BIOINFORMATICS

Vol. 18 no. 1 2002
Pages 100–108

Integrated gene and species phylogenies from unaligned whole genome protein sequences

Gary W. Stuart*, Karen Moffett and Steve Baker

Department of Life Sciences, Indiana State University, Terre Haute, IN 47809, USA

Received on April 2, 2001; revised on July 6, 2001; accepted on July 20, 2001

ABSTRACT

Motivation: Most molecular phylogenies are based on sequence alignments. Consequently, they fail to account for modes of sequence evolution that involve frequent insertions or deletions. Here we present a method for generating accurate gene and species phylogenies from whole genome sequence that makes use of short character string matches not placed within explicit alignments. In this work, the singular value decomposition of a sparse tetrapeptide frequency matrix is used to represent the proteins of organisms uniquely and precisely as vectors in a high-dimensional space. Vectors of this kind can be used to calculate pairwise distance values based on the angle separating the vectors, and the resulting distance values can be used to generate phylogenetic trees. Protein trees so derived can be examined directly for homologous sequences. Alternatively, vectors defining each of the proteins within an organism can be summed to provide a vector representation of the organism, which is then used to generate species trees.

Results: Using a large mitochondrial genome dataset, we have produced species trees that are largely in agreement with previously published trees based on the analysis of identical datasets using different methods. These trees also agree well with currently accepted phylogenetic theory. In principle, our method could be used to compare much larger bacterial or nuclear genomes in full molecular detail, ultimately allowing accurate gene and species relationships to be derived from a comprehensive comparison of complete genomes. In contrast to phylogenetic methods based on alignments, sequences that evolve by relative insertion or deletion would tend to remain recognizably similar.

Availability: Both the program used to convert properly formatted sequence files into sparse n-gram matrices (aacode3) and the program used to generate PHYLIP compatible pairwise distance matrices from the Singular Value Decomposition (SVD) output (cosdist) are available at http://mama.indstate.edu/user/stuart. The SVD package is available at http://www.netlib.org/svdpack/index.html, and the PHYLIP package is available at http://evolution.genetics.washington.edu/phylip.html.

Contact: G-Stuart@indstate.edu

INTRODUCTION

Our ability to generate and store biomolecular sequence information greatly exceeds our ability to ascertain the functions of the biomolecules represented. A good first approximation of function can be obtained directly from primary sequence information by recognizing a sufficient degree of sequence similarity between the molecule in question and one or more sequenced biomolecules of known function (see Koonin *et al.*, 1998). In principle, and to some extent in practice, functional knowledge of a relatively small number of molecules can indirectly facilitate a relatively accurate understanding of the function of a large number of biomolecules. The accuracy of these large scale, inductive estimates of function will depend critically on the accuracy with which similarities between individual biomolecules are recognized and measured and the degree to which multiple similarities (due to multiple functional domains) are accurately represented and resolved. The complexity involved in estimating relatedness between large numbers of biomolecules is enormous and requires the development of innovative computational methods, especially as applied to the interpretation of whole genomes (see Snel *et al.*, 1999; Tekaia *et al.*, 1999; Lin and Gerstein, 2000; Li *et al.*, 2001).

Estimates of relatedness frequently depend upon explicit alignments of similar sequences. Alignment algorithms are intrinsically highly subjective and usually employ cut-off values and gap penalties that are difficult to define. The result of an alignment is consequently, to some extent, already a statement of similarity (greater than a critical score). Of the three major sources of error in molecular phylogenies, inaccurate alignments are considered to be the most significant (Lake and Moore, 1998; Thorne, 2000). Furthermore, once an acceptable alignment is obtained, a high fraction of the original sequence information is sometimes discarded, making the estimation of similarity (and postulated homology) highly restricted and relevant to one or just a few selected

*To whom correspondence should be addressed.

domains. These domains consist almost exclusively of contiguous or nearly contiguous sequence, as most models of sequence evolution used to generate alignments fail to account for insertions and deletions of more than a few bases (Thorne, 2000). Alignments also generally ignore the potential for dissimilar sequences to have similar function (i.e. analogy).

In this report, we describe a general method for generating biomolecular phylogenies using multiple genome datasets of unaligned sequence information. Both gene and species phylogenies are derived from the same analysis. Since specific alignments are not utilized, estimates of sequence relatedness should be relatively unaffected by insertions or deletions of arbitrary size.

