BIOINFORMATICS

Vol. 19 no. 9 2003, pages 1162–1168
DOI: 10.1093/bioinformatics/btg131

The small-world dynamics of tree networks and data mining in phyloinformatics

William H. Piel[1,*,†], Michael J. Sanderson[2] and Michael J. Donoghue[3]

[1]Institute of Evolutionary and Ecological Sciences, Kaiserstraat 63, Leiden University 2311 GP Leiden, The Netherlands, [2]Section of Evolution and Ecology, One Shields Ave, University of California, Davis, CA 95616, USA and [3]Department of Ecology and Evolutionary Biology, PO Box 208106, Yale University, New Haven, CT 06520-8106, USA

Received on May 11, 2002; revised on November 6, 2002; January 7, 2003; accepted on January 14, 2003

ABSTRACT

Motivation: A noble and ultimate objective of phyloinformatic research is to assemble, synthesize, and explore the evolutionary history of life on earth. Data mining methods for performing these tasks are not yet well developed, but one avenue of research suggests that network connectivity dynamics will play an important role in future methods. Analysis of disordered networks, such as small-world networks, has applications as diverse as disease propagation, collaborative networks, and power grids. Here we apply similar analyses to networks of phylogenetic trees in order to understand how synthetic information can emerge from a database of phylogenies.

Results: Analyses of tree network connectivity in TreeBASE show that a collection of phylogenetic trees behaves as a small-world network—while on the one hand the trees are clustered, like a non-random lattice, on the other hand they have short characteristic path lengths, like a random graph. Tree connectivities follow a dual-scale power-law distribution (first power-law exponent ≈1.87; second ≈4.82). This unusual pattern is due, in part, to the presence of alternative tree topologies that enter the database with each published study. As expected, small collections of trees decrease connectivity as new trees are added, while large collections of trees increase connectivity. However, the inflection point is surprisingly low: after about 600 trees the network suddenly jumps to a higher level of coherence. More stringent definitions of 'neighbour' greatly delay the threshold whence a database achieves sufficient maturity for a coherent network to emerge. However, more stringent definitions of 'neighbour' would also likely show improved focus in data mining.

Availability: http://treebase.org

*To whom correspondence should be addressed.
† Current address: Department of Biological Sciences, University at Buffalo, Buffalo, NY 14260, USA

Contact: wpiel@buffalo.edu

INTRODUCTION

The emerging field of phyloinformatics promises not only to synthesize and advance our understanding of evolutionary history, but also to have far-reaching effects in other areas of biology. For example, phylogenetic knowledge is increasingly used as a bridge between functional genomics, evolution, and development—unravelling how genotype becomes phenotype requires that we know phylogeny almost as well as we know genomics (Eisen, 1998; Mizuno *et al.*, 2001). Coincident with this expansion in the use of phylogenies has been an exponential growth in the number of published trees (Pagel, 1997; Sanderson *et al.*, 1993). However, methods of organizing, synthesizing, and data mining this information are non-trivial, and ideas for how to perform these tasks are still in their infancy. The inertia in this field stems from the fact that the building blocks of phylogenetics are complex: trees are discrete structures in which information is stored in the topology of hierarchically nested sets of nodes, in the distances between these nodes, and in the identities of the nodes themselves.

Nonetheless, recent progress has been made in cataloguing and synthesizing phylogenetic data. For example, a handful of databases have been developed to store phylogenetic information (e.g. CladeStore, http://palaeo.gly.bris.ac.uk/cladestore/cladestore.html; GPPRCG, http://ucjeps.herb.berkeley.edu/bryolab/greenplantpage.html; Jungle, http://smiler.lab.nig.ac.jp/jungle/jungle.html; ToL, http://www.tolweb.org; and TreeBASE, http://www.treebase.org). In addition, various supertree methods are now available for synthesizing information from a collection of independently generated trees (Sanderson *et al.*, 1998), and more research in this area is under way. What is still poorly understood is how to assemble and

data mine collections of phylogenetic trees. Accordingly, this paper investigates the dynamics of tree networks in TreeBASE so as to gain insight into methods of amassing and exploring phylogenetic information.

