

Molecular Biology and Evolution

Revised December 12, 2001

**A Comprehensive Vertebrate Phylogeny using Vector
Representations of Protein Sequences from Whole Genomes**

Gary W. Stuart*

Department of Life Sciences, Indiana State University
Terre Haute IN 47809; G-Stuart@indstate.edu

Karen Moffett

Department of Life Sciences, Indiana State University
Terre Haute IN 47809; P-Moffett@indstate.edu

Jeffery J. Leader

Department of Mathematics, Rose-Hulman Institute of Technology
Terre Haute IN 47809; Jeffery.Leader@Rose-Hulman.edu

* corresponding author
email: G-Stuart@indstate.edu
phone: (812) 237-7898
fax: (812) 237-4480

Running Title: Whole Genome Phylogeny using Vector Representation

**Keywords: genomics, mitochondrial DNA, molecular phylogenetics,
molecular systematics, sequence analysis, singular value decomposition**

ABSTRACT

We recently developed a method for producing comprehensive gene and species phylogenies from unaligned whole genome data using singular value decomposition (SVD) to analyze character string frequencies. This work provides an integrated gene and species phylogeny for 64 vertebrate mitochondrial genomes composed of 832 total proteins. In addition, to provide a theoretical basis for the method, we present a graphical interpretation of both the original frequency matrix and the SVD-derived matrix. These large matrices describe high-dimensional Euclidean spaces within which biomolecular sequences can be uniquely represented as vectors. In particular, the SVD-derived vector space describes each protein relative to a restricted set of newly defined, independent axes, each of which represents a novel form of conserved motif, termed a "correlated peptide" motif. A quantitative comparison of the relative orientations of protein vectors in this space provides accurate and straightforward estimates of sequence similarity, which can in turn be used to produce comprehensive gene trees. Alternatively, the vector representations of genes from individual species can be summed, allowing species trees to be produced.

INTRODUCTION

Many molecular phylogenies are based on sequences sampled from only one or a few genes. With the advent of efficient whole genome sequencing, it is now possible to consider building phylogenies based on total genome sequence. Given that most genomes contain millions to billions of sequence characters, standard methods based on character-by-character comparisons performed over ambiguously resolved large scale alignments become impractical. Consequently, several methods for comparing whole genomes have been proposed and practiced with some degree of success. These methods rely largely on some form of data compression in which previously distinct but similar characters are redefined as indistinguishable members of broadly defined groups. Hence, genomes have been compared by dinucleotide content (Nakashima et al., 1998), gene order (Boore and Brown, 1998), categorical gene content (Tekai et al., 1999; Fitzgibbon and House, 1999; Snel et al., 1999; Lin and Gerstein, 2000), or estimates of relative information content (Li et al., 2001).

An alternative approach makes use of data selection rather than data compression. In this case, attempts are made to select the most informative character comparisons for analysis (e.g. Grundy and Naylor, 1999). Whereas these methods are likely to be more rigorous than data compression, they are also more demanding in terms of computational time and resources. However, rapid increases in computational power and whole genome information constantly force a re-evaluation of these methods for comparing biomolecular sequences (Galperin and Koonin, 2000).

In this report, we describe a method for generating phylogenetic trees from whole genome data. Following the conversion of a large multi-genome dataset into a matrix of short character string

frequencies, the matrix analysis method referred to as singular value decomposition (SVD) is applied. The output of the SVD can be used to 1) select the most informative biomolecular sequence characteristics for comparison, and 2) estimate evolutionary distances between sequences using these selected characteristics. This application of SVD to describe biomolecular sequences is similar to its application to large compilations of text documents (Landauer et al, 1998; Berry et al., 1999). The method, as currently designed, operates only on protein sequences, and can be applied to all genome sequences that are accompanied by nearly complete sets of predicted coding regions. The optimized pairwise protein distances obtained following SVD can be used directly to generate a multi-organism gene tree, or combined by species to generate a derived species tree. A distinct advantage of the method is that sequence alignments are not required (Stuart et al., 2001).

In this study, we generate integrated gene and species trees for 832 mitochondrial proteins obtained from 64 whole vertebrate genomes. We also explain the technique in terms of abstract multi-dimensional vector spaces and interpret the SVD-derived independent characteristics as representing novel forms of correlated sequence motifs. This application sets the stage for future analyses involving a variety of more complex whole genome datasets.

MATERIALS AND METHODS

Sequence Data

The complete set of protein sequences from 64 whole mitochondrial genomes was obtained from the NCBI genome database. Species represented in the analysis include the following: *Alligator mississippiensis* (Amis), *Artibeus jamaicensis* (Ajam), *Aythya americana* (Aame), *Balaenoptera musculus* (Bmus), *Balaenoptera physalus* (Bphy), *Bos taurus* (Btau), *Canis familiaris* (Cfam), *Carassius auratus* (Caur), *Cavia porcellus* (Cpor), *Ceratotherium simum* (Csim), *Chelonia mydas* (Cmyd), *Chrysemys picta* (Cpic), *Ciconia boyciana* (Cboy), *Ciconia ciconia* (Ccic), *Corvus frugilegus* (Cfru), *Crossostoma lacustre* (Clac), *Cyprinus carpio* (Ccar), *Danio Rerio* (Drer), *Dasyopus novemcinctus* (Dnov), *Didelphis virginiana* (Dvir), *Dinodon semicarinatus* (Dsem), *Equus asinus* (Easi), *Equus caballus* (Ecab), *Erinaceus europaeus* (Eur), *Eumeces egregius* (Eegr), *Falco peregrinus* (Fper), *Felis catus* (Fcat), *Gadus morhua* (Gmor), *Gallus gallus* (Ggal), *Gorilla gorilla* (Ggor), *Halichoerus grypus* (Hgry), *Hippopotamus amphibius* (Hamp), *Homo sapiens* (Hsap), *Latimeria chalumnae* (Lcha), *Loxodonta africana* (Lafr), *Macropus robustus* (Mrob), *Mus musculus* (Mmus), *Mustelus manazo* (Mman), *Myoxus glis* (Mgli), *Oncorhynchus mykiss* (Omyk), *Ornithorhynchus anatinus* (Oana), *Orycteropus afer* (Oafer), *Oryctolagus cuniculus* (Ocun), *Ovis aries* (Oari), *Paralichthys olivaceus* (Poli), *Pelomedusa subrufa* (Psub), *Phoca vitulina* (Pvit), *Polypterus ornatipinnis* (Porn), *Pongo pygmaeus abelii* (Ppya), *Protopterus dolloi* (Pdol), *Raja radiata* (Rrad), *Rattus norvegicus* (Rnov), *Rhea americana* (Rame), *Rhinoceros unicornis* (Runi), *Salmo salar* (Ssal), *Salvelinus alpinus* (Salp), *Salvelinus fontinalis* (Sfon), *Scyliorhinus canicula* (Scan), *Smithornis sharpei* (Ssha), *Squalus acanthias* (Saca), *Struthio camelus* (Scam), *Sus scrofa* (Sscr), *Talpa europaea* (Teur), *Vidua chalybeata* (Vcha).

