Abstract

Large, recently-available genomic databases cover a wide range of life forms, suggesting opportunity for insights into genetic structure of biodiversity. In this study we refine our recently-described technique using indicator vectors to analyze and visualize nucleotide sequences. The indicator vector approach generates correlation matrices, dubbed Klee diagrams, which represent a novel way of assembling and viewing large genomic datasets. To explore its potential utility, here we apply the improved algorithm to a collection of almost 17000 DNA sequences covering 12 widely-separated animal taxa, demonstrating that indicator vectors for classification gave correct assignment in all 11000 test cases. Indicator vector analysis revealed discontinuities corresponding to species- and higher-level taxonomic divisions, suggesting an efficient approach to classification of organisms from poorly-studied groups. As compared to standard distance metrics, indicator vectors preserve diagnostic character probabilities, enable automated classification of test sequences, and generate high-information density single-page displays. These results support application of indicator vectors for comparative analysis of large nucleotide data sets and raise prospect of gaining insight into broad-scale patterns in the genetic structure of biodiversity.

Introduction

Genetic study of biodiversity has been hampered by the relatively small number of species represented in databases. For example, the largest set of alignable sequences in GenBank (small subunit ribosomal RNA) represents fewer than 21,000 species and the second largest (cytochrome b) includes fewer than 14,000 [1]. This is modest coverage compared to the approximately 1.9 million named species of plants and animals and likely much larger numbers of protozoa, fungi, bacteria, and archaea [2]. Usually, a primary goal of comparative genetic study is assembling a Tree of Life that represents the temporal sequence of evolutionary diversities. As it is computationally difficult to construct a phylogenetic tree for more than a few thousand taxa, most analyses focus on a taxonomically-restricted subset and select a few exemplars from each group (e.g., [3,4]). Beyond computational challenges, potential limitations to tree representations include difficulty in representing discontinuities among species or groups of species, as all taxa are linked in a continuous structure; visualizing horizontal affinities across groups, as taxa within each group are joined in a single branch; and comparing data sets such as from ecological surveys, as branching diagrams challenge visual comparison.

Large, newly-available data sets [5] offer the possibility of studying genetic diversity on a wide scale. In an earlier paper, we described a method for creating “indicator vectors” representative of sets of nucleotide sequences [6]. Our aim is to develop an approach to genetic biodiversity that is computationally efficient and enables quantitative display of affinities at various taxonomic scales. Here we extend and refine this method and first apply it to large-scale differences, using sequences drawn from 12 diverse sets of animal species. On a finer scale we apply this mathematical apparatus to delineate affinities within one of the groups, North American birds, and examine biological implications of discontinuities that appear in structural representations of nucleotide sequence correlations.

Data Preparation

We considered the 648-nucleotide region of COI employed as a standard for distinguishing animal species [5]. Inspection of terminal regions of barcode sequence alignments deposited in the DNA barcode database BOLD http://www.barcodinglife.org showed a high degree of ambiguous or missing nucleotides, presumably reflecting incomplete sequencing runs. To reduce this noise we restricted attention to base pair (bp) positions 100 through 600 in the downloaded alignments, a 501-nucleotide span representing 167 complete codons.

For the correlation analysis of the present framework nucleotide positions that are conserved lead to an uninformative increase in correlation, i.e., these carry no differential information. Among the 16,876 sequences of the 12 groupings considered below, we found that 161 of the 501 positions were conserved (Table 1); for the purposes of this analysis, these were dropped from consideration.

The stretch of 501 nucleotide characters can each be uniquely translated into a digital vector under the nucleotide convention as follows.

\[
\begin{bmatrix}
A \\
C \\
G \\
T
\end{bmatrix} = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix} \text{ (1)}
\]
In schematic form a sequence \( s \) transforms to a vector \( s \) as follows

\[
s = \ldots GATTC \ldots \rightarrow [\ldots, G, A, T, C, \ldots] \rightarrow [\ldots, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, \ldots] = s.
\]  

(2)

There are various metrics for calculating sequence distances based on models of nucleotide substitution. Among these the Hamming distance, \( d_H \), i.e., the number of substitutions required to bring two sequences of like length into agreement, is the freest of additional assumptions. More complex distances distinguish between transitions and transversions, codon positions, and equilibrium based frequencies, as for example \([7–10]\). These forms are based on evolutionary considerations, while for our approach, which is based on the present state of correlations, the Hamming distance is the metric of choice. Each COI sequence thus becomes a vector \( s \) of 2004 entries; after removal of the 161 conserved nucleotides, 1,360 entries remain. The transformation of eq. (2) is not unique. An alternate transformation is

\[
([A \text{ or } T], [C \text{ or } G]) = [\pm 1, \pm 1]
\]  

(3)

which doubles, instead of quadrupling the sequence length as in eq. (2). This does not lead to the desired form of eq. (5) given below. Other alternatives that have been tried also lead to problems.

