9

Some Computational Problems Associated with Horizontal Gene Transfer

Michael Syvanen

It has been over 30 years since the suggestion that horizontal gene transfer (HGT) may have been a factor in the evolution of life entered the literature. Initially these speculations were based on discoveries made in medical microbiology, namely, that genes for resistance to antibiotics were found to move from one bacterial pathogen to another. This discovery was so unexpected and contrary to accepted genetic principles that though it was announced in Japan in 1959 [1,2], it was not generally recognized in the West for another decade. Speculations that HGT may have been a bigger factor in the evolution of life was inviting because it offered broad explanations for a variety of biological phenomena that have interested and puzzled biologists for over the last century and a half. These were problems that had been raised by botanists who have puzzled over the evolution of green plants [3] as well as by paleontologists who recorded macroevolutionary trends [4] in the fossil record that were often difficult to reconcile with the New Synthesis that merged Darwin's thinking with Mendelian genetics. However, outside of the field of bacteriology this exercise did not really attract that much attention until the late 1990s, at which time there was a major influx of data indicating that HGT had been very pervasive in early life. Namely, complete genome sequences began to appear. Simple examination of these sequences showed beyond any doubt that horizontal gene transfer was indeed a major factor in the evolution of modern bacterial, archaeal, and eukaryotic genomes.

Hence, in the past seven years or so, investigations into HGT have moved from the realm of the highly speculative and poorly documented to a robust area of investigation, especially for problems based on computationally intensive studies of genome sequences. A prerequisite to the ability to explore HGT is the ability to distinguish between genomic regions that may have originated from a foreign source (i.e., from a parallel lineage) and genomic regions whose evolutionary history is the result of vertical evolution of that given lineage. In the current review I will go over four areas that pose nontrivial computational problems.

These are: (1) the phylogenetic congruency test, (2) mosaics, (3) distance discrepancy, and (4) nucleotide composition analysis. Even though there is a rich literature concerning phylogenetic incongruities and mosaics, there has been little recent progress on new computational approaches. Therefore this review will focus mainly on the distance discrepancy approach and on atypical nucleotide composition analysis. These latter two approaches have the potential to shed light on some outstanding biological questions. Before going into these problems I will review the concept of common ancestry as a means of introducing the general topic of HGT and to help explain why it is having such profound influence on how we think about biology in general.

THE LAST UNIVERSAL COMMON ANCESTOR

An example of how profoundly the notion of HGT has changed our thinking concerns the concept of the last universal common ancestor (LUCA). This is an idea that was central to the hypothesis that life shared common ancestors. Though the idea of common ancestry remains valid (indeed evidence for common ancestry is everywhere in the sequence of our genes), there is no longer a need to postulate that all life evolved from a single last universal common ancestor. Rather, we can entertain common descent from multiple ancestors.

The notion that all life passed through a single interbreeding bottleneck is still probably believed to be true by most people who think about this problem. The reason is simple. There are many genes involved in information processing (i.e., DNA replication, RNA transcription, and protein synthesis) whose orthologs are found in all three major domains of life. Furthermore, when the sequences of these genes are submitted to phylogenetic analysis they more or less support the following relationship: the Archaea and Eukaryotes define a clade to the exclusion of a bacterial clade and a single line links both of these clades. Figure 9.1A shows this relationship. The figure shows an unrooted tree with four taxa; this happens to be a topology that is susceptible to semi-rigorous statistical analysis (see below). The Archaea/Eukaryote clade, by definition, implies the existence of a common ancestor for these two groups and further we can infer that a point on the line leading to the bacterial clade represents the last common ancestor of all life. Thus we can say that there is empirical support for the existence of the last common ancestor.

I mentioned above that this scenario is more or less supported by the informational genes. The striking finding is that other genes common to the three major kingdoms frequently show exceptions to these relationships. When it comes to the genes for energy metabolism, Eukaryotes and gram-negative bacteria are usually more closely related to one another than they are to the Archaea and other bacteria

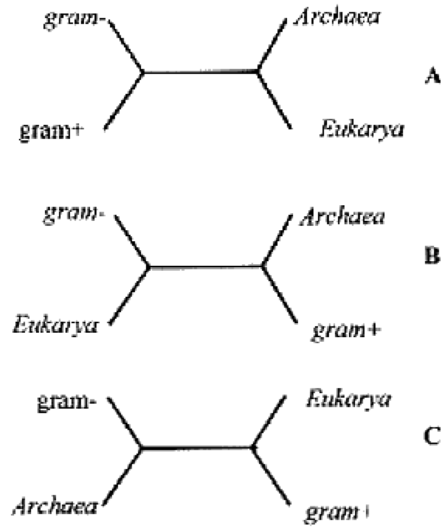

Figure 9.1 Universal tree of life and two alternatives. Bacteria contain many deeply rooted clades; here we include two groups which are shown as the gram (–) or more accurately known as proteobacteria and gram (+) or the low GC gram (+) bacteria. A shows the so-called universal tree that is supported by the rRNA sequences. B shows the relationships found between a very large number of genes involved in metabolism and biosynthesis. C simply shows the remaining 4-taxa relationship which very few genes seem to follow.

(as in figure 9. 1B). These genes are thought to have become associated with the eukaryotic cell through the endosymbiote that eventually gave rise to the mitochondrion [5–7]. In green plants we can also trace the ancestry of many genes involved in carbon fixation, photosynthesis, as well as other metabolic processes to cyanobacteria, the endosymbiote host that gave rise to the chloroplast. For many of the biosynthetic pathways the relevant genes yield even more complex relationships. Thus we have arrived at the current situation that is accepted by most—there remain a few genes (almost all associated with basic genetic informational processing) that reflect an evolutionary history that goes back to some very primitive LUCA, but that superimposed over the remnants of that primitive ancestor in modern genomes are numerous examples of subsequent horizontal gene transfer events.