SYSTEM AND METHODS

Our approach is very similar to one that has been successfully applied in recent attempts to organize large amounts of textual information in such a way as to make that information more easily understandable and accessible. This approach is referred to as 'Latent Semantic Analysis' (LSA) in the fields of psycholinguistics and artificial intelligence, or 'Latent Semantic Indexing' (LSI) when utilized for information retrieval (Landauer and Dumais, 1997; Landauer *et al.*, 1998; Berry *et al.*, 1999). Since biomolecular sequence data can be viewed as a complex written language, individual biomolecules within large datasets can be rigorously compared using large n-gram frequency matrices similar to those used to compare samples of text in LSA. In this case, individual protein sequences, for example, would correspond to 'passages' of text, whereas peptides of a given size could serve as n-gram 'words'. While words in a language text are easy to identify and define, 'words' in proteins are much harder to define. A reasonable definition might be 'peptides with important functions', but these are normally very difficult to identify in large, novel datasets. The problem of defining peptide words can be solved in an operational way by considering all possible combinations of small strings of amino acids. For instance, there are 8000 possible three-letter (tripeptide) words. If four-letter words are chosen, then there are 160 000 possible tetrapeptide words. Regardless of the word/peptide definition used, proteins can be represented by peptide frequency vectors, and conversely, peptides can be represented by protein frequency vectors. However, these simple vector representations fail to account for the potential of two different peptides to have similar functions. For example, two peptides of similar sequence, like LLLL and LLIL, might be relatively interchangeable in the set of homologous proteins in which they appear and may frequently occupy analogous positions within related domains. The simple first-order vector representations described above would not reflect this relationship, as the pattern LLLL is represented independently from the pattern LLIL in the dataset.

Improving vector definitions: singular value decomposition and dimensionality

Following the generation of a high dimensional n-gram frequency matrix representing the dataset in question, the matrix itself is subjected to Singular Value Decomposition (SVD). This analysis serves to decompose the matrix into three component matrices that can be used to reform the original matrix using the relation $A = U\Sigma V^T$. Of the three matrices resulting from the decomposition, the leading matrix, or 'peptide' matrix (U), defines the n-gram peptides of the dataset as relative vectors in an abstract space in which the axes represent newly defined independent characteristics (defined as orthonormal basis vectors in the output matrices). The trailing matrix, or 'protein' matrix (V), defines the proteins as relative vectors using the same set of independent characteristics. The central matrix (Σ) is a diagonal matrix containing the singular values in decreasing order from left to right. The largest singular values identify independent characteristics that provide the strongest contributions to the meaning of peptides and proteins within the dataset, whereas the smaller singular values identify characteristics with less important or even misleading contributions.

By setting the smaller singular values to zero in a process known as 'dimension reduction', estimates of relatedness can be considerably improved. Dimension Reduction (DR) has the effect of providing a least-squares approximation of the peptide and protein definitions in a reduced dimensional space. Empirical studies on large datasets indicate that DR can be optimized to provide a high rate of correct word definition (Landauer and Dumais, 1997). Note that DR not only serves to reduce 'noise' and to minimize conflict in the data, it also has the advantage of greatly reducing the computational complexity of subsequent relatedness estimates (see Section Algorithm). SVD-DR represents a mathematical form of hierarchical understanding, in which a large number of items or concepts can be understood better in a relative sense if they are characterized by a limited number of meaningful characteristics. From a mathematical perspective, choosing an appropriate reduced dimensionality by setting the smaller singular values to zero produces a low rank approximation of the input matrix, $A_k = U_k\Sigma_k V_k^T$, which does not differ appreciably from the original matrix (Berry *et al.*, 1999)

As demonstrated below, accurate global measures of gene and species relatedness can be obtained from the refined vector definitions embodied within the protein matrix obtained as output from SVD-DR. We interpret the individual elements of these protein defining vectors as representing the contribution of specific 'motifs' to the various proteins within the dataset (see Section **Discussion**). Although the model example provided below makes use of prior information about protein family relationships in order to determine an optimal dimensionality, the method is independent of any specific alignment information.

A measure of relatedness: pairwise cosine values

In order to derive estimates of relatedness from the vector definitions of biomolecules obtained via SVD-DR, pairwise cosine values are calculated. Cosines are a standard measure of vector similarity, and their application for this purpose can be understood intuitively as follows. If the angle between two vectors in n-dimensional space is small, then the individual elements of their vectors must be very similar to each other in value, and the calculated cosine derived from these values is near one; if the vectors point in opposite directions, then the individual elements of their vectors must be very dissimilar in value, and the calculated cosine is near minus one. As applied to vectors representing proteins, the same relationships hold and can be interpreted as relative protein similarities. For example, if two proteins differ by only a few amino acids, then their vector representations expressed as a very similar set of peptide frequency elements will be separated by a very small angle in multidimensional space. If, on the other hand, two proteins share little sequence similarity, then their vector representations expressed as a very dissimilar set of peptide frequency elements will be separated by a relatively large angle in the same space.

Cosines are by no means the only possible measure of vector similarity. Two alternatives include simple Euclidean distance (the distance between the two end-points defined by the vectors), and vector length. Length by itself would understandably be a poor measure of protein similarity for most complex datasets, but could contribute greatly to similarity measures when the dataset consists only of small sets of highly conserved proteins. Euclidean distance measures perform reasonably well under some conditions, but since they are largely affected by vector length, they may suffer in this application from the fact that proteins containing multiple copies of a single domain would have much longer vector representations than their one-domain counterparts, and might therefore be judged to be inappropriately dissimilar from them. This example provides a simple rationale for considering vector length to be largely irrelevant and potentially misleading as a measure of similarity. A theoretical and empirical justification for the use of cosines to measure relatedness, especially as it applies to text retrieval and artificial intelligence, can be found in Rehder *et al.* (1998).

