BIOINFORMATICS

Vol. 17 no. 2 2001
Pages 149–154

An information-based sequence distance and its application to whole mitochondrial genome phylogeny

Ming Li[1,*], *Jonathan H. Badger*[1], *Xin Chen*[2], *Sam Kwong*[2], *Paul Kearney*[1] *and Haoyong Zhang*[1]

[1]*Bioinformatics Laboratory, Computer Science Department, University of Waterloo, N2L 3G1, Canada and* [2]*Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

Received on July 19, 2000; revised on October 5, 2000; accepted on October 11, 2000

ABSTRACT

Motivation: Traditional sequence distances require an alignment and therefore are not directly applicable to the problem of whole genome phylogeny where events such as rearrangements make full length alignments impossible. We present a sequence distance that works on unaligned sequences using the information theoretical concept of Kolmogorov complexity and a program to estimate this distance.

Results: We establish the mathematical foundations of our distance and illustrate its use by constructing a phylogeny of the Eutherian orders using complete unaligned mitochondrial genomes. This phylogeny is consistent with the commonly accepted one for the Eutherians. A second, larger mammalian dataset is also analyzed, yielding a phylogeny generally consistent with the commonly accepted one for the mammals.

Availability: The program to estimate our sequence distance, is available at http://www.cs.cityu.edu.hk/~cssamk/gencomp/GenCompress1.htm. The distance matrices used to generate our phylogenies are available at http://www.math.uwaterloo.ca/~mli/distance.html

Contact: mli@wh.math.uwaterloo.ca

INTRODUCTION

The fast advance of worldwide genome sequencing projects has raised a fundamental and challenging question to modern biological science: how do we compare two genomes? In fact, it earned the first position in two recent lists of major open problems in bioinformatics (Koonin, 1999; Wooley, 1999). Existing tools and methods such as multiple alignment and various sequence evolutionary models do not directly apply to complete genomes where such events as rearrangements make traditional full length alignments impossible, and no statistical models currently exist for the evolution of complete genomes.

In the absence of such models, a method which can compute the shared information between two sequences is useful because biological sequences encode information, and the occurrence of evolutionary events (such as insertions, deletions, point mutations, rearrangements, and inversions) separating two sequences sharing a common ancestor will result in the loss of their shared information. Regions of sequences which do not share a common ancestor will not share more information than would be expected at random. Here we present a mathematically rigorous universal distance based on shared algorithmic information; a fully automated and accurate software tool based on such distance to compare two genomes, or two English texts for that matter; and we demonstrate that whole mitochondrial genome phylogenies can be reconstructed automatically from *unaligned* complete mitochondrial genomes by our software.

METHODS

Definition of distance

We begin by defining what we mean by a distance. Without loss of generality, a distance only needs to operate on sequences of 0s and 1s since any sequence can be represented by a binary sequence. We also only consider normalized distance functions d such that $0 \leqslant d(x, y) \leqslant 1$ for all sequences x and y. For a function d to be a 'distance', it must satisfy (a) $d(x, y) > 0$ for $x \neq y$; (b) $d(x, x) = 0$; (c) $d(x, y) = d(y, x)$ (symmetric); and (d) $d(x, y) \leqslant d(x, z) + d(z, y)$ (triangle inequality).

Given two sequences x and y, our new distance $d(x, y)$ is defined as follows

$$d(x, y) = 1 - \frac{K(x) - K(x|y)}{K(xy)}, \qquad (1)$$

*To whom correspondence should be addressed.

where $K(x|y)$ is the conditional Kolmogorov complexity (or algorithmic entropy) of x given y. $K(x|y)$ is defined as the length of the shortest program causing a standard universal computer to output x on input y, and $K(x)$ is defined as $K(x|\epsilon)$, where ϵ is the empty string. See Li and Vitányi (1997) for formal definitions of Kolmogorov complexity and its successful applications in physics, mathematics, and computer science. $K(x|y)$ measures the randomness of x given y. The numerator $K(x) - K(x|y)$ is the amount of information y knows about x. It is a deep theorem in Kolmogorov complexity that $K(x) - K(x|y) \approx K(y) - K(y|x)$ (Li and Vitányi, 1997). That is, the information y knows about x is equal to the information x knows about y, for all x and y. This is called mutual algorithmic information between x and y. The denominator $K(xy)$, being the amount of information in the string x concatenated with y, serves as a normalization factor such that the distance $d(x, y)$ ranges between 0 (when y 'knows' all about x) and 1 (when y 'knows' nothing about x). However, mutual algorithmic information is not itself a distance and it does not satisfy the triangle inequality. Clearly, d satisfies distance conditions (a), (b), (c). It is not obvious that it also satisfies (d). The following theorem answers this. All (in) equalities are modulo an additive $O(\log n)$ term.