**Methods**

**Distances**

If two sequences of length \( N \) disagree at \( K \) positions the Hamming distance is

\[
d_H = K.
\]  

(4)

On the other hand from eq. (2) the square of its Euclidean distance \( d_E \) is

\[
d_E^2 = 2K
\]  

(5)

and therefore

\[
d_H = \frac{d_E^2}{2}.
\]  

(6)

In normalized form this can be written as

\[
\delta_H = \frac{d_H}{N} = \frac{1}{N} \frac{d_E^2}{2} = 1 - \cos \theta
\]  

(7)

which places the sequences vectors measured from a zero origin on the unit sphere and also uniquely associates the correlation coefficient \( \cos \theta \), and the angle \( \theta \), as a consequence of the law of cosines, i.e., the right hand side of eq. (7). \( \delta_H \) is the ratio of substitutions to site number, a customary representation.

Equations (6) & (7) are special cases of a more general recipe for associating a correlation coefficient with a metric. If \( d(\cdot) \) denotes a metric (distance function), then we recall that for elements \( A, B \) & \( C \) by definition the triangle inequality is satisfied

\[
c \leq a + b,
\]  

(8)

where

\[
c = d(A, B), a = d(B, C), b = (A, C).
\]  

(9)

One may then show from eq. (8) that

\[
-1 < \frac{a^2 + b^2 - c^2}{2ab} < +1,
\]  

(10)

which fulfills the requirement of a correlation. And if the ratio in eq. (10) is written as \( \cos \theta \) we obtain the law of cosines.

\[
c^2 = a^2 + b^2 - 2ab \cos \theta
\]  

(11)

In a vector space this is exactly the case. In the construction eq. (7) \( C \) is taken as the origin.

**Indicator Vectors**

For purposes of exposition consider the particular grouping of “Canadian freshwater fish” see Table 2. After the above preparation of sequence data we denote a typical fish sequence by the row vector \( s(F) \). The Canadian fish dataset has 1,324 members. Next we chose \( M \) distinct sequences at random from this set and form the fish set.

\[
G^F = \{s(F)\}, j = 1, \ldots, M.
\]  

(12)

In general if there are \( N \) groupings we consider \( N \) sets \( G^x = \{s(x)\} \), where \( x \) ranges over the \( N \) groupings.

An indicator unit vector \( \psi \) for each set is then determined on the basis that it have a maximal correlation with the selected \( j^{th} \) taxon, and minimal correlation with all other taxa \([6]\). As a simple


| Table 1. Conserved sites in 501-bp sequences used for the 12-group analysis. |
|---|
| 143 | 145 | 146 | 147 | 149 | 151 | 152 | 155 | 157 | 160 | 161 | 170 |
| 172 | 173 | 178 | 179 | 181 | 185 | 188 | 190 | 191 | 193 | 194 | 196 |
| 197 | 199 | 200 | 203 | 208 | 209 | 211 | 212 | 215 | 218 | 221 | 223 |
| 224 | 226 | 227 | 229 | 230 | 232 | 233 | 235 | 236 | 238 | 242 | 245 |
| 247 | 248 | 251 | 256 | 257 | 258 | 260 | 262 | 263 | 268 | 269 | 271 |
| 272 | 274 | 275 | 280 | 283 | 284 | 287 | 290 | 292 | 293 | 295 | 296 |
| 299 | 301 | 302 | 304 | 305 | 307 | 308 | 311 | 314 | 323 | 326 | 332 |
| 335 | 341 | 343 | 346 | 347 | 348 | 350 | 351 | 352 | 353 | 356 | 379 |
| 338 | 383 | 385 | 386 | 388 | 389 | 391 | 392 | 395 | 412 | 413 | 425 |
| 430 | 431 | 434 | 436 | 437 | 438 | 442 | 443 | 445 | 446 | 452 | 454 |
| 455 | 463 | 464 | 469 | 470 | 472 | 473 | 475 | 476 | 479 | 485 | 490 |
| 491 | 493 | 494 | 496 | 497 | 502 | 503 | 509 | 511 | 512 | 514 | 515 |
| 539 | 548 | 551 | 559 | 563 | 566 | 572 | 574 | 575 | 581 | 584 |
| 587 | 590 | 593 | 596 | 599 |
Table 2. COI datasets used in the 12-group analysis.

| No. | BOLD Project                          | Group Designation | No. sequences | No. test sequences |
|-----|---------------------------------------|-------------------|---------------|--------------------|
| 1   | GenBank-Amphibia                       | Amphibians        | 520           | 20                 |
| 2   | Barcoding of Canadian freshwater fishes | Fish              | 1324          | 824                |
| 3   | Bats of Guyana                         | Bats              | 819           | 319                |
| 4   | Birds of North America, General sequences | Birds            | 1688          | 1188               |
| 5   | ACG Generalist Tachinidae              | Flies             | 1981          | 1481               |
| 6   | Hesperiidae of the ACG 1               | Butterflies       | 1581          | 1081               |
| 7   | ACG Microgastrinae                     | Wasps             | 1895          | 1395               |
| 8   | Ants of the World, merged project      | Ants              | 1799          | 1299               |
| 9   | Barcoding the Aphididae                | Aphids            | 666           | 166                |
| 10  | GenBank-Crustacea Malac.-Decapoda       | Crayfish          | 2068          | 1568               |
| 11  | Marine Life, merged project            | Mollusks          | 1652          | 1152               |
| 12  | Genbank Cnidaria                       | Jellyfish         | 883           | 383                |