The above is a good model and it requires good reasons to reject it. To begin, not all of the informational orthologs support the simple phylogenetic pattern outlined above. Even here there are some exceptions. These exceptions have been dealt with in one of two ways. First, in some cases it can be argued that there is insufficient amount of sequence to rigorously support the true clade relationships (i.e., sequence noise or homoplasy is hiding the true pattern), or alternatively, these are informational genes that also have been involved in HGT events. Though some of the cases are still open to debate, there are a number of cases where it is simplest to conclude that some of the informational genes have been involved in HGT events; this is especially true for some of the amino acid-tRNA ligases [8]. Once we reach this point then it is no longer possible to argue that biochemically complex processes such as protein synthesis are too complicated to have their genes being involved in HGT events, a position that was held at least up until 1998. In fact, Woese [9] suggested that there existed in the very primitive cells a less functionally constrained protein synthesis machinery that permitted some HGT events of these components, thereby accounting for the few exceptions. In this formulation a LUCA at least implicitly remains in the model. But evidence for the LUCA is greatly reduced, at least with respect to the number of genes found in modern genomes that can be directly traced back to the LUCA via exclusive vertical evolution. In 1982 it was automatic to assume that because a biochemical process was found in all of modern life, than that process must represent evidence for the one interbreeding population of the LUCA. Now we know that many of the universal biochemical processes have moved horizontally multiple times. Thus today we have a greatly truncated LUCA from what we believed just a decade ago.

When speculating on the nature of the LUCA it is generally accepted that it must have contained the modern universal genetic code since that is a feature shared by all life. However, even if we accept the existence of this LUCA, there are a variety of reasons to believe that the LUCA itself was the product of an evolutionary process that employed horizontal transfer events; this is so especially with respect to the evolution of the genetic code. It is very difficult to see how the modern genetic code could have evolved in a sequential fashion; rather the code must have evolved on separate occasions and become fused into single lineages. This problem is illustrated by considering the case of lysine-tRNA ligase genes found in modern life. All life has two different completely nonhomologous enzymes. If the modern genetic code evolved in a sequential fashion, then we would have to imagine a situation where a lineage that carried one of the two enzymes evolved the second. The raises the question: what selective pressure could possibly account for the emergence of this second enzyme when it already has one? It is much simpler to believe that the lysine enzyme evolved independently in two different lineages, which then fused to give rise to the ancestor of modern life. This is not a radical idea. Of course, if HGT is common to life after the time of LUCA, then it seems not unreasonable to assume that it was common to life before the LUCA. At this point we come to the following model for evolution of life if we try to preserve the LUCA. We have multiple lineages of pre-LUCA life that are linked together by HGT events into a netted or reticulate evolutionary pattern. This leads to the LUCA. The LUCA diversifies into its many modern lineages and then these lineages are again reticulated. We then end up with a topological model that looks like an hourglass, namely, a net above that bottlenecks to the LUCA which then diversifies and yields a net below. At this point the principle of parsimony should kick in. Why encumber our model with this bottleneck? It is not only no longer necessary but is now an exceptional assumption.

There is another reason that we should jettison the LUCA. This has to do with the finding that many of the universal genes, including a number that make up the genetic code, appear to be younger than are the major clades of life. That is, we can be reasonably sure that life forms resembling Archaea, bacteria, and some kind of primitive Eukaryote existed before 1.5 and likely before 2 billion years ago. However, parts of the genetic code are younger than that. The simplest explanation is that the genetic code continued to evolve after modern life diversified. If so, then the only reasonable explanation for this is that these younger members of the genetic code must have achieved their current modern and universal distribution via HGT events. These unexpectedly young genes are young by virtue of their having experienced less divergence than would be expected from certain assumptions of the molecular clock (i.e., a computational problem, see Distance discrepancy section below). In addition, these young genes often seem to display unusual phylogenetic topologies that are observed as the star phylogenies (another computational problem encountered in phylogenetic analysis). Once we accept that something as complex as the genetic code can evolve and spread by HGT events, it strongly suggests that a gene encoding any function could also.

There are deep ideological reasons for believing in a LUCA that explain the reluctance of many to abandon it. In fact this reason is built directly into the most basic model of modern biology, that is, the tree of life. The only figure in Darwin's *Origin of Species* happens to be a tree that inevitably maps back to a single trunk. Indeed the algorithms used in phylogenetic analysis can only find a single trunk, which, of course, is how they are designed. All practicing biologists are aware of the limitations of phylogenetic modeling with its built-in assumptions, but nevertheless these assumptions do cause confusion. For example, let me pose a question and ask how often there was confusion when thinking about mitochondrial Eve? Isn't it a common misperception to think at some point that all of human life could be mapped back to a single woman? When in fact all we can say is that the only surviving remnant of that distant ancestor is her mitochondrial genome, and it is extremely unlikely that any of her other genes survive in any human populations. Because of the phenomena of sexual reproduction and recombination we share genes with multiple ancestors with no need to hypothesize any individual ancestor from whom we have descended. The same reasoning should apply to the evolution of all life; because of the phenomena of horizontal gene transfer we share genes with multiple ancestors with no need to hypothesize individual species from whom we have descended [10].

PHYLOGENETIC CONGRUENCY TEST

Though this is considered the most rigorous method, and has been the most frequently employed, to establish the occurrence of HGT events, it remains very difficult to estimate a level of confidence in resultant findings. This situation has not improved significantly since I last reviewed this topic [11]. The problem lies in the fact that phylogenetic trees are Steiner trees and hence the solution to finding the minimal length tree is np-complete. This means that for large numbers of taxa it is impossible to compare two different topologies and to provide a judgment as to the significance of any differences. This is not to say that there has been little progress on developing new algorithms for searching for phylogenetic trees, just that it remains highly problematical to decide upon competing topologies.

Exact solutions to 4- and 5-taxa trees are possible and there has been some use of 4-taxa trees to determine if two different gene trees, from the same set of taxa, are significantly different. This problem is not too difficult if we simply select an ortholog from four different species and ask if the resulting gene trees are consistent with our expectation based on underlying species phylogeny. I have performed quartet analysis where a simple *t*-test was used to assess the significance of the two trees. In this approach, the number of uniquely shared characters was computed for each of the three unique 4-taxa trees [12]. Zhaxybayeva and Gogarten [13] have applied maximum likelihood and Bayesian probabilities to the 4-taxa problem that give a more rigorous solution to this problem than that provided by the simple *t*-test.

However, the 4-taxa comparison is susceptible to a major artifact, as is any test based upon phylogenetic congruency. One must be wary of the long-branch attraction problem, which arises when the evolutionary rates among the different lineages are highly variable [14]. Long branches attract because of a higher chance that they share unique characters from homoplasy and not homology. In small data sets this can very difficult to assess. Highly unequal rates can be identified provided we have good outgroup taxa against which we can perform a relative rate test [15] and thereby directly assess whether or not the rates of evolution in the various lineages are comparable. If this is known, then we can proceed with the 4-taxa test.

