STUDY NOVEL TECHNIQUES FOR WHOLE GENOME PHYLOGENETIC ANALYSIS

Abstract

Previous research has been proved that phylogenetic trees are useful for the comparative studies between different organism on the basis of the evolution in their DNA or Protein sequences at few specific points. Currently it has been proved that with the help of phylogenetic tree we can predict the evolutionary relationship through the comparative analysis of entire genome.

In this book chapter we focused on various novel techniques for the phylogenetic analysis of complete genome. These experiment target on gene comparison and gene order. Here we discussed several techniques for making these comparisons. In recent time period we were using Maximum Parsimony Method and Distance Methods beside this recently we are using Maximum Likelihood and Bayesian methods are developed.

In this book chapter we discuss all approaches in turn, including their applications and harmful effects and software which make use of them.

Keywords: These experiment target on gene comparison and gene order. Here we discussed several techniques for making these comparisons.

Authors

Reetu Gour
Assistant Professor
Department of Microbiology
IIMT University
Meerut
ritugourbioinfo2@gmail.com

Ayushi Pal
Assistant Professor
Department of Biotechnology
IIMT University
Meerut
ayushipal_slst@iimtindia.net

Prof. Mukta Sharma
Deputy Dean
Department of Microbiology
IIMT University
Meerut
muktavats@yahoo.com

I. INTRODUCTION

Statistical analysis of the relationship between variables has been intensively researched since 1960s.Even though the complete methods and materials available have changed a lot since then, practical tests are still in principle the same (1).

We attempt to analyse characters of our living organism that show difference between few species then we try to create an estimate of the evolutionary tree based on the model to explain the differences or similarities we see (2).

The first experiments were based on morphological characteristics of bacteria and various methods were developed[1]. However, different characters often cause confusion, and since the decision to use morphological characters is somewhat problematic, a more reliable method is needed. Since each disease is defined primarily by the genetic information received from its parents, it is more logical to directly compare the morphological features of genes rather than to identify them (3). Information about the DNA or amino acid sequences of homologous genes has been obtained and many methods have been developed to use this information. Some rely on the original maximum parsimony method, but others are more robust and can not only estimate phylogenetic relationships but also provide distressing levels of confidence. The most approved phylogenies were obtained from the linkage of individual genes (Homologous). However, sometimes this process ends in failure (3). The initial stage researchers observed that phylogenetic relationship between organism was completely different to evolutionary history of genes in which the gene may be present.

This may be due to duplications, deletions, or even horizontal changes in the species' genes (mostly found in prokaryotes), so phylogenies from different genes may appear inconsistent. Second, it is no easy to detect the desirable genes in all three species of interest, but the phylogeny is different enough to be. Determine phylogeny by considering Species (Organism) at genomic level rather than the genes of human (4). This is due to the increasing number of complete genome sequences and the belief that, when considering the genes of a species, the evolutionary history of an organism's entire genome is more reliable than the history of a single gene. In addition, determining the genome phylogeny will form the basis for examining phenomena that affect the analysis of each other, like as redundancy and horizontal interchange of genes (5).

II. EVOLUTIONARY STUDIES AT GENOMIC LEVEL

In this study we will discuss the difference between different organism at genomic level. This difference will calculate on the basis of the evolution at position of specific gene or Gene Content and arrangement of gene in a genome (Gene Order). With the help of this study easily observe the orthologous gene before the comparative analysis of Gene content and order of gene (6).

1. **Prokaryotes:** Prokaryotic genomes are simple and usually consist of a single genome in circular form. Various Prokaryotic species genomes sequencing have been done and Mapping is also has been completely done (7). A recent analysis of prokaryotic gene order has revealed the large differences in gene conservation was predicted in different lineages stages, as well as differences in evolutionary processes; but in most cases the lines between replication history and main content transformations appear to be the same. widespread. Although gene content comparisons have been used to do this, we do not yet know whether comparisons of prokaryotic gene sequences will provide useful phylogenetic information (8). It is increasingly believed that the evolutionary history of prokaryotes cannot be represented by trees due to frequent gene transfer and hybridization. However, recent studies have shown However, many phylogenetic networks may be needed to accurately represent prokaryotic evolution, but there is currently no real consensus on how these should be constructed or what they should represent. See the "Phylogenetic Networks" section below for further discussion (8).

2. **Eukaryotes:** Eukaryotic genomes are more complex than prokaryotic genomes and thus pose a broad challenge to their analysis. Generally speaking, there are more genes, more elements, and more chromosomes. Duplication events result in many copies of genes, resulting in divergence of orthologs (for example, in Yeast we observed approx. 25% redundancy of genes, suggesting an equivalent of 1% of genes for My5). On the other hand, horizontal transfer of genes is less likely in prokaryotes (9).