SVD-based Protein Sequence Representation

Protein sequences were recoded as tetrapeptide frequency values using all possible overlapping tetrapeptides. SVD was performed on the resulting tetrapeptide vs. protein data matrix applying the program "bls-1" from SVDPACK (Berry, 1992). Dimension reduction was accomplished by setting the smaller singular values resulting from SVD equal to zero. The optimal dimension reduction was estimated by evaluating the gene tree output for a given dimension setting as described below.

Vector definitions for all 832 mitochondrial proteins were retrieved from the SVD output and used to calculate pairwise cosine values, which were subsequently converted into a matrix of pairwise evolutionary distances using the formula $d_{ij} = -\ln [(1+\cos\theta)/2]$. This formula is derived from the standard distance formula $d = -\ln S$, and converts a similarity measure ranging from 0 to 1 into a distance measure potentially ranging from 0 to infinity. Corresponding species distances were estimated by summing the 13 vector definitions for each organism, then using the pairwise cosine values between the derived species vectors to estimate evolutionary distance (Stuart et al., 2001).

Tree Generation and Analysis

Distance matrices derived via SVD were converted into phylogenetic trees using the NEIGHBOR program from PHYLIP (Felsenstein, Univ. of Washington). The faster UPGMA option of Neighbor was used to estimate the computationally intensive gene trees (832 OTU's).

Corresponding species trees (64 OTU's) were estimated using the NJ option of NEIGHBOR. Gene trees were evaluated by determining how well individual members of the 13 families of

mitochondrial genes were sorted into uninterrupted family groupings. Adjacent members of the same family were counted as a "group".

RESULTS AND DISCUSSION

Representing proteins and peptides as vectors

Each sequence was recoded initially as a set of tetrapeptide frequency values. Tetrapeptides are used because their frequencies of occurrence in random sequence are very low (about 1 in 160,000); hence, particular tetrapeptides tend to be present within a protein due to their contributions, however slight, to the particular functional role of that protein, not as the result of some random choice. Tripeptides could also be used, but because they appear about 20 times more frequently, they would introduce more "noise" into the analysis. Whereas tripeptides may prove useful with highly diverged sequences (Stuart et al., unpublished), tetrapeptides perform better with the mitochondrial protein dataset used herein. The utility of even larger peptides has yet to be explored, but as the size of the character string (peptide n-gram) grows, so does the likelihood that significant real similarity, even between highly related proteins, will remain undetected.

Tetrapeptides representing a protein include all possible overlapping tetrapeptides. This has two advantages: 1) no important functional peptide present in a sequence will be missed because it is out of phase with the scanning window, and 2) overlapping peptides tend to define a unique order for peptides within the sequence. In contrast, non-overlapping windows would miss certain peptide patterns that are actually present and would provide no information about the order of the peptides

in the sequence. Overlapping, partially redundant information is required to provide unique or nearly unique reconstructions for most sequences. Hence, the overlapping tetrapeptide representation of each protein is in essence a detailed "fingerprint" that serves to uniquely identify and accurately describe the protein (Figure 1).

The end result of recoding the entire dataset in this way is a tetrapeptide frequency matrix in which each protein is represented by a column of peptide frequency values. Since there are 160,000 possible tetrapeptides, each protein is defined by 160,000 values, most of which are zero. For example, the 32 residue string in Figure 1 would be represented by a long column of zeros interrupted rarely by twenty-nine isolated "1's". This column of numbers can be viewed as a vector in an abstract space having 160,000 dimensions (the column space of the tetrapeptide frequency matrix). This provides a highly precise tool for describing proteins and allows application of the extensive mathematical tools of numerical linear algebra.

Vector representations in high-dimensional space provide easily accessible relative measures of relatedness (Figure 2A). For example, any pair of proteins that differ by a single amino acid must also differ by one to four tetrapeptides and would therefore be represented in high-dimensional space by slightly different vectors. Thus, proteins can be recognized as close relatives because their vectors are oriented in space in nearly the same direction. As proteins evolve and diverge, their vector representations in space begin to separate, increasing the measured angle between them, and reducing the calculated cosine of that angle further from its maximum value of unity. Hence, documenting the relatedness of all the sequences in a dataset can be accomplished by arranging all the cosine values into a pairwise similarity matrix, or following a simple conversion, a pairwise

distance matrix. The latter is frequently used as input to one of several standard programs used to generate evolutionary trees (e.g. NEIGHBOR).

Unfortunately, the original peptide frequency matrix frequently provides vector-based protein sequence representations of surprisingly poor quality. This is because the tetrapeptides that define the orthogonal axes of the protein definition space do not appear independently in proteins, and thus should not be equally weighted in their representations. Instead, tetrapeptides tend to appear together as correlated sets within particular protein families in order to satisfy the functional requirements of those proteins. Consequently, particular sets of correlated peptides serve to define particular sets of homologous protein motifs within the dataset. Compared to the individual peptides of which they are composed, motifs are more likely to appear independently in proteins and should therefore provide an improved set of independent characters useful for defining orthogonal basis vectors in a descriptive space.