There is a potential statistical bias in the 4-taxa test, namely, we may identify potential incongruities after examining a gene tree containing a very large number of taxa. For example, let us say that we have a gene tree consisting of 50 taxa and one of the taxa seems significantly displaced. We can pick out the aberrant taxa and compare it to three other selected taxa and perform a test on these four taxa. Let us say that the resulting 4-taxa gene tree is shorter than the one expected from known species relationships and that the difference has a significance of P (either based on the t-test or on maximum likelihood). The problem we now have is that our quartet was selected from a much larger data set. What is the correct value of P? Would it be P times 50? Or P times 5,527,200? (the total number of quartets in the sample) or a value somewhere in between? This is not a simple problem.

Gene and Genome Mosaics

Aside from the phylogenetic congruency test, the finding of mosaics has provided the greatest impetus in the acceptance of HGT, especially in bacterial evolution. Indeed, the finding that different strains of *E. coli* are mosaics of each other led directly to the rejection of the clonal model of *E. coli* populations [16]. Mosaic is sometimes used as a synonym for horizontal gene transfer. But there is a specific analytical process implied by this term as well. Figure 9.2 can illustrate this. Let us consider homologous genomic regions from two different species designated D for the donor and R for the recipient. These regions were derived from a common ancestor but have diverged in their primary sequence. One common horizontal gene transfer event can be the movement of a DNA segment from D into the recipient R followed by a double recombination event (or possibly a gene conversion) to produce a strain that is

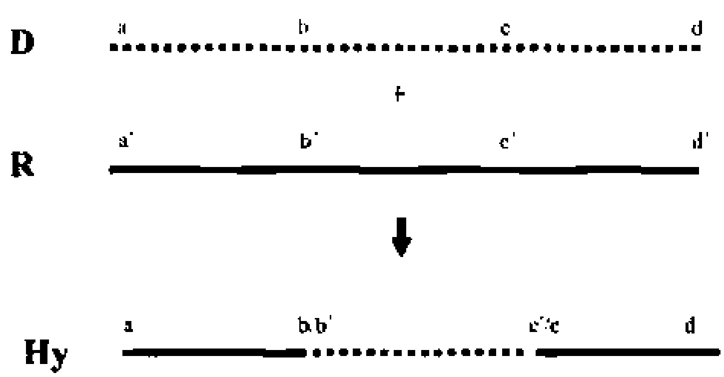

Figure 9.2 Scheme for the formation of a mosaic. Mosaics are created by a simple double recombination event between two homologous, but diverged, DNA sequences designated here as the donor (D sequences a–d) and the recipient (R sequences a'-d') to give rise to the mosaic hybrid (Hy). The two crossover points are at b/b' and c'/c; these points are also referred to as the novel junctions.

a hybrid from D and R. Recombination between two homologous but diverged sequences has been termed "homeologous" recombination [16] while "homologous" recombination involves two identical DNA sequences.

There is no question that homeologous recombination occurs. There are well-documented examples from laboratory studies [17]. It also likely occurs naturally. Some of the more striking examples involve important pathogenicity genes found in bacterial and viral human pathogens. An early example included a penicillin resistance gene found in *Streptococcus viridians* [18]. In this example, sequence R (the sensitive *S. viridians*), sequence D (a resistant *S. pneumoniae*), and the hybrid sequence (the resistant *S. viridians*) were available. Thus a straightforward parsimony argument would suffice to reconstruct a pathway analogous to the one in figure 9.2. In addition, mosaic patterns appear to be common among viruses; it is precisely this type of recombination event that has contributed to much of the variation seen with HIV [19]. A different kind of recombination (called reassortment) has led to the creation of novel human influenza virus strains. Indeed much of the world is now waiting to see if such an event may occur with the agent responsible for the current Southeast Asian "bird flu" that could set off a pandemic among humans.

It is of interest to map the exact recombination crossover points (i.e., b/b′ and c′/c) in figure 9.2. The mosaic problem can be solved using the phylogenetic congruency test where different regions of the mosaic are compared to one another. But if this is done, some sense of where the crossover points are located is needed. Rarely can the exact "novel junctions" be identified, but a target range can be found. A few authors [20,21] have derived statistical tests to help judge the significance of a presumed mosaic, especially with respect to locating inferred crossover points. Conceptually there is an overlap between this problem and the haplotype-mapping problem [22–24] for the simply reason that haplotypes are linkage groups that are defined by recombination units.

Because the recombination events that produced the pathogenic hybrids described above occurred within the past few decades, there were no or few subsequent point mutations that could erase the pattern shown in figure 9.2. These examples are so clear that there is not that much difficulty in identifying the mosaic pattern in the hybrids. However, there are also situations where mosaic patterns of evolution have left an imprint, but where we have only partial data sets and/or degraded data sets. As one possibility, consider a suspected hybrid mosaic in a larger set of homologs where we are missing the putative donor and recipient. Can we identify the mosaic? In principle, this can be accomplished by using the larger data set to establish the expected amount of divergence for each region. Based on my own experience from examining large numbers of aligned sequences, the possible occurrence of mosaics is not rare. The problem is that lacking a possible donor sequence (D), we cannot conclude that the aberrant sequence is due to horizontal gene transfer since there are other mutational mechanisms that can produce an apparent increase in sequence divergence. Nevertheless, it is interesting to identify such regions and there are no currently automated procedures for doing so.

We can also imagine a situation where an ancient homeologous recombination event had occurred, and we encounter highly diverged descendents of the hybrid and donor and/or recipient. At what point can we still identify the mosaic before the mosaic pattern is lost in the evolutionary noise? This is a problem that goes beyond the single issue of horizontal gene transfer. We can imagine homeologous recombination events occurring between paralogs within a genome that result in a protein with a novel function. Modern proteins are certainly the result of fusions and rearrangements of more primitive proteins, and there is considerable interest in reconstructing the pathways by which proteins evolved [25]. Clearly, it seems likely that events similar to those seen in figure 9.2 would have contributed to the evolution of modern proteins. This is a computational problem worthy of continued investigation.

Distance Discrepancy

Distance discrepancy as a means of detecting HGT events remains a potentially powerful but as yet underutilized tool. Accurate determinations of molecular distance have the potential to answer questions about the importance of HGT in evolutionary history in situations where the phylogenetic congruency test is too insensitive. The potential advantage of this tool can be most clearly seen in regard to hypotheses on the importance of HGT in the evolutionary history of higher Eukaryotes—especially multicellular plants and metazoans.