   Firstly, the complete genome of saccharomyces cerevisiae was sequenced. Currently genome of various yeast species related to *Saccharomyces cerevisiae* have been examined. Various analyses have concluded that frequent mutations in small segments, gene duplications and losses and polyploidy events are the main forces driving gene rearrangements. Similar studies have begun to be carried out in the field of animal and plant genomics (10).

3. **Organellar Genomes:** In addition to large, complex nuclear genomes, most eukaryotes also have small, simple mitochondrial genomes that have evolved independently of the nuclear genome, thus providing an additional source of phylogenetic information (10). These genomes contain approximately 35 genes and have been sequenced in many different species. Additionally, all plant species have a chloroplast genome containing 120 genes. (See the MRC website for links to chloroplast and mitochondrial genome sequence databases11.) These "organelle" genomes are highly conserved in terms of gene expression and constitute some of the most widely studied databases (11).

4. **Models of Genome Evolution:** All methods of phylogenetic analysis are based on understanding the mechanisms underlying differences between taxa (12). These can occur, for example, in changes in weights or distance measurements for maximum frugality schemes. Probabilistic methods such as maximum likelihood or Bayesian analysis require good mathematical models (12).

   We use a generalized version of the Nadeau–Taylor model 15 to represent genome evolution. Here, chromosomes are defined in rows or circles around the genes they contain, and orthologous genes are given the same label (13). Various evolutionary events (such as insertions, deletions, duplications, or translocations) can change the sequence or

content of regulation. The precise nature of these events and the rates at which they occur can be determined from our prior knowledge of the genome in question or estimated from our data clocks. Figure 1 shows how genomes are represented and how they have changed through evolutionary events (14). These are all evolution, but only some of them. For example, if we know that there is no significant conservation of genes across genomes, then we should not bother with the model Instead we will consider the model of gene transfer and consider only comparing the content of genes (15). However, studies of genetics can be inaccurate in the absence of complete data, so we will restrict our models to tracing evolutionary patterns and consider comparisons between genetics. Genetics are just one part of our genomes. Often, we don't know much about the actual processes that cause differences in certain data, so our models may not accurately reflect the underlying processes (16). In these cases, we must make sure that our results do not depend on the model we choose (17).

III. PHYLOGENETIC ANALYSIS

Many attempts to reconstruct genome phylogenies have used methods derived from DNA analysis, which are well described by Page and Holmes. How these methods can be applied to the analysis of whole genomes is discussed below(18,19).

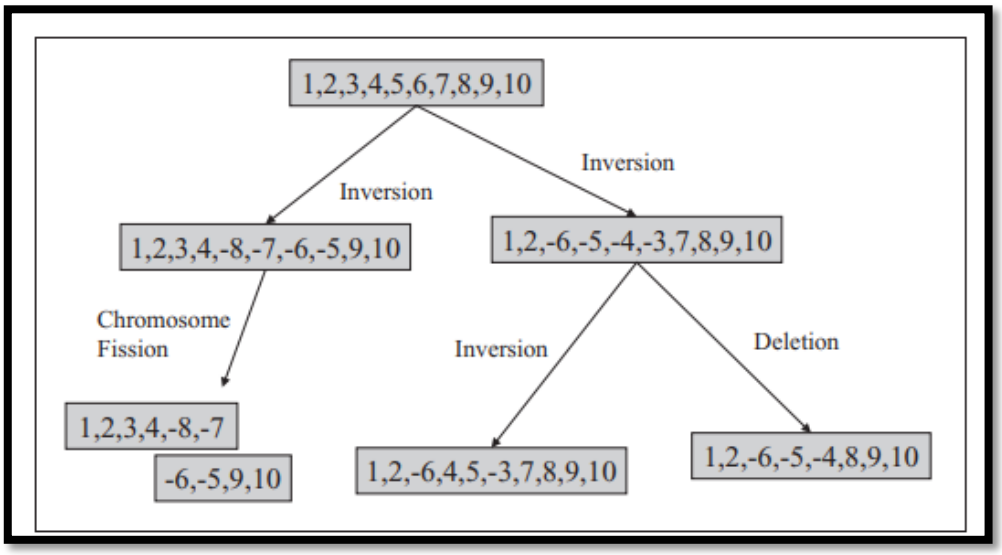

Figure 1: Diagrammatic representation of evolutionary relationship between different species in which negative value is representing the no change in the orientation of gene.

Encoding of Binary Attributes: Early attempts at numerical phylogenetic analysis were based on the analysis of observed dichotomous characters in each species of interest (19).

Therefore, if the genome can be encoded as a series of self-translating boxes, we can resolve genomic phylogenies using a library of techniques designed for such analyses (20).

How can we binary encode observable differences in the genome? The presence or absence of genes or protein families can be easily represented (depending on their

identification),[17] but when we consider genes the answer is not so obvious (21). Researchers have discussed various encodings, including "joining coding," which encodes adjacent genes as symbols, and "relative position coding," in which each gene is replaced by a character (non-binary) that represents its position relative to the genome (22). Researchers used the pairwise neighbour method in "maximum binary encoding" (MPBE), and few other researchers also used this method in the analysis of baculovirus genomes. These analyzes have proven successful because they produce results consistent with those obtained with other methods (23). Figure 2 shows the binary coding of the three compared gene sequences (23,24).