THEOREM 1. *$d(x, y)$ satisfies the triangle inequality, that is, $d(x, z) \leqslant d(x, y) + d(y, z)$.*

PROOF. We need to show:

$$1 - \frac{K(x) - K(x|z)}{K(xz)} \leqslant 1 - \frac{K(x) - K(x|y)}{K(xy)} + 1 - \frac{K(y) - K(y|z)}{K(yz)}.$$

This is equivalent to, by the Symmetry of Information theorem (Li and Vitányi, 1997),

$$\frac{K(z|x) + K(x|z)}{K(xz)} \leqslant \frac{K(y|x) + K(x|y)}{K(xy)} + \frac{K(z|y) + K(y|z)}{K(yz)}.$$

It is sufficient to prove the following two inequalities:

$$\frac{K(x|z)}{K(xz)} \leqslant \frac{K(x|y)}{K(xy)} + \frac{K(y|z)}{K(yz)}$$
$$\frac{K(z|x)}{K(xz)} \leqslant \frac{K(y|x)}{K(xy)} + \frac{K(z|y)}{K(yz)}.$$

The two inequalities are symmetric. To prove the first, let $r = K(x|z)$, $q = K(x|y)$, $p = K(y|z)$. Obviously $r \leqslant p+q$, by elementary facts in Kolmogorov complexity (Li and Vitányi, 1997). Let $r = p + q - \Delta$. Then,

$$\begin{aligned}\frac{K(x|z)}{K(xz)} &\leqslant \frac{r}{K(z) + r} \\ &\leqslant \frac{p + q - \Delta}{K(z) + p + q - \Delta} \\ &\leqslant \frac{p + q}{K(z) + p + q} \\ &\leqslant \frac{q}{K(z) + p + q} + \frac{p}{K(z) + p + q} \\ &\leqslant \frac{q}{K(xy)} + \frac{p}{K(yz)} \\ &= \frac{K(x|y)}{K(xy)} + \frac{K(y|z)}{K(yz)}.\end{aligned}$$

This proves the first inequality. The second is proved symmetrically. □

While it is true that when the quantities of Kolmogorov complexity terms in the above proof are extremely small (non-interesting case), then the above proof needs to be more carefully formulated (Vitányi,P. personal communication). But to maintain clarity, we have chosen not to do that, although a similar proof still works.

UNIVERSALITY

Now, consider any computable distance D. In order to exclude degenerate distances such as $D(x, y) = 1/2$ for all sequences x and y, we limit the number of sequences in a neighborhood of size d. Let us require for each x,

$$|\{y : |y| = n \text{ and } D(x, y) \leqslant d\}| \leqslant 2^{dn}. \qquad (2)$$

Assuming equation (2), we prove the following theorem.

THEOREM 2. *For any computable distance D, there is a constant $c < 2$ such that, with probability 1, for all sequences x and y, $d(x, y) \leqslant cD(x, y)$.*

PROOF. By definition 1 and the Symmetry of Information theorem, as in Theorem 1, we know that, up to $O(\log n)$ factor,

$$d(x, y) = \frac{K(x|y) + K(y|x)}{K(xy)}. \qquad (3)$$

Given D, using the density property in formula 2 and the computable function D, we know that $K(x|y) \leqslant D(x, y)n$, and $K(y|x) \leqslant D(x, y)n$. With probability 1, $K(xy) > n$, noting $|x| = n$ or $|y| = n$. Thus, equation 3 gives

$$d(x, y) = \frac{K(x|y) + K(y|x)}{K(xy)} \leqslant \frac{2D(x, y)n}{n} = 2D(x, y),$$

with probability 1. This proves the theorem with $c \leqslant 2$. □

This demonstrates mathematically the following: if x and y are 'close' according to distance measure D, then they are also 'close' according to our new distance d. In other words, if D reveals some similarity between x and y, so does d.

