91.9068	5.4470	56.2394

Table 26.2 Unweighted and weight contrasts for the two quantitative variables OGT and GC%

Contrast	Unweighted Contrasts		SumV	Weighted Contrasts	
	OGT	GC%		WC_{OGT}	$WC_{GC\%}$
$s_1 - s_2$	−4.0000	−5.0000	4.0000	−2.0000	−2.5000
$s_3 - s_4$	−4.0000	−20.0000	10.0000	−1.2649	−6.3246
$s_5 - s_6$	−4.0000	−10.0000	15.0000	−1.0328	−2.5820
$s_7 - s_8$	−4.0000	−15.0000	9.0000	−1.3333	−5.0000
$x_1 - x_2$	−7.4000	−0.7500	7.6500	−2.6755	−0.2712
$x_3 - x_4$	−6.8444	12.3333	10.4889	−2.1134	3.8082
$x_5 - x_6$	−17.279	4.6218	10.3588	−5.3687	−1.4360

values, are listed in columns 2–4 in Table 26.2.

$$\begin{aligned} C_{s_1-s_2OGT} &= OGT_{s_1} - OGT_{s_2} = 70 - 74 = -4 \\ SumV_{C_{s_1-s_2}} &= v_1 + v_2 = 1 + 3 = 4 \\ C_{x_1-x_2OGT} &= OGT_{x_1} - OGT_{x_2} = 71 - 78.4 = -7.4 \\ SumV_{C_{x_1-x_2}} &= v_{x_1} + v_{x_2} = 3.75 + 3.9 = 7.65 \end{aligned} \tag{26.4}$$

We can now take the third step of obtaining independent weighted contrasts (WC) by dividing each unweighted contrasts by the square root of the associated SumV. For example,

$$\begin{aligned} WC_{s_1-s_2OGT} &= \frac{C_{s_1-s_2OGT}}{\sqrt{Sum\ V_{s_1-s_2}}} = \frac{-4}{\sqrt{4}} = -2 \\ WC_{x_1-x_2OGT} &= \frac{C_{x_1-x_2OGT}}{\sqrt{Sum\ V_{x_1-x_2}}} = \frac{-7.4}{\sqrt{7.65}} = -2.6755 \end{aligned} \tag{26.5}$$

These independent contrasts for OGT thus computed, together with those for GC%, are shown in the last two columns in Table 26.2. Now we need to assess the relationship between WC_{OGT} and $WC_{GC\%}$, specifically whether an increase in OGT will result in an increase in GC%, i.e., whether the two are positively correlated. There are two ways to assess the relationship. The first is parametric by performing a linear regression of $WC_{GC\%}$ on WC_{OGT}, forcing the intercept equal to 0. The reason for a zero intercept is that we do not expect a change in GC% if there is no change in OGT. The resulting slope is 0.4647. The regression accounts for 11.17% of the

variation in $WC_{GC\%}$. The square root of 11.17%, equal to 0.3342, is the correlation coefficient between the two. Of course you may also do a regression of WC_{OGT} on $WC_{GC\%}$, which will result in a slope of 0.2403. These slopes and the correlation coefficients are in the default output in the CONTRAST program in PHYLIP [21]. The relationship between WC_{OGT} and $WC_{GC\%}$, although positive, is not significant ($p = 0.4249$).

One may also assess the relationship between WC_{OGT} and $WC_{GC\%}$ by using nonparametric tests. For example, we expect half of the (WC_{OGT}, $WC_{GC\%}$) pairs to have the same sign (i.e., both positive or both negative) and the other half to have different signs. We observe six pairs to have the same sign and one pair to have different signs (Table 26.2). So we have

$$\chi^2 = \frac{(6-3.5)^2}{3.5} + \frac{(1-3.5)^2}{3.5} = 3.5714 \tag{26.6}$$

With one degree of freedom, the relationship is not significant ($p = 0.05878$).

Although the method of independent contrasts has been available for many years, many studies, even recent ones, still fall into the same trap, as illustrated in Fig. 26.1, of concluding a significant relationship between X and Y without taking the phylogeny into account. A recent claim of a strong relationship between intron conservation and intron number [32] represents one of such studies.

One shortcoming of the method of independent contrasts is that the value of the ancestral state is always somewhere between the two values of the descendents. This implies that it cannot detect directional changes over time. For example, if the ancestor is small in body size and all descendents have increased in body size over time, then the Brownian motion model assumed by the independent contrast method is no longer applicable. In such cases, one should use the generalized least square method [46, 56, 57].

When the method of independent contrasts was applied to the real data to assess the relationship between bacterial OGT and GC% of rRNA stem sequences and between OGT and rRNA stem lengths, the two relationships are both statistically significant [80]. Thus, the selectionist hypothesis is supported, but it accounts for only a very small fraction of variation in the genomic GC% among bacterial species, which calls for an alternative hypothesis for the variation in genomic GC%.

The mutation hypothesis of genomic GC% variation [48, 76, 94, 96] invokes biased mutation in different bacterial species to explain genomic variation in GC%, i.e., GC-rich genomes are the result of GC-biased mutation. One prediction from the mutation hypothesis is that the third codon position should increase more rapidly with the genomic GC% than the first codon position which in turn should have its GC% increase more rapidly with the genomic GC% than the second codon position. The reason for this prediction is that the third codon positions are little constrained functionally because most substitutions at the third codon positions are synonymous. Some nucleotide substitutions at the first codon positions are synonymous, but most are nonsynonymous. All nucleotide substitutions at the second

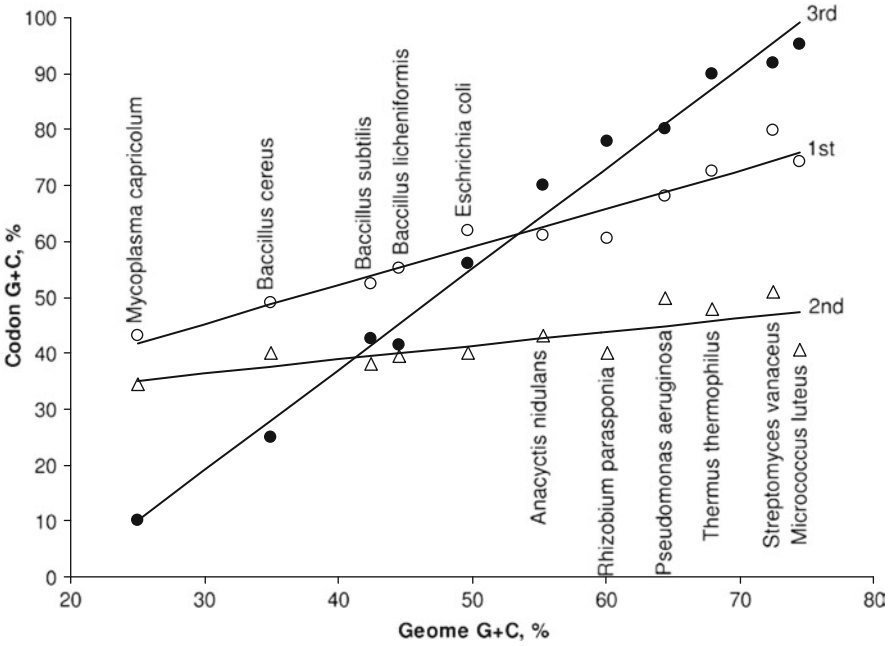

Fig. 26.3 Correlation of GC% between genomic DNA and first, second and third codon positions [48]. While the actual position of the points may be substantially revised with new genomic data (e.g., the GC% for the first, second and third codon positions for *Mycoplasma capricolum* is 35.8%, 27.4%, and 8.8% based on all annotated CDSs in the genomic sequence), the general trend remains the same

codon positions are nonsynonymous and typically involve rather different amino acids [83,91]. The empirical results [48] strongly support this prediction (Fig. 26.3).

The pattern in Fig. 26.3, while consistent with the mutation hypothesis, has resulted in two misconceptions. First, the pattern shown by the third codon position is often interpreted to reflect mutation bias. This interpretation is incorrect because the third codon position is subject to selection by differential availability of tRNA species [16, 82, 86, 88, 90]. We may contrast a GC-rich *Streptomyces coelicolor* and a GC-poor *Mycoplasma capricolum* as an illustrative example. *M. capricolum* has no tRNA with a C or G at the wobble site for four-fold codon families (Ala, Gly, Pro, Thr and Val), i.e., the translation machinery would be inefficient in translating C-ending or G-ending codons. This implies selection in favour of A-ending or U-ending codons and will consequently reduce GC% at the third codon position. This most likely has contributed to the low GC% at the third codon position in *M. capricolum*. In contrast, most of the tRNA genes translating the five four-fold codon families in the GC-rich *S. coelicolor* have G or C at the wobble site, and should favour the use of C-ending or G-ending codons. This most likely has contributed to the high GC% at the third codon position in *S. coelicolor*. The

same pattern is observed for two-fold codon families. The most conspicuous one is the Gln codon family (CAA and CAG). There is only one $tRNA^{Gln}$ gene in *M. capricolum* with a UUG anticodon favouring the CAA codon. In contrast, there are two $tRNA^{Gln}$ in *S. coelicolor*, both with a CUG anticodon favouring the CAG codon. Thus, the high slope for the third codon position in Fig. 26.3 is at least partially attributable to the tRNA-mediated selection. Relative contribution of mutation and tRNA-mediated selection to codon usage has been evaluated in several recent studies [16, 86, 88, 90].

Second, the observation that GC% of the third codon position increases with genomic GC% is sometimes taken to imply that the frequency of G-ending and C-ending codons will increase with genomic GC% or GC-biased mutation [40]. This is not generally true. Take the arginine codons for example. Given the transition probability matrix for the six synonymous codons shown in Table 26.3, the equilibrium frequencies (π) for the six codons are

$$\begin{aligned} \pi_{AGA} &= \frac{1}{2k^2+3k+1} \\ \pi_{AGG} &= \pi_{CGA} = \pi_{CGT} = \frac{k}{2k^2+3k+1} \\ \pi_{CGC} &= \pi_{CGG} = \frac{k^2}{2k^2+3k+1} \end{aligned} \tag{26.7}$$

The three solutions correspond to the number of GC in the codon, with AGA having one, AGG, CGA and CGT having two, and CGC and CGG having three G or C. One may note that the G-ending codon AGG has the same equilibrium frequency as that of the A-ending CGA and the T-ending CGT. Thus, we should not expect A-ending or T-ending codons to always decrease, or G-ending and C-ending codons always increase, with increasing genomic GC% or GC-biased mutation. In fact, according to the solutions in (26.7), AGG, CGA, and CGT will first increase with k until k reaches $\sqrt{2}/2$, and will then decrease with k when $k > \sqrt{2}/2$.