The earliest speculations concerning HGT repeatedly mentioned the explanatory power of HGT for a number of phenomena that had puzzled biologists for over a century [1,2,26]. These include those major episodes of evolution (especially during the emergence of novel structures) that occurred simultaneously and over short periods of time. Such events as the Precambrian radiation or the eutherean radiation have been cited. In addition, the widespread occurrence of parallelism among closely related lineages has recurred throughout the fossil record. If these speculations are correct, it means that HGT events among, for example, the metazoans occur more frequently between closely related lineages than among more distantly related lineages. The phylogenetic congruency test detects movement of genes between highly unrelated lineages, but since it compares topology, it is relatively insensitive to movement between close relatives. In principle, such events will be seen through temporal discrepancies without necessarily giving rise to incongruent phylogenetic topologies (reviewed in [27]).

For example, let us consider a very real possibility concerning metazoan evolution. If the radiation occurred 540 MYA (million years ago), could we detect movement of a gene leading to two modern phyla that occurred 400 MYA? To develop tools that can attack this problem requires good molecular clocks, good calibration points, and accurate species diversification times. At the present time this remains outside of our ability to resolve, but there is no reason in principle that good statistical tools cannot be developed to solve this problem. In what follows I will deal primarily with protein distances, as opposed to nucleotide distances, simply because some of the more interesting problems such as the metazoan radiation, eukaryotic diversification, and vertebrate divergences have occurred over a time period that exceeds the resolving power obtained from an analysis of neutral DNA evolution. We are forced to look at proteins whose molecular clocks have been slowed by functional constraint.

Formally, a molecular distance is the number of mutations (or the number of amino acid replacements) that have occurred since the separation of two genes. A distance measure can be extremely useful because of the existence of molecular clocks, which means there is a possibility of determining a time of separation from the distance. In crude terms we can possibly infer evidence of horizontal gene transfer if the time of divergence of two genes significantly deviates from the time of divergence of two lineages [27]. The use of molecular distances to infer horizontal gene transfer events has not received as much attention as have the phylogenetic congruency test or the deviations in gene composition (see below). It is, however, the belief of this author that many of the more interesting developments in the near future in horizontal gene transfer will emerge from distance studies.

Is There a Molecular Clock? One of the reasons that distance measurements have received minimal notice is a residue of distrust in the notion of the molecular clock. Thirty years ago proponents of the molecular clock argued that not only was there a stochastic clock but also that replacement or substitution rates in different lineages were the same. This latter point is certainly not correct. We now know that substitutions per bp per year vary between different lineages [15]. This fact does not, however, mean that a molecular clock is not operating within a given lineage or that such a clock cannot be calibrated. The relative rate test is again important to ensure that the rates of evolution in the respective lineages are comparable [15]. There are two distance approaches that I would like to consider here: one is what I have called the distance matrix rate test and the other is the use of protein distance ratios. Both of these have the potential to reveal distance discrepancies that may uncover horizontal gene transfer events.

Distance Matrix Rate Test (DMR). This is a test that allows one to compare the rate of divergence of a protein from a large number of species to the rate of evolution of a species being compared. Basically, one plots a whole distance matrix from the gene under consideration against a "standard" distance matrix from the same set of species that presumably represents the genomes of those species. The standard can be the average distance of all of the shared proteins across the genome [28] or it can be a representative gene that is believed to characterize the genome [29]. The advantage of this test is that we need not assume that the rate of evolution in the different lineages is the same (no need for a constant molecular clock) nor do we need to know the phylogenetic topology.

One of the advantages of the DMR test is that we can estimate a confidence level in the discrepancy between a protein distance as compared to the standard. The approach that I took is shown in figure 9.3

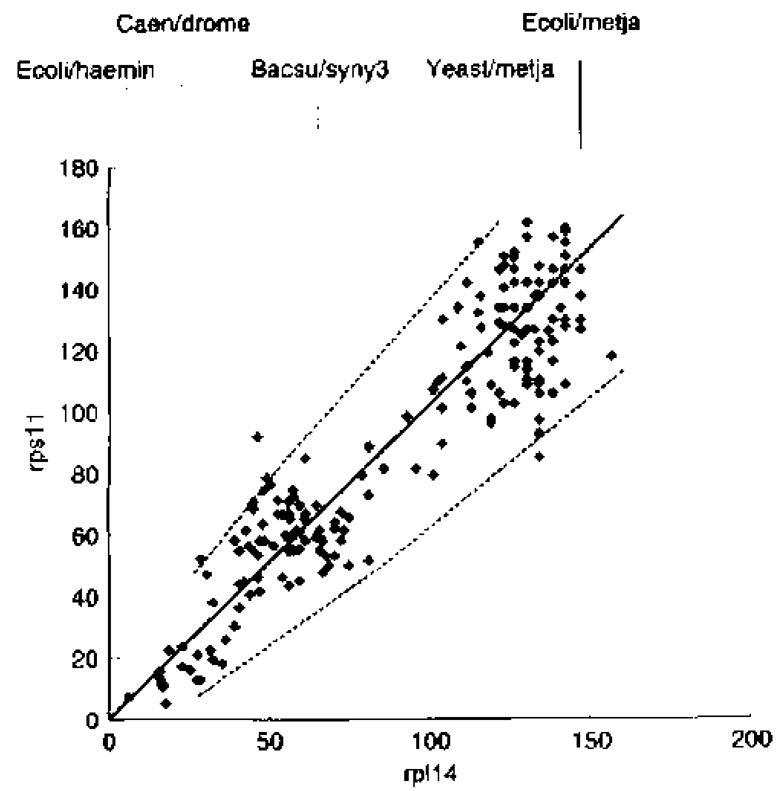

Figure 9.3 Example of a distance matrix rate (DMR) test. Complete distance matrices for both rps11 and rpl14 proteins were computed from the same set of taxa. A distance between taxon X and taxon Y for the rps11 protein is plotted against the distance between taxon X and taxon Y for the rpl14 protein. There are twenty different taxa selected from Bacteria, Eukarya, and Archaea. The location of some representative taxon pairs is at the top of the figure. Ecoli, *Escherichia coli*; haemin, *Haemophilus influenzae*; Cain, *Caenorhabditis elegans*; drome, *Drosophila melanogaster*; Bacsu, *Bacillus subtilis*; syny3, *Synechococcus*; Yeast, *Saccharomyces cerevisiae*; and metja, *Methanococcus jannaschii*.