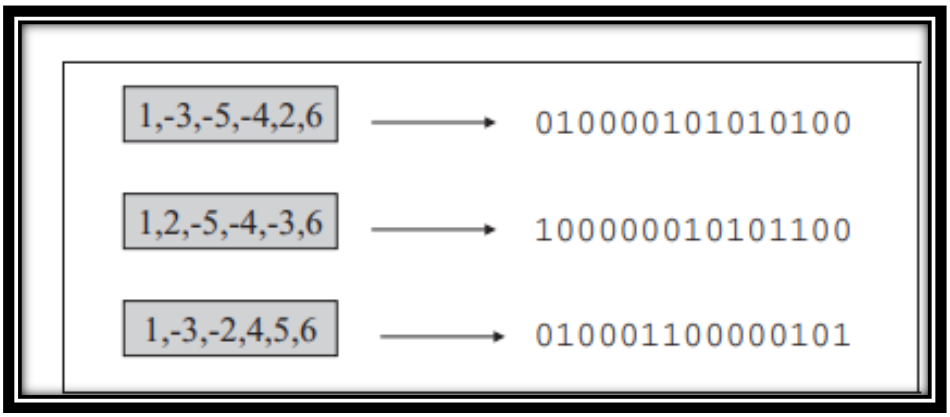

Figure 2: Binary encoding of gene orders in a sequence and this binary encoding sequence representing the absence and presence of pairs of gene. This result can analyse with the help of Standard Phylogenetic Package.

During their research researchers proposed a differential (non-binary) coding in which each gene corresponds to two symbols representing the genes on either side of it (25). This method is good, although the number of possible states for a character is the same as the number of genes in the system, which is often beyond the limits of identification algorithms (26).

However, we must be careful when using encodings of this nature for the following reasons: the obvious seed cannot be closer to more than two seeds, so the state of one behaviour will obviously affect the state of the others (27). We found that not all attributes of the state function correspond to a valid genome. Moreover, if we consider an evolutionary phenomenon such as change, this will both create and destroy (28). Overt behaviour cannot be changed on its own. This means that in order for this method to provide genetic estimates of ancestral nodes, we need to ensure that each of our ancestral states actually corresponds to the valid genome. This would be less of a problem if the gene sequence is not well-composed and the neighbouring gene is not well-composed (29).

The GO Tree software package developed by Bryant handles (among other things) the encoding of genetic information. It is exported in Nexus format 21 and can then be analysed using the popular PAUP phylogeny package (30).

IV. DISTANCE METHODS

One of the most popular ways to create a tree is to combine matrices. The distance between each pair of genomes is found and then we find the tree that fits this distance. In this article we are only interested in finding distance measurements; For a discussion of tree construction, see Page and Holmes. [16] Any distance measurement should relate as closely as possible to the actual evolutionary distance between genomes and should ideally be easy to calculate (31).

In one of the first studies on genome phylogeny (28). used the minimum number of mutation events between two genomes as a measure of distance and applied this to the mitochondrial genomes of various species. It is a type of distant modification and its analogues are often used in gene sequence analysis. Here different evolutionary events can have the same value or different weights depending on how we think they will happen (30). A special type of "exchange" distance is transposition, or distance, which is the minimum number of transformations between two gene sequences. Finding the distance is equivalent to solving the "reverse sorting" problem, and improved algorithms have been successfully proposed to achieve this. Recently developed a linear time algorithm to calculate the inversion distance between two circular genomes (31).

"Breakpoint distance" is also commonly used because it is easy to calculate and is robust to misidentified orthologues (30). It is simply the number of adjacent pairs in a genome that are not adjacent to another genome. Extended this concept to induced breakpoint distance, which can be used to compare genomes with different contents and sizes; first corrected the gene content and then normalized the distance so that the size of the seed appeared. This distance (along with the shared distance) was used to determine the phylogeny of the mitochondrial genomes of various plant and algal species (32). Figure 3 shows how to calculate the induction breakpoint distance for two genomes with different elements (32).

We encountered an unexpected problem when assessing modifications or distances between genomes at unrelated points. If a mutation occurs in a gene, we will not know which form will be deleted when editing the gene content. Sankoff calls the original gene the true standard, and the distance based on preserving the gene but removing excess copies is called the standard distance. Bryant[25] showed that comparing samples remotely is NP-hard (33).

Gene distance points are used to construct prokaryotic phylogenies; The distance used here is the proportion of genes shared by two genomes. This approach was used to construct phylogenies related to prokaryotes and Saccharomyces cerevisiae (33).