The appropriateness of distance measures provided by angles between vectors in Euclidian space can be further examined by considering the standard ‘distance’ qualities: positive definiteness, symmetry, and triangle inequality. If we assume that vector lengths can be profitably ignored, then the acute angle between two vectors is zero if and only if they are equivalent and is positive otherwise (definiteness). One and only one acute angle separates any two vectors in space (symmetry). The angle between any two vectors is always less than or equal to the sum of the angles between each of those vectors and a third vector (triangle inequality). By extension, cosines of angles and the angle-based evolutionary distance formula provided below also exhibit these qualities.

Measuring species relatedness: species vector sums

Protein vectors of reduced dimension derived via SVD-DR should provide precise relative definitions of proteins/genes for the derivation of phylogenetic trees using angle-based distance measures. Such an analysis can be used to provide relative definitions of all the proteins within any given organism (or organelle) for which a complete sequence is known (Figure 1a). The complete set of protein vectors from a given organism, if considered as a group, would likewise provide a precise biomolecular definition of the organism. It follows that if multiple complete genomes are included within a single analysis, then relative species definitions could be obtained by summing the individual protein vectors for each species and determining the angle between the species specific ‘super-vectors’ obtained (Figure 1b). Very similar species are expected to produce Species Vector Sums (SVSs) having very high pairwise cosine values, whereas very different species would exhibit very low SVS cosine values. Species trees based on an optimal, high dimensional comparison of the coding sequences of complete genomes should prove to be very accurate. In addition, such definitions are intrinsically relativistic and flexible; as the available biomolecular sequence databases grow, additional sequence information can be added to the input to produce readjusted phylogenies.

Reasonable alternatives to the use of vector sums to define species are the use of vector averages and the use of normalized vector sums. Vector averaging merely produces shorter vectors having the same relative angle in space, but since only the angle is used as a distance measure, the end result would be identical to that obtained with vector sums. Normalizing the vectors before summation would cause individual proteins to contribute equally to the definition of species; however, longer proteins should on average contain more information about the definition of species than shorter proteins. Vector sums (or averages) have the advantage of causing longer sequences to contribute more heavily to the definition of species.

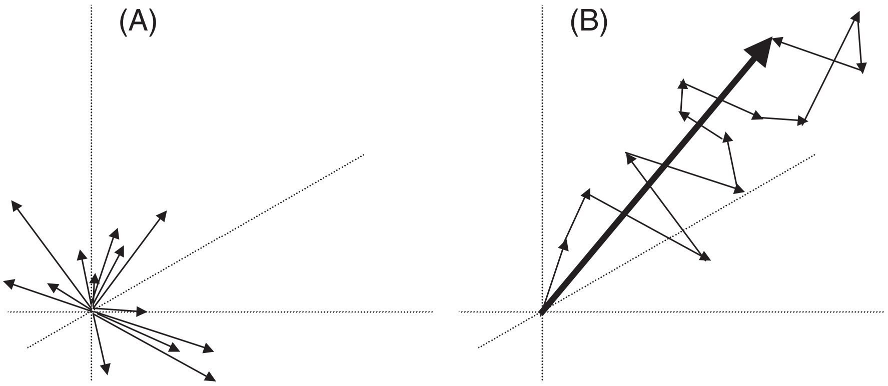

Fig. 1. Hypothetical protein vectors from a mitochondrial genome (A) and the resulting SVS (B). Vectors used in this work use approximately 40 dimensions rather than the 3 shown here.

ALGORITHM

The following steps are used to produce optimized gene and species trees for a given dataset.

Step 1

The input dataset of whole genome protein sequences representing multiple species is converted into a large, sparse n-gram frequency matrix (matrix A). Typically, overlapping 3-grams (tripeptides) or 4-grams (tetrapeptides) are used, as these are relatively rare patterns likely to represent real homology when found in multiple copies. This process results in a unique vector definition for each protein (and each peptide) in the dataset.

Step 2

The large, sparse n-gram frequency matrix in H-B format obtained above is decomposed into singular triplets via SVD (Berry *et al.*, 1999). This is generally the most intensive part of the computation, estimated by Landauer and Dumais (1997) to be of complexity $O(3Dz)$, where D is the number of saved dimensions, and z is the number of non-zero values in the original sparse input matrix. The leading matrix (U) and the trailing matrix (V) derived from this decomposition contain new vector definitions for the peptides and proteins respectively. Unlike the vector definitions of the original matrix, these vector definitions exist in a remodeled space where the axes represent correlated peptide motifs, rather than individual peptides or proteins (Stuart *et al.*, 2001). Initially, an overestimated dimension setting is used. Lower dimension settings are subsequently evaluated as described below.

Step 3

Pairwise cosine values are calculated for each vector combination represented in the SVD-derived protein matrix. These pairwise cosine values are then converted into evolutionary pairwise distances measures using the formula $d_{ij} = -\ln[(1 + \cos\theta)/2]$.

Step 4

The pairwise distance matrix obtained above is used to generate a phylogenetic tree using NEIGHBOR, a popular public domain program from the PHYLIP software suite. These gene trees are generated using the UPGMA option of NEIGHBOR.

Step 5

The gene tree obtained above is evaluated by determining the extent to which homologous proteins are placed within contiguous groups. Steps 2–5 are then repeated until a dimension setting that minimizes the gene grouping error is determined.