(from Syvanen [29]). The distance matrix for the ribosomal protein rpl14 is plotted against that of rps11. The same set of species is present in both. The species include representatives of Bacteria, Archaea, and Eukaryotes. The dotted lines gives an expected 95% confidence level where it is assumed that divergence of the two proteins from the last common ancestor is neutral (hence it is assumed that the rate is determined simply by the genome-wide mutation rate) and further that the amount of functional constraint acting against divergence for each protein is the same in the different lineages. There is no need to assume constant clocks or to know phylogenetic relationships. The two ribosomal proteins were chosen in this example because they were expected to have a very low chance of being involved in HGT events. The linear fit of the data with few points lying outside of the 95% confidence level supports this assumption.

The occurrence of HGT events can be inferred when data points for one of the proteins falls significantly outside of the linear regression. This has been exploited by Novichkov et al. [28], who have successfully used this method to identify very likely HGT events.

It is difficult to provide a rigorous interpretation for the error analysis. Normally with N independent comparisons one could simply calculate the correlation coefficient and covariance in order to test whether a time-dependent stochastic process relates the two variables. This will not work here because the distance matrix values are autocorrelated, that is, N independent sequences will yield $N(N-1)/2$ distances. Thus, any calculated covariance will be artificially low. Therefore, to assess whether or not the replacement process is random, the 95% confidence intervals are given in figure 9.3. In addition to this difficulty, many of lineages may share histories over a considerable period of time, which means there are even not really N independent sequences. This is a problem that also bedevils the relative rate test. Novichkov et al. [28] have used a different approach to analyze error, but the same uncertainties as are found in [29] apply with their approach as well.

Protein Distance Ratios. I have recently developed another approach that exploits aberrant distances in order to identify possible genes that are involved in HGT events. In this approach, distances based upon different proteins within the same genomes are compared. The difficulty with comparing two different proteins is that different levels of functional constraint act upon each, so without knowing in advance the amount of functional constraint a direct comparison of distance will tell us little. However, there is a possibility of using ratios of protein distances that can overcome this problem. This approach requires the sequences of protein orthologs from three different taxa. The approach is illustrated in figure 9.4. In the current example, distances between proteins from the tunicate *Ciona*, humans, and the yeast *Sachharomyces cerevisiae* are used. As is described in figure 9.4, the distances between

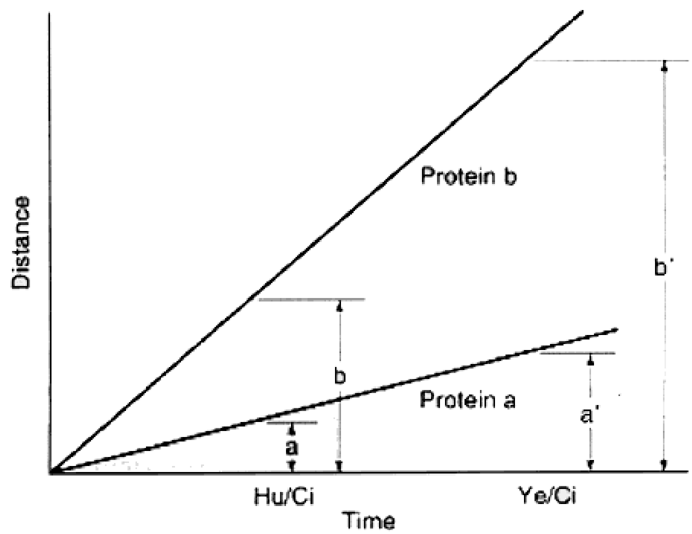

Figure 9.4 Schematic illustrating the distance ratios metric. Evolutionary time is determined by the time of divergence between two lineages. In this example we have two time points—the human (Hu)/*Ciona* (Ci) division and the *S. cerevisiae I* (Ye)/chordate (Ci)division. If distance were some measure of the number of amino acid replacements per amino acid site deduced from pairwise comparisons that is linear with time of divergence, then we would have linear molecular clocks. Protein a and protein b are two different orthologs that differ from each other in their level of functional constraint.

Ciona and human orthologs are normalized to the distance of those proteins to the yeast protein. If three following assumptions are met—that the evolution of each protein within the three lineages reflects the evolutionary history of the underlying species, that the protein distance measure is linear with time of divergence, and that the amount of functional constraint acting upon each protein is the same within the three lineages—then the ratio of the distances of the two proteins should be the same, that is, from figure 9.4 $a/a' = b/b'$. Hence deviations in these ratios will offer evidence of atypical protein evolution.

This exercise was carried out using the complete protein sequence databases for the *Ciona*, human, and yeast genomes. A group of about 200 protein sequences that appeared to be orthologous (i.e., a single copy within the genomes) was chosen. One advantage of using protein ratios in this way is that we can directly test the assumption that the distance measure is linear with time. If a protein distance becomes saturating with time, then we would expect to see the ratio of distances to increase with increasing protein distance. Figure 9.5 shows the result of plotting the ratio of distance against absolute distance. There are a number of competing methods for measuring protein distances available.

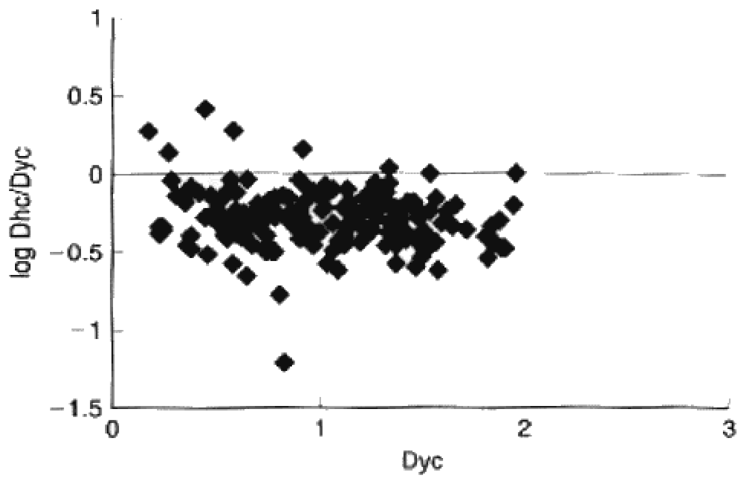

Figure 9.5 Distance ratios are constant with increasing distance. The set of 185 proteins shown were chosen in the following manner. The 6800 protein sequences from the *S. cerevisiae* genome were used as queries in Blast searches against a database that contained the proteins deduced from the human, *Ciona*, *C. elegans*, *Drosophila*, *Arabidopsis*, and *Oryza sativa* genomes. This yielded a list of over a thousand yeast proteins such that at least one copy of the homolog was found in each of the other genomes. This list was reduced by removing the obvious gene families containing parologs, thereby enriching this list for orthologous sets. The chosen genes were aligned, all indels removed, and then the JTT distances determined. The corrected distance between human and *Ciona* (Dhc) was divided by the yeast–*Ciona* distance (Dyc). The logs of the ratios are normally distributed about a constant mean independent of Dyc values.

