J. theor. Biol. (1996) **182,** 505–512

Addition of Darwin's Third Dimension to Phyletic Trees

THEODORA PINOU,†¶ TAMAR SCHLICK,‡ BIN LI,§ AND H. G. DOWLING†¶

†*Biology Department, New York University*, 1009 *Main Building, New York, NY* 10003, ‡*Howard Hughes Medical Institute and New York University, Chemistry Department and Courant Institute of Mathematical Sciences, New York University*, 251 *Mercer Street, New York, NY* 10012 *and* §*the Chemistry Department and Courant Institute of Mathematical Sciences, New York University*, 251 *Mercer Street, New York, NY* 10012, *U.S.A.*

(*Received on* 1 *April* 1996, *Accepted on* 10 *June* 1996)

A three-dimensional (3D) approach for visualizing the phyletic relationship of living animals is proposed and developed as an alternative to current two-dimensional (2D) evolutionary trees. The 3D tree enhances visualization and qualitative analysis since it simultaneously provides topological (tree-structure) and spatial information (based upon genetically measured distances). However, the meaning of the third dimension, particularly its relationship to temporal processes, and further quantitative analyses emerge as open questions. Our method consists of two phases. First, a 3D representation of the genetic relationships of a related group of extant animals is produced using an optimization algorithm developed here. Second, linear connections are added to suggest a visual representation of the differing evolutionary trajectories of the organisms involved on the basis of a 2D tree algorithm. The method is applied to a set of distantly related Caenophidian snakes, and the resulting relationships are analysed. The discussions here are meant to stimulate the generation of 3D trees in the goal of complementing standard 2D views and, perhaps ultimately, improving our classification of evolutionary relationships.

Introduction

The simplification of representing evolutionary lineages by a series of two-dimensional (2D) bifurcating branches is basic to most current tree-building programs (Swofford & Olsen, 1990; Hillis *et al*., 1994) and to the classifications that are based on these trees by phylogenetic taxonomists ("cladists") (Wiley, 1981). Indeed, in his "Origin of Species" (Darwin, 1859) Darwin stated:

> The representation of the groups, as here given in the diagram on a flat surface, is much too simple. The branches ought to have diverged in all directions . . . and it is notoriously not possible to represent in a series, on a flat surface, the affinities which we discover in nature among the beings of the same group. Thus, the natural system is genealogical in its arrangement, like a pedigree. But the amount of modification which the different groups have undergone has to be expressed by ranking them under different so-called genera, subfamilies, families, sections, orders and classes.

Ideally, such relationships involving Euclidean distances are better visualized in the three dimensions familiar to us from everyday life. Many complex spatial relationships that are masked in a flat diagram emerge in higher dimensions. However, in practical terms, it is difficult both to generate such three-dimensional (3D) views and to attribute biological meaning to the added dimension. Here we develop an

‡ Author to whom correspondence should be addressed.
¶ Present address: Department of Biology, Osborn Memorial Laboratory, Yale University, New Haven, CT 06511-8155. Current address for H. G. Dowling: Rendalia Biologists, 1811 Rendalia Motorway, Talladega, AL 35160, U.S.A.

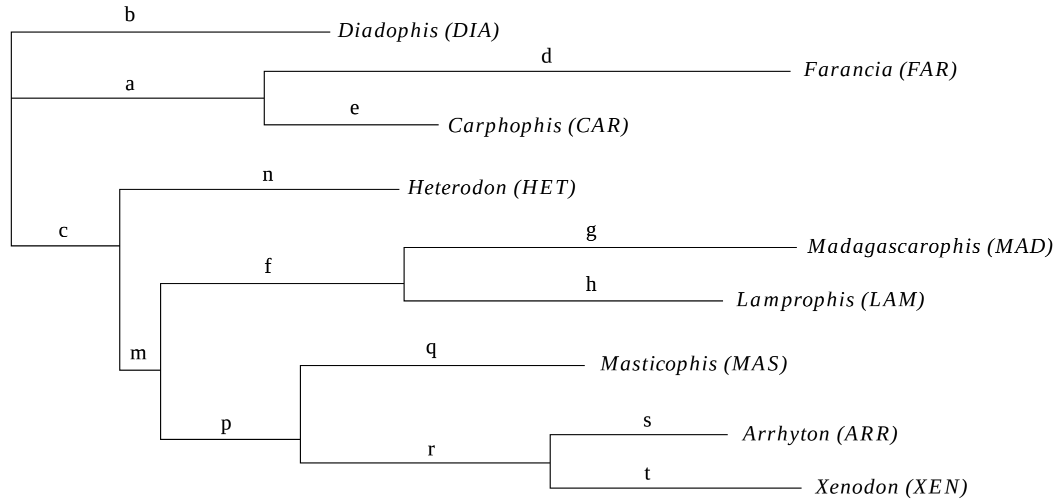

FIG. 1. An unrooted Neighbor-joining tree of the nine genera of snakes, an example of typical 2D computer-generated trees, is used to establish the pattern of evolutionary divergence in our 3D network. The labeled branches indicate the relative branch lengths as calculated by percent divergence and shown in Fig. 2. [a = 3.5; b = 3.5; c = 1.5; d = 7.5; e = 2.0; f = 3.5; g = 5.5; h = 4.5; m = 0.5; n = 4.0; p = 2.0; q = 4.0; r = 3.5; s = 3.5; t = 3.5].

approach for presenting information on evolutionary relationships in three dimensions, thereby allowing a visual representation similar to that of a molecular system. Our work is meant to stimulate further research in this direction by illustrating an application and discussing the new questions that emerge.