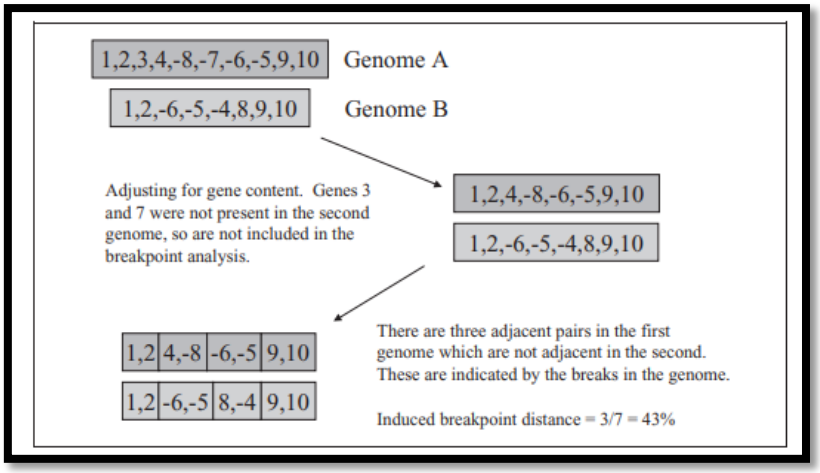

Figure 3: Representation of division between two genomes with unequal count of gene

Prokaryotes and Saccharomyces cerevisiae. Generally, the distance is based on the above limits, meaning they tend to approach a maximum as the actual evolutionary distance grows. For example, in the case of breakpoint distance, this maximum is the number of breakpoints between two genomes. The limit for the induced breakpoint distance is 1 (see Figure 4). Many different models also have a maximum value of the realignment distance (34). Figure 4,5 describes various methods of evolutionary change that allow for a better estimate of the true evolutionary distance. Experiments based on simulated gene sequences show that they can improve the relationship between predictions and distance accuracy and, more importantly, improve the accuracy of phylogeny construction (35). Figure 5 shows a simplified version of the distance measurement described in Figure 4.

Using a comparison of rRNA trees associated with 26 fungal genomes, Keogh et al.[29] determined whether a number of different indices correlated with evolutionary distance (36). They examined the relationship between gene conservation and conservation among all species and S. cerevisiae and found that gene conservation decreases with increasing evolutionary distance (36). However, in many cases there is only a single pair of genes known in S. cerevisiae and orthologs exist in the compared species. Many yeast species need to be sequenced to perform useful phylogenetic analyses (37).

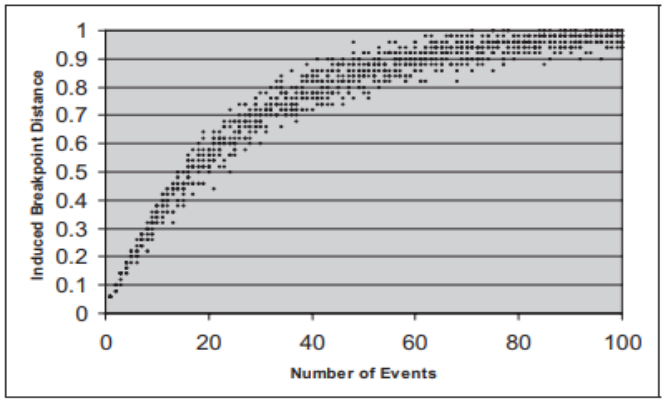

Figure 4: Graphical representation of evolutionary of approx. 100 genes.

1. ***Inslico* Analysis of Phylogenetic Relationship Distance Based Method:** The widely used PHYLIP30 software package contains programs for reconstructing phylogenetic trees from distance matrices. There are many different algorithms that can be used, including neighbour-joining, the Fitch-Margloliash algorithm, and the least-connected algorithm called UPGMA. See PHYLIP documentation for more information (38).

   Several different software packages have recently been released to estimate genome-wide distances. Derange 231 aims to find the weighted tuning distance between two gene sequences. GRIMM Web Server [32] finds the shortest rearrangement distance and provides the best concept of rearrangement between two (possibly multichromosomal) genomes with identical elements (39).

   The SHOT web server provides a complete genome tree only for selected species from the collection, using the sequence of genes or gene content information. There are many options for different transformations and tree building algorithms (40). Almost exclusively in this area the trust can be placed on the inner bones of the tree (see "Measurement Analysis" below) (41).

V. MAXIMUM PARSIMONY

Instead of using the principle of maximum parsimony for binary encodings, it can be applied directly to the entire genome. In this case, we want to find the tree that requires the least number of evolutionary events while accounting for variance (42). We generally limit ourselves to comparing gene sequences because simplicity of gene content can be achieved by binary character coding (see section above) (43).

Finding a small tree (even one with only three genomes) has proven to be NP-challenging, and developing heuristics to find predictive answers is an ongoing research project. Details of the algorithms used to solve this problem and related problems are beyond the scope of this article; see for details (43).