The motifs described above are composed of two classes of correlated peptides: those peptides that appear simultaneously, perhaps as adjacent peptides within a given motif, and those that are interchangeable with other peptides and therefore appear exclusively within a given motif. A comprehensive set of interchangeable peptides accompanied by quantitative estimates of their interchangeability can serve to define a model of sequence evolution. Similar models are frequently embodied in empirically derived PAM tables. Alternative models could be based on the relative interchangeability of di-, tri-, or tetrapeptides (or peptides of any size). Whereas PAM tables have only 20^2 elements, a tetrapeptide version would have $160,000^2$ elements. The initial tetrapeptide frequency matrix described above can be used to derive a similar table representing a model of sequence evolution. Each row of protein frequency values serves to define each tetrapeptide as a

vector in peptide definition space (the row space of the tetrapeptide frequency matrix). Hence, pairwise peptide similarity can be calculated using pairwise cosine values. Peptide similarities derived in this way are somewhat different than those embodied in PAM tables because they encompass both kinds of similarity described above: simultaneous co-occurrence and interchangeability.

The peptide definition space is different from the one used earlier to define proteins, although it is constructed using the same frequency data. If there are 832 protein sequences in the dataset, then the vectors defining each tetrapeptide exist in an 832 dimensional space, which is distinct from the 160,000 dimensional space used for proteins (Figure 2). As in the case of the protein space, the peptide space described by the initial frequency matrix provides a relatively poor definition of peptides. This arises because the proteins that define the orthogonal axes in the space are not all independent; the presence of a given peptide in one member of a protein family generally predicts its presence in all members of that family.

Proteins and peptides would be more accurately defined in high-dimensional space by orthogonal axes representing truly independent characters. We suggest that motifs composed of correlated sets of peptides can provide close approximations to these independent characters (Figure 2C). Motifs defined in this way have several useful properties. As long as the correlated peptides that define the motifs remain unchanged, motifs can be of any size and may contain any number of gaps resulting from sequence insertions or deletions. We refer to these motifs as correlated peptide motifs, or simply, "copep" motifs. These relatively flexible motifs can be contrasted with those defined by the local sequence alignments typically utilized in biomolecular phylogenies. Local alignments define

motifs using models of sequence evolution that fail to account for insertions and deletions of more than a few characters (Thorne et al., 2000).

The SVD: correlated peptide motifs emerge as new orthogonal axes

Improved vector definitions for both proteins and peptides can be obtained by decomposing the initial frequency matrix A into three separate matrices U , Σ , and V using SVD (Figure 3). The three resulting matrices can be used to reconstruct the original matrix using the relation $A=U \Sigma V^T$. From a graphical perspective, the effect of the SVD is to remodel both the peptide space and the protein space such that peptide vectors and protein vectors are represented using a newly derived set of orthonormal axis vectors. Hence, the V matrix is a "protein" matrix that defines proteins as vectors using orthogonal axes that represent derived independent characteristics. The U matrix is a "peptide" matrix that defines peptides as vectors using the same set of independent characteristics. The singular values from the diagonal Σ matrix together with the corresponding rows or columns of the protein and peptide matrices comprise the singular triplets presented in rank order. The triplets with the largest singular values identify the most important independent characteristics serving to define both proteins and peptides in the dataset. As suggested earlier, our interpretation of these independent characteristics is that they represent sequence motifs comprised of correlated sets of peptides, or copep motifs. A preliminary inspection of the first few ranked singular triplets for the mitochondrial dataset analyzed here shows that this is indeed the case. For example, the 11th singular vector describes an ND3 protein motif by defining a particular linear combination of peptides that appears primarily, but not exclusively, as contiguous sequence only within this highly conserved family (not shown).

Since the number of significant copep motifs is far less than the number of possible peptides, the majority of the copep dimensions within the derived protein matrix must have singular values that are nearly zero. In fact, the ranked singular values for the mitochondrial dataset decay rapidly, becoming effectively zero after the first few hundred dimensions. Setting these minor values to exactly zero is a standard and widely used technique that reduces the dimensionality of the vector space description of proteins to a manageable number (hundreds rather than hundreds of thousands).

Reduced dimensionality has the effect of making the problem of vector comparisons computationally approachable. In addition, and perhaps more importantly, it tends to improve the relative accuracy of protein vector definitions by discarding a substantial fraction of the noise (including homoplasy) in the data. If $\sigma_1, \dots, \sigma_r$ are the positive singular values of the tetrapeptide frequency matrix, then the singular vectors associated with any particular singular value (e.g. σ_j) accounts for the fraction

$$\sqrt{\frac{\sigma_j^2}{\sigma_1^2 + \dots + \sigma_r^2}} \quad (1)$$

of the data; choosing the largest singular value, σ_1 , maximizes this value, and choosing the k largest values ($k < r$) explains the fraction

$$\sqrt{\frac{\sigma_1^2 + \dots + \sigma_k^2}{\sigma_1^2 + \dots + \sigma_r^2}} \quad (2)$$

of the data (Berry et al 1999). If the first $k = 100$ singular values derived from our tetrapeptide frequency matrix are sufficiently large, they would represent a subset of motifs that explain nearly the entire matrix, allowing reconstruction of the matrix from a subset of its singular values and singular vectors. The formula for this reconstruction (based on the first k singular triplets) is

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^T. \quad (3)$$

Determining the number (k) of ranked singular values that best serves to separate signal from noise within a dataset is difficult. With the mitochondrial dataset considered here, it is possible to estimate the optimal number of ranked singular values by measuring how well the 832 sequences are grouped into the 13 known mitochondrial gene families (Figure 4). In our example, as the number of included dimensions increases from 10 to about 60, the UPGMA tree calculated from the vector-based pairwise distance matrix steadily improves. At 66, 73, and 74 dimensions, an optimum of 16 groupings is observed in the tree. At higher dimensions, the tree begins to deteriorate again. Our conclusion is that this 832 protein dataset is best described by 64 to 80 copep motifs distinguishable within the 13 mitochondrial protein families (roughly 5-6 per protein); that is, we should retain the motifs defined by the singular vectors associated with the largest 64 to 80 singular values as useful "signal" and discard the remainder (out of 832) as distracting "noise" which has degraded the integrity of the data.