Table 26.3 Transition probability matrix for the six synonymous arginine codons, with α for transitions (C↔ T and A ↔ G), β for transversions, and k modeling AT-biased mutation ($0 \le k \le 1$) or GC-biased mutation ($k > 1$). We ignore nonsynonymous substitutions because nonsynonymous substitution rate is often negligibly low compared to synonymous rate. The diagonal is constrained by the row sum equal to 1

	CGT	CGC	CGA	CGG	AGA	AGG
CGT		$k\alpha$	β	$k\beta$	0	0
CGC	α		β	β	0	0
CGA	β	$k\beta$		$k\alpha$	β	0
CGG	β	β	α		0	β
AGA	0	0	$k\beta$	0		$k\alpha$
AGG	0	0	0	$k\beta$	α	

One may ask why the phylogeny-based comparison was not used for characterizing the relationship between codon GC% and genomic GC% in the 11 species in Fig. 26.3. The reason is that the two variables change very fast relative to the divergence time among the studied species, i.e., phylogenetic relatedness among the 11 species is a poor predictor of the codon GC% or genomic GC%. That genomic GC% has little phylogenetic inertia is generally true in prokaryotic species [93]. In such cases, one may assume approximate data independence and perform a phylogeny-free analysis. Another study that leads to insight into the relationship between UV exposure and GC% in bacterial genomes [73], which may be the first comparative genomic study, is also not phylogeny-based.

26.3 DNA Methylation, CpG Dinucleotide Frequencies and GC Content

CpG deficiency has been documented in a large number of genomes covering a wide taxonomic distribution [15, 35–37, 53]. DNA methylation is one of the many hypotheses proposed to explain differential CpG deficiency in different genomes [10, 62, 77]. It features a plausible mechanism as follows. Methyltransferases in many species, especially those in vertebrates, appear to methylate specifically the cytosine in CpG dinucleotides, and the methylated cytosine is prone to mutate to thymine by spontaneous deamination [23, 44]. This implies that CpG would gradually decay into TpG and CpA, leading to CpG deficiency and reduced genomic GC%. Different genomes may differ in CpG deficiency because they differ in methylation activities, with genomes having high methylation activities exhibiting stronger CpG deficiency than genomes with little or no methylation activity.

In spite of its plausibility, the methylation-deamination hypothesis has several major empirical difficulties (e.g., [15]), especially in recent years with genome-based analysis (e.g., Goto et al. 2000). For example, *Mycoplasma genitalium* does not seem to have any methyltransferase and exhibits no methylation activity, yet its genome shows a severe CpG deficiency. Therefore, the CpG deficiency in *M. genitalium*, according to the critics of the methylation-deamination hypothesis, must be due to factors other than DNA methylation.

A related species, *M. pneumoniae*, also devoid of any DNA methyltransferase, has a genome that is not deficient in CpG. Given the difference in CpG deficiency between the two Mycoplasma species, the methylation hypothesis would have predicted that the *M. genitalium* genome is more methylated than the *M. pneumoniae* genome, which is not true as neither has a methyltransferase. Thus, the methylation hypothesis does not seem to have any explanatory power to account for the variation in CpG deficiency, at least in the Mycoplasma species.

These criticisms are derived from phylogeny-free reasoning. When phylogeny-based comparisons are made, the Mycoplasma genomes become quite consistent with the methylation hypothesis [85]. First, several lines of evidence suggest that the common ancestor of *M. genitalium* and *M. pneumoniae* have methyltransferases

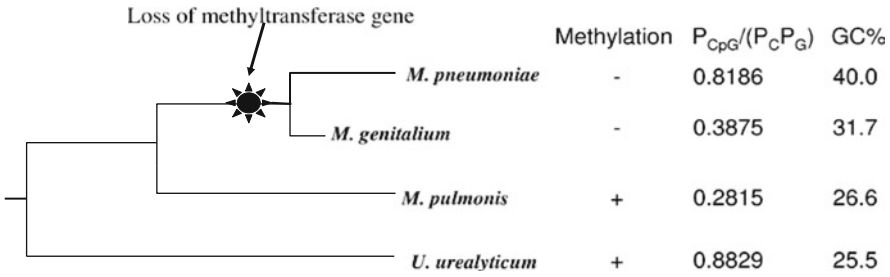

Fig. 26.4 Phylogenetic tree of Mycoplasma pneumoniae, *M. genitalium*s, and their relatives, together with the presence (+) or absence (−) of CpG-specific methylation, $P_{CpG}/(P_C P_G)$ as a measure of CpG deficiency, and genomic GC%. *M. pneumoniae* evolves faster and has a longer branch than *M. genitalium*

methylating C in CpG dinucleotides, and should have evolved strong CpG deficiency and low genomic GC% as a result of the specific DNA methylation. Methylated m^5C exists in the DNA of a close relative, *Mycoplasma hyorhinis* [61], suggesting the existence of methyltransferases in *M. hyorhinis*. Methyltransferases are present in *Mycoplasma pulmonis* which contains at least four CpG-specific methyltransferase genes [17]. Methylatransferases are also found in all surveyed species of a related genus, Spiroplasma [52]. These lines of evidence suggest that methyltransferases are present in the ancestors of *M. genitalium* and *M. pneumoniae*.

Second, the methyltransferase-encoding *M. pulmonis* genome is even more deficient in CpG and lower in genomic GC% than *M. genitalium* or *M. pneumoniae*, consistent with the methylation hypothesis (Fig. 26.4). It is now easy to understand that, after the loss of methyltransferase in the ancestor of *M. genitalium* and *M. pneumoniae* (Fig. 26.4), both genomes would begin to accumulate CpG dinucleotides and increase their genomic GC%. However, the evolutionary rate is much faster in *M. pneumoniae* than in *M. genitanlium* based on the comparison of a large number of protein-coding genes [85]. So *M. pneumoniae* regained CpG dinucleotide and genomic GC% much faster than *M. genitalium*. In short, the Mycoplasma data that originally seem to contradict the methylation hypothesis actually provide strong support for the methylation hypothesis when phylogeny-based genomic comparisons are made.

One might note that *Ureaplasma urealyticum* in Fig. 26.4 is not deficient in CpG because its $P_{CpG}/(P_C P_G)$ ratio is close to 1, yet its genomic GC% is the lowest. Has its low genomic GC% resulted from CpG-specific DNA methylation? If yes, then why doesn't the genome exhibit CpG deficiency? It turns out that *U. urealyticum* has C-specific, but not CpG-specific, methyltransferase, i.e., the genome of *U. urealyticum* is therefore expected to have low CG% (because of the methylation-mediated $C \rightarrow T$ mutation) but not a low $P_{CpG}/(P_C P_G)$ ratio. The methyltransferase gene from *U. urealyticum* is not homologous to that from *M. pulmonis*.

26.4 Comparative Genomics and Comparative Methods for Discrete Characters

A genome typically encodes many genes. The presence or absence of certain genes, certain phenotypic traits and environmental conditions jointly represent a major source of data for comparative genomic analysis. These binary data are best analyzed by comparative methods for discrete data.

A total of 896 bacterial genomes and 63 archaea genomes have been made available for research through Entrez as of May 21, 2009. In addition to genomic GC that can be computed as soon as the sequences are available, each sequencing project also delivers a list of genes in the sequenced genome, identified by one of two categories of methods, i.e., by checking against the 'gene dictionary' through homology search, e.g., BLAST [1, 2] or by computational gene prediction, e.g., GENSCAN [13, 14]. The availability of such annotated genoes facilitates the large-scale comparative genomics illustrated in Fig. 26.5.

The comparison in Fig. 26.5, albeit in a very small scale, can immediately lead to interesting biological questions. First, *Escherichia coli* and *Klebsiella pneumoniae* have genes coding proteins for lactose metabolism, but others do not. This leads to at least three possible evolutionary scenarios. First, lactose-metabolizing function may be absent in the ancestor A (Fig. 26.5), but (1) gained along lineage B and lost in lineage F and G or (2) gained independently along lineage E and lineage H (e.g., by lateral gene transfer or LGT). The third possible scenario is that the function is present in the ancestor A, but lost in all species except for lineages E and H.

If lactose-metabolizing genes are frequently involved in LGT, then we should expect the gene tree built from the lactose operon genes to be different from the

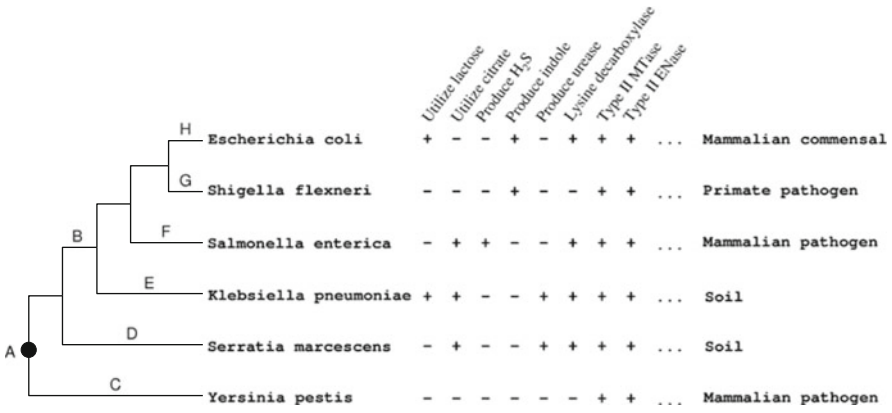

Fig. 26.5 Phylogeny-based comparative bacterial genomics, with +/− indicating the presence/absence of gene-mediated functions. Modern bacterial comparative genomics typically would have thousands of columns each representing the presence/absence of one gene function as well as many environmental variables of which only a habitat variable is shown here. Modified from Ochman et al. [54]

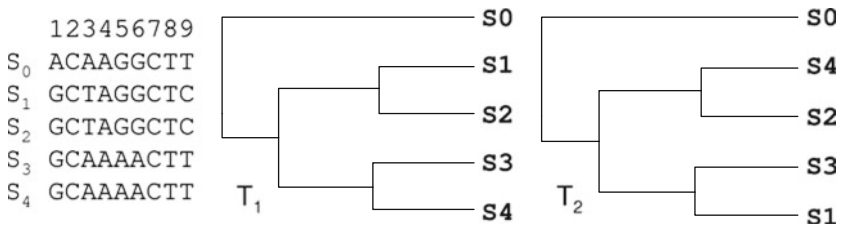

Fig. 26.6 DNA sequence data for significance tests of two alternative topologies

Table 26.4 Phylogenetic incongruence tests with maximum likelihood (ML) and maximum parsimony (MP) methods. lnL_1 and lnL_2 are site-specific log-likelihood values based on the F84 model and T_1 and T_2 (Fig. 26.6), respectively, and NC_1 and NC_2 are the minimum number of changes required for each site given T_1 and T_2, respectively

Site	ML		MP	
	lnL_1	lnL_2	NC_1	NC_2
1	−4.0975	−4.0990	1	1
2	−2.0634	−2.7767	0	0
3	−5.1147	−7.7335	1	2
4	−1.9481	−2.6238	0	0
5	−3.2142	−5.0875	1	2
6	−3.2142	−5.0875	1	2
7	−2.0634	−2.7767	0	0
8	−2.3938	−3.2626	0	0
9	−3.1090	−3.8572	1	2

species tree, which is typically approximated by a tree built from many housekeeping genes. Is the lactose operon gene tree significantly different from the species tree?