In the current study I tested five of them (data not shown). Shown in figure 9.5 is the one that gave the best result, that is, the slope of the curve was closest to zero. This distance is based on the JTT matrix [30], though the Dayhoff PAM measures worked reasonably well also. Simple distances, Poisson corrected distances, and Kimura protein distances gave significantly nonzero slopes.

Figure 9.6 shows the data from the ordinate in figure 9.5 plotted as a simple histogram. The log of distance ratios is given because the direct ratios are not normal, though there are reasons to believe they should be lognormal. In figure 9.6 we can see that the data roughly approximates a normal curve but that there are many outlying points. In fact the distribution is highly overdispersed with about 10% of the sequences lying outside of a normal distribution. The genes represented by these outliers are candidates for possible horizontal gene transfer events. This study remains unfinished at this point. Though we have a tool here that allows identification of sequences that are atypical, once identified there is a difficulty with concluding that an HGT event

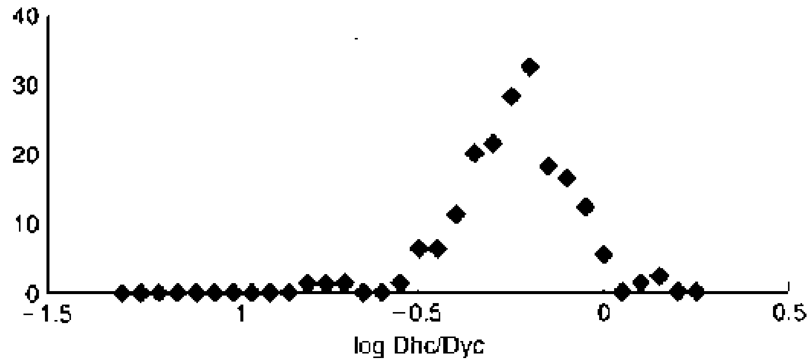

Figure 9.6 Distance ratios distribution. The distance for each of the proteins between human and *Ciona* (Dhc) was divided by the distance between the two metazoa and *S. cerevisiae* (Dyc) and the distribution of the log of this ratio is shown.

had occurred. This is due to the very real possibility that the three proteins are not an orthologous set but that a paralog is present. Many of the early claims of HGT (reviewed in [11]) turned to artifactual because of the inclusion of parologous sequences. The possibility of selecting for a paralog seems high with a screen that looks at hundreds of genes as described here. Further analysis of these examples will be required before we can conclude that the highly dispersed points in figure 9.6 are evidence for HGT events.

In summary, distance ratios provide another metric that allows us to detect HGT events. Though this work remains incomplete, there are two internal controls that offer encouragement. One is that the log ratios versus distance plot in figure 9.5 is flat, thus supporting the notion that distances upon which we are basing the ratios are linear with time. Another factor that is encouraging is that the peak of the curve in figure 9.6 lies at 0.44, which places the *Ciona*/human (two chordates) divergence time at about 440 million years ago (assuming a fungal–metazoan divergence of 1 BYA). This is after the Cambrian radiation of 540 MYA, as it likely should be.

ATYPICAL NUCLEOTIDE COMPOSITION (THE ANC ISLANDS)

In the previous three sections the procedures described for identifying HGT events rely on the comparisons of orthologs from multiple species. This section describes an approach where deviations in nucleotide composition are used to identify foreign genes [31,32]. The genes identified by this criterion fall into a number of different categories. These include phages, plasmids, insertion sequences, and other mobile genetic elements. In addition, we can include a class of genomic regions that have been called, in different contexts, pathogenicity islands, plasticity islands, or accessory gene regions. These classes range from those genes that are clear parasites to selfish genes to those that can be considered accessory and even some that are essential. The notion of accessory genes has been around for many years either implicitly or explicitly [33]. They are DNA elements that are often associated with mobile genetic elements. They allow the organism to exploit some highly specialized ecological niche that may require only temporary unions of genes that can be easily lost when no longer useful. Mobility becomes important to long-term gene survival when selection for an associated trait is lost. Because accessory genes include a large variety of genetic elements, I will collectively refer to them as the atypical nucleotide composition (ANC) genes or ANC islands when multiple genes are clustered.

When we refer to atypical compositions we refer to deviations from within a given genome since different species have their own unique composition. There are different evolutionary forces responsible for these unique genome compositions. It is clear that major bacterial assemblages vary in their GC content, from a low of 25% to as high as 75%. Even for bacteria with similar GC content, synonymous codon use can differ. There appear to be biases in nearest neighbor frequencies (dinucleotide frequencies) or even biases in oligonucleotide frequencies of up to eight [34,35]. Biases in longer oligonucleotides are probably caused by other mechanisms than just GC content and codon bias. Positive selection against certain restriction nuclease sites or other DNA metabolism factors could lead to different bacteria obtaining different localized compositions.

Thus, for a variety of reasons, the distribution of oligonucleotides will be skewed from random and that skewing can provide a unique signature for a given bacterium. In most bacteria, the composition signature that uniquely characterizes the genome applies to about 85% of the genes on average, while the remainder of genes seem to have compositions governed by different rules [36,37]. It was originally proposed that these atypical genes could be the result of horizontal gene transfer (HGT) with the atypical gene carrying the signature of a foreign donor genome. According to this postulate the donor is so remotely related to the recipient that its genome composition is significantly different. In fact, in recent years this explanation has become so widely accepted that the finding of atypical regions has been considered a measure of HGT. A strong prediction of the remote donor hypothesis for ANC is the existence of a donor species whose genome composition at the time of HGT reflects this atypical composition. Because such remote donor species have not been found, the hypothesis remains on shaky grounds.

In this review I suggest a revision of the remote donor HGT hypothesis for atypical nucleotide composition and will go into a detailed discussion of the need to consider alternatives. I will continue to maintain that the atypical genes identified in the above studies have most likely been involved in HGT events. The revision that I propose is that the atypical composition observed does not reflect the genome composition of the donor species, but rather reflects the property of gene mobility per se.