Datasets used to calculate and test group indicator vectors.
doi:10.1371/journal.pone.0009266.t002

but useful illustration consider N sequence vectors \( x_i, i = 1, ... , N \), say one representative from each of \( N \) groups, or each an average of each group. We then seek \( \mathbf{v} \), the \( i^{th} \) indicator vector such that

\[
C_i = (s'_i, \mathbf{v})^2 - \langle (s'_i, \mathbf{v})^2 \rangle_{i \neq l}
\]

is a maximum,

\[
||\mathbf{v}||^2 = 1
\]

where \( <> \) signifies the average. It is straightforward to show that under the reasonable assumption that if \( \{s'_i\} \) are linearly independent then the criterion function \( C_i \) has a positive maximum and that it is determined as the eigenvector with the largest (positive) eigenvalue of

\[
s's_i \mathbf{v} = -\frac{1}{N-1} \sum_{j \neq i} s'_j \mathbf{v} = \lambda \mathbf{v}.
\]

One consequence of the particular criterion for choosing the \( \mathbf{v} \) is that it provides a natural structural representation expressed as auto- and cross-correlations, given by

\[
S_{ij} = (\mathbf{v}, \mathbf{v}); \ i, j = 1, ... , N
\]

and referred to as the structure matrix. We also define the diversity matrix as given by

\[
D_{ij} = \langle (G^j), (G^i) \rangle.
\]

This notation denotes the mean over all \( M(M-1)/2 \) inner products pairs of the members of \( G^j \) with those of \( G^i \), which thus gives a depiction of within and among group correlations.

A fixed number of members, \( M \), in the sets \( G \) confers equal weights on each of the taxa. These may be considered as the “training set,” for the indicator vector and the remaining sequences are used as a “test set.” There is reason to make \( M \) relatively small in initial calculations. Once past the testing stage there may be reason to take \( M \) as large as possible within the restriction of equal weightings.

Probabilities

Another consequence of embedding a character sequence into a vector space, eq. (2), is that the average of an ensemble of sequences \( \{s_j\} \) can be defined as

\[
\mathbf{s} = \langle s_j \rangle
\]

Which through the inverse operation of eq. (2) furnishes the probability of occurrence of \((A,T,C,G)\) at each nucleotide position and that

\[
p(A) + p(T) + p(C) + p(G) = 1,
\]

which is a consequence of eq. (2).

Conservation of Probability. Eq. (4) allows us to regard the 4-vectors as specifying the probabilities of the associated symbols. We now demonstrate that this property is inherited by the indicator vectors, i.e., its 4-vectors sum to unity. To see this define

\[
U = p\left[ \begin{array}{cccccccc} 1 & 1 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & 1 & 1 & 1 \end{array} \right]
\]

where \( p \), the number of rows, is also the number of bps. Multiplication of (15) by \( U \) yields

\[
Us' = Us = \sum_{j \neq i} Us'_j \mathbf{v} = \lambda (Us) \mathbf{v},
\]

but

\[
Us = u = \left[ \begin{array}{c} 1 \\
1 \\
\vdots \\
1 \end{array} \right] p
\]
Figure 1. Correlations among indicator vectors for 12 animal groups. (A) Structure matrix eq. (16). (B) Diversity matrix eq. (17). Numerical forms of matrices given in Table 3. Differing color bar scales in (A) and (B) are used to emphasize off diagonal resemblance between matrices. doi:10.1371/journal.pone.0009266.g001

Table 3. Numerical representations of Figure 1A and 1B, respectively.