REMARK 1. *Our distance $d(x, y)$ naturally takes care of some tricky problems which plague other stochastic measures. For example, some methods might tend to cluster all sequences with high $G + C$ content together. Since $d(x, y)$ only measures the shared information, the base composition bias gets naturally self canceled in the $K(x) - K(x|y)$ calculation (in our approximation, which will be described in the next section, this is done by arithmetic encoding). Therefore, unlike many stochastic measures, our distance does not require that the sequences obey the stationarity condition. This may also be viewed as a consequence of the universality statement.*

RESULTS

Whole genome phylogeny

With many genomes already sequenced, imminent completion of the human genome and the completion of many other sequencing projects on the horizon, whole genome analysis and especially pairwise genome comparison is a great challenge in genomics (Koonin, 1999; Wooley, 1999). Already there have been proposals to compare genomes using gene order (Boore and Brown, 1998) and gene content (Fitz-Gibbon and House, 1999; Snel *et al.*, 1999). Such comparisons are time consuming as they require gene identification. These distances, together with G + C content; edit distance; and reversal and rearrangement distances (Kececioglu and Sankoff, 1995; Hannenhalli and Pevzner, 1995; Nadeau and Sankoff, 1998) compare genomes using only partial genome information whereas our new distance uses all genome information. The transformation distance (Varre *et al.*, 1998) and compression distance (Grumbach and Tahi, 1994) are essentially defined as $K(x|y)$ which is badly asymmetric, and so, is not a distance. Only when all sequences are random, which is not the case for DNA sequences do these two distances coincide with our new distance. In fact, mathematically, all above distances can be formulated as special cases of our new distance.

Kolmogorov complexity can be thought of as the ultimate lower bound of all measures of information and cannot be computed in the general case (Li and Vitányi, 1997). Therefore, our new distance must be approximated.

For this purpose, we use the program *GenCompress* (Chen *et al.*, 2000) which is currently the best compression program for DNA sequences, achieving the best compression ratios for benchmark DNA sequences, to heuristically approximate $K(x|y)$, $K(x)$, and $K(xy)$. *GenCompress* finds approximate matches (hence edit distance becomes a special case), approximate reverse complements, among other things, with arithmetic encoding when necessary. *GenCompress* was implemented by Chen *et al.* (2000) and can be downloaded from our website. It is capable of compressing in a day the complete, 34 megabase human chromosome 22, (achieving 12% compression).

To test our theory, we have performed several experiments on the construction of whole genome phylogenies using our new distance, all with very encouraging results. Two such experiments follow.

Phylogeny of Eutherian orders

It has been debated which two of the three main groups of placental mammals are more closely related: Primates, Ferungulates, and Rodents. This is because by the maximum likelihood method, some proteins support the (Ferungulates, (Primates, Rodents)) grouping while other proteins support the (Rodents, (Ferungulates, Primates)) grouping (Cao *et al.*, 1998). Cao *et al.* aligned 12 concatenated mitochondrial proteins from the following species: rat (*Rattus norvegicus*), house mouse (*Mus musculus*), grey seal (*Halichoerus grypus*), harbor seal (*Phoca vitulina*), cat (*Felis catus*), white rhino (*Ceratotherium simum*), horse (*Equus caballus*), finback whale (*Balaenoptera physalus*), blue whale (*Balaenoptera musculus*), cow (*Bos taurus*), gibbon (*Hylobates lar*), gorilla (*Gorilla gorilla*), human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), pygmy chimpanzee (*Pan paniscus*), orangutan (*Pongo pygmaeus*), Sumatran orangutan (*Pongo pygmaeus abelii*), using opossum (*Didelphis virginiana*), wallaroo (*Macropus robustus*) and platypus (*Ornithorhynchus anatinus*) as the outgroup, and built the maximum likelihood tree to confirm the grouping (Rodents, (Primates, Ferungulates)). Using the complete mitochondrial genomes of these species we computed our new distance $d(x, y)$ between each pair of species x and y and constructed a tree (Figure 1) using the neighbor joining (Saitou and Nei, 1987) program in the MOLPHY package (Adachi and Hasegawa, 1996). The tree is identical to the maximum likelihood tree of Cao *et al.*

Because neighbor-joining is sometimes distrusted (Hillis *et al.*, 1994; Kuhner and Felsenstein, 1994), to further corroborate this grouping we applied our own hypercleaning program (Bryant *et al.*, 2000) to the same distance matrix and obtained the same tree. The hypercleaning program constructs an evolutionary tree using the edges best supported by all possible four taxa subtrees (commonly called 'quartets'). Thus using the new information-theoretic distances derived from the *complete* mtDNA genomes we have re-confirmed the hypothesis of (Rodents, (Primates, Ferungulates)). The distance matrix can be found at our website (see the abstract).