A number of lines of evidence can be offered to support the gene mobility hypothesis for the ANC islands. Let us begin with the sequence composition of phages, prophages, plasmids, and insertion sequences that are invariably atypical when compared to the hosts that carry them. This, of course, has been attributed to their likely residence in remote donors with different nucleotide compositions. The problem, however, is that the sequences of these particular elements have been the subject of extensive study for over three decades without any hint of remotely related donors. For example, let us consider the lambdoid phages that have been extensively surveyed but so far have been encountered only within a narrow group of gamma-proteobacteria, the so-called enterics [38,39]. The lambdoid phages are a closely related group of phages that have similar genome organizations and can give rise to hybrids between different members of the group. The enteric bacteria as a group, however, do not differ in their genome composition. For example, *E. coli* and *Salmonella* have virtually identical GC compositions and codon usages [40]. A similar pattern is seen with plasmids and insertion sequences. That is, with a few spectacular exceptions, plasmids and insertion sequences seem to have limited host ranges (i.e., they are found within a group that generally have the same genome compositions) but their sequences themselves show the atypical composition. Thus there is no independent evidence for remote species donors, rather the alternative seems to be the case; the apparent donor species most likely have the same genome composition as do the recipients.

There are other puzzling patterns in the nature of these genes with atypical composition that are difficult to reconcile with the idea of remote species donors. There are many genes with aberrant GC content found in the chromosome of *E. coli*. It turns out that over 90% of those that deviate significantly from *E. coli*'s GC content of 0.5 are greatly enriched for AT. It was difficult to explain this asymmetry as being due to the donor species distribution. I suggested in 1994 [11] an alternative explanation that a bias in favor of AT is consistent with a gene which in the course of its life cycle was frequently submitted to homeologous recombination events. This predicts a family of mobile genes whose integration in new recipient chromosomes relies on homeologous recombination [17] as opposed to the site-specific integration events associated with prophages, inserted plasmids, and insertion sequences.

Evidence that such a class of genes is found in the ANC islands has recently been described. Koski et al. [41] argued against the use of atypical compositions as an indicator of horizontally transferred genes after a detailed analysis of the types of genes that Lawrence and Ochman [42] had identified as HGT candidates because of ANC. One of the apparent problems that Koski et al. noted was that 135 of the 747 genes classified as horizontally transferred in *E. coli* turned out to have positional orthologs in the bacterium *Salmonella*. If both *E. coli* and *Salmonella* have the same positional ortholog, then this strongly suggests that the common ancestor to these two strains also carried this ortholog. However, it appears the these two bacteria diverged about 100 million years ago, whereas Lawrence and Ochman [42] had already estimated that after 100 million years of vertical evolution a gene's composition would be expected to have converged to that of the host genome. They call this amelioration. Thus the apparent dilemma: common ancestry suggests that these 135 genes were present in the lines leading to *E. coli* and *Salmonella* 100 million years ago, but sequence composition suggests that these genes were introduced into *E. coli* considerably more recently. The explanation that I am offering for this class of genes is not that these genes are not involved in HGT events, as argued by Koski et al. [41], but rather that they are mobile genes that can be lost but also can reestablish themselves in enteric chromosomes via homeologous recombination events. Such a mechanism of transfer would necessarily preserve positional orthology and, according to the current hypothesis, such recombination events select for the atypical nucleotide composition.

The phenomena of homeologous recombination between homologous genetic regions of *E. coli* and *Salmonella* have been well documented. Mutant strains of *Salmonella* missing the methyl-directed mismatch repair pathway recombine with an *E. coli* donor [17]. It is probably not coincidence that wild strains of *E. coli* are frequently encountered that have lost this mismatch repair pathway even though such mutants have a serious growth disadvantage when compared to wild-type strains. Evidence that recombination events have occurred naturally between these two lineages is seen with apparent concerted evolution of the elongation factor Tu, *tufA* and *tufB* loci [40]. Another fact that supports frequent recombination events between diverged but homologous sequences is the finding that one very important class of genes found on one of *E. coli's* ANC islands has a highly mosaic pattern [43].

As mentioned above, there is so far no direct evidence for the remote species origins for the ANC islands. There is one recent study that illustrates the size of this problem. Nakamura et al. [36] surveyed the genomes of 116 bacteria for nucleotide composition and found that on average 14% of the genes had atypical compositions, which was represented by 1357 gene clusters. They furthermore probed their 116-genome database with these 1357 gene clusters to see if the codon bias in one of them would show a match to any of the other 115 genomes and thereby possibly locate the donor. The power of the technique was shown in that they did in fact identify one donor–recipient pair among those 1357 gene clusters. This pair, however, turned out to be an artifact; the strain of *Neisseria meningitides* that had been sequenced, happened to carry erythromycin-resistant *Staphylococcus* plasmid genes that had been unsuspectingly cloned into the *Neisseria* species [44], and this was the gene that Nakamura et al. [36] identified. Thus, in this fairly large survey, we can conclude that so far no clear case of naturally occurring remote species HGT events have been identified using genome difference analysis.

The reason for presenting detailed arguments for the mobility hypothesis as an explanation for the ANC islands is that it leads to testable predictions that can be revealed by future computations. In addition, it gives rise to certain expectations with respect to the chemical properties of DNA and molecular mechanisms of genetic recombination. But this goes well beyond the scope of this review.

Horizontal gene transfer is an indisputable fact. In general terms, types of genes have been divided into two classes on the basis of transfer frequency: informational genes and operational genes. The accessory genes found on the ANC islands should be included as a third category:

$$\text{Informational} \rightarrow \text{Operational} \rightarrow \text{Accessory}$$

And moving from left to right the likelihood of the genes being involved in horizontal gene transfer seems to increase dramatically.

REFERENCES

1. Ochiai, K., T. Yamanaka, K. Kimura and O. Sawada. Inheritance of drug resistance, and its transfer between *Shigella* strains and between *Shigella* and *E. coli* strains. *Hihon Iji Shimpor*, 1861:34, 1959, in Japanese.
2. Akiba, T., K. Koyama, Y. Ishiki, S. Kimura and T. Fukushima. On the mechanism of the development of multiple-drug-resistant clones of *Shigella*. *Japanese Journal of Microbiology* 4:219–27,1960.
3. Went, F. W. Parallel evolution. *Taxon*, 20:197–226,1971.
4. Reanney, D. Extrachromosomal elements as possible agents of adaptation and development. *Bacteriological Reviews*, 40:552–90,1976.
5. Golding, G. B. and R. S. Gupta. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Molecular Biology and Evolution*, 12:1–6,1995.
6. Gogarten, J. P., W. F. Doolittle and J. G. Lawrence. Prokaryotic evolution in light of gene transfer. *Molecular Biology and Evolution*, 19:2226–38,2002.
7. Doolittle, W. F. Lateral genomics. *Trends in Cell Biology*, 9:M5–8,1999.
8. Brown, J. R. and W. F. Doolittle. Gene descent, duplication, and horizontal transfer in the evolution of glutamyl- and glutaminyl-tRNA synthetases. *Journal of Molecular Evolution*, 49:485–95,1999.
9. Woese, C. The universal ancestor. *Proceedings of the National Academy of Sciences USA*, 95:6854–9,1998.