Dimension reduction frequently fails to produce trees with the minimum number of monophyletic branches (in this case, 13). Apparently not all of the homoplasy within datasets can be separated from useful information. Perhaps some of the copep motifs that help to define certain protein families contain subgroups of peptides that are also present in other, less related protein families by chance, or even by convergent evolution (see Figure 1 for an example). As a result, some less related proteins may group with the wrong family, depending on which dimension setting (i.e. which set of copep motifs) is used in the analysis.

Integrated gene and species trees for 64 vertebrates based on whole mitochondrial genome sequence.

The 832 member gene tree at 74 dimensions is summarized in Figure 5. This tree is one of three in which a minimal number of groupings (16) was observed. In this tree, all sequences from a given gene family that appear within monophyletic groups have been compressed into a single branch for efficient presentation. Only CYTB, ND3, and ND5 fail to form single monophyletic branches containing all 64 sequences. In addition, both CYTB and ND5 have only one sequence each that fails to group with their remaining family members, while ND3 is separated into one mammalian branch (31 genes) and one non-mammalian branch (33 genes). Although the deeper branches within this gene tree may reflect possible evolutionary relationships between certain mitochondrial gene families, we place little significance on these distant relationships, as all these sequences, by default, appear in the tree independent of their degree of relatedness. The relationships of the more similar sequences within each family are more useful.

Many of the branches of the gene tree of Figure 5, if presented in uncompressed form, would provide a distinct, single gene hypothesis of the evolutionary relationship between the species represented. These multiple hypotheses can be synthesized into a global hypothesis by deriving a species tree based on vector representations of all the proteins in the dataset. The species tree is produced from a pairwise comparison of a set of derived species vectors, which are formed by vector addition using only protein vectors from the same organism. The accuracy of the derived species tree would therefore be directly dependent on the accuracy with which the individual proteins are represented as vectors in the reduced dimensional definition space.

The species tree generated at 74 dimensions was nearly identical to the consensus tree derived from the 17 individual species trees obtained using dimension settings ranging from 64 to 80 (Figure 6). These trees were similar in size and structure to a recently generated 69 member species tree (Pollack et al., 2000), placing the vast majority of species into well accepted groupings. Within the mammals, perissodactyls, carnivores, and cetartiodactyls are grouped together as expected (Xu et al., 1996; Arnason et al., 2000; Murphy et al., 2001). In addition, these three groups, which form the ferungulates (+ cetacea), cluster with the mole (Teur) and the bat (Ajam), as observed in recent independent analyses (Mouchaty et al., 2000; Nikaido et al., 2000). These groups and the remaining eutherian mammals are rooted in this portion of the tree by the non-eutherian mammals, within which the monotremes are seen to be more closely aligned with marsupials (Janke et al., 1997). The overall structure of the mammalian portion of the tree is very similar to that derived in an analysis focused on mammalian mitochondrial genomes by the same method (Stuart et al., 2001). There is a somewhat surprising tendency of elephants to group with primates. Since few organisms likely to be close relatives of the elephant (e.g. Dnov) are represented in our dataset, and since elephants and primates both branch near the root of recently published mammalian trees (Murphy et al., 2001), the association of elephants with primates seen in our tree is perhaps not too unreasonable.

In the non-mammalian portion of the consensus tree, the birds, the reptiles, and the fish tend to form distinct groups. The avian subtree is monophyletic, and its overall structure is very similar to that provided in a recent biomolecular analysis of avian relationships, placing the passeriforms (e.g. Cfru) at the avian root (Harlid et al., 1998). In contrast, the reptiles are presented as polyphyletic, but only because a single species (Eegr, a lizard) appears on a separate branch. Rooting the tetrapod portion of the tree is the lungfish (Pdol). Notably, all the species above this lung-bearing fish in the

tree have lungs, whereas all the species below are fish without lungs. Finally, osteichthyes (bony fish) are grouped separately from chondrichthyes (cartilaginous fish), one of which (Rrad) was arbitrarily chosen as the root of the tree. Interestingly, the coelacanth (Lcha) is well separated from its fellow sarcopterygian fish, the lungfish; instead it groups with the bichir (Porn) deep within the bony fish. The overall phylogeny of fish, including the relationship between cartilaginous fish and bony fish, is currently uncertain. Our analysis adds another potential solution to those offered previously (Rasmussen and Arnason., 1999; Curole and Kocher, 1999).

The 17-fold consensus tree (Figure 6) was obtained using dimension settings within the optimization well ranging from 64 to 80 (Figure 4). We interpret the included dimensions to the left of this well to be associated with "signal" and included dimensions to the right to be associated with "noise". Within the well, however, it is difficult to separate useful information from misinformation. Consequently, dimension settings within the range 64 to 80 may be equally valid. This interpretation is reinforced by the fact that the consensus tree over this range is nearly identical to the species tree obtained at 74 dimensions. This observation also suggests that both trees are relatively stable approximations of the true tree. Only a few branches in the consensus tree appear in less than 50% of the trees; one of these calls into question the monophylogeny of rodents, which is currently an unresolved issue (see Kramerov et al., 1999, Reyes et al., 2000). Other poorly supported branches question the monophylogeny of cetartiodactyls, the specific grouping of bats with cetartiodactyls (although the grouping with ferungulates is stable), the specific grouping of rabbits and rodents with ferungulates, and the sister group relationship between birds and mammals. The latter instability may be due to a tendency for reptiles and birds to group together as sauropods. In fact, this grouping was specifically observed in the tree generated at 74 dimensions (not shown).