One of the first attempts focused on finding “divergence phylogenies,” that is, trees that minimized the total number of points between all contiguous areas. This is easier to find than the maximum parsimony tree (although it is still NPhard) because the distance itself is easier to calculate than the distance itself. The BPAnalysis33 package uses a heuristic that was developed to solve the phylogeny breakpoint problem but proved too slow to be useful. GRAPPA (44) speeds up the computation by several orders of magnitude, so that entire chloroplast and mitochondrial genomes can be analysed efficiently without significant loss of resolution problems. GOTree20 also finds summary phylogenies, but is not limited to identifying genomes with identical gene content (like GRAPPA) (45).

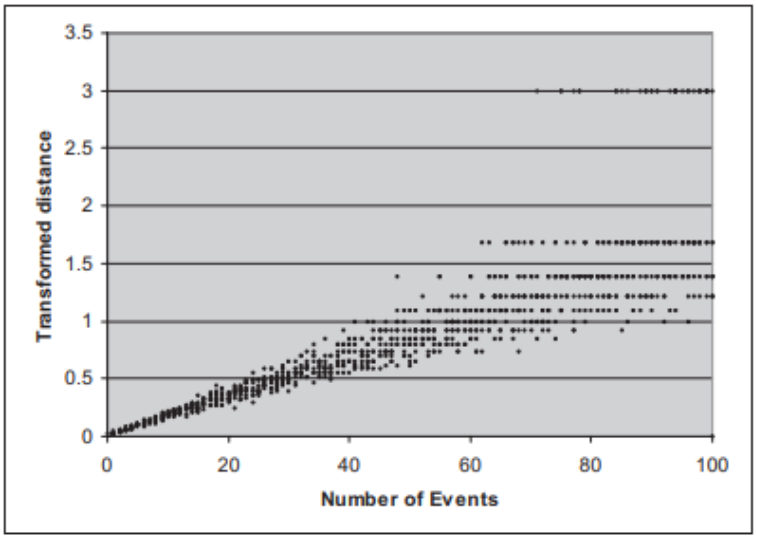

Figure 5: Diagrammatic representation of the effects of transforming the distances between different species.

Recent efforts have been directed towards finding the true maximum parsimonious (relative to evolutionary events) tree, and this is now possible thanks to advances in algorithms and computing power. The MGR algorithm proposed by Borque and Pevzner14 provides a solution to this problem and can be used on web servers (45). Thanks to the improvement of the distance between differences mentioned previously, the new version of GRAPPA can also find distance-based parsimonious trees (45). Parsimony methods are considered more reliable than distance methods when reconstructing trees, but take longer (some distances are O(N2) in number of species, while finding parsimony trees is NP-hard) (46). Additionally, the parsimony method provides us with all updates, not an estimate of the tree topology. In particular, they provide estimates of the genetic order and the content of ancestral nodes. However, it is difficult to evaluate the accuracy of the solution using parsimony: many methods give the best tree but give little or no information about whether there are other trees with good balance or near-good scores (47).

1. **Maximum Likelihood:** Felsenstein proposed a method to find maximum likelihood (ML) phylogenies of DNA or protein, and this method can be used to examine entire genomes. This is one of the best methods based on the general concept of statistical probability; In other words, the highest probability tree is the tree that gives the highest probability of the genome analysis result according to our defined criteria. Therefore, we need a way to calculate the probability of a particular evolutionary parameter, given the gene content and order of the extant species. We can then use the similarity measure to evaluate the phylogenetic signature (48).

   In principle, Felsenstein's approach can be applied to this new problem. However, the number of possible chromosomes at the nodes of the phylogenetic tree weakens this possibility. They developed a method to estimate these trees by limiting the number of states allowed for nodes. However, the accuracy of this approach has not yet been tested. A recent study (J. Dicks (unpublished) showed that taking into account the chromosomal status of internal nodes may not be important for ML tree estimation. Therefore, rapid

evolution of genome-wide trees may be possible in the future with machine learning (48). However, this increase in speed will come at the expense of information about the most likely chromosomal evolutionary pathways in the data set, which is one of the most interesting areas of focus. Dicks used sparse integration and maximum likelihood in the CHROMTREE software for one or more chromosomes (49).

2. **Method of Invariants:** Researchers offered a unique approach in the context of DNA sequence studies and was adapted by Sankoff and Blanchette for genetic studies (48). This approach has the advantage of not relying on long-term assumptions and therefore variable costs.

   The process is done by defining a function of the genetic order of each leaf, which is not equal to the real tree (and whose probability is zero), but will not be zero in other trees. The tree can then check the value of the parameter function, which can be measured using the specified data. To explain this scientist proposed a stochastic model of genetic evolution similar to the Jukes-Cantor model of DNA mutation. The model is simple and does not react to the type of evolutionary events that lead to genetic changes (49).

   Sankoff and Blanchet note that their hopes for a good prediction of function are not consistent, which would require large data (relative to genome size). Simulation studies show that the method performs poorly for small genomes, but performance increases rapidly as genome length increases. This approach is thought to be particularly interesting when larger genomes (possibly eukaryotic nuclear genomes) are being processed because the computational problem does not increase as the length of the genome under analysis increases (48,49).