We use immunological distances (ID) that have been obtained by the quantitative micro-complement fixation technique (MC'F) for a set of distantly related Caenophidian snakes (Dowling & Jenner, 1988) to develop and apply our algorithm for generating a 3D view of phyletic relationships. Snakes are particularly suited for such extensive genetic analysis, since their "stripped-down" morphology allows little room for obvious modifications, and their fossil record is too limited to provide detailed phylogenetic information.

The first component of our algorithm combines distance geometry and nonlinear optimization techniques to compute a 3D illustration of points in space. This tree represents global intergenetic distances among the snake species analysed. The second component uses the Neighbor-joining tree-building algorithm to process further these spatial relationships, thereby associating a linkage-diagram suggesting a path to the common ancestor. Thus, the connecting of these points to make a network or a tree is a secondary procedure that does not affect their interrelations as positioned on the landscape. These patterns can be based on discrete character data obtained from morphological or DNA sequencing techniques, or distance data obtained from immunological techniques (MC'F), which we use here. The position of the taxa on the landscape, however, must be determined by distance data, inasmuch as character data cannot be represented as a comparative quantitative unit among taxa.

The suggestion of using 3D trees in taxonomy is not new (Sokal & Rohlf, 1981). Embedding a distance matrix in three dimensions is also a well-known problem in statistics, and a mathematically elegant method in case an exact representation exists (Neumaier, 1981, 1990). However, the practical implementation of this idea when an exact representation does not exist remains a challenge. We hope that the algorithm (and software) developed here will be used to explore this direction further to aid biological interpretations.

The method presented here produces an evolutionary tree embedded in a 3D space. Clearly, the visual

TABLE 1
Pairwise distance data for the nine snake species†

	1	2	3	4	5	6	7	8	9
1 (MAD)	0	70	105	113	80	116	61	114	88
2 (MAS)		0	68	56	70	119	74	63	82
3 (DIA)			0	82	56	91	90	93	66
4 (ARR)				0	63	109	112	36	84
5 (HET)					0	104	73	76	73
6 (FAR)						0	130	133	56
7 (LAM)							0	99	92
8 (XEN)								0	92
9 (CAR)									0

Pairwise distances, in units of Immunological Distance (ID), are given for distantly related snakes, currently thought by some to belong in the same family [McDowell, 1987].

† Abbreviations used: MAD: Madagascarophis; MAS: Masticophis; DIA: Diadophis; ARR: Atthyton; HET: Heterodon; FAR: Farancia; LAM: Lamprophis; XEN: Xenodon; CAR: Carphophis.