The phylogenetic trees in TreeBASE are stored as distinct entities and are primarily searchable by querying taxonomic names that identify either the internal nodes of a tree or the leaves of a tree (Piel *et al.*, 2001; Sanderson *et al.*, 1994). This approach does not always succeed in finding all relevant trees because seldom do the authors fully annotate all relevant internal nodes, and seldom is the taxon sampling of relevant trees sufficiently dense as to guarantee a hit by the query. To overcome this problem, TreeBASE has implemented a tool called 'tree surfing,' which finds neighbouring trees by searching on all taxa found in a starting tree (Piel *et al.*, 2001). The taxa in the set of found trees can then be used to find even more trees, etc. Each subsequent iteration produces a new set of trees that is yet another degree of separation from the starting tree. In this fashion, a neighbourhood of trees can be assembled and explored with far more success than what would result from a series of simple taxon name queries. This approach is a promising way of exploring phyloinformatic data, especially once the database has caught up with the exponential growth in published phylogenies, and hence once the sheer number of trees would otherwise overwhelm a user's ability to recover information successfully if limited to simple taxon queries.

While this concept has promise, it is unclear whether tree surfing will continue to function properly once TreeBASE has expanded to include a significantly greater number of trees. For example, will the diameter of the network (as measured by the 'characteristic path length', which is the average of the degree of separation between pairs of trees) expand with TreeBASE's growth until traversing the data is excessively laborious; or will it implode into a tight ball such that traversing the database is too instantaneous? Will TreeBASE ever achieve a fully connected network, or will there always be a certain fraction of disconnected 'satellite' networks? Recent interest in disordered networks has demonstrated how analysis of the small-world dynamics in a network can help to answer these questions (Amaral *et al.*, 2000; Barbási and Albert, 1999; Strogatz, 2001; Watts, 1999; Watts and Strogatz, 1998). Here we examine the dynamics of tree networks in TreeBASE so as to assess tree surfing as a means of mining phyloinformatic data.

ANALYSES

Small-world networks

Do networks of trees assembled from the literature behave like small-world networks? Small-world networks are defined as that subset of disordered networks having two seemingly paradoxical properties: on the one hand their connections are locally regular and non-random; on the other hand, any two vertices (i.e. nodes) can be linked through just a few edges (i.e. connections) by way of other vertices (Amaral *et al.*, 2000; Watts, 1999). The former can be estimated by using a measure of cliquishness (such as a clustering coefficient, C, which is the fraction of edges in each vertex's neighbourhood that actually exist, averaged over all vertices) to compare between the actual network and a randomly permuted network (Watts and Strogatz, 1998). The latter can be measured in several ways: one way is to ask whether the average shortest distance between all pairs of vertices increases logarithmically with the size of the network (Amaral *et al.*, 2000; Bollobás, 1985); another way is to compare the characteristic path length (L) between the actual network and a randomly permuted one (Watts and Strogatz, 1998).

Simulated island growth

We simulated the growth of TreeBASE by randomly creating 50 subsets of 490 studies, and then estimating L by averaging the minimum number of degrees of separation among 100 pairs of randomly selected trees from within the largest interconnected network. Since TreeBASE is a relational database where trees are stored in a table with a many-to-many relation with another table that stores the taxa, determining the minimum number of degrees of separation between two trees is a simple matter of executing a series of joins between the two tables. One degree of separation is counted each time a new selection of trees is created based on a selection of taxa made by a join from the previous selection of trees.

The simulated growth of TreeBASE appears to approximate a positive logarithmic relationship between the number of trees in the largest island (i.e. the largest single network of interconnected trees) and the characteristic path length of the network (Fig. 1). Since the growth of the island's diameter slows relative to the growth of the database the ease of traversing the network remains manageable, and this effect is one of the prime advantages of having a network with small-world properties. In contrast, the growth of island diameter in a regular lattice simulation (in which the number of trees and distribution of tree size are the same) continues growing linearly, while that of a randomly rewired version of the database drops off, collapsing network traversal (Fig. 1).