Suppose we have the sequence data (Fig. 26.6) from housekeeping genes, a species tree (T_1) and a lactose operon gene tree (T_2). We wish to test whether T_1 is significantly better than T_2 given the housekeeping gene sequences, with the null hypothesis being that T_2 is just as good as T_1. Both the maximum parsimony (MP) and the maximum likelihood (ML) methods have been used for such significance tests.

For the ML method, we compute the log-likelihood (lnL) for each of the nine sites (Fig. 26.6) given T_1 and T_2, respectively (lnL_1 and lnL_2 for T_1 and T_2, respectively, Table 26.4). A simple numerical illustration of computing site-specific lnL can be found in Xia [66, pp. 279–280]. A paired-sample t-test can then be applied to test whether mean lnL_1 is significantly different from mean lnL_2. For our data in Table 26.4, $t = 4.107$, $DF = 8$, $p = 0.0034$, two-tailed test). So we reject the null hypothesis and conclude that the lactose operon gene tree (T_2) is significantly worse than the species tree (T_1). A natural explanation for the phylogenetic incongruence is LGT.

For the MP method, we compute the minimum number of changes (NC) for each site given T_1 and T_2 (Fig. 26.6), respectively (NC_1 and NC_2 for T_1 and T_2,

respectively, Table 26.4). A simple numerical illustration of computing site-specific NC can be found in Xia [66, pp. 272–275]. We can then perform a paired-sample t-test as before to test whether mean NC_1 is significantly smaller than NC_2, in one of three ways. The first is to use the entire nine pairs of data, which yields t $= -2.5298$, DF $= 8$, p $= 0.0353$, and a decision to reject the null hypothesis that T_1 and T_2 are equally good at the 0.05 significance level, i.e., T_1 is significantly better than T_2. Second, we may use only the five polymorphic sites in the paired-sample t-test, which would yield t $= -4$, DF $= 4$, and p $= 0.0161$. This leads to the same conclusion. The third is to use only the four informative sites which is however inapplicable in our case because we would have four NC_1 values all equal to 1 and four NC_2 values all equal to 2, i.e., the variation in the difference is zero.

When the phylogenetic incongruence test is applied to real lactose operon data, it was found that the lactose operon gene tree is somewhat compatible to the species tree, and the case for LGT is therefore not strong [74]. This suggests the possibility that the lactose operon was present in the ancestor, but has been lost in a number of descendent lineages. In contrast, the urease gene cluster, which is important for long-term pH homeostasis in the bacterial gastric pathogen, *Helicobacter pylori* [63, 92], generate genes trees significantly different from the species tree (unpublished result). This suggests that the urease gene cluster is involved in LGT and has implications in emerging pathogens. For example, many bacterial species pass through our digestive system daily, and it is conceivable that some of them may gain the urease gene cluster and become acid-resistant, with the consequence of one additional pathogen for our stomach.

The second type of biological questions one can derive from Fig. 26.5 is functional association between genes. We note that Type II ENase (restriction endonuclease) is always accompanied by the same type of MTase (methyltransferase) recognizing the same site (Fig. 26.5). Patterns like this allow us to quickly identify enzymes that are partners working in concert. ENase cuts the DNA at specific sites and defends the bacterial host against invading DNA phages. MTase modifies (methylates) the same site in the bacterial genome to prevent ENase from cutting the bacterial genome. Obviously, ENase activity without MTase is suicidal, so the two must both be present. This also explains why the activity of many ENases depends on S-adenosylmethionine (AdoMet) availability. AdoMet always serves as the methyl donor for MTase. Without AdoMet, the restriction sites in the host genome will not be modified even in the presence of MTase because of the lack of the methyl donor, and ENase activity will then kill the host. So it is selectively advantageous for ENase activity to depend on the availability of AdoMet. Although rare, MTase can be present without the associated ENase. For example, *E. coli* possesses two unaccompanied MTases, Dam and Dcm. Some bacteriophages carry one or more MTases to modify their own genome so as to nullify the hostile action of the host ENases.

Sometimes one may find the presence of orthologous genes in different species but the function associated with the gene is missing in some species. Such is the case of ERG genes involved in sterol metabolism. Many species, including *Drosophila melanogaster* and *Caenorhabditis elegans*, share orthologous genes

involved in de novo sterol synthesis [78], but *D. melanogaster* and *C. elegans* have lost their ability to synthesize sterols de novo, although their ERG orthologs are still under strong purifying selection revealed by a much lower nonsynonymous substitution rate than the synonymous substitution rate. Further microarray studies demonstrated a strong association between the orthologs of ERG24 and ERG25 in *D. melanogaster* and genes involved in ecdysteroid synthesis and in intracellular protein trafficking and folding [78]. This suggests that the ERG genes in *D. melanogaster* have diverged and evolved new functions.

Another example in which a phylogenetic backdrop facilitates the study of evolutionary mechanisms involves the translation initiation. All molecular biology textbooks tell us that prokaryotes use the matching of the Shine-Dalgarno (SD) sequence in the mRNA and the anti-SD sequence in the small subunit rRNA to locate the translation initiation site, whereas eukaryotes use the Kozak initiation consensus to locate the translation initiation site. This would constitute a great piece of evidence for creationists to argue for independent creation. However, it is possible that the ancient organisms may have evolved these two translation initiation recognition mechanisms in parallel, and both might have contributed to the accurate localization of the translation initiation site. It is remarkable that some ancient lineages of prokaryotes living in deep sea hydrothermal vents still retain both mechanisms (unpublished results).

Mapping genes and gene functions to a phylogeny has revealed the loss of an essential single-copy *Maelstrom* gene in fish, and a plausible explanation is that the essential function has been fulfilled by a non-homologous gene [97]. Such findings that a specific molecular function can be performed by evolutionarily unrelated genes suggest a fundamental flaw in research effort to identify the minimal genome by identifying shared orthologous genes [47]. The rationale for such an approach is this. Suppose a minimal organism needs to perform three essential functions designated x, y, z, and three different genes, designated A, B, C, encode products that perform these three functions. If we have a genome (G1) with five genes A, B, C, D, E and another genome (G2) with four genes A, B, C, F, with genes of the same letter being orthologous, then shared orthologous genes between G1 and G2 are A, B, C which would be a good approximation of the minimal genome. In reality, it is possible that G1 = {A, D, E} for functions x, y, z and G2 = {A, C, F} for functions x, y, z. Both are already minimal genomes, but the intersection of G1 and G2 is only A which is a severe underestimation of a minimal genome. Creating a cell with such a 'minimal' genome is doomed to fail.

The third type of questions one can derive from Fig. 26.5 is the association between gene function and environmental variables. Note that *Klebsiella pneumoniae* and *Serratia marcescens* produce urease (Fig. 26.5). Both species can generate acids by fermentation leading to acidification of their environment. The presence of urease, which catalyzes urea to produce ammonia, can help maintain cytoplasmic pH homeostasis and allow them to tolerate environmental pH of 5 or even lower. Thus, comparative genomics can help us understand gene functions in particular environmental conditions.

Urease gene cluster serves as one of the two key acid-resistant mechanisms in the bacterial pathogen *Helicobacter pylori* in mammalian stomach, with the other mechanism being a positively charged cell membrane that alleviates the influx of protons into cytoplasm. The latter mechanism is established by comparative genomics between *H. pylori* and its close relatives as an adaptation to the acidic environment in the mammalian stomach [92].

The second and the third type of questions involve the same statistical problem, i.e., the identification of association either between two genes (e.g., between a type II ENase and a type II MTase) or between a gene and an environmental variable (e.g., between urease production and the habitat). A statistician without biological background might use a 2×2 contingency table (i.e., $N_{+/+}$, $N_{+/-}$, $N_{-/+}$, $N_{-/-}$) and Fisher's exact test to identify the association between two columns without taking the phylogeny into consideration. However, such an approach can lead to both false negatives and false positives. Fig. 26.7 illustrates the association study of two pairs of genes. Ignoring the phylogeny will lead to a significant association between genes *ORC3* and *CIN3*. However, the data points are not independent as the superficial association could be caused by only two consecutive gene–gain events (Fig. 26.7) and all the seven '11' could then the consequence of shared ancestral characters.

A phylogeny-based comparative analysis [7, 55] characterizes the state transition by a Markov chain, and uses a likelihood ratio test to detect the presence of

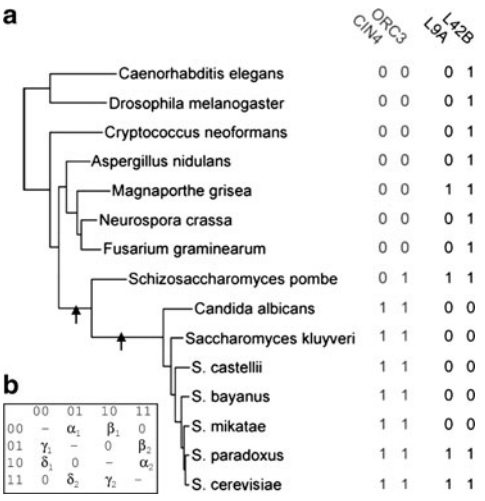

Fig. 26.7 Comparative methods for discrete binary characters. The presence and absence (designated by 1 and 0, respectively) of four genes are recorded for each species (**a**). The two black arrows indicate a gene–gain event. The instantaneous rate matrix (**b**), with notations following Felsenstein [22], shows the relationship among the four character designation, i.e., 00 for both genes absent, 01 for the absence of gene 1 but presence of gene 2, 10 for the presence of gene 1 but absence of gene 2, and 11 for both genes present. The diagonals are constrained by each row sum equal to 0. Modified from Barker and Pagel [7]

association between genes or between a gene function and an environmental condition. Two genes, each with two states (presence/absence), have four possible joint states and eight rate parameters ($\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2, \gamma_1$ and γ_2) to be estimated from the data (Fig. 26.7). When the gain or loss of one gene is independent of the other gene, then $\alpha_1 = \alpha_2, \beta_1 = \beta_2, \delta_1 = \delta_2, \; and \; \gamma_1 = \gamma_2$, with only four rate parameters to be estimated. Thus, we compute the log-likelihood for the eight-parameter and the four-parameter model given the tree and the data, designated lnL_8 and lnL_4, respectively, and perform a likelihood ratio test with test statistic being $2(lnL_8 - lnL_4)$ and four degrees of freedom.