Step 6

The SVD-derived protein vectors obtained at the optimal dimension setting are summed for each species to produce ‘super-vectors’ representing each species in the dataset. An optimized species tree is then produced following a procedure analogous to that described in Steps 3 and 4 above. The more rigorous Neighbor Joining (NJ) option of NEIGHBOR is used to build species trees.

IMPLEMENTATION

We chose a subset of 34 completely sequenced mammalian mitochondrial genomes available from NCBI as a pilot dataset for several reasons. First, this dataset is relatively small and simple, coding for only 442 proteins within 13 well delineated families. Due to its relative simplicity, this small dataset is expected to be highly correlated and relatively ‘noiseless’. Second, this dataset is curated by the NCBI and is expected to be very accurate. Third, the orthologous relationships of the proteins within

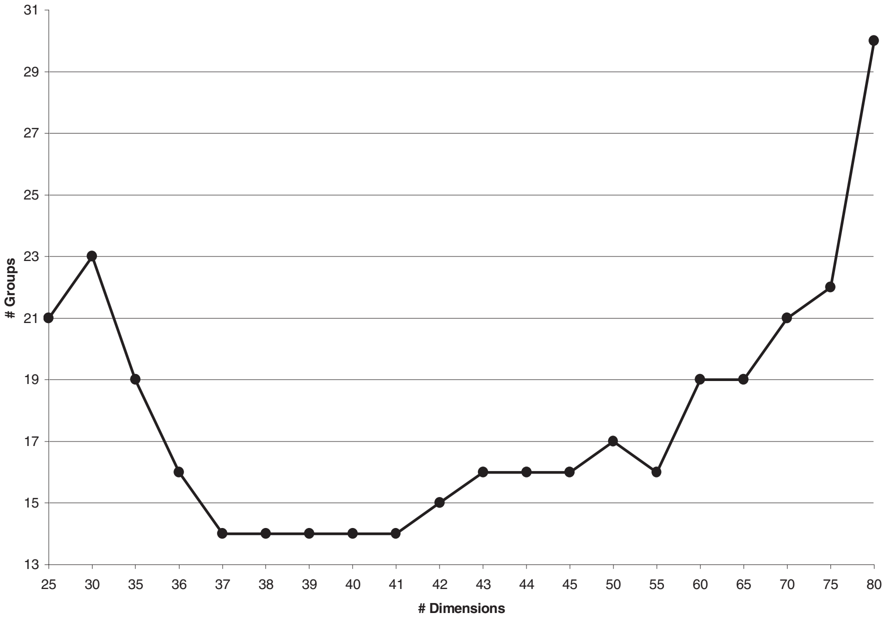

Fig. 2. Optimizing dimensions, Individual UPGMA trees derived using from 25 to 80 SVD-based dimensions were examined for the optimal grouping of 13 families of mitochondrial proteins from 34 mammalian genomes (442 total proteins). Values along the y-axis were determined by manually counting the number of 'groups' within each tree. A group is defined as an unbroken collection of immediately adjacent proteins, all of which are members of a single protein family. Settings between 37 and 41 dimensions produced a similar set of optimized trees containing only one extra group (14 rather than 13). See Figure 3 for an example at 38 dimensions.

each of the 13 families of mitochondrial proteins are obvious and undisputed. Hence the correct gene tree is ideally expected to consist of 13 branches containing 34 proteins each. Fourth, mitochondrial sequences are frequently used to generate metazoan phylogenies, hence the species trees to be derived in our analyses can be compared to those generated by other methods. In particular, the 34 species used here correspond exactly to those used to derive a recently published phylogeny of mammals (Reyes *et al.*, 2000).

An optimized phylogeny of mammalian mitochondrial proteins

The 34 genome dataset was first used to produce a 160 000 word by 442 protein data matrix that represented all of the proteins in the dataset in terms of their overlapping tetrapeptide frequencies. Following SVD-DR, pairwise cosine values were calculated for each pair of proteins and used to generate a distance matrix for input into the NEIGHBOR program of PHYLIP. Using the UPGMA option of NEIGHBOR, a series of gene trees was produced that corresponded to a series of different reduced dimension settings ranging from 25 to 80. For making these very large gene trees, the rapid UPGMA option was chosen over the NJ option as a balance between speed and rigor. Each of the resulting trees was then evaluated by determining how well related genes were placed into contiguous groups (Figure 2). Dimension settings between 37 and 41 appeared to produce a roughly equivalent set of trees in which all the 442 proteins were placed into just 14 groups, one more than the 13 groups expected in a perfect tree. A detailed inspection of these 5 trees revealed that in every case, a variable number of ND5 sequences were misplaced within a given tree (not shown). Although all these trees represented a reasonably accurate global hypothesis of sequence relatedness, the tree at 38 dimensions was particularly attractive, as only three members of the ND5 family were misplaced as a single group in the tree (Figure 3). In contrast, the other 4 trees contained between 11 and 14 misplaced ND5 sequences. In any case, the branch structure of these trees indicates that the vast majority of proteins were well recognized and placed within complete 34 member groups. A detailed version of the 34 member ND2 branch is provided as an example (Figure 4). The topology of this branch can be interpreted as a 'single gene's opinion' of the evolutionary relationships of the organisms represented. Other subtrees formed by other mitochondrial gene families may support alternative relationships (see Cao *et al.*, 1998).