Cliquishness and the diameter of the network

Small-world networks, as opposed to purely random networks, are locally cliquish while maintaining a relatively small network diameter regardless of the network's size (Watts, 1999; Watts and Strogatz, 1998). To evaluate this tendency in TreeBASE, we generated a randomly

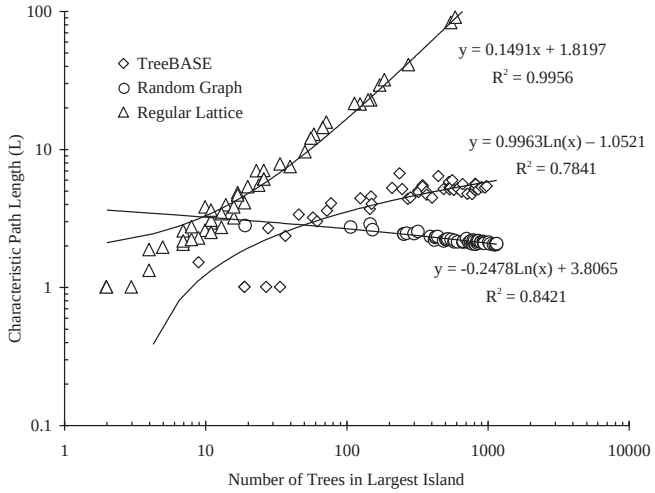

Fig. 1. Simulated growth of TreeBASE. Selecting at random and without replacement from among the 490 studies in TreeBASE created 50 databases, wherein the size of each database was also determined at random. For each replicate, the largest island of interconnected trees was measured for its characteristic path length (L), by taking the average for the minimum number of degrees of separation between 100 pairs of randomly selected trees. Using the same distribution of numbers of taxa per tree as in TreeBASE, a randomly permuted version (Random Graph) and a linearly connected version (Regular Lattice) were analyzed for comparison. The characteristic path length in TreeBASE appears to grow logarithmically with the number of vertices and is intermediate to a random graph and regular lattice as would be expected from a small-world network.

permuted version of the database so as to compare the values for the actual characteristic path length (which estimates the diameter of the network) and clustering coefficient (a measure of cliquishness) with those of a random graph. The 10 825 taxa for the 989 trees that make up the largest island in TreeBASE were permuted. In this way the actual and randomized databases had the same number of trees and each tree had the equivalent number of taxa, yet taxa in each tree of the permuted database were assigned at random. The clustering coefficient was calculated according to Watts and Strogatz (1998) as the average over all vertices (i.e. trees) of the fraction of possible edges in each vertex's neighbourhood (i.e. connections) that actually exist:

$$C = \frac{\sum_{i=1}^{N} \frac{2v_i}{k_i(k_i-1)}}{N}$$

wherein k_i is the number of neighbours of the ith vertex; v_i is the number of edges among all such neighbours; and N is the number of vertices.

The results show comparatively similar network dynamics between TreeBASE's main tree island and other well-studied systems (Table 1). Special properties in these other systems, notably in the case of the film actor's network, are thought to be affected by preferential attachment of new vertices in a growing network, extinction of old vertices, and vertex saturation (Barbási and Albert, 1999). As with these other networks, we see in TreeBASE a greater relative difference between the cliquishness of the actual and random networks than between the characteristic path lengths. The relatively high sense of cliquishness ($C_{\text{actual}} = 0.813$) presumably occurs because the same *combinations* of taxa in one tree are more likely to occur in neighbouring trees as compared to unrelated trees. This effect emerges because underlying the production of phylogenetic trees is a single, enormous tree of life. In contrast, the unconstrained rewiring of a random network has an absence of cliquishness ($C_{\text{rand}} = 0.182$).