3. **Bayesian Analysis:** Another method based on the Bayesian inference model to solve the genome-wide phylogeny problem was proposed (45). In the general Bayesian framework, each measurement, observable or unobservable, is considered a random parameter with an associated probability. If the ratio associated with each parameter is known (or can be calculated), then it is just mathematical information to give the distribution of the parameter we need, given the values of the parameters. This is done using Bayes theorem (46).

   With an appropriate model of how the genome is analysed based on the unknown (and some prior classification for each unknown variant of interest), we can explore the classification of unknown abnormalities based on the analysis of each genome (47).

   But tree topology is not a simple code, and neither is the genome itself. Therefore, the probability distribution can become very complex and require Monte Carlo simulations to solve. Presented a simple case based on gene ordering and used this method to find the relevance of phylogenies to qualitative data (48).

4. **Statistical Analysis:** Phylogenetic analysis of organisms is not fair in all cases, but many tests and methods have been prepared and used to test trees formed by molecular sequences. Many of these methods can be used to study genomic trees if there is a way to transfer new data from old data. If we have a set of independently evolving sources (as in molecular arrays), nonparametric bootstrapping will work; where it is taken as basis and

replaced with the original data to create a new data set. This approach can be easily adapted to any type of data that can be expressed in a suitable format (see Binary Coding section above). But it is difficult to see how gene order could be changed (49).

The pocket knife method can be used to combine genes. Here we sample part of the text without modification to get a new set of changes, even though it is clear that the recycled data is not the same size as the original data. The confidence level provided by SHOT package [45] uses this model to sample three-quarters of orthologous families to generate new data (50).

If we have a sample of what the data looks like, entire gene sequences can be resampled using parameter bootstrapping, which is an option for resampling any dataset. We can create new data using trees estimated from old data, but these should be approached with caution, as using the predictions and models we show new errors versus using standard (non-parametric) bootstrap systems This error may not occur (51).

5. **Phylogenetic Studies:** All the analyses we have discussed so far assume that phylogenetic trees adequately represent evolutionary history. This is not necessary, especially when dealing with whole genomes, especially prokaryotic genomes. In this case the most general acyclic network should be used to represent the transition (51).

Now, segmentation plot [46] represents phylogenetic inconsistency by showing that the same data sets support potentially incompatible species groupings. These can also be used to identify common occurrences in individual molecular sequences and can be used to represent areas where the phylogenetic tree is inaccurate (50).

Legendre and Makarenov developed a method to add more edges to adjacent joins to create a “mesh graph” and thus improve the fit of the distance matrix. They are used to describe many different things. Makarenov developed the T-Rex software to create network graphs from distance matrices (51).

For further information, a special issue of the Journal of Taxonomy is devoted to the topic of evolutionary relationships between representations of networks, although often not in the context of whole genome evolution (52,53).

VI. FUTURE ASPECTS

Various experiments have been proved that use genome-level data for the analysis of evolutionary relationship by using different methods as discussed above. Comparisons of gene expression are becoming more common and are being done across many different databases. Comparison of genes raises broader questions, and although techniques for gene sequence analysis are available, their benefits has far away been restricted to small size genomes Binary Coding, usually bacterial or endosymbiotic genomes.

Most of the above methods require the genome to be encoded as a set of known genes, but all known genomes are now exceptions to this rule (especially in the case of eukaryotic nuclear genomes). Perhaps the most important thing is to focus on the information we currently have or will have in the future. These will be in various states of completion,

ranging from whole genome alignments (mostly bacterial, prokaryotic, or endosymbiotic genomes, although some may include eukaryotic nuclear genomes) to a section or map of various species.

Gene expression and analysis of gene expression rely on the success of identifying orthologous genes in most taxa and large evolutionary gaps that have been confounded by the evolution of genes. Identification of orthologs is usually done by performing all-to-all BLAST or other similar searches and grouping them together by cluster analysis based on the results. This method is not error-prone, so it is important that genome-wide phylogenetic analysis methods are robust to missing or inaccurate data. They also need to be robust to the use of inappropriate models because the mechanisms of genome evolution are not understood. Researchers addressed this issue during GRAPPA testing.

Further studies are needed to evaluate the value of phylogenetic information obtained from genome-scale comparisons. More testing is needed to determine the best way to remove it. Additionally, to increase the likelihood that phylogenetic measures will be useful, we must conduct more rigorous analysis to determine their accuracy and reliability.