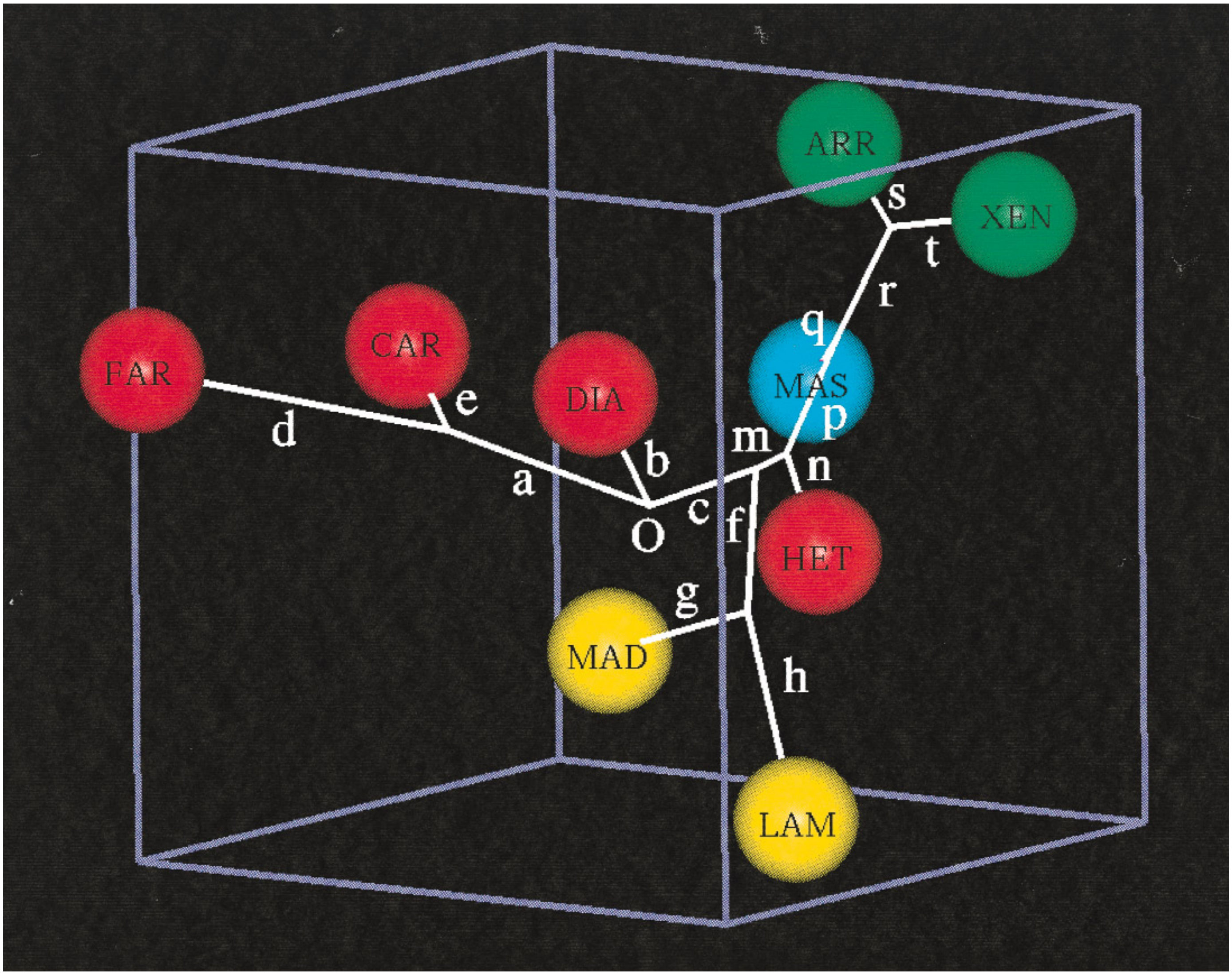

FIG. 2. A 3D representation of the complex immunological interrelations of nine snake genera. The taxa are represented as colored spheres and their positions were determined on the basis of each taxon's distance from all other taxa through the use of our algorithm (Appendix A). A cube is inserted around the figure to offer perspective. The colors of the spheres are based upon morphological similarities and differences: Green = Arrhyton, Xenodon (Neotropical snakes); Yellow = Lamprophis, Madagascarophis (Ethiopian snakes); Red = Carphophis, Diadophis, Farancia, Heterodon (relict Nearctic snakes); Blue = Masticophis (advanced Nearctic racer, a recent entrant).

The spheres are connected back to the Origin (O) through use of the Neighbor-joining tree (Fig. 1) based upon percentage-divergences in AIDs, thus forming a complex tree. However, linkage can be resolved by any tree-building method considered appropriate.

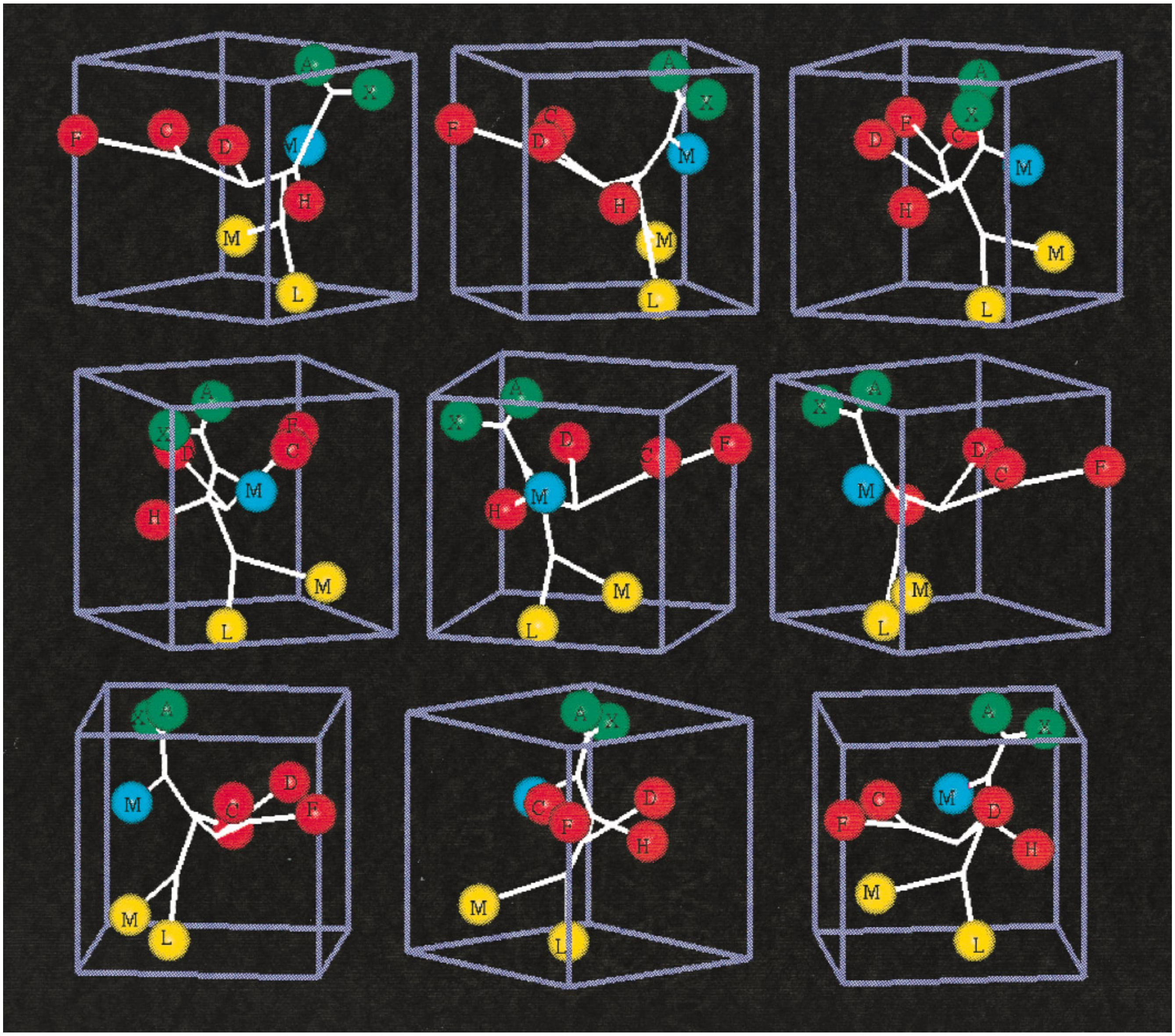

FIG. 3. Nine rotated representations of the 3D tree. The image at the upper left is the same view as in Fig. 2. It is rotated 40 degrees around the *x*-*y* plane of the cube sequentially from left to right and top to bottom, with the right bottom image the last. Note that the Neotropical (green) genera remain together, as do the Ethiopian (yellow) genera, and that both pairs are distinct from all other taxa. By contrast, only two of the Nearctic relicts (red), Carphophis and Farancia retain similar attitudes, the other relicts (Diadophis, Heterodon), and Masticophis (blue) reveal trajectories that differ from one another and from all other snakes.