I illustrate the computation of lnL_8 by using a simpler tree with only four operational taxonomic units or OTUs (Fig. 26.8). The joint states, represented by binary numbers 00, 01, 10 and 11, correspond to decimal numbers 0, 1, 2 and 3 which will be used to denote the four states in some equations below. The likelihood for the eight-parameter model is

$$L_8 = \sum_{z=0}^{3} \sum_{y=0}^{3} \sum_{x=0}^{3} \pi_z P_{zx}(b_6) P_{x0}(b_1) P_{x3}(b_2) P_{zy}(b_5) P_{y0}(b_3) P_{y3}(b_4) \qquad (26.8)$$

Equation 26.8 may seem to suggest that we need to sum 3^4 terms. However, the amount of computation involved is greatly reduced by the pruning algorithm [19]. To implement this algorithm, we define a vector L with elements L(0), L(1), L(2), and L(3) for every node including the leaves. L for leaf i is defined as

$$L_i(s) = \begin{cases} 1, & if \; s = S_i \\ 0, & otherwise \end{cases} \qquad (26.9)$$

L for an internal node with two offspring (o_1 and o_2) is recursively defined as

$$L_i(s) = \left[\sum_{k=0}^{3} P_{sk}(b_{i,o_1}) L_{o_1}(k) \right] \left[\sum_{k=0}^{3} P_{sk}(b_{i,o_2}) L_{o_2}(k) \right] \qquad (26.10)$$

where b_{i,o_1} means the branch length between internal node i and its offspring o_1, and P_{sk} is the transition probability from state s to state k computed from the rate matrix (Fig. 26.7b). For example, b_{x,S_1} (branch length between internal node x and its offspring S_1) is b_1 in Fig. 26.8. The computation involves finding the eight rate parameters that maximize L_8. As there is no analytical solution, the maximizing algorithm will simply try various rate parameter values and evaluate L_8 repeatedly until we converge on a set of parameter values that result in maximum L_8. Many such algorithms are well explained and readily available in source code [60].

While the equations might be confusing to some, the actual computation is quite simple. With only four OTUs, $S_1 = S_3 =$ '00' and $S_2 = S_4 =$ '11' (Fig. 26.8), the likelihood surface is quite flat and many different combination of the rate parameters can lead to the same maximum L_8. In fact, the only constraint on the rate parameters

is high rates from states 01 and 10 to states 00 and 11 (i.e., large $\delta_1 + \gamma_1 + \alpha_2 + \beta_2$) and low rates from states 00 and 11 to states 01 and 10 (i.e., small $\delta_2 + \gamma_2 + \alpha_1 + \beta_1$). This should be obvious when we look at the four OTUs in the tree (Fig. 26.8), with only 00 and 11 being observed at the leaves. This implies that 01 and 10 should be transient states, quickly changing to 00 or 11, whereas 00 and 11 are relatively conservative stable states. One of the rate matrices that approaches the maximum L_8 is

$$Q = \begin{bmatrix} & 00 & 01 & 10 & 11 \\ 00 & -16.47 & 13.15 & 3.32 & 0 \\ 01 & 1.10 & -135653.97 & 0 & 135652.87 \\ 10 & 1816.49 & 0 & -20308.04 & 18491.54 \\ 11 & 0 & 18.30 & 207.21 & -225.52 \end{bmatrix} \tag{26.11}$$

The rate of transition from states 01 and 10 to states 00 and 11 is 644.5 times greater (The true rate should be infinitely greater) than the other way round, which implies that we will almost never observe 01 and 10 states. The transition probability matrices with branch lengths of 0.1 and 0.3, which are computed as e^{Qt}, where t is

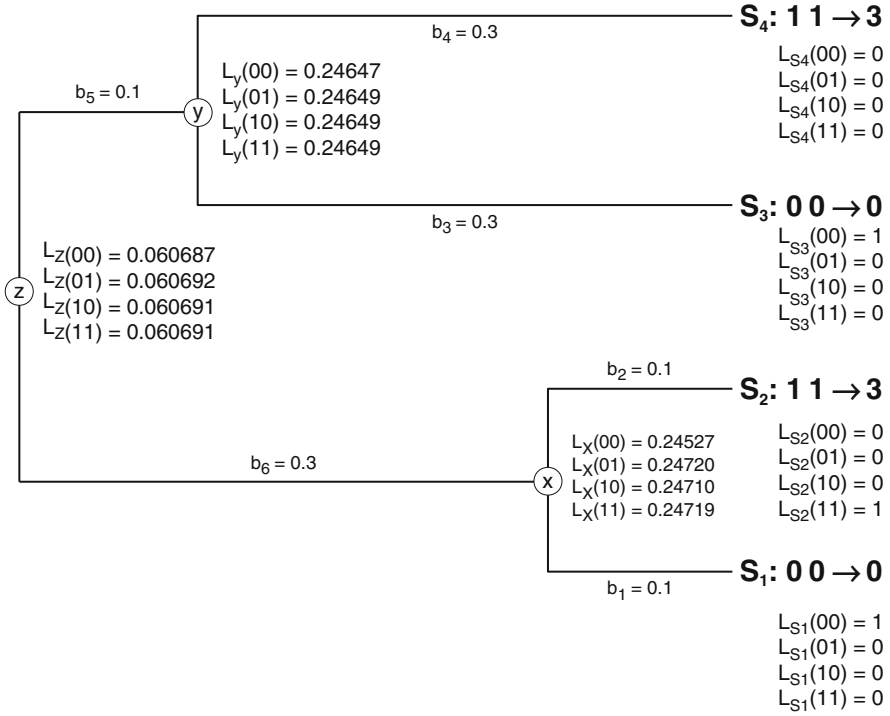

Fig. 26.8 Four-OTU tree with branch lengths (b_1–b_6) for illustrating likelihood computation. The L vectors are computed recursively according to (10)–(11)

the branch length, are, respectively,

$$P(0.1) = \begin{bmatrix} & 00 & 01 & 10 & 11 \\ 00 & 0.54616 & 0.00011 & 0.00467 & 0.44908 \\ 01 & 0.51459 & 0.00011 & 0.00499 & 0.48038 \\ 10 & 0.51738 & 0.00011 & 0.00496 & 0.47759 \\ 11 & 0.51458 & 0.00011 & 0.00499 & 0.48034 \end{bmatrix}$$
$$P(0.3) = \begin{bmatrix} & 00 & 01 & 10 & 11 \\ 00 & 0.53145 & 0.00011 & 0.00482 & 0.46377 \\ 01 & 0.53144 & 0.00011 & 0.00482 & 0.46382 \\ 10 & 0.53144 & 0.00011 & 0.00482 & 0.46382 \\ 11 & 0.53144 & 0.00011 & 0.00482 & 0.46382 \end{bmatrix} \quad (26.12)$$

We can now compute L_8 by using the pruning algorithm. First, L_{S1}–L_{S4} are straightforward from (26.9) and shown in Fig. 26.8. L_x and L_y are computed according to (26.10), e.g.,

$$\begin{aligned} L_x(00) &= P_{00,00}(0.1)P_{00,11}(0.1) = 0.54616 \times 0.44908 = 0.24527 \\ L_x(01) &= 0.51459 \times 0.48038 = 0.24720 \\ L_x(10) &= 0.51738 \times 0.47759 = 0.24710 \\ L_x(11) &= 0.51458 \times 0.48037 = 0.24719 \end{aligned} \quad (26.13)$$

Similarly, $L_y(00)$, $L_y(01)$, $L_y(10)$, and $L_y(11)$ are computed the same way and have values 0.24647, 0.24649, 0.24649, and 0.24649, respectively. Similarly, L_z is also computed by applying (26.9), e.g.,

$$\begin{aligned} L_z(00) &= AB = 0.246207 \times 0.246487 = 0.060687, \ \textit{where} \\ A &= [P_{00,00}(b_6)L_x(00) + P_{00,01}(b_6)L_x(01) + P_{00,10}(b_6)L_x(10) \\ &\quad + P_{00,11}(b_6)L_x(11)] = 0.246207 \\ B &= [P_{00,00}(b_5)L_y(00) + P_{00,01}(b_5)L_y(01) + P_{00,10}(b_5)L_y(10) \\ &\quad + P_{00,11}(b_5)L_y(11)] = 0.246487 \end{aligned} \quad (26.14)$$

$L_z(01)$, $L_z(10)$, and $L_z(11)$ are 0.060692, 0.060691, and 0.060691, respectively. The final L_8 is

$$\begin{aligned} L_8 &= \sum_{k=0}^{3} \pi_k L_z(k) = 0.060687 \times 0.5 + 0.060691 \times 0.5 = 0.060689 \\ \ln(L_8) &= -2.802 \end{aligned} \quad (26.15)$$

where we used the empirical frequencies for π_k, although π_k could also be estimated as a parameter of the model. Note that states 01 and 10 are not observed, and π_{01} and π_{10} are assumed to be 0 in (26.15).

The computation of $ln(L_4)$ is simpler because only four rate parameters need to be estimated, and is equal to -5.545. If quite a large number of OTUs are involved, then twice the difference between the two log-likelihood, designated $2\Delta lnL$, follows approximately the χ^2 distribution with 4 degrees of freedom. If we could assume large-sample approximation in our case, then $2\Delta lnL = 5.486$, which leads to $p = 0.241$, i.e., the eight-parameter model is not significantly better than the four-parameter model. Such a result is not surprising given the small number of OTUs.

With this phylogeny-based likelihood approach, Barker et al. [6] found that the superficial association between genes *CIN4* and *ORC3* is not significant, although Fisher's exact test ignoring the phylogeny would produce a significant association between the two genes. Similarly, genes *L9A* and *L42B* were found to be significantly associated based on the phylogeny-based likelihood approach, although Fisher's exact test ignoring the phylogeny would suggest a lack of the association. In this particular case, *L9A* and *L42B* are known to be functionally associated and *CIN4* and *ORC3* are known not be functionally associated. Ignoring the phylogeny would have produced both a false positive and a false negative. Phylogeny-based comparative methods for continuous and discrete methods have been implemented in the freely available software DAMBE [84, 95] at http://dambe.bio.uottawa.ca.

One difficulty with the comparative methods for the continuous and discrete characters is what branch lengths to use because different trees, or even the same topology with different branch lengths, can lead to different conclusions. One may need to explore all plausible trees to check the robustness of the conclusion.

Modern comparative genomic studies may often involve the functional association of thousands of genes or more. With N genes, there are $N(N-1)/2$ possible pairwise associations and $N(N-1)/2$ tests of associations. There are $N(N-1)(N-2)/6$ possible triplet associations. So it is necessary to consider the topic of how to control for error rates in multiple comparisons.

26.5 Controlling for Error Rate in Multiple Comparisons

There are two approaches for adjusting type I error rate involving multiple comparisons, one controlling for familywise error rate (FWER), and the other controlling for the false discovery rate (FDR) [51]. While FWER methods are available in many statistical packages and covered in many books, there are few computational tutorials for the FDR in comparative genomics, an imbalance which I will try to compensate below.

The difference between the FDR and FWER is illustrated in Table 26.5, where N_{12} denotes the number of null hypotheses that are true but rejected (false positives). FWER is the probability that N_{12} is greater or equal to 1, whereas FDR is the expected proportion of $N_{12}/N_{.2}$, and defined to be 0 when $N_{.2} = 0$. Thus, FDR is a less conservative protocol for comparison, with greater power than FWER, but at a cost of increasing the likelihood of obtaining type I errors.