Table 1. Characteristic path length L and clustering coefficient C for four actual networks compared to randomized versions in which the number of vertices and the average number of edges per vertex are the same. Data for the network of film actors, the power grid, and the neuronal network of *C.elegans* were published in Watts and Strogatz (1998)

Network	Vertices	L_{actual}	L_{rand}	C_{actual}	C_{rand}
Film actors	225 226	3.65	2.99	0.79	0.00027
Power grid	4 941	18.7	12.4	0.08	0.005
C.elegans	282	2.65	2.25	0.28	0.05
TreeBASE	989	5.11	2.00	0.813	0.182

The distribution function of connectivities

The dynamics of large networks can be studied by examining the distribution function of connectivities, seen as a curve representing the cumulative distribution of connections. The cumulative distribution is the running sum of the frequencies or probabilities of connecting to k neighbouring vertices, and this is plotted against the number of vertices, k (Amaral *et al.*, 2000; Barbási and Albert, 1999). Networks that maintain a stationary, scale-free power-law distribution arise as a consequence of the way in which they grow—where new vertices connect preferentially to vertices that are already more popular (Barbási and Albert, 1999). Amaral *et al.* (2000) show that broad-scale networks with a sharp cut-off to their power law regime and single-scale networks with a fast decaying tail are also encountered. The tendency of networks to take on any of these three characteristics is thought to arise from differences in certain limiting factors that affect the growth of the network.

Two important limiting factors that have been identified in other systems are: (1) the ageing of vertices, such as deceased actors in a growing network of actors; and (2) the

increased costs of adding links to popular vertices, such as the limited growth space at an airport (Amaral *et al.*, 2000). In principle, trees do not age, but the popularity of taxa that link trees could decrease with time. As might be expected, species such as *Drosophila melanogaster* and *Saccharomyces cerevisiae* are among the most popular in TreeBASE, probably because so much of their genome has already been sequenced that authors of new trees are tempted to include them as outgroups. To the extent that the genomes of other taxa will become better represented in GenBank, previously popular taxa may become less frequently used and their relative popularity will wane. This aging will cause trees to lose their popularity gradually, and we would expect a scale-free regime to shorten. Likewise, to the extent that TreeBASE's staff seeks out trees rather than relying on passive voluntary submission, it is likely that the staff would prefer to add trees that are not already represented in the database. This preference would, in effect, increase the cost-to-benefit ratio of adding trees with tight connections to pre-existing popular trees. Such increasing costs could also narrow the scale-free range, and perhaps introduce a Gaussian or exponential dip to the tail. Nonetheless, popular taxa (e.g. *D.melanogaster*, *Homo sapiens*, *S.cerevisiae* etc.) are still well represented in GenBank, and their impact on phylogenetics is likely to continue. Additionally, most submissions to TreeBASE enter passively through the voluntary efforts of authors and journal editors. Therefore, it is unclear to what degree the popularity of taxa and selectivity of staff will affect network dynamics.

We examined the distribution function of connectivities by graphing the running sum of the frequencies that trees connect to other trees on log–log and semi-log plots. Remarkably, the result is an unusual dual-scale power-law regime (Fig. 2B) that does not match any of the classes of networks documented by Amaral *et al.* (2000). Like an arm with an elbow joint, our tree network shows a sudden change in the distribution function for trees with more than 25 neighbours. The exponent of the first distribution, $\alpha - 1 \approx 0.87$, $\alpha \approx 1.87$ and the second distribution, $\alpha - 1 \approx 3.82$, $\alpha \approx 4.82$, compares with $\alpha = 2.3$ for the actor network; $\alpha = 2.1$ for the www network; and $\alpha = 4$ for the power grid data, as reported in Barbási and Albert (1999).

We surmised that the dual scale distribution could be an artefact because there are, on average, 2.6 trees per submission, and that while these trees are neighbours, as they share almost all their taxa in common, they nonetheless cannot be considered independent instances of neighbouring trees. To exclude the effect of multiple trees, we pruned the database to retain just one tree per submission and then re-analyzed the network. The result still showed a dual-scale power-law regime, but with the first log–log slope being steeper ($4.6652k^{-0.997}$) and the second one flatter ($2485k^{-2.8814}$) than in the previous analysis. While the difference between the dual-scales is less pronounced when superfluous trees are excluded, the dual-scale nature of this function is by no means eliminated, and the inflection point at about 25 neighbours remains the same.