TABLE 2
Computed three-dimensional tree

(a) *Coordinates*

Snake	*x*	*y*	*z*
1	17.31	−2.29	−1.03
2	65.16	29.30	23.61
3	39.74	101.25	5.62
4	98.15	65.78	23.22
5	29.87	74.05	44.32
6	23.69	88.88	−71.54
7	−4.72	19.74	43.22
8	95.52	64.11	54.92
9	45.27	61.91	−36.11

(b) *Quality of Fit*†

Pair distance	Value	% Error
1–2	62.41	8
1–3	106.15	1
1–4	108.42	4
1–5	89.67	10
1–6	115.42	1
1–7	54.12	8
1–8	116.86	2
1–9	78.31	9
2–3	78.40	12
2–4	49.18	9
2–5	60.63	10
2–6	119.67	1
2–7	73.20	1
2–8	55.80	8
2–9	70.89	11
3–4	70.57	12
3–5	48.33	10
3–6	79.77	10
3–7	100.18	9
3–8	83.19	8
3–9	57.62	10
4–5	71.94	11
4–6	122.71	11
4–7	114.46	2
4–8	31.85	6
4–9	77.57	5
5–6	116.96	10
5–7	64.40	9
5–8	67.24	9
5–9	82.79	11
6–7	136.95	5
6–8	147.53	9
6–9	49.48	8
7–8	110.24	9
7–9	102.81	10
8–9	104.00	11

† The numbers refer to the snake species as given in Table 1. The same units in Table 1 are also used.

aspect is better since a 3D tree offers simultaneously infinitely many 2D views (projections). Thus, spatial relationships (similarities and dissimilarities) among subgroups of species can be more easily examined from a multifurcating tree perspective (Mayr, 1974). Both O'Hara (1993) and Darwin (1859) noted that the determination of a phylogeny and the construction of a classification are two separate activities. In this report we are interested only in discerning a phylogeny, not in establishing a classification of the group considered.

However, what can we say about the 3D tree quantitatively? Three fundamental questions emerge.

(1) *What is the meaning of the third dimension?* Our tree is obtained by embedding genetic distances (in terms of amino acid sequences in albumin) in three dimensions, which is visually more attractive. However, the third dimension cannot be directly connected to another parameter, such as time. Certainly, the distance data themselves may reflect temporal evolutionary changes, but this cannot be immediately inferred from the 3D tree structure. Therefore, at present we view the 3D tree to be more useful for morphological descriptions than for use as a descriptor of molecular clocks.

(2) *How should the species in the 3D tree be connected?* The optimization algorithm presented below produces only Cartesian coordinates for all the species involved. Therefore, a linkage step (equivalent to tree building) when used is a secondary and separate aspect of the tree generation. For linkage, distance or character data, or a combination of the two, can be used. This combination of 3D representation and linkage might provide insight into the historical phylogeny of an animal by depicting both a pattern (topology) and relative directions of evolutionary changes.

(3) *Is the 3D approach better than the conventional 2D representation?* This is a difficult question. It is our hope that the third dimension provides a visual framework that permits qualitative examination of inter-relationships among groups of taxa that might be more difficult to appreciate from a flat representation. Therefore, a 3D tree might be particularly useful when a question arises regarding the particular subgroup classification for a family of related species (Borowsky *et al*., 1995). Many statistical tests that are applied to distance data can be used for analysis of the 3D tree. Moreover, the energy function used in the optimization is a measure of the overall fitness of the 3D topology to the data, with the distribution of deviations serving as an additional fit criterion (e.g., four distances are within 2% deviation of the measured genetic distance, six distances have 5% error, and so on). In fact, the obtained distance matrix [see Table 2(b)] might be useful to biologists in comparison with the original matrix. As we will discuss below, the optimization can only offer a local, rather than global, solution in any case for this nonlinear problem.

Having said this, some attractive features of the 3D approach might also be summarized. (1) No a priori assumptions are necessary for tree generation, as in two dimensions, regarding tree topology, for example. (2) The 3D description is more stable to additions of species than a 2D tree. That is, the positions in space (coordinates) of the species relative to one another will not drastically change as the number of taxa compared increases. This is because the reciprocal genetic distances will not change between two previously-located species. As the number of taxa increases, most modifications necessary to accommodate the additional distances would be local, and would not disrupt the relative positions of the distantly-related taxa. (3) The 3D tree offers simultaneously many 2D views and therefore might facilitate analysis of interrelationships among subgroups of species.

Materials and Methods

DATA COLLECTION

For our application we use a sample of nine Caenophidian (modern) snakes (Dowling & Jenner, 1988). Five of these (*Carphophis*, *Diadophis*, *Farancia*, *Heterodon*, *Masticophis*) are of Nearctic (temperate North American) derivation, two (*Arrhyton*, *Xenodon*) are Neotropical (West Indies and South America), and two (*Lamprophis*, *Madagascarophis*) are Ethiopian (Africa and Madagascar, respectively). The data consist of reciprocal immunological distances (antibody–antigen) between the albumins (Albumin Immunological Distances: AIDs) of the nine taxa (Table 1). As mentioned above, detailed information on extinct species of snakes is not available, as it is for mammals; most snake fossils consist of isolated vertebrae, which can show only minimal degrees of variation. Thus, this tree is constructed on the basis of the genetic relationships of living species.