|      | Amphibians | Fish | Bats | Birds | Flies | Butterflies | Wasps | Ants | Aphids | Crayfish | Mollusks | Jellyfish |
|------|------------|------|------|-------|-------|-------------|-------|------|--------|----------|----------|----------|
| (A)  | 1.0000     | 0.8060 | 0.6799 | 0.7221 | 0.2215 | 0.2201 | 0.1594 | 0.1754 | 0.1920 | 0.3220 | 0.3252 | 0.2894   |
|      | 0.8060     | 1.0000 | 0.7146 | 0.7260 | 0.2169 | 0.2209 | 0.1021 | 0.1432 | 0.1587 | 0.3199 | 0.2898 | 0.2706   |
|      | 0.6799     | 0.7146 | 1.0000 | 0.7458 | 0.2477 | 0.2662 | 0.1729 | 0.1792 | 0.1754 | 0.3498 | 0.3565 | 0.2872   |
|      | 0.7221     | 0.7260 | 0.7458 | 1.0000 | 0.1900 | 0.1806 | 0.0787 | 0.1372 | 0.1284 | 0.3841 | 0.2514 | 0.2225   |
|      | 0.2215     | 0.2169 | 0.2477 | 0.1900 | 1.0000 | 0.6219 | 0.4446 | 0.3790 | 0.3526 | 0.5245 | 0.4358 | 0.3034   |
|      | 0.2201     | 0.2209 | 0.2662 | 0.1806 | 0.6219 | 1.0000 | 0.4708 | 0.4775 | 0.3794 | 0.4120 | 0.3648 | 0.2397   |
|      | 0.1594     | 0.1021 | 0.1729 | 0.0787 | 0.4446 | 0.4708 | 1.0000 | 0.5160 | 0.4222 | 0.3455 | 0.3363 | 0.2616   |
|      | 0.1754     | 0.1432 | 0.1792 | 0.1372 | 0.3790 | 0.4775 | 0.5160 | 1.0000 | 0.4753 | 0.2803 | 0.2062 | 0.1844   |
|      | 0.1920     | 0.1587 | 0.1754 | 0.1284 | 0.3526 | 0.4222 | 0.4753 | 1.0000 | 0.1980 | 0.2540 | 0.2043 | 0.0403   |
|      | 0.3220     | 0.3199 | 0.3498 | 0.3841 | 0.5245 | 0.4120 | 0.3455 | 0.2803 | 1.0000 | 0.4127 | 0.2409 |         |
|      | 0.3252     | 0.2898 | 0.3565 | 0.2514 | 0.4358 | 0.3648 | 0.3363 | 0.2062 | 0.2540 | 1.0000 |         |         |
|      | 0.2894     | 0.2706 | 0.2872 | 0.2225 | 0.3034 | 0.2397 | 0.2516 | 0.1844 | 0.2043 | 0.2409 | 0.3756 | 1.0000   |

|      | 0.6826     | 0.6633 | 0.6545 | 0.6567 | 0.5603 | 0.5500 | 0.4977 | 0.4793 | 0.4885 | 0.5547 | 0.5577 | 0.4916   |
|      | 0.6633     | 0.7144 | 0.6715 | 0.5464 | 0.5413 | 0.4682 | 0.4611 | 0.4656 | 0.5494 | 0.5447 | 0.4842 |         |
|      | 0.6545     | 0.6686 | 0.7433 | 0.6877 | 0.5783 | 0.5731 | 0.5055 | 0.4905 | 0.4877 | 0.5726 | 0.5767 | 0.5031   |
|      | 0.6567     | 0.6715 | 0.6877 | 0.7600 | 0.5295 | 0.5191 | 0.4453 | 0.4527 | 0.4433 | 0.5616 | 0.5279 | 0.4680   |
|      | 0.5603     | 0.5464 | 0.5783 | 0.5295 | 0.8623 | 0.7314 | 0.6482 | 0.5874 | 0.5971 | 0.6521 | 0.6366 | 0.5323   |
|      | 0.5500     | 0.5413 | 0.5731 | 0.5191 | 0.7314 | 0.8306 | 0.6480 | 0.6083 | 0.6000 | 0.6148 | 0.6034 | 0.5021   |
|      | 0.4977     | 0.4682 | 0.5055 | 0.4453 | 0.6482 | 0.6480 | 0.8185 | 0.6034 | 0.6025 | 0.5632 | 0.5663 | 0.4897   |
|      | 0.4793     | 0.4611 | 0.4905 | 0.4527 | 0.5874 | 0.6083 | 0.6034 | 0.6920 | 0.5857 | 0.5198 | 0.4936 | 0.4360   |
|      | 0.4885     | 0.4656 | 0.4877 | 0.4433 | 0.5971 | 0.6000 | 0.6025 | 0.5857 | 0.6871 | 0.5039 | 0.5144 | 0.4487   |
|      | 0.5347     | 0.5494 | 0.5726 | 0.5616 | 0.6521 | 0.6148 | 0.5632 | 0.5198 | 0.5039 | 0.6820 | 0.5828 | 0.4826   |
|      | 0.5577     | 0.5447 | 0.5767 | 0.5279 | 0.6366 | 0.6034 | 0.5563 | 0.4936 | 0.5144 | 0.5828 | 0.7585 | 0.5286   |
|      | 0.4916     | 0.4842 | 0.5031 | 0.4680 | 0.5323 | 0.5021 | 0.4897 | 0.4360 | 0.4487 | 0.4826 | 0.5286 | 0.6680   |

doi:10.1371/journal.pone.0009266.t003
as to be reasonable.) Therefore each indicator vector can be regarded the calculation of the above distances. It is the distance matrix between the taxons which proves the assertion. (This proof depends specifically on regarding an unknown bp as \(\{1/4,1/4,1/4,1/4,1/4\}\), which we deem to be reasonable.) Therefore each indicator vector can be regarded as \(p\) quartets of probability in the four possible symbols.