Note that this analysis requires that all sequences, even significantly unrelated ones, be placed into specific relationships within the tree. Although many of the

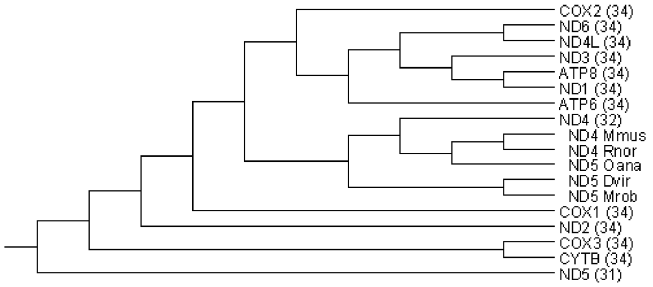

Fig. 3. Compressed version of the topological UPGMA tree derived from 442 mitochondrial proteins represented as vectors at 38 dimensions. Only one of the 13 families of sequences is not represented in the tree as a contiguous group of 34 members. Species origins for some sequences are specified by a four letter designation (first letter of genus name, first three letters of species name). Of the 442 sequences in the tree, only 3 of the ND5 sequences, those from non-eutherian mammals, fail to associate with the remaining members of its family, grouping instead with the ND4 family.

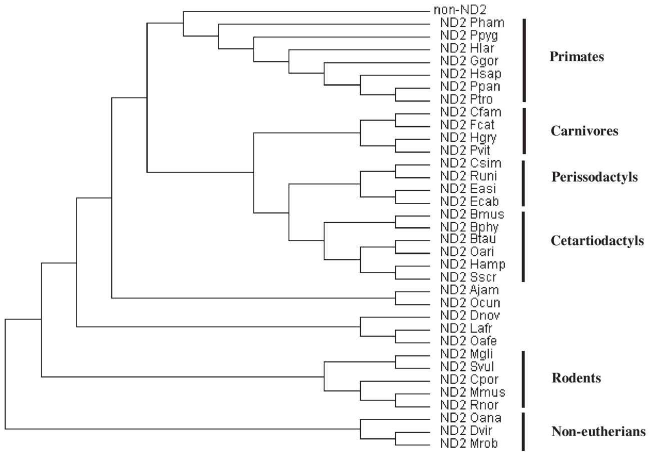

Fig. 4. The 34 member ND2 branch of the UPGMA tree at 38 dimensions summarized in Figure 3. This tree was arbitrarily rooted using the ND2 sequences from the non-eutherian mammals. Species origins are specified by a four letter designation (first letter of genus name, first three letters of species name).

13 mitochondrial gene families may share a significant number of important characteristics and may even be to some extent functionally similar, the specific relationships postulated by the deepest branches of these trees are unlikely to be meaningful.

An integrated gene to species phylogeny of mammals

In order to generate the species tree, all 13 vectors from each organism were added to make ‘SVSs’ (or supervectors) that represent species in 38 dimensional space. Again, following the calculation of a cosine-based distance matrix, a tree was generated using the NJ option of NEIGHBOR (Figure 5). The NJ option is generally considered to be more rigorous than the faster UPGMA option used above for generating gene trees, but since the species trees had far fewer nodes, tree building was reasonably rapid. The non-eutherian mammals (monotreme and marsupials) were chosen as the conventional species to root the tree (Janke *et al.*, 1997). The primates, rodents, cetartiodactyls, carnivores, and perissodactyls are some of the well recognized mammalian lineages that appear as uninterrupted groupings within this tree as well as in previously published trees (e.g. Xu *et al.*, 1996; Janke *et al.*, 1997). Overall, this tree is topologically very similar to one generated by Reyes *et al.* (2000) using the same set of species in a maximum likelihood analysis of mitochondrial gene sequences (Figure 6). Both trees, as well as other published results, support a close relationship between whales and hippos (Arnason *et al.*, 2000), between bats and ferungulates (Nikaido *et al.*, 2000), and between squirrels and other rodents (Reyes *et al.*, 2000). The most obvious difference between the two trees is that the murid rodents of the latter tree appear to diverge before the primates, and are therefore separated from the non-murid rodents. Other analyses also suggest