The antibodies and sera to be compared were obtained by standard methods (Maxson & Maxson, 1990). Each AID determined by MC'F is approximately equal to one amino acid change in the albumin molecule (Maxson & Maxson, 1986). Although the percent deviation from reciprocity in our raw data was less than the acceptable 100%, the data were scaled (Cronin & Sarich, 1975) for antisera which consistently overestimated or underestimated the distances obtained. These scaled distances gave a slightly better overall fit to the 3D algorithm. The mean of each pair of reciprocal distances was considered to be the best estimate of the number of amino acids which had changed since the two compared species had diverged.

These means were converted to percent divergence by using the previously determined (Benjamin *et al.*, 1984) protein size of albumin of approximately 580 amino acids: % divergence $= 100 * \mathrm{ID}/580$. These values were then used to construct Fig. 1 by the Neighbor-joining algorithm (Saitou & Nei, 1987) that was suggested by Schubert *et al.*, (1993) as appropriate for estimating patterns of genetic divergence (Fig. 1), and to compute the coordinates (Table 2) for 3D figures (Figs 2 and 3) constructed by our algorithm.

MODELING

The mathematical problem for computing a 3D arrangement in space can be stated as follows. Given a set of n species and an associated pairwise distance matrix, find the coordinates for all the species in three dimensions that match those pairwise distances as closely as possible. The pairwise data come from experimental measurements, for example from distances of homologous proteins (or genes) in two different species. Although we would like to satisfy those constraints exactly, this is usually impossible (see below); we thus seek the best possible approximation. This class of problems is central to distance geometry (Crippen, 1991; Crippen & Havel, 1988), a field with wide applications in chemistry and biology. A classic problem involves calculating the 3D molecular configuration subject to nuclear magnetic resonance data (inter-proton distances).

While it is easy to visualize a static molecular configuration in space, it is more difficult to picture evolutionary interrelationships in three dimensions: the spatial relationships (in terms of immunological distances) help differentiate the amino-acid changes in the protein albumin among the taxa compared (Maxson, 1992). Clearly, only closely related organisms will share a close relationship in space. Moreover, because the position of each taxon in space is based upon its distance to every other taxon compared, a temporal pattern of divergence might be approximated from a tree-building algorithm which utilizes distance data, such as the Neighbor-joining method (Tateno *et al.*, 1994). When branches connect the taxa in the 3D image, the vectors represent the taxa's interrelated historical trajectories. Therefore, closely related groups should share not only a small difference between their amino acid sequence, but also a similar direction in their trajectories.

The following, more precise, mathematical problem can be formulated. We are given for n species a set of measured distances $\{\delta_{ij}^0\}$ for $i, j = 1, \ldots, n$, where δ_{ij}^0

is the distance between specimens i and j. We would like to find a set of n 3D points $\{\mathbf{x}_i\}$, $i = 1, \ldots, n$ which approximates the pairwise data in some way. More compactly, we donate by $\mathbf{X}$ the collective vector of $3n$ components that lists the positions of each specimen in turn. We require that the set of inter-species data, $\{d_{ij}\}$, at $\mathbf{X}$, where $d_{ij}(\mathbf{X}) = \|\mathbf{x}_i - \mathbf{x}_j\|$ is measured in the standard Euclidean norm, fit the measurements within some margin of error:

$$l_{ij} \leqslant d_{ij}(\mathbf{X}) \leqslant u_{ij}, \quad i, j = 1, \ldots, n. \tag{1}$$

The measured values $\{\delta^0_{ij}\}$ lie between these lower and upper bounds:

$$l_{ij} \leqslant \delta^0_{ij} \leqslant u_{ij}, \quad i, j = 1, \ldots, n. \tag{2}$$

Why do we seek some approximate, rather than exact, solution? This kind of problem is typically overdetermined. Thus, we can expect at best a good overall fit. For n species, the matrix of pairwise distances contains $(n(n-1))/2$ unique non-zero elements (in the strict lower or upper triangle); in contrast, the Cartesian vector $\mathbf{X}$ contains $3n - 6$ independent variables for $n > 3$, as three degrees of freedom are removed for rigid-body translation and rotation invariance. Thus, more constraints than degrees of freedom follow for $n > 4$, and current optimization techniques can only provide a locally optimal solution. This is not unlike the generation of 2D phylogenetic trees, where various a priori assumptions must be made (e.g., additivity) in designing the algorithm that produces a certain form of tree.

Our approach to find an optimal solution for the data of snake species consists of several components: (1) Formulation of an energy function that describes the quality of fit. (2) Generation of reasonable starting structures. (3) Minimization of the objective function by a rapidly-converging Newton method for nonlinear functions. (4) Repeated projections/minimizations of the solution vector onto the bounds to optimize the fit to the data within the prescribed error bounds. We include steps 2 and 3 above in Phase II of our algorithm. The algorithm combines strategies in nonlinear optimization, distance geometry, and molecular mechanics. While it is heuristic, the overall form of the final configuration (Fig. 2) is the best we found through many trials and variations in parameters and starting configurations. For nearby solutions (less optimal fit), very similar patterns emerged, so we believe that our 3D image is a key representative structure for the data given here. Full details of the algorithm are given in Appendix A.

Results and Discussion

The 2D Neighbor-joining tree (Fig. 1) shows three distinct sister-relationships: a Neotropical Arrhyton-Xenodon clade, an Ethiopian Lamprophis-Madagascarophis clade, and a Carphophis-Farancia clade among the five Nearctic genera. An apparent relationship between the North American Masticophis and the Neotropical clade also is suggested, although Masticophis is immunologically distant from the Neotropical members. The three clades and all of the remaining Nearctic taxa are shown to be only distantly related to one another. The topology of this tree, based upon percentage-divergence information, is identical to those based directly upon either raw or scaled AID data.