Table 26.5
Cross-classification of N tests of hypothesis

H_0	Reject	
	No	Yes
TRUE	N_{11}	N_{12}
FALSE	N_{21}	N_{22}
Subtotal	$N_{.1}$	$N_{.2}$

Table 26.6 Illustration of the BH [8] and BY [9] procedures in controlling for FDR, with 15 sorted p values taken from Benjamini and Hochberg [8]

i	p	$p_{critical.BH.i}$	$p_{critical.BY.i}$
1	0.0001	0.00333	0.00100
2	0.0004	0.00667	0.00201
3	0.0019	0.01000	0.00301
4	0.0095	0.01333	0.00402
5	0.0201	0.01667	0.00502
6	0.0278	0.02000	0.00603
7	0.0298	0.02333	0.00703
8	0.0344	0.02667	0.00804
9	0.0459	0.03000	0.00904
10	0.324	0.03333	0.01005
11	0.4262	0.03667	0.01105
12	0.5719	0.04000	0.01205
13	0.6528	0.04333	0.01306
14	0.759	0.04667	0.01406
15	1	0.05000	0.01507

The FDR protocol works with a set of p values. For example, with 10 genes, there are 45 pairwise tests of gene associations, yielding 45 p values. The FDR protocol is to specify a reasonable FDR (typically designated by q) and find a critical p (designated $p_{critical}$) so that a p value that is smaller than $p_{critical}$ is considered as significant, otherwise it is not. The q value is typically 0.05 or 0.01. Two general FDR procedures, Benjamini-Hochberg (BH) and Benjamini-Yekutieli (BY), are illustrated below.

Suppose we have a set of 15 sorted p values from testing 15 different hypotheses (Table 26.6). The Bonferroni method uses α /m (where m is the number of p values) as a critical p value ($p_{critical.Benferroni}$) for controlling for FWER. We have m = 15. If we take $\alpha = 0.05$, then $p_{critical.Benferroni} = 0.05/15 = 0.00333$ which would reject the first three hypotheses with the three smallest p values.

The classical FDR approach [8], now commonly referred to as the BH procedure, computes $p_{critical.BH.i}$ for the ith p value (where the subscript BH stands for the BH procedure) as

$$p_{critical.BH.i} = \frac{q \cdot i}{m} \tag{26.16}$$

where q is FDR (e.g., 0.05), and i is the rank of the p value in the sorted array of p values (Table 26.6). If k is the largest i satisfying the condition of $p_i \leq p_{critical.BH.i}$, then we reject hypotheses from H_1 to H_k. In Table 26.6, k = 4 and we reject the first

four hypotheses. Note that the fourth hypothesis was not rejected by $p_{critical.Bonferroni}$ but rejected by $p_{critical.BH.4}$. Also note that $p_{critical.Bonferroni}$ is the same as $p_{critical.BH.1}$.

The FDR procedure above assumes that the test statistics are independent. A more conservative FDR procedure has been developed that relaxes the independence assumption [9]. This method, now commonly referred to as the BY procedure, computes $p_{critical.BY.i}$ for the i_{th} hypothesis as

$$p_{critical.BY.i} = \frac{q \cdot i}{m \sum_{i=1}^{m} \frac{1}{i}} = \frac{p_{critical.BH.i}}{\sum_{i=1}^{m} \frac{1}{i}} \tag{26.17}$$

With m = 15 in our case, $\sum 1/i = 3.318228993$. Now k (the largest i satisfying $p_i \leq p_{critical.BY.i}$) is 3 (Table 26.6). Thus, only the first three hypotheses are rejected. The BY procedure was found to be too conservative and several alternatives have been proposed [25]. For large m, $\sum 1/i$ converges to $ln(m) + \gamma$ (Euler's constant equal approximately to 0.57721566). Thus, for m = 10,000, $\sum 1/i$ is close to 10. So $p_{critical.BY}$ is nearly 10 times smaller than $p_{critical.BH}$.

One may also obtain empirical distribution of p values by resampling the data. For studying association between genes or between gene and environmental factors, one may compute the frequencies of states 0 (absence) and 1 (presence) for each gene (designated f_0 and f_1, respectively) and reconstitute each column by randomly sampling from the pool of states with f_0 and f_1. For each resampling, we may carry out the likelihood ratio test shown above to obtain p values. If we have generated 10,000 p values, then the 500th smallest p value may be taken as the critical p value. Note that all the null hypotheses from resampled data are true. So FDR and FWER are equivalent. This is easy to see given that FDR is defined as the expected proportion of $N_{12}/N_{.2}$ (Table 26.5) and FWER as the probability that N_{12} (Table 26.5) is greater or equal to 1. As we cannot observe N_{ij}, we use n_{ij} to indicate their realized values. When all null hypotheses are true, $n_{22} = 0$ and $n_{12} = n_{.2}$. Now if $n_{12} > 0$, then FDR = $E(n_{12}/n_{.2}) = 1$, and FWER = $P(n_{12} \geq 1)$ is naturally also 1. If $n_{12} = 0$, then FDR = 0 (Recall that FDR is defined to be 0 when $n_{.2} = 0$), and FWER = $P(n_{12} \geq 1)$ is also 0 [8])

26.6 Comparative Viral Genomics: Detecting Viral Recombination

There are two major reasons to study recombination. The first is that it is biologically interesting. For example, different strains of viruses often recombine to form new strains of recombinants leading to host-jumping or resistance to antiviral medicine, posing direct threat to our health. The second reason is that recombination is the source of many evils in comparative genomics and molecular evolution as it can generate rate variation among sites and among lineages and distort phylogenetic relationships [43]. We may be led astray without controlling for the effect of recombination in comparative genomic analysis.

Detecting viral recombination and mapping recombination points represent important research themes in viral comparative genomics [68]. This is often done in two different situations. The first is to address whether one particular genome (typically the one causing human health concerns, designated hereafter as R) is the result of viral recombination from a set of N potential parental strains (designated hereafter as P_i, where i $= 1, 2, \ldots,$ N). Graphic visualization methods such as Simplot [45] and Bootscan [67] , as well as the phylogenetic incongruence test, are often used in this first situation.

In the second situation, one does not know which one is R and which ones are P genomes. One simply has a set of genomic sequences and wishes to know whether some are recombinants of others. This is a more difficult problem. Many methods have been developed to solve the problem, and have been reviewed lucidly [31]. I will include here only what has been left out in the review, i.e., the graphic methods (Simplot and Bootscan) for the first situation and the compatibility matrix methods for the second. The compatibility matrix methods are among the most powerful methods for detecting recombination events.

26.6.1 Is a Particular Genome a Recombinant of N Other Genomes?

Given a sequence alignment, compute genetic distances $d_{R,Pi}$ (between R and P_i) along a sliding window of typically a few hundred bases. If we have a small d_{R,P_i} and a large d_{R,P_k} for one stretch of the genome, but a large d_{R,P_i} and a small d_{R,P_k} for another stretch of the genome, then a recombination likely occurred. This method, with visualization of the d values along the sliding windows, is known as Simplot [45]. Its disadvantage is that it does not generate any measure of statistical confidence.

I will illustrate the Simplot procedure by using HIV-1M genomes in an A-J-cons-kal153.fsa file [68]. HIV-1 has three groups designated M (main), O (outgroup) and N (non-M and non-O), with the M group further divided into A-D and F-K subtypes. The A-J-cons-kal153.fsa contains consensus genomic sequences for subtypes A, B, C, D, F, G, H, and J, as well as the KAL153 strain which may be a recombinant of two of the subtypes.

The result of applying the Simplot procedure is shown in Fig. 26.9. The genetic distance used is a simultaneously estimated (SE) distance based on the F84 model [89]. Note that $d_{KAL153,A}$ is relatively small and $d_{KAL153,B}$ relatively large up to site 2,601, after which $d_{KAL153,A}$ becomes large and $d_{KAL153,B}$ small until site 8,701. After site 8,701, $d_{KAL153,A}$ again becomes small and $d_{KAL153,B}$ large (Fig. 26.9). The simplest interpretation is that KAL153 is a recombinant between an A-like strain and a B-like strain. The two sites at which KAL153 changes its phylogenetic affinity (i.e., 2,601 and 8,701) may be taken as the recombination sites.

One may ask what the interpretation would be if B is missing from the data. The interpretation unavoidably would be that KAL153 is a recombinant between

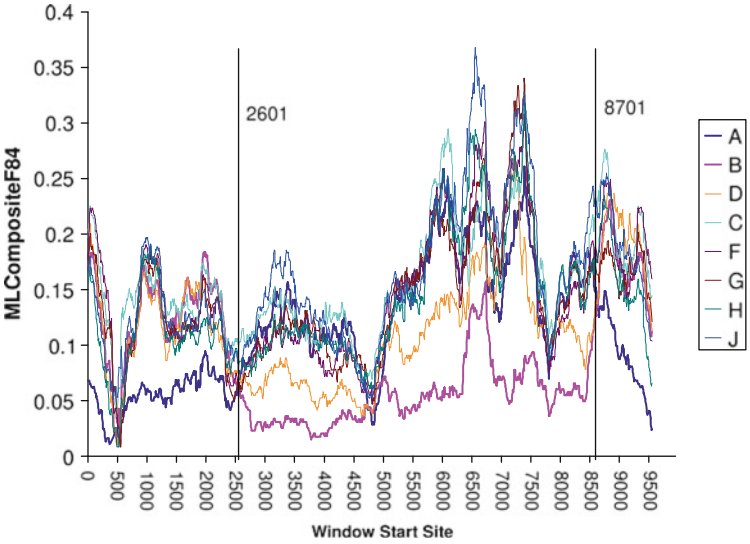

Fig. 26.9 Genetic distance between the query sequence (KAL153) and the consensus subtype sequences (A–J). MLCompositeF84 [89] is a simultaneously estimated distance based on the F84 model. KAL153 is genetically close to A before window start site at 2,601 and after window start site 8,701, but becomes close to B between window start sites 2,601 and 8,701. Output from DAMBE [84, 95]

an A-like strain and a D-like strain (Fig. 26.9). This interpretation is still reasonable because subtypes B and D are the most closely related phylogenetically. However, if A is missing from the data set, then the recombination event would become difficult to identify.

One might also note a few locations where the HIV-1 viral genomes are highly conserved across all included subtypes. Biopharmaceutical researchers typically would use such comparative genomic method to find conserved regions as drug targets or for developing vaccines against the virus.

One shortcoming of the Simplot method is that it does not produce any measure of statistical confidence. Given the stochastic nature of evolution, the distance of a sequence to other homologous sequence will often fluctuate. So the interpretation of patterns in Fig. 26.9 is associated with much uncertainty. Two approaches have been developed to overcome this shortcoming, one being the Bootscan method [67, 68], and the other is the phylogenetic incongruence test mentioned before.