Tree Construction

A customary practice is to express sequence separations as distances, which play a role in the construction of trees. It is straightforward to show the connection of distances to the correlations contained in eq. (16) and of eq. (17). In fact it directly follows from eq. (7) that

\[
U \bar{v} \propto \bar{u}
\]

(23)

which proves the assertion. (This proof depends specifically on regarding an unknown bp as \(\{1/4,1/4,1/4,1/4,1/4\}\), which we deem to be reasonable.) Therefore each indicator vector can be regarded as \(p\) quartets of probability in the four possible symbols.

Thus amphibians have the highest relationships with the others in this set. The next \(5 \times 5\) block representing Class Insecta, are ordered by relationship as above. The diversity matrix eq. (17) quantifies the degree of diversity within and among data sets (Figure 1B), and has an impressionistic similarity to the structure matrix of unitary indicator vectors (Figure 1A). The diagonal elements of Figure 1B illustrate the high internal diversity of amphibians, ants, crayfish, and jellyfish, and relative lack of internal diversity for flies, butterflies, wasps, and aphids. Numerical equivalents of Figure 1 are given in Table 3. Lack of diversity might be consistent with these data being drawn from single families or subfamilies. Diversity as defined by (24) introduces an objective measure of diversity based on variance.

We applied the indicator vector approach at a finer scale, analyzing differences within the dataset of North American birds, which contained 1,693 sequences representing 558 species. As a compromise between a large \(M\) and a large test set, we chose \(M = 3\), giving 262 admissible species and 471 test sequences. With the input ordered alphabetically by taxonomic genus, the resulting structure matrix appears to be disordered with small regions of high correlation (Figure 3A). When arranged in a taxonomic order representing phylogenetic relationships [11] (Table 4), these correlations coalesced into a coherent picture (Figure 3B), which could be viewed as taxonomy organizing the structure matrix according to closeness of correlations. Discontinuities in the correlation among North American birds, evident as "boxes" or "blocks" in the color matrix, corresponded to avian taxonomic divisions (Figure 4). Most of the blocks represented families, with some blocks corresponding to lower (genera) or higher (suborder) groupings (Figure 4).

Among the 471 test bird sequences, there were 16 apparently incorrect assignments distributed among 4 species pairs (Junco phaeonotus/J. hyemalis; Anas platyrhynchos/A. rubripes; Larus smithsonianus/L. glaucescens; Sphyrapicus ruber/S. nuchalis). In the first instance each sequence set of \(M = 3\) were identical so that the indicator

\[
R_i = \sum_{j \neq i} S_{ij}.
\]

(26)

Results

We first considered 12 animal groups, using COI sequences deposited in BOLD taxon-specific projects (Table 2). In all cases analysis was restricted to sequences of sufficient length, and excluded those containing excessive blank positions.
vectors were also identical. In the remaining cases the indicator vectors were close but not equal reflecting the fact that the defining sequence sets shared some identical members. While such singular behavior is revealed by the present algorithm, these sets of species were previously noted to be indistinguishable by COI barcode [12].

As indicated in eq. (24) the structure matrix $S^i$ can be directly associated with a matrix of inter-species distances $D^i$. Since such a matrix can be made the basis of tree constructions we can apply the neighbor joining (NJ) algorithm of Saitou and Nei [13] to $D^i$. Using consistency arguments [14,15], Bryant [16] has demonstrated that the NJ construction is a unique clustering algorithm of the distance matrix [17]. Since the distance matrix is based on genomic distances, and not on evolutionary hypotheses, we can view the resulting NJ tree as intrinsic to the data. The species ordering according to this tree produces the structure matrix shown in Figure 5. This demonstrated the same set of clusters as seen in Figure 4; only the order of clusters differed. Thus at this level of resolution the indicator vector approach to classification coupled with NJ provides a self-generating ranking that is in general agreement with established taxonomy. Figure 6 compares the NJ tree that emerges from the structure matrix with the tree that derives from the averaged Hamming distance matrix between species, is equivalent to the diversity matrix (17).

**Discussion**

This paper describes a mathematical approach to comparative analysis of nucleotide sequences using digital transformation in vector space. We term the resulting structure matrices “Klee diagrams”, in acknowledgement of the geometric paintings of artist Paul Klee (see Figure 7). This approach is of general utility and could be applied to any set of aligned sequences. In this study we explore its potential by analyzing a large, diverse set of DNA barcodes, the short segment of mitochondrial COI gene employed as a standard for identification of animal species (6). The resulting Klee diagrams display the structure of present-day mitochondrial genetic diversity, a “macroscopic” view of the products of evolution [18,19]. This approach is akin to a distance metric (see Methods), and in fact the matrix of indicator vector correlations can be used to generate an NJ tree (Figure 6).