The 3D information is provided in Figs 2 and 3. Each of the taxa is represented as a colored sphere placed in relation to its % AID to all other taxa by our algorithm. Figure 2 shows the complexity of relationships revealed by this technique in a "fixed" position. It shows three distinct branches from the central Origin (O), with some of the Nearctic genera apparently occupying somewhat intermediate positions. This figure indicates clearly, however, that the Ethiopian (yellow) and Neotropical (green) genera are at opposite poles of relationship and that none of the Nearctic genera is closely related to either clade.

Figure 3 begins at the upper left with the same view as Figure 2. Each subsequent view shows the relationships revealed in a 40-degree turn, with the final one at the lower right. They show that whereas the pairs of genera of Neotropical and of Ethiopian origin remain together and separate from other taxa; this is not true of all Nearctic genera. In particular Masticophis (blue M), Diadophis (red D), and Heterodon (red H) show entirely separate patterns of relationship to one another as well as to all of the others. They also show that no single (2D) view can resolve the complex relationships shown here.

The addition of the 3D view to the pattern of immunological distances suggests the complex and different "evolutionary trajectories" of the taxa. Even though these trajectories are here visualized only in terms of immunological distance measurements (rather than genomic, morphological, or ecological differences), it is evident (Fig. 2) that those taxa which were believed to be closely related on morphological and distributional bases (Dowling & Jenner, 1988; Pinou, 1993) follow similar directional paths. They cluster around one another, and occupy distinct regions separate from those that may be roughly equidistant (immunologically) from a common ancestor, but which have followed different evolutionary trajectories (directional paths) as shown by the differences in their relative positions on the "landscape".

Such a distinction in the relationship of Masticophis (blue M) to Arrhyton and Xenodon (green) is particularly illuminated by the 3D approach (Fig. 3). Although the Neighbor-joining figure (Fig. 1) suggests that Masticophis is most closely related to the Neotropical clade, its independent trajectory, particularly well shown in images 3, 4, and 7 (Fig. 3), contradicts such a suggested relationship in spite of the relatively shorter AID between Masticophis and this clade. The 3D views (Fig. 3) also suggest the independent evolutionary trajectories of the other Nearctic snakes. Only Farancia and Carphophis retain similar associations in the rotation of the figure. By contrast, Diadophis and Heterodon reveal trajectories that not only differ from one another but also from all other taxa.

Conclusions

The optimization algorithm presented here for computing 3D trees from distance data can be easily used and applied to problems of phylogeny. Although new problems regarding the interpretation of the third dimension emerge as discussed in the introduction, especially the connection to temporal processes, the additional dimension provides a useful visual, complementary tool for analysing relationships among related species. The inherent difficulty of compressing a multidimensional and multifurcate phylogeny into a bifurcate, 2D format has led to many approaches for construction of evolutionary trees (Swofford & Olsen, 1990; Hillis *et al*., 1994) and to various selection criteria for the "correct" tree among the many candidates generated by computer programs (Hedges, 1992).

In theory, the 3D dendrogram can better serve as one of the foundations from which a classification of the taxa involved can be derived. When a speciation event occurs, each of the species loses its genetic contact with the other and develops its own evolutionary trajectory (Frost *et al*., 1992). If plotted as a dendrogram (tree), the resulting relationship between the two species and the (real or presumed) ancestor can be accurately represented on a plane as three connecting points. The angle may indicate the degree of separation between the two sister species, and the plane thus defined is equivalent to the evolutionary trajectory of the species involved. If a third species evolves from the same ancestral stock, however, inasmuch as it is on a trajectory independent of the others, it is unlikely that its trajectory will be the same as that of the first two. Thus, it cannot be indicated on the same plane, thereby complicating the problem of accurate tree delineation.

Biological classification should be treated as a hypothesis that also serves as an organized reference system for information storage and retrieval. For this reason, a single phylogenetic tree should not be used as the sole basis of a classification. However, by incorporating a trajectory, a 3D dendrogram might illuminate differences among the taxa which otherwise remain obscure in a conventional 2D tree. This 3D framework suggests analysis of the complex relationships among taxa by a combination of morphological, physiological, geographical, ecological, and additional genetic data. Ultimately, such comprehensive analyses might be used to derive more sophisticated evolutionary classifications.

New applications of our 3D tree algorithm are now underway with regards to fresh-water fish†. Further studies are necessary for understanding the type of information that 3D trees might be useful for and the relation between spatial and temporal processes.

† We invite interested users to contact T. Schlick by e-mail (schlick@nyu.edu) for other applications of the optimization code to generation of 3D trees from given distance data.

We thank Dr. Carla A. Hass for her critical review of this manuscript, Dr. Suse Broyde for introducing the group of authors to one another and inspiring this collaboration, and Mr. Edward Friedman, of the Academic Computing Facility at New York University, for assistance with the graphics. The immunological work was supported by the Department of Biology at New York University and was conducted in the laboratory of Dr. Linda R. Maxson at Pennsylvania State University, University Park. T. S. is an investigator of the Howard Hughes Medical Institute.

REFERENCES

Benjamin, D. C., Berzofsky, J. A., East, I. J., Gurd, F. R. N., Hannum, C. & Leach, S. J. *et al*. (1984). The antigenic structure of proteins: a reappraisal. *Annu. Rev. Immunol*. **2,** 67–101.

Borowsky, R. L., McClelland, M., Cheng, R. & Welsh, J. (1995). Arbitrarily primed DNA fingerprinting for phylogenetic reconstruction in vertebrates: the *Xiphophorous* model. *Mol. Biol. Evol*. **12,** 1022–1032.