The Bootscan method also takes a sliding window approach, but bootstraps the sequences to find the number of times each P_i has the smallest distance to R. The application of the bootscan method to the HIV-1M data (Fig. 26.10) shows that A is closest to KAL153 for almost all resampled data up to site 2,601, after which B becomes the closest to KAL153 until site 4,801. At this point A again becomes the closest to KAL153, albeit only briefly and with limited support. After site 5,051, B again becomes the closest to KAL153 until site 8,701 after which A again becomes

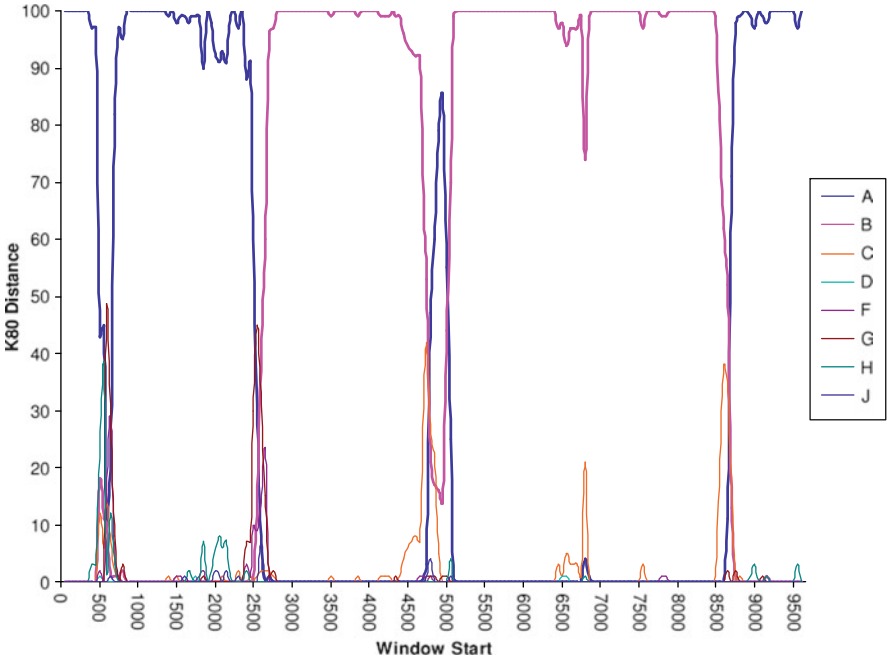

Fig. 26.10 BootScan output from scanning the HIV-1M sequences with KAL153 as the query. Output from DAMBE [84, 95], with window size being 400 nt and step size being 50 nt. DAMBE implements many other distances including the GTR distance and several simultaneously estimated distances suitable for highly diverged sequences

the closest to KAL153 (Fig. 26.10). The result suggests that there might be two recombination events.

The Simplot and the Bootscan procedures work well with highly diverged parental sequences, e.g., when the parental sequences belong to different subtypes as in our examples above. However, they are not sensitive when the parental sequences are closely related. This is true for most of the conventional methods for detecting recombination.

The second method for confirming KAL153's phylogenetic affinity reflected by changes in the genetic distance to other HIV-1M genomes (Fig. 26.9) is the phylogenetic incongruence test. The result in Fig. 26.9 allows us to partition the aligned genomic sequences into two sets, one consisting of the segment from 2,601 and 8,630 (hereafter referred to MIDDLE), and the other made of the rest of the sequences (hereafter referred to as TAILS). The phylogenetic tree for the eight subtypes of HIV-1M is shown in Fig. 26.11. A new HIV-1M genome suspected to be a recombinant, such as Kal153, may be phylogenetically grafted onto any one of the positions indicated by the numbered arrows (Fig. 26.11), creating 13 possible unrooted trees referred hereafter as $T_1, T_2, \ldots, T_13$, respectively, with the subscript number corresponding to the numbers in the arrow in Fig. 26.11). From results in

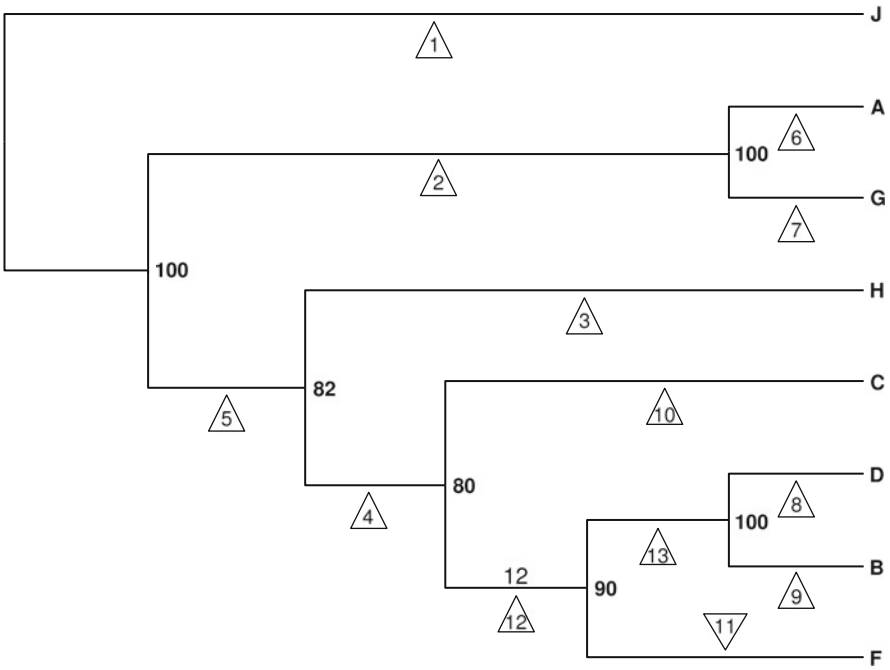

Fig. 26.11 Phylogenetic tree of the eight HIV-1M subtype genomes, with percentage bootstrap support indicated at each internal node. The numbered arrows indicate branches to which KAL153 can be granted to generate a new tree

Fig. 26.9, we can already infer that T_6 should be supported by the TAILS data set and T_9 should be supported the MIDDLE data set. However, will the support be significant against other alternative trees?

The result of phylogenetic tests (Table 26.7) shows that the TAILS data set strongly support T6 (grouping KAL153 with subtype A) but the MIDDLE data set strongly support T9 (grouping KAL153 with subtype B). This suggests that KAL153 is very highly likely to be a recombinant from subtypes A and B.

The use of the MIDDLE and TAILS for the phylogenetic incongruence test might be criticized for having fallen into a sequential testing trap [75]. A sliding-window approach together with the control for the false discover rate may be statistically more defendable.

26.6.2 General Methods Based on the Compatibility Matrix

In the set of four sequences in Fig. 26.12a, there are three possible unrooted trees labeled T_1, T_2 and T_3. Except for site 49, all sites are compatible with each other because they all support T_1. In contrast, site 49 supports T_3. In the classical population genetics with the infinite alleles model [38] where each mutation is unique and not reversible, site 49 would be considered as resulting from

Table 26.7 Statistical tests of 13 alternative trees, based on the TAILS and MIDDLE data sets

Data	Tree	lnL[a]	$\triangle lnL$[b]	$SE(\triangle)$[c]	T	pT[d]	pSH[e]	$pRELL$[f]
TAILS	6	−15046.0	0.000	0.000				1.000
	2	−15223.6	−177.587	28.579	6.214	0.000	0.000	0.000
	7	−15225.4	−179.382	28.092	6.385	0.000	0.000	0.000
	1	−15279.4	−233.325	34.684	6.727	0.000	0.000	0.000
	5	−15287.2	−241.162	34.013	7.090	0.000	0.000	0.000
	3	−15334.1	−288.028	38.281	7.524	0.000	0.000	0.000
	4	−15341.0	−294.930	38.052	7.751	0.000	0.000	0.000
	10	−15373.2	−327.121	40.059	8.166	0.000	0.000	0.000
	12	−15379.0	−332.934	39.987	8.326	0.000	0.000	0.000
	11	−15423.2	−377.209	42.205	8.938	0.000	0.000	0.000
	13	−15424.7	−378.629	41.968	9.022	0.000	0.000	0.000
	9	−15592.2	−546.125	48.274	11.313	0.000	0.000	0.000
	8	−15598.1	−552.052	47.741	11.563	0.000	0.000	0.000
MIDDLE	9	−23875.2	0.000	0.000				1.000
	13	−24086.1	−210.934	30.721	6.866	0.000	0.000	0.000
	8	−24091.5	−216.388	30.005	7.212	0.000	0.000	0.000
	12	−24398.1	−522.909	47.870	10.924	0.000	0.000	0.000
	10	−24535.3	−660.101	54.873	12.030	0.000	0.000	0.000
	4	−24553.5	−678.299	54.061	12.547	0.000	0.000	0.000
	3	−24623.9	−748.766	56.714	13.202	0.000	0.000	0.000
	5	−24627.3	−752.148	56.671	13.272	0.000	0.000	0.000
	1	−24652.2	−776.994	57.503	13.512	0.000	0.000	0.000
	2	−24653.3	−778.099	57.767	13.470	0.000	0.000	0.000
	7	−24749.9	−874.732	61.169	14.300	0.000	0.000	0.000
	6	−24753.4	−878.281	61.246	14.340	0.000	0.000	0.000

[a] log-likelihood of each tree.
[b] differences in log-likelihood between tree i and the best tree.
[c] standard error of $\triangle lnL$.
[d] P value for paired-sample t-test (two-tailed).
[e] P value with multiple-comparison correction [72].
[f] RELL bootstrap proportions [39].

recombination because mutations, being unique and not reversible by definition with the infinite alleles model, could not produce the pattern in site 49. In other words, parallel convergent mutations in different evolutionary lineages (homoplasies) are not allowed in the infinite allele model.

The infinite alleles model is not applicable to nucleotide sequences where each site has only four possible states that can all change into each other. So we need to decide whether site 49 in Fig. 26.12a can be generated by substitutions without involving recombination. In general, sequence-based statistical methods for detecting recombination share one fundamental assumption (or flaw) that we have only two alternatives, homoplasy or recombination, to explain polymorphic site patterns in a set of aligned sequences. If we reject the homoplasy explanation, then we arrive at the conclusion of recombination which is aptly named a backdoor conclusion [29]. Such a backdoor conclusion is ultimately not as satisfying as empirical demonstrations of recombination. For example, statistical detection of

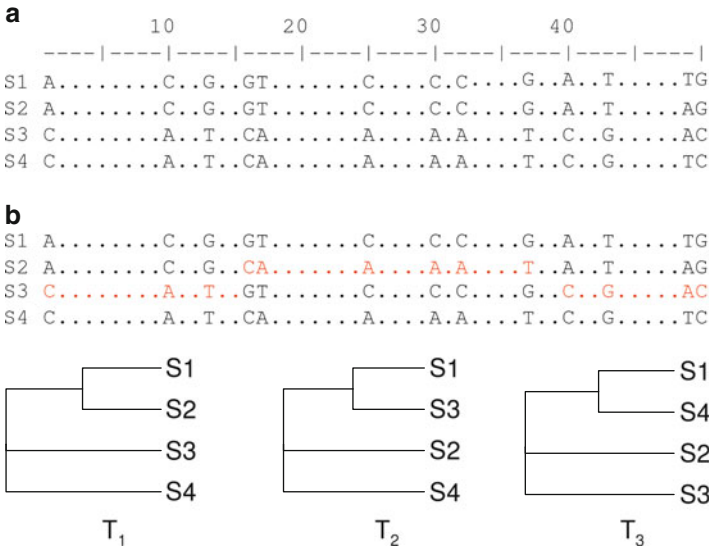

Fig. 26.12 Two sets of aligned nucleotide sequences for illustrating the compatibility-based method for detecting recombination events. (**a**) Four sequences without recombination. (**b**) Four sequences with recombination between S2 and S3, indicated by the switching of colored nucleotides. Dots indicate monomorphic sites

recombination involving mammalian mitochondrial genomes have been reported numerous times, but only an empirical demonstration [41] convinced the skeptical majority.