CONCLUSIONS

We have used the method described herein to generate an integrated gene and species phylogeny for 64 vertebrates using sequence information from complete mitochondrial genomes. The optimized gene tree at 74 dimensions placed the vast majority of the 832 genes of the dataset into the expected 13 groups, each of which usually contained all 64 homologs. The optimized species tree at 74 dimensions and the 17-fold consensus species tree spanning the "optimization well" between 64 and 80 dimensions are nearly identical, and both may represent reasonably good approximations of the true tree. This SVD-based phylogenetic analysis also serves to identify correlated peptides as motifs that define a limited number of axes in an integrated high-dimensional space describing both peptides and proteins. These "copep" motifs are very flexible in that their identities are relatively insensitive to multiple insertions or deletions during protein evolution as long as the correlated peptides that define the motifs remain substantially unaltered. Although our method neither produces nor utilizes specific alignments, it could be used subsequently to help generate alignments by defining motifs as the basis for those alignments. The resolution of this particular analysis may be limited by the relatively small number of independent characteristics (64-80 copep motifs) extracted from the 13 gene families of mitochondria. Many more independent characteristics would likely be available from larger datasets containing whole bacterial or nuclear genomes.

Optimal dimensions were estimated by reference to prior categorical information concerning family memberships; no prior alignment information of any kind was used as input. Although many datasets may lack prior categorical information, such information is commonly available for even very large datasets (e.g. the COG database: <http://www.ncbi.nlm.nih.gov/COG/>). Nevertheless, the

development of a procedure whereby optimal dimensions can be approximated without reference to prior information would represent an important advance (e.g. Ding, 1999).

ACKNOWLEDGEMENTS

We gratefully acknowledge the help and advise of Thomas Landauer on the interpretation of the SVD. We also thank the multi-institutional "Complexity" group (Indiana State, Rose-Hulman Inst. of Tech., Saint Mary-of-the-Woods C., and Ivy Tech) for many interesting discussions on related issues. Special thanks to Steve Baker for providing programming expertise and a reliable computing environment.

LITERATURE CITED

Arnason, U., A. Gullberg, S. Gretarsdottir, B. Ursing, and A. Janke. 2000. The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. *J. Mol. Evol.* 50:569-78.

Berry, M. W., Z. Drmac, and E. R. Jessup. 1999. Matrices, vector spaces, and information retrieval. *SIAM Review* 41:335-62.

Berry, M. W. 1992. Large scale singular value computations. *Inter. J. of Supercomputer Appl.* 6:13-49.

Boore, J. L. and W. M. Brown. 1998. Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genet. Dev.* 8:668-74.

Curole, J. P., and T. D. Kocher. 1999. Mitogenomics: digging deeper with complete mitochondrial genomes. *Trends Eco. Evol.* 14:394-8.

Ding, C. H. Q. 1999. A Similarity-based Probability Model for Latent Semantic Indexing. *Proc. of 22nd ACM SIGIR Conference (SIGIR'99)*:59-65.

Fitz-Gibbon, S. T. and C. H. House. 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucl. Acids Res.* 271:4218-22.

Galperin, M. Y. and E. V. Koonin. 2000. Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.* 186:609-13.

Grundy, W.N. and G.J. Naylor. 1999. Phylogenetic inference from conserved sites alignments. *J. Exp. Zool.* 285:128-39.

Harlid, A., A. Janke and U. Arnason. 1998. "The complete mitochondrial genome of *Rhea americana* and early avian divergences. *J. Mol. Evol.* 46:669-79.

Janke, A., X. Xu, and U. Arnason. 1997. The complete mitochondrial genome of the wallaroo *Macropus robustus*. and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. *Proc. Natl. Acad. Sci. USA* 94:1276-81.

Kramerov, D., N. Vassetzky, and I. Serdobova. 1999. The evolutionary position of dormice Gliridae. in Rodentia determined by a novel short retroposon. *Mol. Biol. Evol.* 165:715-7.

Landauer, T. K., P.W. Foltz, and D. Laham. 1998. Introduction to latent semantic analysis. *Discourse Proc.* 25:259-84.

Li, M., J.H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang. 2001. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics* 17:149-54.

Lin, J., and M. Gerstein. 2000. Whole-genome trees based on the occurrence of folds and orthologs, implications for comparing genomes on different levels. *Genome Res.* 10:808-18.

Mouchaty, S. K., A. Gullberg, A. Janke and U. Arnason. 2000. The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences. *Mol. Biol. Evol.* 17:60-67.

Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614-18.

Nakashima H., M. Ota, K. Nishikawa, and T. Ooi. 1998 . Genes from nine genomes are separated into their organisms in the dinucleotide composition space. *DNA Res.* 5:251-9.

Nikaido, M., M. M. Harad, Y. Cao, M. Hasegawa, and N. Okada. 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.* 51:318-28.

Pollack, D. D., J. A. Eisen, N. A. Doggett, and M. P. Cummings. 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. *Mol. Biol. Evol.* 17:1776-88.

Rasmussen, A. S., and U. Arnason. 1999. "Molecular studies suggest that cartilaginous fishes have a terminal position in the piscine tree." *Proc. Natl. Acad. Sci. USA* 96:2177-82.

Reyes, A. C., C. Gissi, G. Pesole, F. M. Catzeflis, and C. Saccone. 2000. Where do rodents fit? Evidence from the complete mitochondrial genome of *Sciurus vulgaris*. *Mol. Biol. Evol.* 17:979-83.

Snel, B., P. Bork, and M. A. Huynen. 1999. Genome phylogeny based on gene content. *Nat. Genet.* 21:108-10.

Stuart, G.W., K. Moffet and S. Baker. 2002. Integrated gene and species phylogenies from unaligned whole genome sequence. *Bioinformatics* 18:100-108.