Burkert, U. & Allinger, N. L. (1982). *Molecular Mechanics*. Washington, D.C.: American Chemical Society. ACS Monograph 177.

Crippen, G. M. (1991). Chemical distance geometry: Current realization and future projection. *J. Math. Chem*. **6,** 307–324.

Crippen, G. M. & Havel, T. F. (1988). *Distance Geometry and Molecular Conformation*. New York: John Wiley & Sons.

Cronin, J. E. & Sarich, V. M. (1975). Molecular systematics of the new world monkeys. *J. Hum. Evol*. **4,** 357–375.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Dowling, H. G. & Jenner, J. V. (1988). Snakes of Burma: Checklist of reported species and bibliography. *Smithsonian Herpetol. Inform. Serv.* **No. 76.**
Frost, D. R., Kluge, A. G. & Hillis, D. M. (1992). Species in contemporary herpetology: Comments on phylogenetic inference and taxonomy. *Herpetol. Rev.* **23,** 46–54.
Glunt, W., Hayden, T. L., Shelling, J. G., Ward, D. J., & Wells, C. (1994). Applications of weighting and chirality strategies for distance geometry algorithms. *J. Math. Chem.* **15,** 353–366.
Hedges, S. B. (1992). The number of replications needed for accurate estimation of the bootstrap *P* value in phylogenetic studies. *Mol. Biol. Evol.* **9,** 366–369.
Hillis, D. M., Huelsenbeck, J. P. & Cunningham, C. W. (1994). Application and accuracy of molecular phylogenies. *Science*, **264,** 671–677.
Maxson, L. R. (1992). Tempo and pattern in anuran speciation and phylogeny: An albumin perspective. In: *Herpetology*: *Current Research on the Biology of Amphibians and Reptiles*, (Adler, K., ed.) Proceedings of the First World Congress of Herpetology pp. 41–57. Oxford, OH: Society for the Study of Amphibians and Reptiles.
Maxson, L. R. & Maxson, R. D. (1990). Proteins II: Immunological techniques. In: *Molecular Systematics*, (Hillis, D. M. & Moritz, C., eds) pp. 127–155. Sunderland, Massachusetts: Sinauer Associates
Maxson, R. D. & Maxson, L. R. (1986). Microcomplement fixation: A quantitative estimator of protein evolution. *Mol. Biol. Evol.* **3,** 375–388.
Mayr, E. (1974). Cladistic analysis or cladistic classification? *Zool. Syst. Evol. Forsh.* **12,** 94–128.
McDowell, S. B. (1987). Systematics. In: *Snakes*: *Ecology and Evolutionary Biology*, (Siegel, R. A., Collins, J. T., & Novak, S., eds) pp. 3–50. New York: McGraw Hill
Neumaier, A. (1981). Distance matrices, dimension and conference graphs. *Indigationes Math.* **43,** 385–391.
Neumaier, A. (1990). Derived eigenvalues of symmetric matrices, with applications to distance geometry. *Linear Algebra and its Applications*, **134,** 107–120.
O'Hara, R. J. (1993). Systematic generalizations, historical fate, and the species problem. *Syst. Biol.* **42,** 231–246.
Pinou, T. (1993). *Relict Caenophidian snakes of North America*. PhD thesis New York University, New York.
Saitou, N. & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425.
Schlick, T. (1987). Modeling and Minimization Techniques for Predicting Three-Dimensional Structures of Large Biological Molecules. PhD thesis New York University, New York.
Schlick, T. & Fogelson, A. (1992a). TNPACK—A truncated Newton minimization package for large-scale problems: I. Algorithm and usage. *ACM Trans. Math. Softw.* **14,** 46–70.
Schlick, T. & Fogelson, A. (1992b). TNPACK—A truncated Newton minimization package for large-scale problems: II. Implementation examples. *ACM Trans. Math. Softw.* **14,** 71–111.
Schlick, T. & Overton, M. L. (1987). A powerful truncated Newton method for potential energy functions. *J. Comp. Chem.* **8,** 1025–1039.
Schubert, F. R., Nieselt-Struwe, K. & Gruss, P. (1993). The antennapedia-type homobox genes have evolved from three precursors separated early in metazoan evolution. *Proc. Natl. Acad. Sci., U.S.A.* **90,** 143–147.
Sokal, R. R. & Rohlf, F. J. (1981). *Biometry*: *the Principles and Practice of Statistics in Biological Research*. 2nd Edn. San Francisco, CA: W. H. Freeman and Company.
Swofford, D. L. & Olsen, G. J. (1990). Phylogeny reconstruction. In: *Molecular Systematics*, (Hillis, D. M. & Moritz, C., eds) pp. 411–501. Sunderland, MA: Sinauer Associates.
Tateno, Y., Takezaki, N. & Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11,** 261–277.
Wiley, E. O. (1981). *Phylogenetics*: *The Theory and Practice of Phylogenetic Systematics*. New York, NY: John Wiley & Sons.

APPENDIX A

The Numerical Optimization Procedure

We formulate the objective function as the sum of weighted squared distance deviations:

$$E^k(\mathbf{X}) = \sum_{i<j} \omega_{ij}^k \,(d_{ij}(\mathbf{X}) - \delta_{ij}^k)^2, \qquad \text{(A.1)}$$

where d_{ij} is the Euclidean distance between particles i and j in the configuration $\mathbf{X}$, δ_{ij} is a target distance, and ω_{ij} is a weight. A weighted sum is used since we would like to incorporate relative, rather than absolute, errors for each distance; each ω_{ij} should depend on the associated difference between the upper and lower bounds, $u_{ij} - l_{ij}$, so that differing magnitudes among the elements as well as differing accuracies in the measurements could be incorporated. (Only the first is relevant to this work but the second may be important for similar applications with a larger set of species.) The purpose of the superscript k on the energy function, weights, and targets of eqn (A.1) will become evident below; they basically indicate that our objective function is modified as the algorithm proceeds.