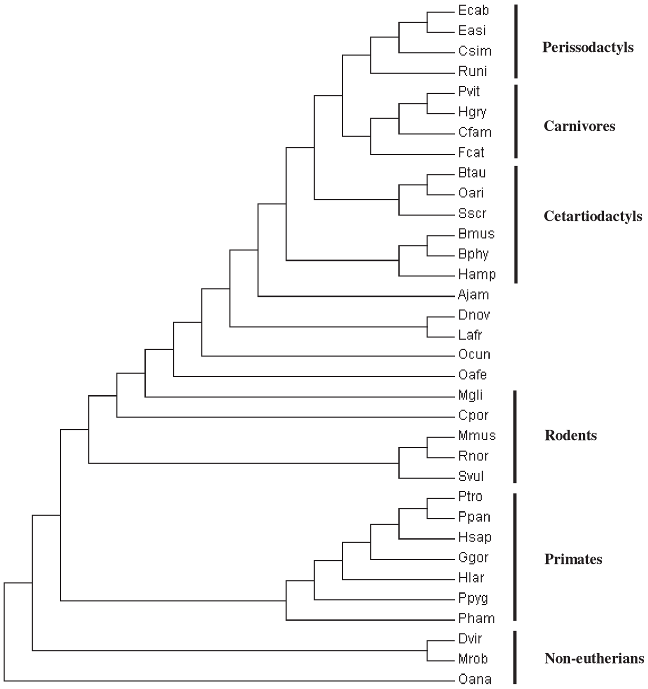

Fig. 5. Topological NJ species tree for 34 mammals derived from SVSs of 442 mitochondrial sequences represented as vectors at 38 dimensions. The corresponding gene tree is shown in Figure 3.

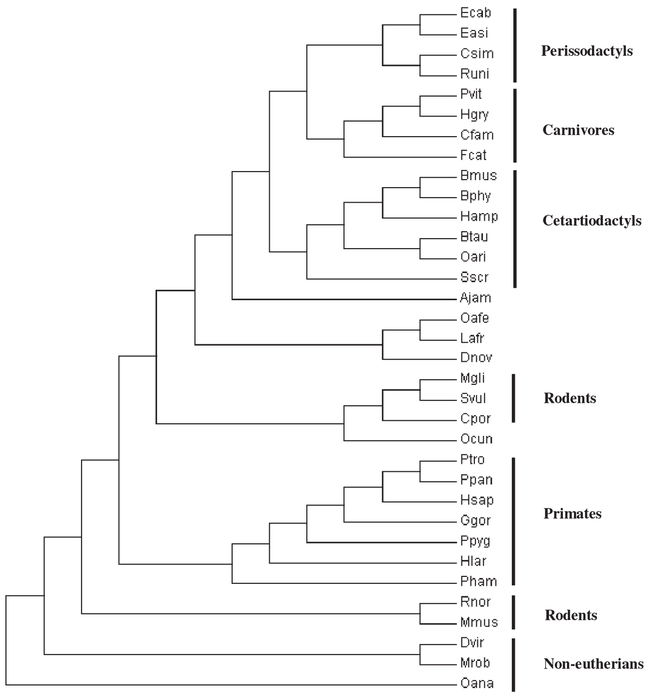

Fig. 6. Topological ML species tree redrawn form Reyes *et al.* (2000). Though similar to the tree presented in Figure 5, rodents are indicated as polyphyletic, with the murid rodents branching earlier than primates.

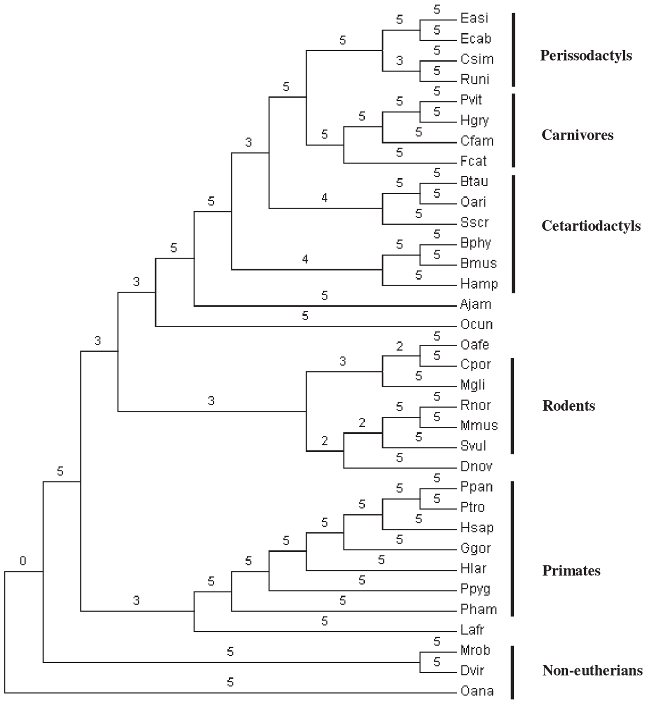

Fig. 7. Consensus NJ species tree derived from the 5 optimal species trees obtained using between 37 and 41 dimensions. The number of individual trees within which particular taxa or groups of taxa were found is indicated above the branches. Only groups including Dnov, Oafe, and Svul within the 'rodent' branch were found in less than half the trees.

that at least some rodents diverged prior to primates (see Grundy and Naylor, 1999). Both primate and rodent mitochondrial genomes are believed to have evolved rapidly and could conceivably root together deep in the tree for this reason alone (Gissi *et al.*, 2000). However, a recent comprehensive molecular analysis of mammalian phylogeny suggests a close relationship between rodents and primates (Murphy *et al.*, 2001).

Although rodent monophyly is an attractive hypothesis supported by our analysis at 38 dimensions, a consensus tree generated by combining the 5 'optimal' analyses between 37 and 41 dimensions suggests that rodent relationships are not well resolved. Several relationships within the large branch that contains the rodents appear in only 2 out of 5 trees (Figure 7). This branch of the consensus tree also contains the non-rodents armadillo (Dnov) and aardvark (Oafe). Another relatively uncertain relationship involves the elephant (Lafr), which groups with the primates in 3 out of 5 trees. In general, however, the remaining relationships are reasonably well supported and coincide well with the trees presented in Figures 5 and 6.