10. Zhaxybayeva, O. and J. P. Gogarten. Cladogenesis, coalescence and the evolution of the three domains of life. *Trends in Genetics*, 20:291,2004.
11. Syvanen, M. Horizontal gene transfer: evidence and possible consequences. *Annual Review of Genetics*, 28:237–61,1994.*
12. Syvanen, M. On the occurence of horizontal gene transfer among an arbitrarily chosen group of 26 genes. *Journal of Molecular Evolution*, 54: 258–66,2002.*
13. Zhaxybayeva, O. and J. P. Gogarten. An improved probability mapping approach to assess genome mosaicism. *Genomics*, 4:37,2003.
14. Felsenfeld, J. Evolutionary trees from DNA sequence. *Journal of Molecular Evolution*, 17:368–76,1981.
15. Li, W. -H., M. Tanimura and P.M. Sharp. An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *Journal of Molecular Evolution*, 25:330–42,1987.
16. Milkman, R. Transduction, restriction and recombination patterns in *Escherichia coli*. *Genetics*, 139:35–43,1995.
17. Rayssiguier, C., D. S. Thaler and M. Radman. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. *Nature*, 342:396–401,1989.
18. Dowson, C. G., A. Hutchison, N. Woodford, A. P. Johnson, R. C. George and B. G. Spratt. Penicillin-resistant viridans streptococci have obtained altered penicillin-binding protein genes from penicillin-resistant strains of *Streptococcus pneumoniae*. *Proceedings of the National Academy of Sciences USA*, 87:5858–62,1990.
19. Korber, B., C. Brander, B. F. Haynes, R. Koup, J. P. Moore, B. D. Walker and D. I. Watkins. http://hiv-web.lanl.gov/content/hiv-db/CRFs/CRFs.html at http://hiv-web.lanl.gov/content/immunology/.
20. Maynard-Smith, J. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution*, 34:126–9,1992.
21. Kececioglu, J. and G. Gusfield, Reconstructing a history of recombinations from a set of sequences. In *Proceedings of the 5th ACM-SIAM Symposium on Discrete Algorithms*, pp. 471–80, 1994.
22. Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly and D. Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296:2225–9,2002.
23. Eskin, E., E. Halperin and R. M. Karp. Efficient reconstruction of haplotype structure via perfect phylogeny. *Journal of Bioinformatics and Computational Biology*, 1:1–20,2003.
24. Gusfield, D, S. Eddhu and C. Langley. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. *Journal of Bioinformatics and Computational Biology*, 21:173–213,2004.
25. Tani, T., Y. Takahashi, S. Urushiyama, Y. Oshima, M. Go and P. Schimmel, eds. in *Tracing Biological Evolution in Protein and Gene Structures*. Elsevier, Amsterdam, 1995.
26. Syvanen, M. Cross-species gene transfer: implications for a new theory of evolution. *Journal of Theoretical Biology*, 112:333–43,1985.*
27. Syvanen, M. Molecular clocks and evolutionary relationships: possible distortions due to horizontal gene flow. *Journal of Molecular Evolution*, 26:16–23,1987.*

28. Novichkov, P. S., M. V. Omelchenko, M. S. Gelfand, A. A. Mironov, Y. I. Wolf and E. V. Koonin. Genome-wide molecular clock and horizontal gene transfer in bacterial evolution. *Journal of Bacteriology*, 186:6575–85, 2004.
29. Syvanen, M. Rates of ribosomal RNA evolution are uniquely accelerated in eukaryotes. *Journal of Molecualar Evolution*, 55:85–91,2002.*
30. Jones, D. T., W. R. Taylor and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Computational and Applied Bioscience*, 8:275–82,1992.
31. Lawrence, J. G. and H. Ochman. Molecular archaeology of the *Escherichia coli* genome. *Proceeding of the National Academy of Sciences USA*, 95:9413–17,1998.
32. Garcia-Vallve, S., E. Guzman, M. A. Montero and A. Romeu. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Research*, 31:187–9,2003.
33. Court, D. and A. Oppenheim. Phage lambda's accessory genes. In R. Hendrix, J. Roberts, F. Stahl and R. Weisberg (Eds.), *Lamdba II*. Cold Spring Harbor Press, Cold Spring Harbor, N.Y., 1983.
34. Pride, D. T., R. J. Meinersmann, T. M. Wassenaar and M. J. Blaser. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Research*, 13:145–58,2003.
35. Tsirigos, A. and I. Rigoutsos. A new computational method for the detection of horizontal gene transfer events. *Nucleic Acids Research*, 16:922–33,2005.
36. Nakamura, Y., T. Itoh, H. Matsuda and T. Gojobori. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, 36, 760–6, 2002.
37. Ochman, H., J. G. Lawrence and E. A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405:299–304,2000.
38. Hendrix, R. W. Bacteriophage genomics. *Current Opinion in Microbiology*, 6:506–11,2003.
39. Hendrix, R. W. Bacteriophage lambda: the genetic neighborhood. In R. Calendar (Ed.), *The Bacteriophages* (pp. 409–47). Oxford University Press, New York, 2006.
40. Sharp, P. M. Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. *Journal of Molecular Evolution*, 33:23–33,1991.
41. Koski, L. B., R. A. Morton and G. B. Golding. Codon bias and base composition are poor indicators of horizontally transferred genes. *Molecular Biology and Evolution*, 18:404–12,2001.
42. Lawrence, J. G. and H. Ochman. Amelioration of bacterial genomes: rates of change and exchange. *Journal of Molecular Evolution*, 44:383–97,1997.
43. Denamur, E., G. Lecointre, P. Darlu, O. Tenaillon, C. Acquaviva, C. Sayada, I. Sunjevaric, R. Rothstein, J. Elion, F. Taddei, M. Radman and I. Matic. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell*, 103:711–21,2000.
44. van Passel, M., A. Bart, Y. Pannekoek and A. van der Ende. Phylogenetic validation of horizontal gene transfer? *Nature Genetics*, 36:1028, 2004.

*These papers are available in pdf format at http://www.vme.net/hgt/