As compared to standard distance metrics with neighbor-joining, indicator vectors preserve character probabilities that distinguish sequence sets, enable automated classification of test sequences, and generate high-information density displays without constraints of tree diagrams. Regarding the latter point, as one example, the 12-group Klee diagram displays affinity among flies and crayfish, a finding which might be of interest for further exploration, and yet this sort of horizontal similarity is not represented in the NJ tree diagram, shown in Figure 8. Discontinuities in indicator vector correlations, evident as blocks in Klee diagrams, corresponded to branches in the tree; for example, in North American bird matrix, these blocks represent families, genera, and sub-orders (Figures 4, 5). These results, generated with a small sample of world birds, suggest that this approach might be usefully applied to generate a classification for poorly-studied groups by combining DNA barcodes with indicator vector analysis. Such a classification could be refined when additional morphologic, ecological, and genetic study was available.

The results so far suggest natural discontinuities, or fractures, in the genetic structure of biodiversity, at least as reflected in animal mitochondrial genomes. In quantitative terms, blocks represent higher correlation within than among sets of sequences. Further study will help determine the nature of underlying mitochondrial differences, for instance whether species- and family-level blocks, for example, reflect differences in coding or non-coding positions. The present-day discontinuities seen in Klee diagrams may not be evident from a historical perspective, such as in a phylogenetic tree which links all forms in a continuous structure. It is of interest to reconcile these two perspectives, namely the continuous nature of evolution with the fractures in present-day genetic biodiversity;
|   | Scientific Name |   | Scientific Name |   | Scientific Name |
|---|----------------|---|----------------|---|----------------|
| 1 | Anser albifrons | 67 | Pandion haliaetus | 133 | Selasphorus rufus |
| 2 | Chen caeruleus | 68 | Accipiter striatus | 134 | Megaceryle torquata |
| 3 | Branta bernicla | 69 | Accipiter cooperii | 135 | Megaceryle alcyon |
| 4 | Branta hutchinsii | 70 | Accipiter gentilis | 136 | Chloroceryle americana |
| 5 | Branta canadensis | 71 | Buteo swainsoni | 137 | Melanerpes lewis |
| 6 | Aix sponsa | 72 | Falco columbarius | 138 | Melanerpes formicivorus |
| 7 | Anas strepera | 73 | Gallinula chloropus | 139 | Melanerpes carolinus |
| 8 | Anas americana | 74 | Fulica americana | 140 | Sphyrapicus thyroideus |
| 9 | Anas rubripes | 75 | Grus americana | 141 | Sphyrapicus varius |
| 10 | Anas platyrynchos | 76 | Pluvialis dominica | 142 | Sphyrapicus nuchalis |
| 11 | Anas discors | 77 | Charadrius semipalmatus | 143 | Sphyrapicus ruber |
| 12 | Anas clypeata | 78 | Charadrius melodus | 144 | Picoidea nuttallii |
| 13 | Anas acuta | 79 | Haematopus bachmani | 145 | Picoidea villosus |
| 14 | Anas carolinsis | 80 | Actitis macularius | 146 | Picoidea albolavarius |
| 15 | Aythya valisineria | 81 | Tringa glareola | 147 | Picoidea dorsalis |
| 16 | Aythya americana | 82 | Limnodromus griseus | 148 | Colaptes auratus |
| 17 | Aythya collaris | 83 | Gallinago dominica | 149 | Contopus sordidulus |
| 18 | Aythya fuligula | 84 | Scolopax minor | 150 | Empidonax flaviventris |
| 19 | Aythya marila | 85 | Phalaropus lobatus | 151 | Empidonax alnorum |
| 20 | Aythya affinis | 86 | Rissa tridactyla | 152 | Empidonax trallii |
| 21 | Somateria fischeri | 87 | Larus ridibundus | 153 | Empidonax minimus |
| 22 | Somateria spectabilis | 88 | Larus atricilla | 154 | Empidonax hammondii |
| 23 | Somateria mollissima | 89 | Larus heermanni | 155 | Empidonax difficilis |
| 24 | Historionicus histronicus | 90 | Larus canus | 156 | Pyrocephalus rubinus |
| 25 | Melanitta fusca | 91 | Larus occidentalis | 157 | Myiarchus ludovicianus |
| 26 | Melanitta nigra | 92 | Larus californicus | 158 | Myiarchus cinereus |
| 27 | Clangula hyemalis | 93 | Larus smithsonianus | 159 | Myiarchus tyrannulus |
| 28 | Bucephala albeola | 94 | Larus fuscus | 160 | Pitangus sulphuratus |
| 29 | Bucephala clangula | 95 | Larus glaucescens | 161 | Myiodynastes luteiventris |
| 30 | Bucephala islandica | 96 | Onychoprion aleuticus | 162 | Lanius ludovicianus |
| 31 | Lophodytes cucullatus | 97 | Thalasseus maximus | 163 | Vireo griseus |
| 32 | Mergus merganser | 98 | Thalasseus sandvicensis | 164 | Vireo solitarius |
| 33 | Mergus serrator | 99 | Thalasseus elegans | 165 | Vireo huttoni |
| 34 | Perdix perdix | 100 | Storchararius pomarinus | 166 | Vireo philadelphicus |
| 35 | Bonasa umbellus | 101 | Storchararius parasiticus | 167 | Vireo olivaceus |
| 36 | Centrocercus urophasianus | 102 | Storchararius longicaudus | 168 | Vireo flaviviridis |
| 37 | Falcipennis canadensis | 103 | Uria aalge | 169 | Cyanocitta cristata |
| 38 | Lagopus lagopus | 104 | Alca torda | 170 | Aphelocoma californica |
| 39 | Lagopus muta | 105 | Cepphus grylle | 171 | Gymnorhinus cyanocephalus |
| 40 | Lagopus leuco | 106 | Brachyramphus marmoratus | 172 | Nucifraga columbiana |
| 41 | Dendragapus obscurus | 107 | Brachyramphus brevirostris | 173 | Pica nuttallii |
| 42 | Tymanuchus phasianellus | 108 | Cerorhinca monocerata | 174 | Corvus caurinus |
| 43 | Tymanuchus pallidicinctus | 109 | Fratercula arctica | 175 | Corvus cornix |
| 44 | Meleagris gallopavo | 110 | Zenaida macroura | 176 | Tachycineta bicolor |
| 45 | Oreonyx pictus | 111 | Columbina inca | 177 | Poecile gambeli |
| 46 | Gavia pacifica | 112 | Columbina passerina | 178 | Poecile gambeli |
| 47 | Gavia adamsii | 113 | Myiopitta monachus | 179 | Poecile sclateri |
| 48 | Podiceps grisegena | 114 | Coccyzus erythrophthalmus | 180 | Poecile rufescens |
| 49 | Fulmarus glacialis PS-1 | 115 | Crotaphaga ani | 181 | Poecile cincta |
| 50 | Puffinus creatopus | 116 | Tyto alba | 182 | Auriparus flaviceps |
| 51 | Puffinus carneipes | 117 | Megascops kennicottii PS-1 | 183 | Sitta canadensis |
these might be viewed, respectively, as “time-like” and “space-like”. One may speculate on the relation of such jump phenomena to adaptive radiations and the punctuated equilibrium model of evolution [20]. It may be possible to make useful observations for time-like behavior from space-like behavior as was done through the ergodic theory of statistical physics, [21].