DISCUSSION

We have shown that it is possible to build accurate gene and species trees with a novel pattern based method utilizing all possible overlapping tetrapeptide words (4-grams). This is essentially a distance method that uses total genome data as input, but does not generate or require explicit sequence alignments. Both gene trees and species trees derive from the same analysis, providing straightforward biomolecular species (organelle) definitions based on total genome content. Although this work focused on tetrapeptide patterns within the dataset, both larger and smaller n-grams are feasible. In fact, a similar analysis using tripeptides produces gene and species trees that are nearly as accurate as those of the tetrapeptide analysis presented here (not shown). The application of larger n-grams and multiple size n-grams are currently being explored.

No rigorous statistical evaluation of the trees generated using this method is currently available. As we have done here, a rough idea of the stability of optimized trees can be obtained by generating consensus trees from the set of trees present in or near the 'optimization well' revealed in the dimensional analysis phase (see Figure 2). Although only a few trees derived from this analysis would likely be considered optimized, trees to be included in the consensus could be derived from a greatly expanded number of dimensions, potentially producing a somewhat stronger tree statistic.

The vector sum method used here to compare species depends critically on accurate relative vector definitions for each protein. If each protein in each species is defined in relative vector space with the highest possible accuracy, then the sum of these vectors should most accurately define a given species. Logically, then, an optimal dimension setting would be one that at least allows the method to group each and every protein in the dataset with all other closely related proteins (i.e. in effect 'recognizes' proteins as members of a given family). With 'known' problems like the mitochondrial dataset used here, the optimal dimension can be estimated by examining the extent to which the known orthologs cluster into 13 well defined families. Although a relatively broad 'well' of optimal dimensions was observed, the species trees produced from within that well were consistently similar (see Figure 7), indicating that the information extracted in the analysis was consistent in its support for the derived relationships.

The optimal number of dimensions to use with other datasets may depend on many factors, including the size of the dataset, the nature of the sequences, the overall relatedness of the sequences, the repetitiveness of the sequences, and the size of the n-grams. For larger and more complex 'unknown' datasets, the optimal dimension could, to some extent, be estimated by comparison to previously published hypotheses of relatedness concerning the sequences involved. The database of COGs (Clusters of Orthologous Genes) maintained by the NCBI (Tatusov *et al.*, 2000) may prove useful in this regard; an optimal dimension setting should produce an optimal clustering of the members of identified COGs within the gene tree. However, 'tuning' the trees in this way would tend to reinforce the importance of contiguous groups of n-grams (i.e. larger motifs corresponding to alignments), since COGs are based primarily on local alignments. Thus, the method would become indirectly dependent on these alignments.

Although our method currently requires prior knowledge of family relationships for optimization, it is not the case that merely using information about family relationships is the same as using alignment data. The family information used for optimization was categorical information only. No information about alignment was transferred to the method, hence the method does not require any specific alignment information (beyond tetrapeptide matches). While it is true that family relationships frequently derive from alignments, it is also true that family information can be obtained without alignments. For instance, gene family relationships in mammalian mitochondrial genomes can be assigned based on relative position in the genome alone. From this perspective, even indirect alignment information was unnecessary in our example. Logically, of course, alignments should generally be obtainable once the family relationships are known. Family grouping information was used in our method only to uncover an optimal dimensionality. If dimensionality was directly calculable from the SVD output (perhaps one using excessive dimensions), then family information would not be required *a priori*, but would instead be obtained as an output of the analysis (i.e. from the gene tree produced). Mathematical methods for estimating dimensionality directly from sparse matrix data are currently being investigated (Ding, 1999), and may be adaptable for our use.

Recent work (Stuart *et al.*, 2001, unpublished results) indicates that each of the singular vectors obtained via SVD-DR represents an idealized peptide motif that is conserved within particular protein families of the dataset. Each singular vector is an orthonormal basis vector for the remodeled definition space, and as such precisely defines a given motif in terms of the contribution provided by each of the large number of possible peptide vectors. Thus, peptides represented by vectors with large projections on a given singular vector will have contributed substantially to the definition of that singular vector, and will therefore be seen to have a relatively high degree of conservation within the motif defined by that singular vector. Motifs defined by singular vectors that correspond to larger singular values contain specific sets of peptides that co-occur with a relatively high frequency in the dataset. Although many of these co-occurring peptides are adjacent and can be aligned within a given motif, adjacency and alignment are not required. Since the motifs are defined by peptide co-occurrence only, some of the individual peptides of a motif may be separated by a variable number of amino acid residues. From an evolutionary perspective, this would allow insertions and deletions to have little impact on observed similarity.

