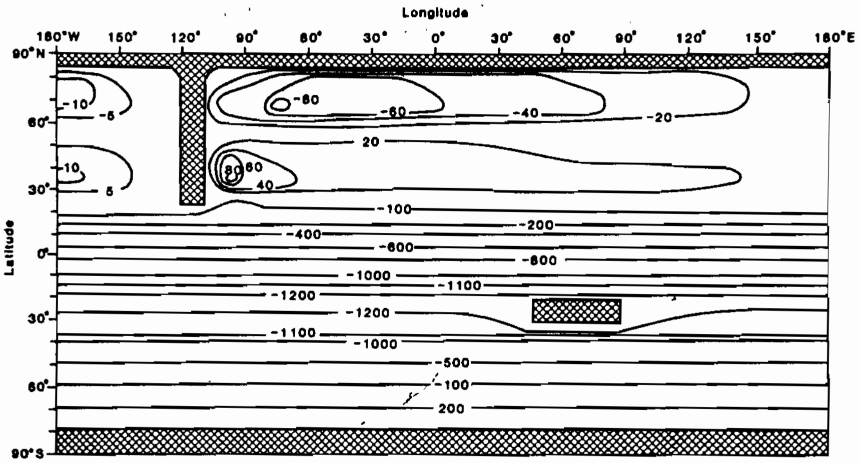

Fig. 1 Streamlines for the 12-year integration of the Archaean global climate model (in units of Sverdrups). The hypothesized land configuration is also shown (cross-hatched); we assume a peninsula (Northern hemisphere) and an island (Southern hemisphere) only. The small polar land areas are required by the OGCM to avoid numerical singularities.

one of the two radiative solutions to the paradox of equable early Archaean temperatures. Once formed, a near-global ocean does not freeze, despite incident solar radiation of only 70% of the present-day intensity. On the contrary, equatorial temperatures lie in the range 13–16 °C with ice extending to ~58° in the mean annual case. Thus both types of radiative solutions to the paradox are viable, and indeed the geophysical evidence points to both a near-global ocean and greatly enhanced levels of atmospheric CO_2.

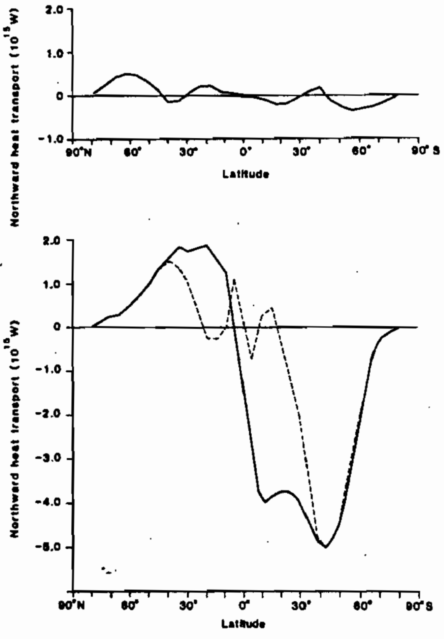

Fig. 2 The Archaean total northward heat transport simulated by the Bryan–Cox–Semtner[3] model (upper) can be compared with present-day values calculated by Meehl *et al.*[5] with the same ocean general circulation model (lower). In the lower figure, the solid line represents total transport and the dashed line indicates diffusive transport. (The comparison should probably be made using the diffusive transport appropriate to the modern ocean, because the total transport is influenced strongly by the intense transport of the land-locked boundary currents, which cannot be present in the Archaean ocean as postulated.)

We acknowledge NERC for provision of computer time.

A. Henderson-Sellers
*School of Earth Sciences,
Macquarie University,
North Ryde,
New South Wales 2109, Australia*

B. Henderson-Sellers
*School of Information Systems,
University of New South Wales,
PO Box 1, Kensington,
New South Wales 2033, Australia*

Loss of information in genetic distances

Sir—We certainly cannot dispense with DNA sequences when reconstructing evolutionary trees. Diamond[1] is correct in identifying 'convergence' as one of the important features for tree reconstruction — the tree should not change as more data are collected. But his discussion of the merits of DNA sequences versus DNA/DNA hybridization omits at least one important factor.

The omission is the fact that information is lost in the conversion of DNA sequences to a distance matrix. This loss occurs whether distance matrices are derived from known sequences, or by DNA/DNA hybridization. This conversion is not invertible.

Part of the problem has been discussed previously[2] when calculating the compression ratio (the ratio of the average number of distinct sets of sequence data, to the number of similarity matrices). So with nine taxa and 20 characters (each of which would have four states), there are, on average, at least $10^{14.9}$ distinct sets of sequences for each similarity matrix[2].

This figure is only a lower bound on the information loss because matrices that fail to meet the triangle inequality have not been excluded. It is difficult to calculate the exact proportion that fail the triangle inequality but a lower bound on the loss is easily calculated. For $n=3$ taxa the proportion of similarity matrices that fail the triangle inequality tends to one half. For c characters it is $(3\times{}^cC_3/c^3)$. For larger n, a much greater proportion fail. The result for three taxa gives a lower bound for larger n by using the maximum number of triangles (with independent edges) in a complete graph on n points (a Steiner triple system).

Examples of average information loss ($\log_{10}\Gamma(n,c)$)

c	$\log_{10}\Gamma(9,c)$	c	$\log^{10}\Gamma(9,c)$
12	3.3	100	178.0
20	18.5	500	798.9
50	79.9	1,000	1,390.4

For example, with $c=20$ there are, on average, at least $10^{18.5}$ distinct sets of sequences for each distance matrix. The information lost in converting to distances cannot be recovered.

Combining the calculations from the previous paragraphs gives a lower bound on the information loss. Let $\Omega(n,c)$ be the set of essentially distinct sequence spaces on n taxa, four colours (character states) and length c. ('Essentially distinct' means that two sets of data which differ only by a permutation of the sites or by a permutation of character states across sites, are considered identical.)

All matrices derived from sequence data will meet the restrictions $0\leqslant d_{i,j}=d_{j,i}\leqslant c$ and the triangle inequality $d_{i,j}\leqslant d^{i.k}+d_{j.k}$ for all taxa, i,j,k, where d is the 'distance' between two sequences. Let $D(n,c)$ denote the set of $n\times n$ matrices satisfying these conditions.

Let $\Gamma(n,c)=|\Omega(n,c)|/|D(n,c)|$. We define this ratio to be the average information loss. In Table 1 we calculate some values for $\log_{10}\Gamma(n,c)$ for different lengths c on four character states for $n=9$ taxa. Examples of the effect of this information loss have been given before[2].

The reconstruction of evolutionary trees need not be such a controversial field. There are times when distance methods, accompanied by an appropriate analysis[3], can give a reliable result. Cladists should accept these cases. That method, however, has not given a clear result with the difficult case of resolving the human, chimpanzee, gorilla trichotomy[3]. In such cases we need more information than is contained in the distance matrix. DNA sequences will be necessary for the hard problems.

M.A. Steel
M.D. Hendy
Mathematics and Statistics Department,
D. Penny
*Botany and Zoology Department,
Massey University,
Palmerston North, New Zealand*

1. Diamond, J.M. *Nature* **332**, 685–686 (1988).
2. Penny, D. *J. theor. Biol.* **96**, 129–142 (1982).
3. Felsenstein, J. *J. molec. Evol.* **26**, 123–131 (1987).