As currently developed, our approach is limited to complete sets of homologous sequences, rather than overlapping sets of incomplete data as are often used in phylogenetic inference. In addition, the present analysis employing COI shares problems inherent to mitochondrial biology, including maternal inheritance, introgression, hybridization, male-biased dispersal patterns, and recent speciation among others [22]; most of these are likely to apply only at the fine-scale level of distinguishing closely-related species. As noted, the indicator method is of general utility and could readily be applied to longer sequences or concatenated multi-gene alignments without substantially increasing computation time, which might address some of these limitations. In this regard, it of interest to compare indicator vector affinities using mitochondrial and nuclear genes in puzzling cases that appear to represent convergent evolution [23].

Although the output is different, it may be revealing to compare the efficiency of the indicator vector approach to that of phylogenetic treebuilding programs. Due to computational demands, data sets in analyses beyond 1000 species are exceptional (e.g., [24–26]) and calculation times for larger studies are typically several CPU-months. The largest published phylogenetic tree includes 73,060 eukaryote taxa [1] and took 2.5 months with 16 processors, and the next largest analyzed 13,533 plant taxa [27]. The present study ranks with the largest

![Figure 4](image_url)

**Figure 4.** Annotated structure matrix of 262 North American bird species arranged in taxonomic order reflecting phylogenetic relationships. Representational fractures define “boxes” which correspond to taxonomic divisions. doi:10.1371/journal.pone.0009266.g004
biodiversity analyses in terms of number of organisms, and is at least two orders of magnitude faster. For example, the case of 12 animal groups deals with almost 17,000 sequences and required times of roughly 10–20 minutes on an ordinary desktop computer. This suggests the potential for analyzing the largest datasets available, including, for example, BOLD (> 700,000 sequences) http://www.barcodinglife.org, NCBI Influenza Virus Resource (> 50,000 complete genomes) http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html, or Los Alamos HIV Sequence Database (http://www.hiv.lanl.gov).

In addition to animals, cytochrome c oxidase is present in plants, protozoa, fungi, and some bacteria, which raises the prospect of insight into broad-scale patterns in the genetic structure of biodiversity. Also, the methodology as present here applies to nucleotide sequences of any sort and so might usefully be applied to a variety of questions.