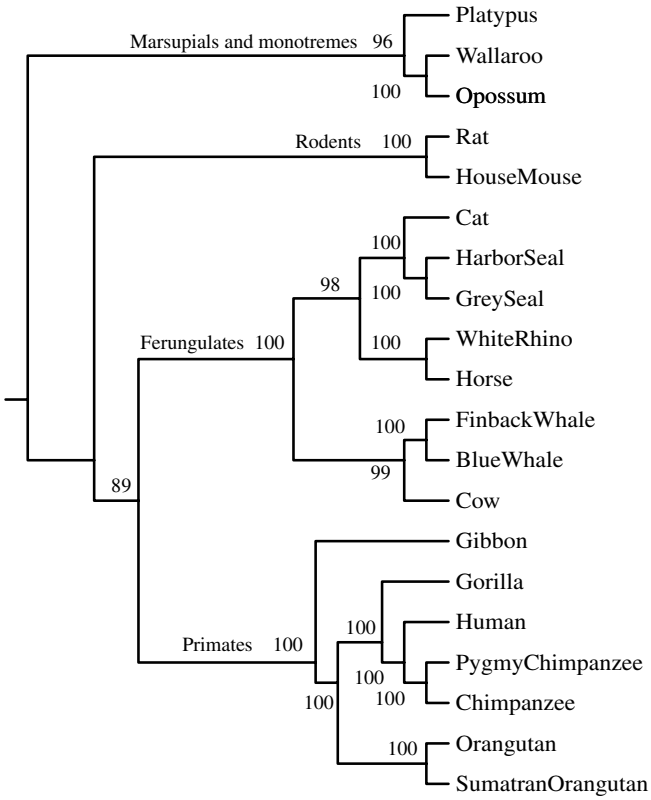

Fig. 1. The evolutionary tree built from the complete mammalian mtDNA sequences of the taxa analyzed in Cao *et al.* (1998). The numbers associated with each clade are the percentage of quartets supporting the grouping according to the hypercleaning algorithm (Bryant *et al.*, 2000), and can be interpreted in a similar fashion to bootstrap values.

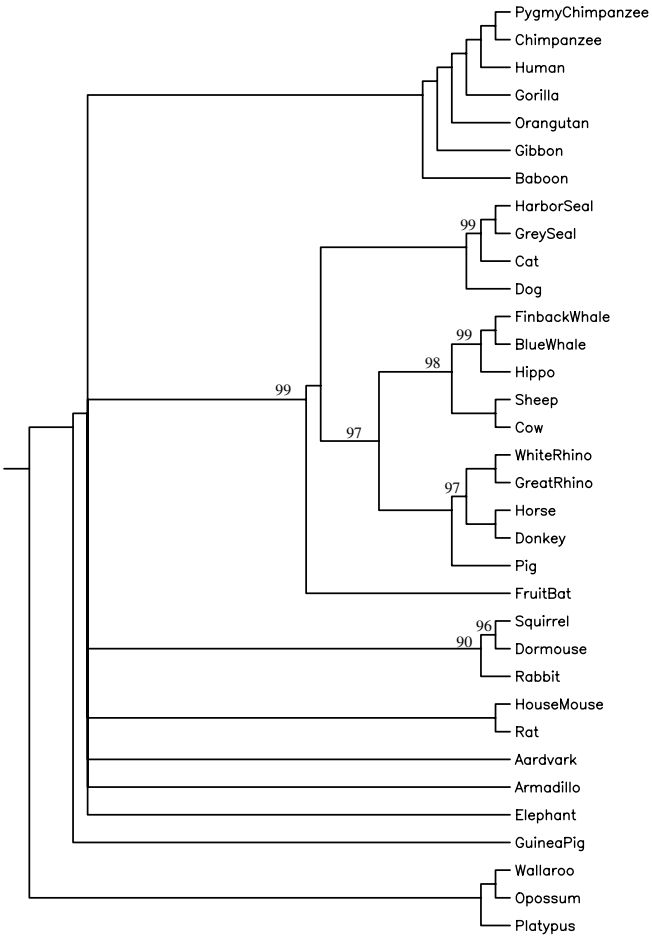

Fig. 2. The evolutionary tree built from the complete mammalian mtDNA sequences of the taxa analyzed in Reyes *et al.* (2000). The numbers associated with each clade are the percentage of quartets supporting the grouping according to the hypercleaning algorithm (Bryant *et al.*, 2000), and can be interpreted in a similar fashion to bootstrap values. Only values less than 100 are shown.