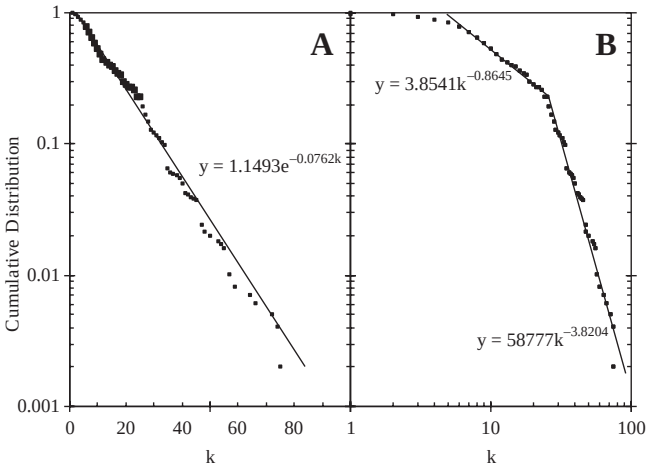

Fig. 2. The distribution function of connectivities for the main island in TreeBASE. These graphs plot the cumulative sum of the frequencies with which a tree connects to k other trees. For this island of trees, $N = 989$ vertices and average connectivity $(K) = 14.99$. (**A**) Linear-log plot more or less approximates a straight line, suggesting an exponential decay in the probability of connectivities. (**B**) Log–log plot of the same data. A scale-free distribution would show a single power-law decay and the points would approximate a straight line. For tree networks, however, there are clearly two distinct scales, each with its own power law.

The emergence of an island network

Aside from the dynamics that go on within a tree network, it is also important for us to know how the main island network is expected to grow with the growth of the database. Indeed, supertree algorithms for assembling the tree of life can only function once all phylogenetic trees are, in some way, connected within a single tree network. Yet assembling the tree of life is like putting together an enormous puzzle using millions of small puzzle-pieces. How many puzzle-pieces are needed before all pieces can be linked, or before a skeleton of the whole picture starts to emerge? In other words, when are trees most disconnected and when will they be fully connected? How do the criteria for 'neighbour' affect the emergence of a single island network?

We examined these issues by creating eight databases of different sizes using the respective trees of randomly selected studies from TreeBASE. The neighbourhood network for each tree of a given database was followed

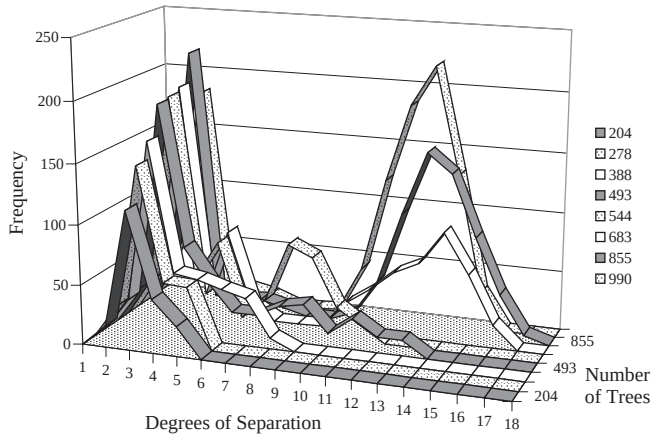

Fig. 3. The formation of a main island network of interconnected trees. The distribution of the minimum degrees of separation between every tree and the farthest tree in its island is presented for each of eight randomly selected collections of trees. A main super-island network of trees is fully separated from the remaining satellite islands with a database of 500 trees. A database of about 850 trees has a main island network that is most diffuse, with some trees distant by up to 17 degrees of separation. By about 1000 trees, the main island network is tightening, and new trees are more likely to join the main island than they are to form independent satellite islets.