REFERENCES

[1] Felsenstein, J. (1982), ‘Numerical methods for inferring evolutionary trees’, Q. Rev. Biol., Vol. 57(4), pp. 379–404.
[2] Suyama, M. and Bork, P. (2001), ‘Evolution of prokaryotic gene order: genome rearrangements in closely related species’, Trends Genet., Vol. 17(1), pp. 10–13.
[3] Wolf, Y. I., Rogozin, I. B., Grishin, N. V. and Koonin, E. V. (2002), ‘Genome trees and the tree of life’, Trends Genet., Vol. 18(9), pp. 472–479.
[4] Wolf, Y. I., Rogozin, I. B., Grishin, N. V. et al. (2001), ‘Genome trees constructed using five different approaches suggest new major bacterial clades’, BMC Evol. Biol., Vol. 1(:) 8.
[5] Gu, Z., Steinmetz, L. M., Gu, X. et al. (2003), ‘Role of duplicate genes in genetic robustness against null mutations’, Nature, Vol. 421, pp. 63–66.
[6] Huynen, M. A., Snel, B. and Bork, P. (2001), ‘Inversions and the dynamics of eukaryotic gene order’, Trends Genet., Vol. 17(6), pp. 304–306.
[7] Seoighe, C., Federspiel, N., Jones, T. et al. (2000), ‘Prevalence of small inversions in yeast gene order evolution’, Proc. Natl Acad. Sci. USA, Vol. 97, pp. 14433–14437.
[8] Fischer, G., Neuveglise, C., Durrens, P. et al. (2001), ‘Evolution of gene order in the genomes of two related yeast species’, Genome Res., Vol. 11(12), pp. 2009–2019.
[9] Wong, S., Butler, G. and Wolfe, K. H. (2002), ‘Gene order evolution and paleopolyploidy in hemiascomyte yeasts’, Proc. Natl Acad. Sci. USA, Vol. 99(14), pp. 9272–9277.
[10] Pevzner, P. and Tesler, G. (2003), ‘Genome rearrangements in mammalian evolution: lessons from human and mouse genomes’, Genome Res., Vol. 13(1), pp. 37–45. 11. URL: http://www.hgmp.mrc.ac.uk/ GenomeWeb/organelle-gen-db.html
[11] Sankoff, D., Leduc, G., Antonie, N. et al. (1992), ‘Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome’, Proc. Natl Acad. Sci. USA, Vol. 89, pp. 6575–6579.
[12] Cosner, M., Jansen, R., Moret, B. et al. (2000), ‘An empirical comparison of phylogenetic methods on chloroplast gene order data in campanulaceae’, in Sankoff, D. and Nadeau, J., Eds, ‘Comparative Genomics’, Kluwer Academic Publishers, Dordrecht, pp. 99–121.