Tekaia, F., A. Lazcano, and B. Dujon. 1999. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* 9:550-7.

Thorne, J. L. 2000. Models of protein sequence evolution and their applications. *Curr. Opin. Genet. Dev.* 10:602-5.

Xu, X., A. Janke, and U. Arnason. 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.* 13:1167-73.

Figure 1

**MANYCATSLIKERICEWHENMICEHAVE
LICE
MANY ATSL KERI EWHE MICE AVEL
ANYC TSLI ERIC WHEN ICEH VELI
MYCA SLIK RICE HENM CEHA ELIC
YCAT LIKE ICEW ENMI EHAV LICE
CATS IKER CEWH NMIC HAVE**

Figure 2

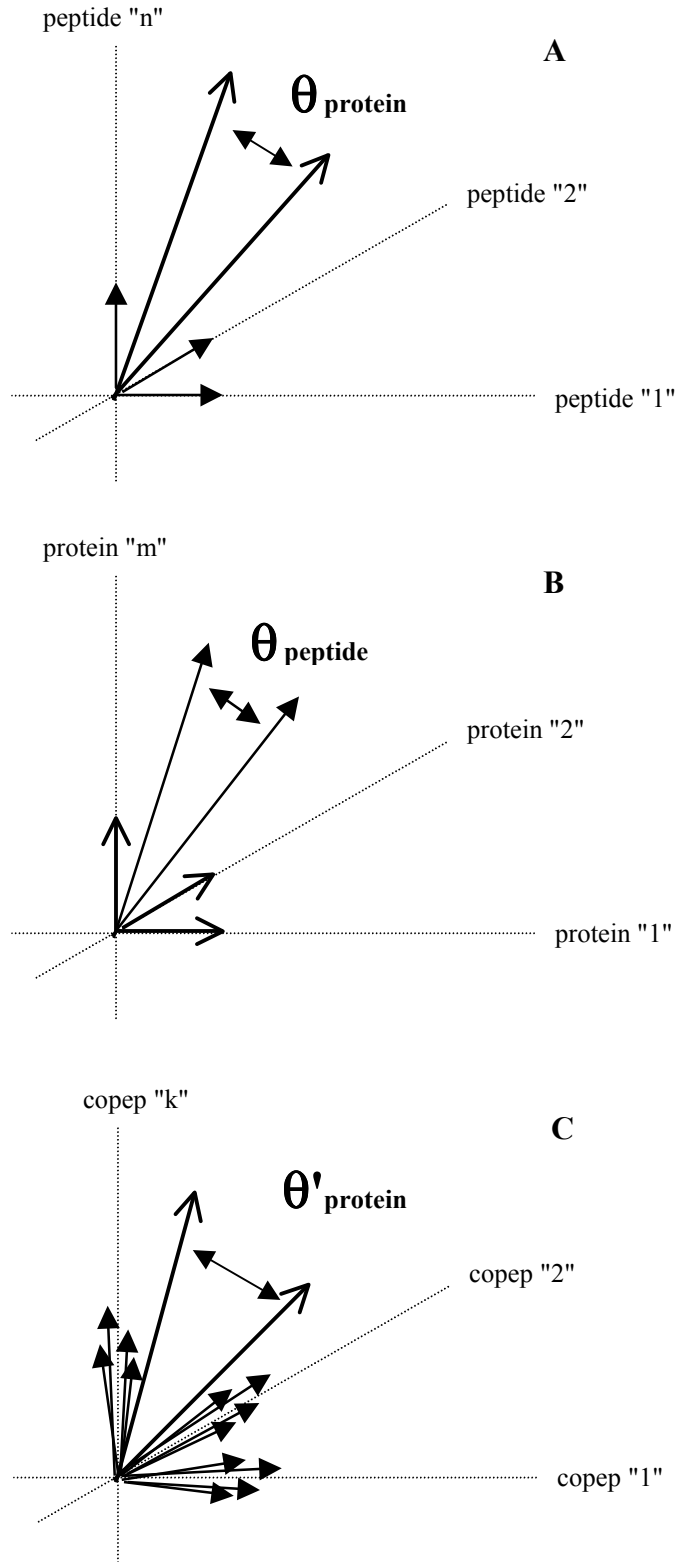


Figure 3

$$A = U * \Sigma * V^T$$

* * * * *		a 0 0 0 0		* * * * *
* * * * *		0 b 0 0 0		* * * * *
* * * * *	*	0 0 c 0 0	*	* * * * *
* * * * *		0 0 0 d 0		* * * * *
* * * * *		0 0 0 0 e		* * * * *
* * * * *		0 0 0 0 0		
U		Σ		V ^T
(n x n)		(n x m)		(m x m)