From the point of view of accuracy, density of information and assimilation it would seem compelling that any properly ordered distance matrix should be viewed as a Klee diagram. It may be that the focus on evolution and therefore trees impeded this direction. In this connection we note that the distance matrix for a species count of \( N \) contains \( N(N-1)/2 \) distances and for large \( N \) a tree-building algorithm cannot accommodate this number of condi-

---

**Figure 5.** Annotated structure matrix of 262 North American bird species according to NJ tree ranking.
doi:10.1371/journal.pone.0009266.g005

**Figure 6.** Comparison of NJ trees based on the structure matrix right, and on the diversity matrix, left.
doi:10.1371/journal.pone.0009266.g006

**Figure 7.** Flora on Sand by Paul Klee.
doi:10.1371/journal.pone.0009266.g007
tions, and an increasing number of larger and larger errors occur with increasing N. Klee diagrams accurately display distances for any species count.

An important advance in the present treatment derives from the vectorization of nucleotide sequences, (1), which has been accomplished with the exact preservation of Hamming distances. Advantages flow from a vector space framework, an example of which is the optimization procedure leading to the indicator vectors. Another consequence is that bps occupation is rigorously transformed to the probability of occurrence of the four nucleotides, which opens the possibility of introducing information theory into these considerations.

Acknowledgments
We thank Jesse Ausubel, Ehud Kaplan, and Bruce Knight for helpful comments. Thanks also to Ellen Paley for her careful and patient preparation of this manuscript.

Author Contributions
Conceived and designed the experiments: LS MYS YZ. Performed the experiments: MYS YZ. Analyzed the data: LS MYS YZ. Contributed reagents/materials/analysis tools: LS YZ. Wrote the paper: LS MYS.

References
1. Goloboff P, Catalano S, Miranda J, Szumik C, Arias J, et al. (2009) Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics 25: 1–29.
2. Chapman A (2009) Numbers of Living Species in Australia and the World. Canberra: Australia: Australian Biological Resources Study, 2nd ed. edition. http://www.environment.gov.au/biodiversity/abs/other/species-numbers/2009/pdfs/shalw-2nd-complete.pdf.
3. Caterino M, Reed R, Kuo M, FAH S (2001) A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). Syst Biol 50: 106–127.
4. Hackett S, Kimball R, Reddy S, Bowie R, Braun E, et al. (2008) A phylogenetic study of birds reveals their evolutionary history. Science 320: 1763–1768.
5. Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
6. Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
7. Kimura K (1980) A simple method for estimating evolutionary rates of base substitutions through studies of nucleotide sequences. J Mol Evol 16: 111–120.
8. Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. Mol Biol Evol 9: 678–687.
9. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.
10. Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
11. (2009) American Ornithologists' Union Check-list of North American Birds, 7th edition and supplements. http://www.aou.org/checklist/north/print.php.
12. Kerr K, Stoeckle M, Dove C, Weigt L, Francis C, et al. (2007) Comprehensive DNA barcode coverage of North American birds. Mol Ecol Notes 7: 535–543.
13. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406–425.
14. Atteson K (1999) The performance of neighbor-joining methods of phylogenetic reconstruction. Algorithmica 25: 251–278.
15. Gascuel O (1994) A note on Sattath and Tversky's Saitou and Nei's, and Studier and Keppler's algorithms. Mol Biol Evol 11: 961–963.
16. Bryant D (2005) On the uniqueness of the selection criterion in neighbor-joining. J of Classification 22: 3–15.
17. Mihaescu R, Levy D, Pachter L (2009) Why neighbor-joining works. Algorithmica 54: 1–24.
18. de Rosnay J (1979) The Macroscope: A New World Scientific System Harper & Row.
19. Ausubel J (2009) A botanical macroscope. Proc Natl Acad Sci USA 106: 12569.
20. Eldredge N, Gould S (1972) Punctuated equilibrium: An alternative to phyletic gradualism. San Francisco: Freeman Cooper Company.
21. Petersen K (1990) Ergodic Theory (Cambridge Studies in Advanced Mathematics) Cambridge University Press.
22. Moritz C, Cicero C (2004) Dna barcoding: promise and pitfalls. PLoS Biology 2: e354.
23. Castoe T, de Koning A, Kim HM, Gu W, Noonan B, et al. (2009) Evidence for an ancient adaptive episode of convergent molecular evolution. Proc Natl Acad Sci USA 106: 8996.
24. Kallersjo M, Farris J, Chase M, Bremer B, Fay M, et al. (1988) Simultaneous parsimony jackknife analysis of 2538 rbcL DNA sequences reveals support for
major clades of green plants, land plants, seed plants and flowering plants. Plant Systematics Evol 211: 259.

25. McMahon M, Sanderson M (2006) Phylogenetic supermatrix analysis of genbank sequences f from 2228 papilionoid legumes. Syst Biol 55: 811.

26. Sanderson M (2006) Paloverde: an OpenGL 3D phylogeny browser. Bioinformatics 22: 1004–1006.

27. Smith S, Beaulier J, Donoghue M (2009) Mega-phylogeny approach from comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol Biol 9: 37.