The simple asymmetric measure $K(x|y)$ leads to a wrong tree using the same data and programs, as expected (data not shown). The gene order (Boore and Brown, 1998) and gene content (Fitz-Gibbon and House, 1999; Snel *et al.*, 1999) approaches, while yielding symmetric distances, have the disadvantage of requiring laborious human analysis of the sequences and also are unlikely to provide enough information to distinguish closely related species such as the above data set. Our method is fully automatic and utilizes the information contained in noncoding regions in addition to the information contained in the genes.

To further assure our result, we have extracted the coding regions only from mtDNAs of the above species, and performed the same computation. We have obtained the same tree.

Phylogenetic position of the rodents

We also analyzed a larger dataset derived from the 34-taxa mitochondrial genome phylogeny in Reyes *et al.* (2000). This dataset included 19 of the 20 taxa in Cao *et al.* (1998) and the additional 15 taxa: aardvark (*Orycteropus afer*), armadillo (*Dasypus novemcintus*), baboon (*Papio hamadryas*), dog (*Canis familiaris*), donkey (*Equus asinus*), dormouse (*Glis glis*), elephant (*Loxodonta africana*), fruit bat (*Artibeus jamaicensis*), great rhino (*Rhinoceros unicornis*), guinea pig (*Cavia porcellus*), hedgehog (*Erinaceus europaeus*), hippo (*Hippopotamus amphibius*), pig (*Sus scrofa*), rabbit (*Oryctolagus cuniculus*), sheep (*Ovis aries*), and squirrel (*Sciurus vulgaris*). This denser, more diverse, and more controversial dataset yielded a distance matrix, which, when analyzed by neighbor-joining and hypercleaning, resulted in two somewhat different phylogenies. The consensus of these two phylogenies is presented in Figure 2.

While this consensus agrees with the overall structure of the phylogeny presented in Reyes *et al.* (2000), including the possible nonmonophyleticity of the rodents, several discrepancies exist. In particular, our method groups the pig with the perisodactyls rather than with the cetartiodactyls, the carnivores form a ferungulate outgroup, and the guinea pig groups with neither the murid nor the nonmurid rodents. However, neither of these latter two discrepancies are unreasonable hypotheses. The outgroup status of the carnivores has been suggested by Graur *et al.* (1997), and the phylogenetic position of the guinea pig is one of the most controversial topics in systematic biology (Graur *et al.*, 1991; D'Erchia *et al.*, 1996; Cao *et al.*, 1997; Sullivan and Swofford, 1997; Reyes *et al.*, 1998).

CONCLUSIONS

Our goal in this paper is not to confirm or refute previous phylogenetic studies but rather to bring a new methodology and a new tool to the comparative genomics research community. Our new method for whole genome comparison and phylogeny does not require gene identification nor any human intervention, in fact, it is totally automatic. It is mathematically well-founded being based on general information theoretic concepts. It works when there are no agreed upon evolutionary models, as further demonstrated by the successful construction of a chain letter phylogeny (manuscript in preparation by Bennett,C.H., Li,M., and Ma,B.), and when individual gene trees do not agree (e.g. Cao *et al.*, 1998) as is the case for genomes. Another possible use of our method is as an evaluator of alignments, as alignments with little shared information are unlikely to yield meaningful phylogenies by any method. Although a possible criticism of our method is that it is based on information theory rather than a biological model, it is worth stressing that the alignment algorithms that biologists use today are in fact also based on information theory, although less rigorously.

The method depends on the ability to efficiently and sharply approximate shared information, as this information is not computable. Our preliminary experiments have shown that our estimation approach is fruitful. However, to improve our results we need a better conditional entropy estimator of DNA sequences, as distances between highly divergent sequences tend to be similar to each other by our current estimator, making deep branches in phylogenies difficult to resolve. Still, our experiments have demonstrated clearly that the method is useful even with our current estimator. With a sharper conditional entropy estimator, we believe that this method will accurately recover whole organismal genome phylogenies as well.