If we are happy with the fundamental assumption above that we have only two alternatives to discriminate between, then the method based on a compatibility matrix is both powerful and computationally fast. With a set of aligned sequences, two sites are compatible if and only if they both support the same tree topology. We only need to consider informative sites, i.e., sites featuring at least two states each of which is represented by at least two sequences. Non-informative sites are always compatible with other sites and need not be considered.

A pairwise compatibility matrix, or just compatibility matrix for short, lists whether sites i and j are compatible. The compatibility matrices for the two set of sequences in Fig. 26.12, one experiencing no recombination (Fig. 26.12a) and the other experiencing recombination involving the segment between informative sites 16–39 (Fig. 26.12b) are shown in Table 26.8. Two points are worth highlighting. First, sites that share the same evolutionary history are expected to be more compatible than those that do not (e.g., when the shared ancestry is disrupted by recombination). Note more 0's (compatible sites) in the upper triangle for sequences without recombination than in the lower triangle for sequences with recombination involving informative sites 16–39 (Table 26.8). Second, recombination tends to create similar neighbors in the compatibility matrix. Note the blocks of 1's and 0's in the lower triangle in Table 26.8. This similarity among neighbors has been

Table 26.8 Pairwise compatibility matrices, with 0 for compatible sites and 1 for incompatible sites, for aligned sequences in Fig. 26.12a (*upper triangle*) without recombination and those in Fig. 26.12b (*lower triangle*) with recombination between informative sites 16–39

Site	1	10	13	16	17	25	30	32	37	40	43	49	50
1		0	0	0	0	0	0	0	0	0	0	1	0
10	0		0	0	0	0	0	0	0	0	0	1	0
13	0	0		0	0	0	0	0	0	0	0	1	0
16	1	1	1		0	0	0	0	0	0	0	1	0
17	1	1	1	0		0	0	0	0	0	0	1	0
25	1	1	1	0	0		0	0	0	0	0	1	0
30	1	1	1	0	0	0		0	0	0	0	1	0
32	1	1	1	0	0	0	0		0	0	0	1	0
37	1	1	1	0	0	0	0	0		0	0	1	0
40	0	0	0	1	1	1	1	1	1		0	1	0
43	0	0	0	1	1	1	1	1	1	0		1	0
49	1	1	1	1	1	1	1	1	1	1	1		1
50	0	0	0	1	1	1	1	1	1	0	0	1	

characterized by the neighbor similarity score (NSS) which is the fraction of neighbors sharing either 0 (compatible) or 1 (incompatible). NSS is the basis of a number of methods for detecting recombination events [11, 34, 58, 59, 81] because its significance can be easily assessed by reshuffling the sites and recomputing NSS many times. The clumping of the compatible and incompatible sites in the compatibility matrix also suggests the possibility of mapping the recombination points. For example, one may infer from the compatibility matrix for the four sequences in Fig. 26.12b (lower triangle in Table 26.8) that the 5'-end recombination point is between informative sites 13 and 16, and that the 3'-end recombination point is between informative sites 37 and 40.

The compatibility matrix approach can be refined in two ways. First, when sequences are many, one will have some sites that are highly incompatible with each other as well as some sites that are only slightly incompatible with each other. The compatibility matrix approach lumps all these sites as incompatible sites, resulting in loss of information. Second, neighboring sites in a set of aligned sequences are expected to be more compatible with each other than with sites that are far apart. These two refinements were included in a recent study [12] that uses a refined incompatibility score (RIS) and the PHI statistic based on RIS. This new method appears much more sensitive than previous ones based on empirical applications [12, 65].

26.7 Summary

With the increasing availability of genomic sequences, comparative genomics has expanded rapidly and contributed significantly to our understanding of how mutation, recombination and natural selection have jointly governed the evolutionary

process. Comparative genomic analysis, aided by the phylogeny-based comparative methods, has resulted in improved detection of (1) functional association between genes and between genes and environment which is essential for understanding the origin and maintenance of the genetic components of biodiversity, (2) lateral gene transfer in prokaryotes and (3) recombination events and recombination sites. Development of comparative genomics has also motivated the research in statistics such as those controlling for the false discovery rates. Comparative genomics has dramatically changed the way of how regulatory sequence motifs are discovered, leading to the active development of phylogenetic footprinting which will be covered in the next chapter. What is particularly worth pointing out is that powerful and sophisticated software packages have been developed to facilitate research in comparative genomics.

Acknowledgements I thank J. Felsenstein and M. Pagel for identifying ambiguities and errors in the manuscript and for their many suggestions to improve the manuscript. S. Aris-Brosou, Y. B. Fu and G. Palidwor, as well as two anonymous reviewers, provided comments and references. I am supported by the Strategic Research, Discovery and Research Tools and Instrument Grants of Natural Science and Engineering Research Council of Canada.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*, 403–410.
2. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang Z., M., & Lipman, D.J. (1997). Gapped blast and psi-blast: A new generation of protein database search programs. *Nucleic Acids Research*, *25*, 3389–3402.
3. Argos, P., Rossmann, M.G., Grau, U.M., Zuber, A., Franck, G., & Tratschin, J.D. (1979). Thermal stability and protein structure. *Biochemistry (Moscow)*, *18*, 5698–5703.
4. Aris-Brosou, S., & Xia, X. (2008). Phylogenetic analyses: A toolbox expanding towards Bayesian methods. *International Journal of Plant Genomics*, *2008*, DOI 10.1155/2008/683509
5. Ballester, R., Marchuk, D., Boguski, M., Saulino, A., Letcher, R., & Wigler, M. (1990). The nf1 locus encodes a protein functionally related to mammalian gap and yeast ira proteins. *Cell*, *63*, 851–859.
6. Barker, D., Meade, A., & Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics*, *23*, 14–20.
7. Barker, D., & Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Computational Biology*, *1*, e3.
8. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*, 289–300.
9. Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple hypothesis testing under dependency. *The Annals of Statistics*, *29*, 1165–1188.
10. Bestor, T.H., & Coxon, A. (1993). The pros and cons of dna methylation. *Current Biology*, *6*, 384–386.

11. Brown C.J., Garner, E.C., Dunker, A.K, & Joyce, P. (2001). The power to detect recombination using the coalescent. *Molecular Biology and Evolution*, *18*, 1421–1424.
12. Bruen, T.C., Philippe, H., & Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics*, *172*, 2665–2681.
13. Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic dna. *Journal of Molecular Biology*, *268*, 78–94.
14. Burge, C.B., & Karlin, S. (1998). Finding the genes in genomic dna. *Current Opinion in Structural Biology*, *8*, 346–354.
15. Cardon, L.R., Burge, C., Clayton, D.A., Karlin, S. (1994). Pervasive CpG suppression in animal mitochondrial genomes. *Proceedings of the National Academy of Sciences*, *91*, 3799–3803.
16. Carullo, M., & Xia, X. (2008). An extensive study of mutation and selection on the wobble nucleotide in trna anticodons in fungal mitochondrial genomes. *Journal of Molecular Evolution*, *66*, 484–493.
17. Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., et al. (2001). The complete genome sequence of the murine respiratory pathogen mycoplasma pulmonis. *Nucleic Acids Research*, *29*, 2145–2153.
18. Dalgaard, J.Z., & Garrett, R.A., (1993). Archaeal hyperthermophile genes. In M. Kates, D. J. Kushner, & A. T. Matheson (Eds.), *The biochemistry of Archaea (Archaebacteria)*. Amsterdam: Elsevier.
19. Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*, 368–376.
20. Felsenstein, J. (1985). Phylogenies and the comparative method. *American Natural*, *125*, 1–15.
21. Felsenstein, J. (2002). *PHYLIP 3.6 (phylogeny inference package)*. Seattle: Department of Genetics, University of Washington.
22. Felsenstein, J. (2004). *Inferring phylogenies*. Sunderland, Massachusetts: Sinauer.
23. Frederico, L.A., Kunkel, T.A., & Shaw, B.R. (1990). A sensitive genetic assay for the detection of cytosine deamination determination of rate constants and the activation energy. *Biochemistry (Moscow)*, *29*, 2532–2537.
24. Galtier, N., & Lobry, J.R. (1997). Relationships between genomic g+c content, rna secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, *44*, 632–636.
25. Ge, Y., Sealfon, S.C., & Speed, T.P. (2008). Some step-down procedures controlling the false discovery rate under dependence. *Statistica Sinica*, *18*, 881–904.
26. Gordon, J.L., Byrne, K.P., & Wolfe, K.H. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genetics*, *5*(5), e1000,485. DOI 10.1371/journal.pgen.1000485
27. Goto M., Washio T., Tomita M. (2000). Causal analysis of CpG suppression in the Mycoplasma genome. Microbial and Comparative Genomics, 5, 51–58.
28. Harvey, P.H., & Pagel, M.D. (1991). *The comparative method in evolutionary biology*. Oxford: Oxford University Press.
29. Hey, J. (2000). Human mitochondrial dna recombination: can it be true? *Trends in Ecology and Evolution*, *15*, 181–182.
30. Hurst, L.D., & Merchant, A.R. (2001). High guanine-cytosine content is not an adaptation to high temperature: A comparative analysis amongst prokaryotes. *Proceedings of the Royal Society B*, *268*, 493–497.
31. Husmeier, D., & Wright, F. (2005). Detectign recombination in DNA sequence alignments. In D. Husmeier, R. Dybowski, & S. Roberts (Eds.), *Probabilistic modeling in bioinformatics and medical informatics* (p. 504). London: Springer.
32. Irimia, M., Penny, D., & Roy, S.W. (2007). Coevolution of genomic intron number and splice sites. *Trends Genetics*, *23*, 321.
33. Jacob, F. (1988). *The statue within: an autobiography*. New York: Basic Books, Inc.
34. Jakobsen, I.B., & Easteal, S. (1996). A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *Computer Applications in the Biosciences*, *12*, 291–295.