Figure 4

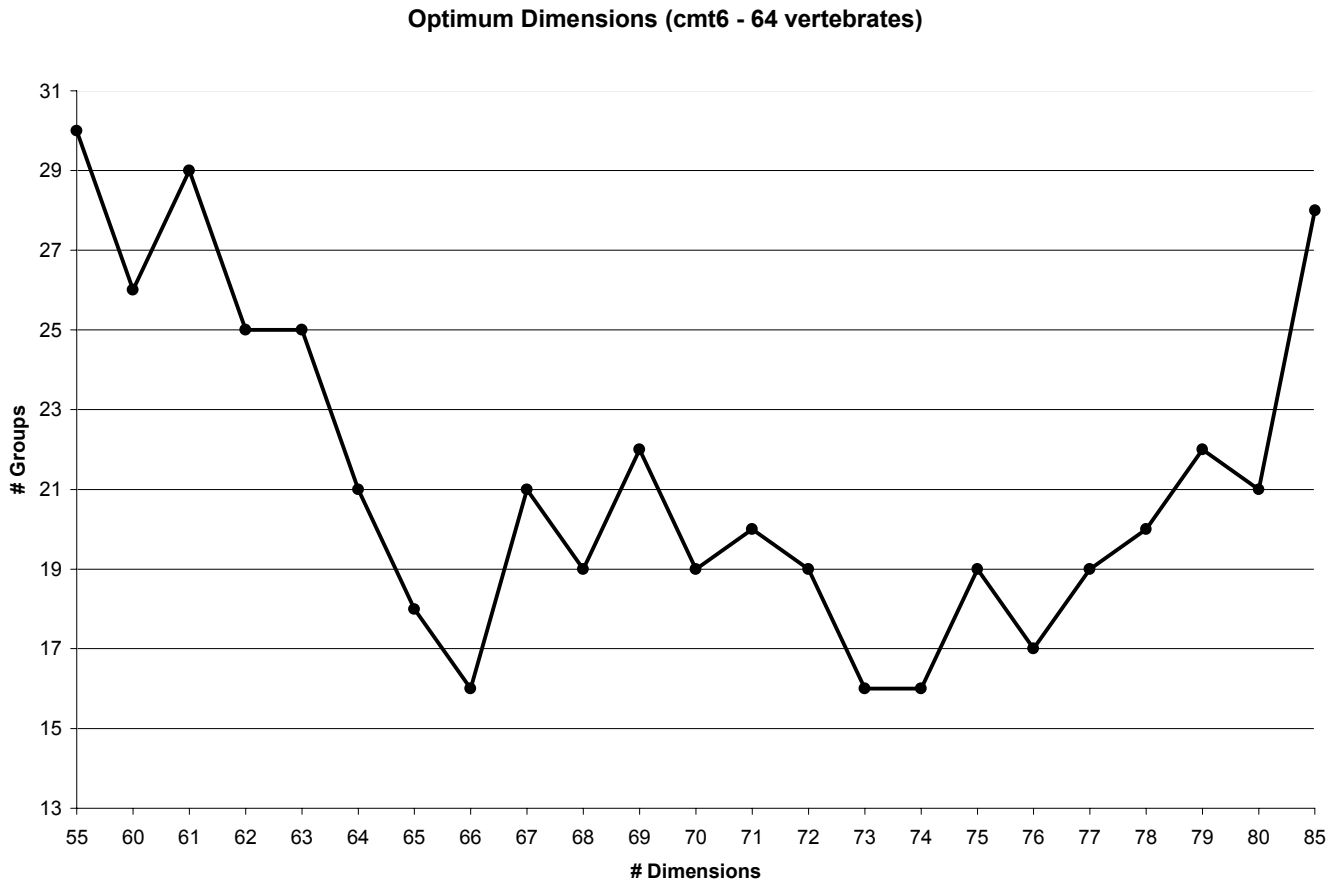


Figure 5

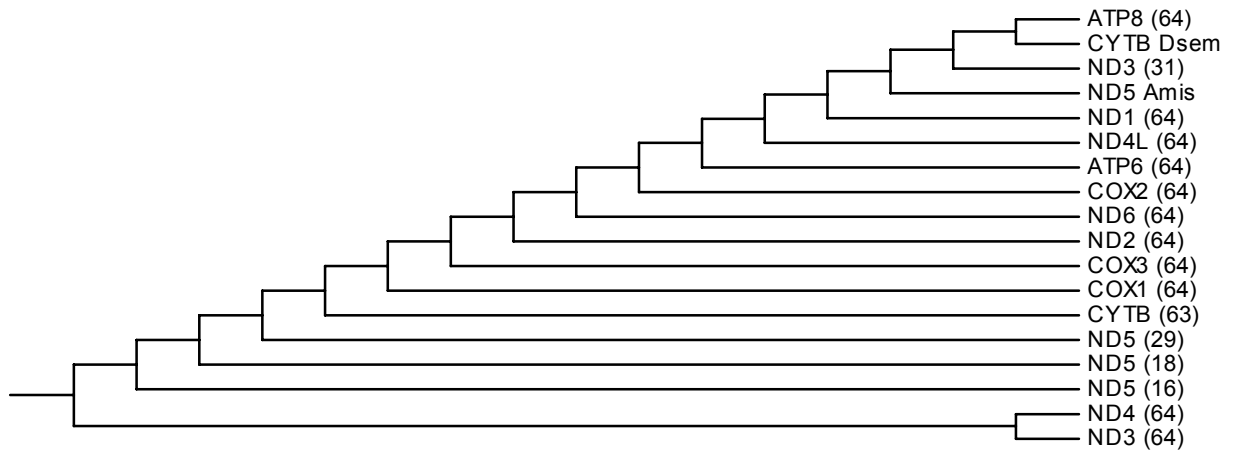


Figure 6

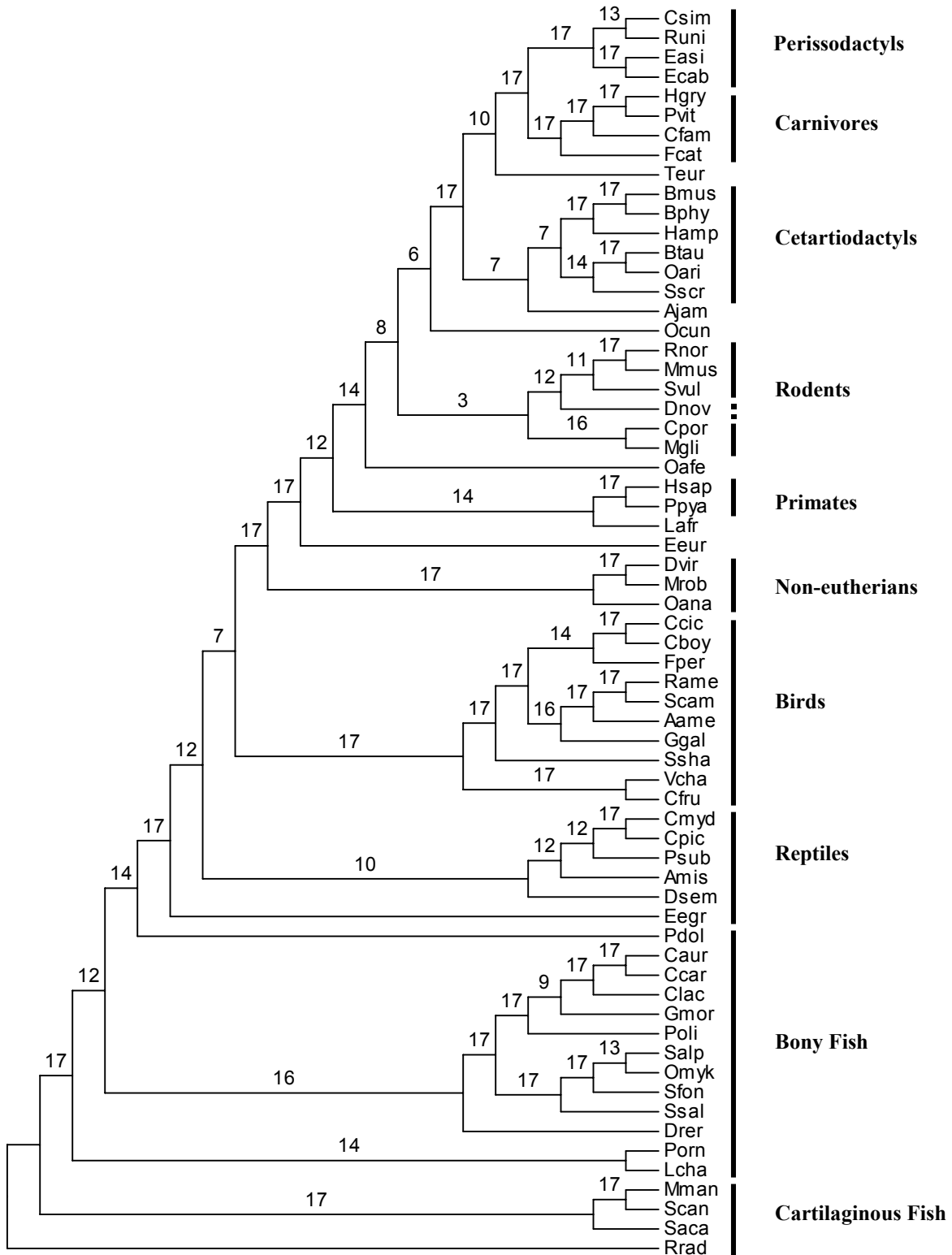


Figure 1: Recoding protein strings as overlapping sets of peptide words. Twenty-nine overlapping tetrapeptides are derived from a 32 residue string. The unique set of overlapping peptides dictates the original order of residues in the string such that there is only one possible reconstruction using each tetrapeptide only once in three residue overlaps. Some "misleading" alternative peptide words like "ERIC" have no meaning or function in this string, but are likely to have meaning or function in other strings and could represent potential homoplasies embedded within conserved motifs.

Figure 2: Multi-dimensional vectors and cosine based relative similarity. 3-D approximations of derived vector spaces are depicted. (A) Two similar proteins represented in "protein space" using information provided by a peptide vs. protein frequency matrix. The cosine of the angle provides a quantitative estimate of similarity: values of 1, 0, and -1 would indicate identical sequences, uncorrelated sequences, and anticorrelated sequences, respectively. (B) An alternative "peptide space" that reveals similar peptides using protein axes can be derived from the same peptide vs. protein frequency matrix. (C) In the SVD-remodeled k -dimensional space, correlated peptides have vectors that are clustered. Groups of correlated peptides conveniently define alternative axes uncovered by SVD. Each axis represents a "copep" motif that contributes to the functional meaning of individual peptides and proteins mapped within the derived definition space.

Figure 3: Decomposing matrices using the SVD. The original peptide frequency matrix is decomposed into three matrices that when multiplied together recreate the original matrix (A) having n (peptide) rows and m (protein) columns. Setting the singular values "d" and "e" to zero would effectively reduce the rank and dimensionality of the peptide (U) and protein (V) matrices to 3. If the discarded columns and rows (shown in gray) represent information that is misleading