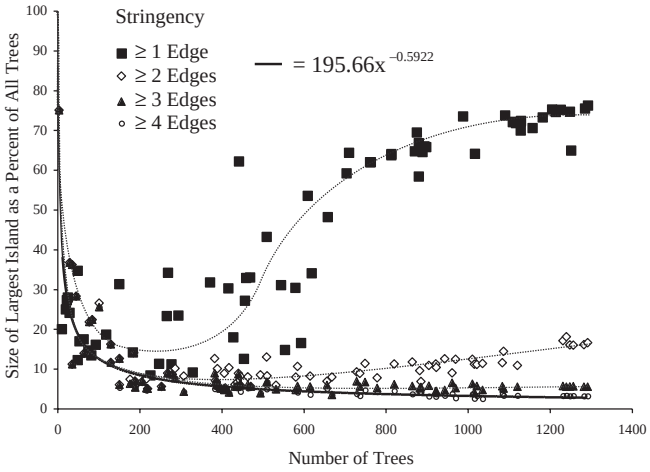

Fig. 4. Largest island of trees as a percentage of all trees for simulated databases of different sizes. The largest island size was calculated for each randomly generated database and plotted as a percentage of all trees in the database. The same was done using more stringent definitions of neighbour, from two or more connecting taxa to four or more connecting taxa. A power regression line ($R^2 = 0.85$) is shown for those databases with a stringency of four or more taxa. The dashed lines are drawn by hand to illustrate the trends in the data.

until a final degree of separation was reached when no more trees could be found by any subsequent iteration. The distribution of this maximum chain of interconnected trees is shown for each of the eight databases (Fig. 3). In all databases, we see evidence for numerous, small disconnected islets, where each is traversable in four or fewer degrees of separation. As the database grows in size, the number of islets continues to grow, but in addition several large islands start to emerge. Between 544 and 683 trees, the major islands have grouped into a single main island that takes as much as 16 degrees of separation to traverse (Fig. 3). At about 850 trees, the main island has reached its most thinly interconnected form, with some trees requiring 17 degrees of separation to cover the entire network. Subsequently, at 990 trees, it would appear that new trees are more likely to join with the main island than they are to from lonely, disconnected islets. This pattern causes the main island to tighten, with the most distant trees being separated by no more than 15 degrees. The database would appear to show that there is a critical point when a skeleton tree suddenly synthesizes with remarkably few trees.

We examined the growth of the main island more closely as a percentage relative to all trees in the database (solid squares in Fig. 4). Initially, a database with only one tree would, of course, have a main island that represents 100% of the database. But as trees are added, the largest island as a percentage of all trees reaches a minimum at about 250 trees (Fig. 4). Between 500 and 700 trees, we see a sudden inflection point in the curve, where the largest island grows from about 25% to over 60% of all trees in the database. Subsequent to this, the growth of the curve begins to slow, and by 1200 trees the main island holds about 80% of the database. Presumably the curve will eventually reach 100%—assuming that all trees are connectable since all life is related. However, realistically there will always be some trees that fail to connect with other trees because of various oddities, such as peculiar or unusual spelling of taxon names. Therefore, although this graph (Fig. 4) does not allow us to predict the number of trees required to achieve a complete network, it does demonstrate some interesting critical values. Specifically, a database of 250 trees is at its least connected state; and a database should have more than 600 trees in order to cross a critical threshold in tree connectivity. This behaviour is equivalent to the sudden coalescence among components of a random graph (Erdös and Rényi, 1960).

As is evident in the graph (Fig. 4), increasing the stringency of the definition of neighbour greatly affects the growth of the main island. Tree surfing with a stringency of level two indicates that at least two taxa must be in common between two trees for them to be considered neighbours; level three means at least tree or more taxa,

and so forth. For a stringency of level two, although the largest island as a proportion of all trees is steadily increasing, it is still far behind the growth of the curve for stringency level one, and it does not yet appear to have passed its own inflection point. In fact, the largest island has only 214 trees in a database of 1292. Under such conditions it appears that it will take a database of more than double the size as is required for level one before a fully interconnected network is achieved. Tree surfing with a higher stringency has the advantage of largely avoiding spurious connections that result from trees with unusually distant outgroups. We can imagine that for certain data mining purposes, higher stringency levels would be desirable despite the fact that these levels demands that the database be better populated.