35. Josse, J., Kaiser, A.D., & Kornberg, A. (1961). Enzymatic synthesis of deoxyribonucleic acid vii. frequencies of nearest neighbor base-sequences in deoxyribonucleic acid. *The Journal of Biological Chemistry*, *236*, 864–875.
36. Karlin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: A genomic signature. *Trends in Genetics*, *11*, 283–290.
37. Karlin, S., & Mrazek, J. (1996). What drives codon choices in human genes. *The Journal of Biological Chemistry*, *262*, 459–472.
38. Kimura, M., & Crow, A.J.F (1964). The number of alleles that can be maintained in a finite population. *Genetics*, *49*, 725–738.
39. Kishino, H., & Hasegawa, M. (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, *29*, 170–179.
40. Kliman, R.M., & Bernal, C.A. (2005). Unusual usage of agg and ttg codons in humans and their viruses. *Gene*, *352*, 92.
41. Kraytsberg, Y., Schwartz, M., Brown, T.A., Ebralidse, K., Kunz, W.S., Clayton, D.A., et al. (2004). Recombination of human mitochondrial dna. *Science*, *304*, 981.
42. Kushiro, A., Shimizu, M., & Tomita, K. I. (1987). Molecular cloning and sequence determination of the tuf gene coding for the elongation factor tu of thermus thermophilus hb8. *European Journal of Biochemistry*, *170*, 93–98.
43. Lemey, P., & Posada, D. (2009). Introduction to recombination detection. In P. Lemey, M. Salemi, & A. M. Vandamme AM, *The phylogenetic handbook* (2nd ed.). Cambridge: Cambridge University Press.
44. Lindahl, T. (1993). Instability and decay of the primary structure of dna. *Nature*, *362*, 709–715.
45. Lole, K.S., Bollinger, R.C., Paranjape, R.S., Gadkari, D., Kulkarni, S.S., Novak, N.G., et al. (1999). Full-length human immunodeficiency virus type 1 genomes from subtype c-infected seroconverters in india, with evidence of intersubtype recombination. *The Journal of Virology*, *73*, 152–160.
46. Martins, E.P., & Hansen, T.F. (1997). Phylogenies and the comparative method: A general approach to incorporating phylogenetic information into the analysis of interspecific data. *The American Naturalist*, *149*(4), 646–667.
47. Mushegian, A.R., & Koonin, E.A. (1996). Minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *93*, 10268–10273.
48. Muto, A., & Osawa, S. (1987). The guanine and cytocine content of genomic dna and bacterial evolution. *Proceedings of the National Academy of Sciences*, *84*, 166–169.
49. Nakashima, H., Fukuchi, S., & Nishikawa, K. (2003). Compositional changes in rna, dna and proteins for bacterial adaptation to higher and lower temperatures. *The Journal of Biochemistry (Tokyo)*, *133*, 507–513.
50. Nei, M., & Kumar, S. (2000). *Molecular evolution and phylogenetics*. New York: Oxford University Press.
51. Nichols, T., & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: A comparative review. *Statistical Methods in Medical Research*, *12*, 419–446.
52. Nur, I., Szyf, M., Razin, A., Glaser, G., Rottem, S., & Razin, S. (1985). Procaryotic and eucaryotic traits of dna methylation in spiroplasmas (mycoplasmas). *The Journal of Bacteriology*, *164*, 19–24.
53. Nussinov, R. (1984). Doublet frequencies in evolutionary distinct groups. *Nucleic Acids Research*, *12*, 1749–1463.
54. Ochman, H., Lawrence, J.G., & Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, *405*, 299–304.
55. Pagel, M. (1994). Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society London B: Biological Sciences*, *255*, 37–45.
56. Pagel, M. (1997). Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, *26*, 331–348.

57. Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, *401*, 877–884.
58. Posada, D. (2002). Evaluation of methods for detecting recombination from dna sequences: Empirical data. *Molecular Biology and Evolution*, *19*, 708–717.
59. Posada, D., & Crandall, K.A. (2001). Evaluation of methods for detecting recombination from dna sequences: Computer simulations. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 13757–13762.
60. Press, W.H., Teukolsky, S.A., Tetterling, W.T., & Flannery, B.P. (1992). *Numerical recipes in C the art of scientifi computing* (2nd edn.). Cambridge: Cambridge University Press.
61. Razin, A., & Razin, S. (1980). Methylated bases in mycoplasmal dna. *Nucleic Acids Research*, *8*, 1383–1390.
62. Rideout, W.M.I., Coetzee, G.A., Olumi, A.F., & Jones, P.A. (1990). 5-methylcytosine as an endogenous mutagen in the human ldl receptor and p53 genes. *Science*, *249*, 1288–1290.
63. Sachs, G., Weeks, D.L., Melchers, K., & Scott, D.R. (2003). The gastric biology of helicobacter pylori. *Annual Review of Physiology*, *65*, 349–369.
64. Saenger, W. (1984). *Principles of nucleic acid structure*. New York: Springer.
65. Salemi, M., Gray, R.R., & Goodenow, M.M. (2008). An exploratory algorithm to identify intrahost recombinant viral sequences. *Molecular Phylogenetics and Evolution*, *49*, 618.
66. Salemi, M., & Vandamme, A.-M. (eds.) (2003). *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*. Cambridge University Press.
67. Salminen, M.O., Carr, J.K., Burke, D.S., & McCutchan, F.E. (1995). Identification of breakpoints in intergenotypic recombinants of hiv type 1 by bootscanning. *AIDS Research and Human Retroviruses*, *11*, 1423–1425.
68. Salminen, M., & Martin, D. (2009). Detecting and characterizing individual recombination events. In P. Lemey, M. Salemi, A. M. Vandamme (Eds.), *The phylogenetic handbook* (2nd ed.). Cambridge: Cambridge University Press.
69. Sankoff, D. (2009). Reconstructing the history of yeast genomes. *PLoS Genetics*, *5*, e1000,483.
70. Sankoff, D., & El-Mabrouk, N. (2002). Genome rearrangement. In T. Jiang, Y. Xu, & M. Q. Zhang (Eds.), *Current topics in computational molecular biology*. Cambridge: MIT.
71. Schluter, D., Price, T.D., Mooers, A.Ø., & Ludwig, D. (1997). Likelihood of ancestor states in adaptive radiation. *Evolution*, *51*, 1699–1711.
72. Shimodaira, H., & Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, *16*, 1114–1116.
73. Singer, C.E., & Ames, B.N. (1970). Sunlight ultraviolet and bacterial dna base ratios. Science, *170*, 822–826.
74. Stoebel, D.M. (2005). Lack of evidence for horizontal transfer of the lac operon into escherichia coli. *Molecular Biology and Evolution*, *22*, 683–690.
75. Suchard, M.A., Weiss, R.E., Dorman, K.S., & Sinsheimer, J.S. (2002). Oh brother, where art thou? a bayes factor test for recombination with uncertain heritage. *The Systems Biology*, *51*, 715–728.
76. Sueoka, N. (1964). *On the evolution of informational macromolecules*. New York: Academic.
77. Sved, J., & Bird, A. (1990). The expected equilibrium of the cpg dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the United States of America*, *87*, 4692–4696.
78. Vinci, G., Xia, X., & Veitia, R.A. (2008). Preservation of genes involved in sterol metabolism in cholesterol auxotrophs: Facts and hypotheses. *PLoS ONE*, *3*, e2883.
79. Wang, H.C., & Hickey, D.A. (2002). Evidence for strong selective constraint acting on the nucleotide composition of 16s ribosomal rna genes. *Nucleic Acids Research*, *30*, 2501–2507.
80. Wang, H.C., Xia, X., & Hickey, D.A. (2006). Thermal adaptation of ribosomal rna genes: A comparative study. *Journal of Molecular Evolution*, *63*, 120–126.
81. Wiuf, C., Christensen, T., & Hein, J. (2001). A simulation study of the reliability of recombination detection methods. *Journal of Molecular Evolution*, *18*, 1929–1939.
82. Xia, X. (1998). How optimized is the translational machinery in escherichia coli, salmonella typhimurium and saccharomyces cerevisiae? *Genetics*, *149*, 37–44.

83. Xia, X. (1998). The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes. *Journal of Molecular Evolution*, *15*, 336–344.
84. Xia, X. (2001). *Data analysis in molecular biology and evolution*. Boston: Kluwer Academic Publishers.
85. Xia, X. (2003). Dna methylation and mycoplasma genomes. *Journal of Molecular Evolution*, *57*, S21–S28.
86. Xia, X. (2005). Mutation and selection on the anticodon of trna genes in vertebrate mitochondrial genomes. *Gene*, *345*, 13–20.
87. Xia, X. (2007). Molecular phylogenetics: Mathematical framework and unsolved problems. In U. Bastolla, M. Porto, H. E. Roman, & M. Vendruscolo (Eds.), *Structural approaches to sequence evolution* (pp. 171–191).
88. Xia, X. (2008). The cost of wobble translation in fungal mitochondrial genomes: Integration of two traditional hypotheses. *BMC Evolutionary Biology*, *8*, 211.
89. Xia, X. (2009). Information-theoretic indices and an approximate significance test for testing the molecular clock hypothesis with genetic distances. *Molecular Phylogenetics and Evolution*, *52*, 665–676.
90. Xia, X., Huang, H., Carullo, M., Betran, E., & Moriyama, E.N. (2007). Conflict between translation initiation and elongation in vertebrate mitochondrial genomes. *PLoS ONE*, *2*, e227.
91. Xia, X., & Li, W.H. (1998). What amino acid properties affect protein evolution? *Journal of Molecular Evolution*, *47*, 557–564.
92. Xia, X., & Palidwor, G. (2005). Genomic adaptation to acidic environment: Evidence from helicobacter pylori. *The American Naturalist*, *166*, 776–784.
93. Xia, X., Wang, H.C., Xie, Z., Carullo, M., Huang, H., & Hickey, D.A. (2006). Cytosine usage modulates the correlation between cds length and cg content in prokaryotic genomes. *Molecular Biology and Evolution*, *23*, 1450–1454.
94. Xia, X.H, Wei, T., Xie, Z., & Antoine, D. (2002). Genomic changes in nucleotide and dinucleotide frequencies in pasteurella multocida cultured under high temperature. *Genetics*, *161*, 1385–1394.
95. Xia, X., & Xie, Z. (2001). Dambe: Software package for data analysis in molecular biology and evolution. *Journal of Heredity*, *92*, 371–373.
96. Xia, X., & Yuen, K.Y. (2005). Differential selection and mutation between dsdna and ssdna phages shape the evolution of their genomic at percentage. *BMC Genetics*, *6*, 20.
97. Zhang, D., Xiong, H., Shan, J., Xia, X., & Trudeau, V. (2008). Functional insight into maelstrom in the germline pirna pathway: A unique domain homologous to the dnaq-h 3-5 exonuclease, its lineage-specific expansion/loss and evolutionarily active site switch. *Biology Directorate*, *3*, 48.