The form of $E(\mathbf{X})$ is identical to a harmonic bond potential as used in molecular mechanics calculations (Burkert & Allinger, 1982). The corresponding first and second derivatives are straightforward to calculate (Schlick, 1987), and thus this function can be subjected to a Newton minimization algorithm. Any available Newton minimizer for non-linear functions can be applied to this problem, and we choose for the calculations reported here the truncated-Newton package (Schlick & Overton, 1987; Schlick & Fogelson, 1992a,b), well suited to explore conformational regions with many minima, maxima, and saddle points.

When there are no available data regarding a 3D configuration, a first objective is to generate a reasonable starting structure for the optimization method. We employ here the following strategy, which might be a useful technique in general. Since for nine species, there are 36 unique pairwise distances (see Table 1) but only 21 independent Cartesian coordinates, we select a subgroup G1 of 21 distance pairs which have good reciprocity values: 1–2, 1–5, 1–7, 1–9, 2–4, 2–6, 2–7, 2–9, 3–4, 3–5, 3–6, 3–9, 4–5, 4–7, 4–8, 5–9, 6–7, 6–8, 6–9, 7–9, and 8–9. We then set the weights for the G1 pairs as $\omega_{ij}^k = S_1/(\delta_{ij}^0)^2$, where $S_1 = 100$; for the remaining 15 distances (group G2), we set the weights to zero. This setting yields an initial objective function as a sum of squared relative errors for group G1 distances. Relative errors are important here, as our distances vary between 56 and 133. An initial configuration for the nine species is chosen by distributing nine points in three parallel lines, in such a way that each neighboring pair of lines is in a perpendicular orientation.

Minimization with the target function defined above produced a final energy value of zero, as all 21 distances in G1 were satisfied exactly. Errors of up to 150% occurred for the distances in G2, not included in the target function. Gradually, we then increased the weights for the G2 pairs by setting $\omega_{ij}^k = S_2/(\delta_{ij}^0)^2$, where S_2 was increased slowly from 10^{-4} to 100 (S_1 remained at 100 for G1). Each k setting was followed by a minimization of $E^k(\mathbf{X})$ from the previous configuration. This procedure produced a final configuration in which the largest distance violation was 20%. The largest deviations occurred for distance pairs 2–3 (19%), 2–9 (15%), 3–5 (15%), and 5–7 (13%). This completes Phase I of our algorithm.

Phase II begins as no further improvements could be made by varying S_1 and S_2. Our minimization/projection procedure of Phase II adopts a similar strategy to that proposed by Hayden and co-workers (Glunt *et al.*, 1994). The goal of this procedure is to fit the solution as well as possible to the 3D region specified by the lower and upper bounds of each distance value. In other words, we seek to optimize the fit. In this work, we specify 10% margins of error, as dictated by the data collection procedure. The projection strategy involves changing the targets $\{\delta_{ij}^k\}$ at each step k: when a certain distance $d_{ij}(\mathbf{X}^k)$ is outside its permitted range, the corresponding target is set to the nearest bound (so it is approached upon minimization); when $d_{ij}(\mathbf{X}^k)$ lies within the bounds, the target retains the current value.

To accelerate convergence, the algorithm also modifies the weights at each step so that they reflect the magnitude of the error, or the 'distance' to the nearest bound. With such stepwise weight adjustments and target resettings, we found rapid convergence to the solution reported here with our minimizer. This phase of the solution can now be summarized as follows.

Projection/Minimization Algorithm

(a) Let $\mathbf{X}^0, \{\delta_{ij}^0\}, \{l_{ij}\}, \{u_{ij}\}$ be given for $i, j = 1, \ldots, n$, where

$\mathbf{X}^0 =$ an approximation to the solution;
$\delta_{ij}^0 =$ the measure target distance for pair i, j (Table 1);
$l_{ij}, u_{ij} =$ lower and upper bounds, respectively, for measurement i, j (in this work, $l_{ij} = (0.9)\delta_{ij}^0$, $u_{ij} = (1.1)\delta_{ij}^0$);
$w_{ij}^0 = S/(\delta_{ij}^0)^2$ where S is a constant ($S = 100$ in this work).

(b) for $k = 1, 2, \ldots,$ until $E^k(\mathbf{X})$ is sufficiently small:

1. Minimize the function:

$$\mathrm{E}^k(\mathbf{X}) = \sum_{i<j} \omega_{ij}^k \, [d_{ij}(\mathbf{X}) - \delta_{ij}^k]^2;$$

Set $k \leftarrow k + 1$ and $\mathbf{X}^k$ to the minimum of $E^{k-1}(\mathbf{X})$.

2. Update the targets of the objective function above:

$$\delta_{ij}^k = \begin{cases} d_{ij}(\mathbf{X}^k) & \text{if } l_{ij} \leqslant d_{ij}(\mathbf{X}^k) \leqslant u_{ij} \\ l_{ij} & \text{if } d_{ij}(\mathbf{X}^k) < l_{ij} \\ u_{ij} & \text{if } d_{ij}(\mathbf{X}^k) > u_{ij} \end{cases}.$$

3. Update the weights of the target function:

$$\omega_{ij}^k = S(1 + \epsilon_j)/(\delta_{ij}^k)^2,$$

$$\epsilon_{ij} = \max\{0, d_{ij}(\mathbf{X}^k) - u_{ij}, l_{ij} - d_{ij}(\mathbf{X}^k)\}.$$

4. Go to step 1.