[13] Borque, G. and Pevzner, P. A. (2002), 'Genome-scale evolution: Reconstructing gene orders in the ancestral species', Genome Res., Vol. 12, pp. 26–36.
[14] Nadeau, J. and Taylor, B. (2984), 'Lengths of chromosomal segments conserved since divergence of man and mouse', Proc. Natl Acad. Sci. USA, Vol. 81, pp. 814–818.
[15] Page, R. D. M. and Holmes, E. C. (1998), 'Molecular Evolution: A Phylogenetic Approach', Blackwell Science, Oxford.
[16] Herniou, E. A., Luque, T., Chen, X. et al. (2001), 'Use of whole genome sequence data to infer baculovirus phylogeny', J. Virol., Vol. 75, pp. 8117–8126.
[17] Gallut, C., Barriel, V. and Vignes, R. (2000), 'Gene order and phylogenetic information', in Sankoff, D. and Nadeau, J., Eds, 'Comparative Genomics', Kluwer Academic Publishers, Dordrecht, pp. 123–132.
[18] Bryant, D. (2000), 'A lower bound for the breakpoint phylogeny problem', in 'Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching', Springer, pp. 234–247.
[19] Bryant, D., 'GOTree reference manual'(URL: http://www.math.mcgill.ca/bryant/ GoTree/).
[20] Maddison, D. R., Swofford, D. L. and Maddison, W. P. (1997), 'NEXUS: An extensible file format for systematic information', System. Biol., Vol. 46(4), pp. 590–621.
[21] Swofford, D., 'PAUP – Phylogenetic Analysis Using Parsimony' (URL: http:// paup.csit.fsu.edu/).
[22] Bader, D. A., Moret, B. M. E. and Yan, M. (2001), 'A linear-time algorithm for computing inversion distance between signed permutations with an experimental study', J. Comput. Biol., Vol. 8(5), pp. 483–491.
[23] Sankoff, D., Deneault, M., Bryant, D. et al. (2000), 'Chloroplast gene orders and the divergence of plants and algae, from the normalized number of induced breakpoints', in Sankoff, D. and Nadeau, J. 'Comparative Genomics', Kluwer Academic Publishers, Dordrecht, pp. 89–98.
[24] Bryant, D. (2000), 'The complexity of calculating exemplar distances', in Sankoff, D. and Nadeau, J. 'Comparative Genomics', Kluwer Academic Publishers, Dordrecht, pp. 207–212.
[25] Snel, B., Bork, P. and Huynen, M. A. (1999), 'Genome phylogeny based on gene content', Nature Genet., Vol. 21, pp. 108–110.
[26] Caprara, A. and Lancia, G. (2000), 'Experimental and statistical analysis of sorting by reversals', in Sankoff, D. and Nadeau, J. 'Comparative Genomics', Kluwer Academic Publishers, Dordrecht, pp. 171–189.
[27] Moret, B. M. E., Wang, L.-S., Warnow, T. and Wyman, S. K. (2001), 'New approaches to reconstructing phylogenies from gene order data', Bioinformatics, Vol. 17, pp. S165–S173.
[28] Keogh, R. S., Seoighe, C. and Wolfe, K. (1998), 'Evolution of gene order and chromosome number in Saccharomyces, Kluyveromyces and related fungi', Yeast, Vol. 14, pp. 443–457.
[29] Felsenstein, J., 'PHYLIP – Phylogenetic Inference Package' (URL: http:// evolution.genetics.washington.edu/phylip/ phylip.html).
[30] Blanchette, M., 'Derange 2' (URL: http:// www.cs.washington.edu/homes/blanchette/ software.html).
[31] Tesler, G. (2002), 'GRIMM: Genome rearrangements web server', Bioinformatics, Vol. 18(3), pp. 492–493.
[32] Blanchette, M., 'BPAnalysis' (URL: http:// www.cs.washington.edu/homes/blanchette/ software.html).
[33] Moret, B. M. E., Bader, D. A., Warnow, T. and Yan, M. (2001), 'A new implementation and detailed study of breakpoint analysis', inProceedings of the 6th Pacific Symposium on Biocomputing (PSB2001)', World Scientific Pub., pp. 583–594.
[34] URL: http://www.cs.ucsd.edu/groups/ bioinformatics/MGR/
[35] Felsenstein, J. (1981), 'Evolutionary trees from DNA sequences: A maximum likelihood
[36] approach', J. Mol. Evol., Vol. 17, pp. 368–376.
[37] Dicks, J. (1999), 'Comparative mapping and phylogeny', DPhil thesis, University of Oxford.
[38] Dicks, J. (2000), 'Chromtree: Maximum likelihood estimation of chromosomal phylogenies', in Sankoff, D. and Nadeau, J. 'Comparative Genomics', Kluwer Academic Publishers, Dordrecht, pp. 333–342.
[39] Dicks, J., 'CHROMTREE' (URL: http:// jic-bioinfo.bbsrc.ac.uk/bioinformaticsresearch/software/CHROMTREE/).
[40] Lake, J. A. (1987), 'A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony', Mol. Biol. Evol., Vol. 4(2), pp. 167–199.
[41] Sankoff, D. and Blanchette, M. (1999), 'Phylogenetic invariants for genome rearrangements', J. Comput. Biol., Vol. 6, pp. 431–445.
[42] Larget, B., Simon, D. L. and Kadane, J. B. (2002), 'Bayesian phylogenetic inference from animal mitochondrial data', J. R. Statist. Soc. B, Vol. 64, pp. 1–13.

[43] Efron, B., Halloran, E. and Holmes, S. (1996), 'Bootstrap confidence levels for phylogenetic trees', Proc. Natl Acad. Sci. USA, Vol. 93, pp. 13429–13435.
[44] Felsenstein, J. (1985), 'Confidence limits onphylogenies: An approach using the bootstrap', Evolution, Vol. 39, pp. 783–791.
[45] Korbel, J. O., Snel, B., Huynen, M. A. and Bork P. (2002), 'Shot: A web server for the construction of genome phylogenies', Trends Genet., Vol. 18(3), pp. 158–162.
[46] Bandelt, H. J. and Dress, A. W. M. (1992), 'Split decomposition: A new and useful approach to phylogenetic analysis of distance data', Molecular Phylogenet. Evol., Vol. 1(3), pp. 242–252.
[47] Legendre, P. and Makarenov, V. (2002), 'Reconstruction of biogeographic and evolutionary networks using reticulograms', Systematic Biol., Vol. 51(2), pp. 199–216.
[48] Makarenov, V. (2001), 'T-rex: Reconstructing and visualizing phylogenetic trees and reticulation networks', Bioinformatics, Vol. 17(7), pp. 664–668.
[49] Legendre, P. (2000), 'Special section on reticulate evolution', J. Classification, Vol. 17, pp. 153–195.
[50] Montague, M. G. and C. A. H. III (2000), 'Gene content phylogeny of herpesviruses', Proc. Natl Acad. Sci. USA, Vol. 97, pp. 5334– 5339.
[51] Tekaia, F., Lazcano, A. and Dujon, B. (1999), 'The genomic tree as revealed from whole proteome comparisons', Genome Res., Vol. 9(6), pp. 550–557.
[52] Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. (2000), 'Biological Sequence Analysis', Cambridge University Press, Cambridge.
[53] Sankoff, D. and Nadeau, J. (Eds) (2000), 'Comparative Genomics', Kluwer Academic Publishers, Dordrecht.