(singular value “d”), insignificant (singular value “e”), and/or irrelevant (singular value “0”), then the vector definitions of peptides and proteins are improved. Adapted from Berry et al. (1999).

Figure 4: Estimating optimum dimensions. UPGMA trees (832 proteins) were constructed using pairwise distance values derived via SVD at the specified dimension setting. Each tree was evaluated based on the extent to which the members of the 13 known types of mitochondrial proteins in vertebrate mitochondria were placed correctly into groups consisting only of adjacent members of a single type. Although trees at 66, 73, and 74 dimensions had the fewest groups (16), trees within the approximate optimization well between 64 and 80 dimensions may be equally valid.

Figure 5: Gene tree for 832 mitochondrial proteins from 64 vertebrates at 74 dimensions. Only ND3, ND5, and CYTB fail to form monophyletic branches, indicating that optimal dimension reduction removes many but not all homoplasies. The three adjacent branches of ND5 near the base of the tree were counted as a single "group", though this group was not monophyletic.

Figure 6: Consensus species tree derived from 17 trees describing 64 species at 64-80 dimensions. The number of trees in which a given cluster is observed is shown above the branch leading to that cluster. For each tree, vectors representing the protein taxa described in Figure 5 were summed within species to generate species vectors. Pairwise cosine values calculated between species vectors were used to generate a distance matrix, which was then converted into a neighbor-joining species tree via PHYLIP-NEIGHBOR. The single tree obtained at 74 dimensions differed from the 17-fold consensus tree shown here in only three respects: 1) Porn and Lcha were placed together as basal members of the branch containing the majority of bony fish, 2) Ocun was placed between the

non-murid rodents and a group containing the murid rodents plus Svul and Dnov, and 3) birds and reptiles (excluding Eegr) grouped together as sauropods.