ACKNOWLEDGEMENTS

We thank Ford Doolittle, Brian Golding, Masami Hasegawa, and Huaichun Wang for providing very useful information, and Tao Jiang, Pavel Pevzner, David Sankoff and Huaichum Wang. We especially thank Paul Vitányi, who pointed out several loopholes in an earlier version of this paper, and suggested the new formulation of Theorem 2. A referee suggested Reyes *et al.* (2000) data to us. J.H.B. was supported by a CITO grant. X.C., S.K., and M.L. were supported in part by a CityU research grant 7000 875, P.K. was supported by NSERC Research Grant 160321 and a CITO grant, M.L. was also supported in part by NSERC Research Grant OGP0046506, a CITO grant, and an NSERC Steacie Fellowship. H.Z. was supported by NSERC Research Grants OGP0046506 and 160321.

REFERENCES

Adachi,J. and Hasegawa,M. (1996) MOLPHY version 2.3: programs for molecular phylogenetics based on maximum likelihood. *Comput. Sci. Monogr. Inst. Stat. Math.*, **28**, 1–150.

Boore,J.L. and Brown,W.M. (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.*, **8**, 668–674.

Bryant,D., Berry,V., Kearney,P., Li,M., Jiang,T., Wareham,T. and Zhang,H. (2000) A practical algorithm for recovering the best supported edges of an evolutionary tree. In *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*.

Cao,Y., Janke,A., Waddell,P.J., Westerman,M., Takenaka,O., Murata,S., Okada,N., Pääbo,S. and Hasegawa,M. (1998) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.*, **47**, 307–322.

Cao,Y., Okada,N. and Hasegawa,M. (1997) Phylogenetic position of guinea pigs revisited. *Mol. Biol. Evol.*, **14**, 461–464.

Chen,X., Kwong,S. and Li,M. (2000) A compression algorithm for DNA sequences based on approximate matching. In Shamir,R., Miyano,S., Istrail,S., Pevzner,P. and Waterman,M. (eds), *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)*. Association for Computing Machinery, Tokyo, Japan, pp. 107.

D'Erchia,A.M., Gissi,C., Pesole,G., Saccone,C. and Arnason,U. (1996) The guinea pig is not a rodent. *Nature*, **381**, 597–599.

Fitz-Gibbon,S.T. and House,C.H. (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Res.*, **27**, 4218–4222.

Graur,D., Gouy,M. and Duret,L. (1997) Evolutionary affinities of the order Perissodactyla and the phylogenetic status of the superordinal taxa Ungulata and Altungulata. *Mol. Phylogenet. Evol.*, **7**, 195–200.

Graur,D., Hide,W.A. and Li,W.-H. (1991) Is the guinea pig a rodent? *Nature*, **351**, 649–652.

Grumbach,S. and Tahi,F. (1994) A new challenge for compression algorithms: genetic sequences. *J. Info. Process. Manage.*, **30**, 875–866.

Hannenhalli,S. and Pevzner,P. (1995) Transforming cabbage into turnip. In *Proceedings of the 27th ACM Symposium on Theory of Computing*, pp. 178–189.

Hillis,D., Huelsenbeck,J.P. and Swofford,D.L. (1994) Hobgoblin of phylogenetics? *Nature*, **369**, 363–364.

Kececioglu,J. and Sankoff,D. (1995) Exact and approximation algorithms for the inversion distance. *Algorithmica*, **13**, 180–210.

Koonin,E.V. (1999) The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, **15**, 265–266.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468.

Li,M. and Vitányi,P. (1997) *An Introduction to Kolmogorov Complexity and its Applications*. 2nd ed, Springer, New York.

Nadeau,J.H. and Sankoff,D. (1998) Counting on comparative maps. *Trends Genet.*, **14**, 495–501.

Reyes,A., Gissi,C., Pesole,G., Catzeflis,F.M. and Saccone,C. (2000) Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Mol. Biol. Evol.*, **17**, 979–983.

Reyes,A., Pesole,G. and Saccone,C. (1998) Complete mitochondrial DNA sequence of the fat dormouse, *Glis glis*: further evidence of rodent paraphyly. *Mol. Biol. Evol.*, **15**, 499–505.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Snel,B., Bork,P. and Huynen,M.A. (1999) Genome phylogeny based on gene content. *Nature Genet.*, **21**, 108–110.

Sullivan,J. and Swofford,D.L. (1997) Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mammal. Evol.*, **4**, 77–86.

Varre,J.-S., Delahaye,J.-P. and Rivals,E. (1998) The transformation distance: a dissimilarity measure based on movements of segments. In *German Conf. Bioinformatics*. Koel, Germany.

Wooley,J.C. (1999) Trends in computational biology: a summary based on a RECOMB plenarylecture. *J. Comput. Biol.*, **6**, 459–474.