DISCUSSION

Analysis of network dynamics is one approach to understanding the behaviour of complex systems in biology. This method is increasingly being used in bioinformatics, such as in improving models of disease propagation (Liljeros *et al.*, 2001; Newman, 2002), evaluating protein and DNA networks (Fraser *et al.*, 2002; Jeong *et al.*, 2001; Rzhetsky and Gomez, 2001), and data mining gene chip expression patterns (del Rio *et al.*, 2001) Recently much attention is specifically applied to the behaviour of growing networks (Klemm and Eguíluz, 2002; Xulvi-Brunet and Sokolov, 2002). Similarly, phyloinformatic databases store a growing network of complex and unusual data types that can benefit from novel approaches to data mining, such as network analysis. In TreeBASE the 'tree surfing,' function uses iterative joins as a means of traversing and exploring the data. The network dynamics of an interconnected island of trees is critical to the performance of this method. By running Monte Carlo simulations with tree networks and by closely examining of the data in TreeBASE, we have shed some light on the nature of these tree networks.

Central to the success of tree surfing is the fact that networks of trees demonstrate small-world dynamics (Fig. 1 and Table 1). Despite the apparent non-random aspect to published phylogenies—collectively, networks of trees allow the data miner to traverse the entire database in short order regardless of the size of the database. The implication of this result is that even if phylogenies for all the world's taxa were represented in the database, a method for assembling a generalized picture for the complete tree of life is still manageable if it is based on a tree surfing approach.

The distribution function of connectivities can tell us about factors that influence the nature and growth of networks (Amaral *et al.*, 2000). In TreeBASE, the function of connectivities shows a hitherto undocumented distribution type in which a distinct dual-scale is evident (Fig. 2). We demonstrated that this dual-scale is in part due to the artefactual presence of numerous alternative tree topologies submitted by authors. Removing this effect causes the angle between the two power regression lines to increase, but not sufficiently to eliminate their dual nature. It is unclear what other factors could explain their continued presence; possibly the inclusion of distant outgroups in certain trees may be a contributing cause.

The manner with which a phyloinformatic network matures has some notable characteristic features. If we assume that TreeBASE is a fair reflection of a reasonably random collection of trees taken from the literature, we can draw some conclusions about tree networks in general. When a database reaches about 500 trees, there begins a process of rapid fusion in the network accompanied by a clear and distinct separation between a main island of trees and the remaining disconnected islets (Fig. 3). By about 900 trees, this island network has reached its most diffuse state—after that the island diameter shrinks as additional trees provide alternative pathways that shorten minimum distances within the network. These results suggest that phyloinformatic databases should store at least 1000 trees before there is a reasonable skeleton of the tree of life with which to build on. Specifically, a database of 250 trees is at its least coherent and most disconnected state, while an abrupt inflection point in the maturation of the network occurs at about 600 trees (Fig. 4).

However, the assumption that TreeBASE contains a fair sampling of trees is tenuous, and clearly our observations are more germane in terms of how biologists unearth phylogenetic knowledge than they are in terms of how nature herself evolves. A disproportionate number of trees were initially taken from plants, and more recently the databases have shifted its emphasis toward fungi. The database has not yet included some very large trees that have recently been published. Moreover, even if it were a fair sampling of trees, the biologists publishing these trees undoubtedly are biased in how they sample the earth's organisms. At best, TreeBASE is only a sampling of the effective or working number of species in the world, not the true number—since only a subset of the true number is available for biologists to study.