In limited but relatively complex datasets, many proteins may lack both close paralogs and close orthologs. Vector representations for these proteins are more likely to be erroneous, as their position in vector space would be based on weak, perhaps insignificant peptide co-occurrence frequencies. These potentially 'poorly defined' vectors could add noise to the derivation of species trees that follows vector addition. Larger datasets are expected to improve estimates of relatedness by providing multiple versions of homologous proteins to be compared. For example, a 64 species vertebrate mitochondrial dataset appears to produce a robust phylogeny that corresponds well with accepted phylogenetic theory (Stuart *et al.*, 2001). We have also begun work on a 44 species bacterial genome COG dataset, as well as a complete bacterial genome dataset representing less than 20 genomes. With respect to the latter, we expect the bacterial species definitions obtained using this method to appropriately reflect species-specific gene duplications and deletions. Furthermore, lateral gene transfer events are likely to be recognized and incorporated into the global definition of species.

Assuming that methods for finding optimal dimensionality without *a priori* reference to gene family information are adaptable to our analyses, then a single SVD-DR can be used to simultaneously provide: (1) precise descriptions of all significantly conserved motifs within a dataset of protein sequences; (2) protein/gene family relationships in the form of a gene tree; and (3) species family relationships in the form of a species tree. This method may then be useful in the future to help annotate newly sequenced whole genomes by providing functional predictions for virtual proteins based on rigorous but flexible estimates of relative similarity to all the proteins present in a large number of other genomes, some of which may be well described. We therefore would like to propose the methods and ideas presented in this paper as a reasonable basis for the development of an 'intellectual database of genome comparisons', the absence of which was lamented in a recent editorial (Koonin, 1999).

ACKNOWLEDGEMENTS

We are grateful for the advice and support of the interdisciplinary 'Complexity' group comprised of members from Indiana State University, Rose-Hulman Institute of Technology, Saint-Mary-of-the-Woods College, and Ivy Tech State College of Terre Haute. This work would not have been accomplished without the insightful encouragement of Thomas Landauer.

REFERENCES

Arnason,U., Gullberg,A., Gretarsdottir,S., Ursing,B. and Janke,A. (2000) The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. *J. Mol. Evol.*, **50**, 569–578.

Berry,M.W., Drmac,Z. and Jessup,E.R. (1999) Matrices, vector spaces, and information retrieval. *SIAM Rev.*, **41**, 335–362.

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

Ding,C.H.Q. (ed) (1999) A similarity-based probability model for latent semantic indexing. *Proceedings of 22nd ACM SIGIR Conference (SIGIR'99)*. pp. 59–65.

Gissi,C., Reyes,A., Pesole,G. and Saccone,C. (2000) Lineage-specific evolutionary rate in mammalian mtDNA. *Mol. Biol. Evol.*, **17**, 1022–1031.

Grundy,W.N. and Naylor,G.J. (1999) Phylogenetic inference from conserved sites alignments. *J. Exp. Zool.*, **285**, 128–139.

Janke,A., Xu,X. and Arnason,U. (1997) The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl Acad. Sci. USA*, **94**, 1276–1281.

Koonin,E. (1999) Why genome analysis? *Trends Genet.*, **15**, 131.

Koonin,E.V., Tatusov,R.L. and Galperin,M.Y. (1998) Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.*, **8**, 355–363.

Lake,J.A. and Moore,J.E. (1998) Phylogenetic analysis and comparative genomics. *Trends Guide to Bioinformatics, Trends Journal Supplement 1998*, 22–23.

Landauer,T.K. and Dumais,S.T. (1997) A solution to Plato's problem, the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psych. Rev.*, **104**, 211–240.

Landauer,T.K., Foltz,P.W. and Laham,D. (1998) Introduction to latent semantic analysis. *Discourse Process.*, **25**, 259–284.

Li,M., Badger,J.H., Chen,X., Kwong,S., Kearney,P. and Zhang,H. (2001) An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, **17**, 149–154.

Lin,J. and Gerstein,M. (2000) Whole-genome trees based on the occurrence of folds and orthologs, implications for comparing genomes on different levels. *Genome Res.*, **10**, 808–818.

Murphy,W.J., Eizirik,E., Johnson,W.E., Zhang,Y.P., Ryder,O.A. and O'Brian,S.J. (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.

Nikaido,M.M., Harada,M., Cao,Y., Hasegawa,M. and Okada,N. (2000) Monophyletic origin of the order chiroptera and its phylogenetic position among mammalia, as inferred from the complete sequence of the mitochondrial DNA of a Japanese megabat, the Ryukyu flying fox (*Pteropus dasymallus*). *J. Mol. Evol.*, **51**, 318–328.

Rehder,B., Schriener,M.E., Wolfe,M.B.W., Laham,D., Landauer,T.K. and Kintsch,W. (1998) Using latent semantic analysis to assess knowledge: some technical considerations. *Discourse Process.*, **25**, 337–354.

Reyes,A.C., 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.

Stuart,G.W., Moffet,K. and Leader,J.J. (2001) A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes, submitted.

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

Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Tekaia,F., Lazcano,A. and Dujon,B. (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, **9**, 550–557.

Thorne,J.L. (2000) Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.*, **10**, 602–605.

Xu,X., Janke,A. and Arnason,U. (1996) The complete mitochondrial DNA sequence of the greater indian rhinoceros, *Rhinoceros unicornis*, and the phylogenetic relationship among carnivora, perissodactyla, and artiodactyla. *Mol. Biol. Evol.*, **13**, 1167–1173.