The stringency of our definition of neighbour has a profound impact on our ability to traverse a network and on the ability of a network to form in the first place. Under the minimum stringency of one taxon connecting two trees, a database of 1300 trees has an island representing over 80% of trees, compared with 20 and 5% for stringencies of two and three taxa respectively (Fig. 4). However, ultimately a more stringent definition will likely produce better results for use with supertree algorithms and synthetic methods for inferring the tree of life. For example, matrix representation with parsimony

requires an overlap of at least two taxa among all trees.

Our vision for the future of data mining in phyloinformatics is the following: imagine a diffuse three-dimensional shape of points representing individual trees in which the arrangement of these points in space is determined by tree-to-tree distances, as measured by their degrees of separation under various levels of stringency. By manipulating this cloud of points in virtual space, the biologist can explore and select collections of trees that focus on certain segments of the tree of life. These collections of trees could then be subjected to various synthetic tools, such as supertree algorithms, and from that a coherent pattern of organic evolution would emerge. Methods specifically having to do with how tree-to-tree distances are estimated would stem from what we have learned about the network dynamics of trees, as reported herein.

ACKNOWLEDGEMENTS

We are indebted to J.A.J.Metz for his thoughts on the subject and his comments on the manuscript.

REFERENCES

Amaral,L.A.N., Scala,A., Barthélémey,M. and Stanley,H.E. (2000) Classes of small-world networks. *Proc. Natl Acad. Sci. USA*, **97**, 11 149–11 152.

Barbási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Bollobás,B. (1985) *Random Graphs*. Academic Press, London.

del Rio,G., Bartley,T.F., del-Rio,H., Rao,R., Jin,K-L, Greenberg,D.A., Eshoo,M. and Bredesen,D.E. (2001) Mining DNA microarray data using a novel approach based on graph theory. *FEBS Lett.*, **509**, 230–234.

Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.*, **8**, 163–167.

Erdös,P. and Rényi,A. (1960) On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, **5**, 17–61.

Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C. and Feldman,M.W. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.

Jeong,H., Mason,S.P., Barbási,A.-L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Klemm,K. and Eguíluz,V.M. (2002) Growing scale-free networks with small-world behavior. *Phys. Rev. E*, **65**, 057102.

Liljeros,F., Edling,C.R., Amaral,L.A.N., Stanley,H.E. and Åberg,Y. (2001) The web of human sexual contacts. *Nature*, **411**, 907–908.

Mizuno,H., Tanaka,Y., Nakai,K. and Sarai,A. (2001) ORIGENE: gene classification based on the evolutionary tree. *Bioinformatics*, **17**, 167–173.

Newman,M.E.J. (2002) Spread of epidemic disease on networks. *Phys. Rev. E*, **66**, 016128.

Pagel,M. (1997) Inferring evolutionary processes from phylogenies. *Zoologica Scripta*, **26**, 331–348.

Piel,W.H., Donoghue,M.J. and Sanderson,M.J. (2001) TreeBASE: a database of phylogenetic information. *2nd International Workshop of Species 2000*. National Institute for Environmental Studies, Tsukuba, Japan.

Rzhetsky,A. and Gomez,S.M. (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics*, **17**, 988–996.

Sanderson,M.J., Baldwin,B.G., Bharathan,G., Campbell,C.S., Von Dohlen,C., Ferguson,D., Porter,J.M., Wojciechowski,M.F. and Donoghue,M.J. (1993) The rate of growth of phylogenetic information, and the need for a phylogenetic database. *Syst. Biol.*, **42**, 562–568.

Sanderson,M.J., Donoghue,M.J., Piel,W.H. and Eriksson,T. (1994) TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am. J. Bot.*, **81**, 183.

Sanderson,M.J., Purvis,A. and Henze,C. (1998) Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution*, **13**, 105–109.

Strogatz,S.H. (2001) Exploring complex networks. *Nature*, **410**, 268–276.

Watts,D.J. (1999) *Small worlds: The dynamics of networks between order and randomness*. Princeton University Press, Princeton, NJ.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **394**, 440–442.

Xulvi-Brunet,R. and Sokolov,I.M. (2002) Evolving networks with disadvantageous long-range connections. *Phys. Rev. E*, **66**, 026118.