Characterization of DNA Primary Sequences by a New Similarity/Diversity Measure Based on the Partial Ordering

R. Todeschini,[†,*] V. Consonni,[†] A. Mauri,[†] and D. Ballabio[‡]

Milano Chemometrics & QSAR Research Group, Department of Environmental Sciences, University of Milano−Bicocca, P.za della Scienza 1, 20126 Milano, Italy, and Department of Food Science and Technologies, University of Milan, via Celoria, 2-20133 Milano, Italy

Received March 20, 2006

The similarity/diversity measures play a fundamental role in library searching, virtual screening, and quantitative structure−activity relationship/quantitative structure−property relationship modeling as well as in genomics and proteomics. In this paper, a new similarity/diversity measure is proposed as a new approach for the analysis of sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements. This methodology can be applied for evaluating molecular similarity/diversity, using sets of sequential descriptors, and for evaluating the similarity between spectra, sensor arrays, and other sequential data such as DNA and protein sequences. The new proposed distance (weighted standardized Hasse distance) is evaluated between pairs of Hasse matrices derived from the classical partial-ordering rules. It can be naturally standardized, thus allowing the interpretation of these distances as absolute values (e.g., percentage) and deriving simple similarity and correlation indices. A simple example is taken to highlight the behavior of the new similarity/diversity measure on DNA sequences taken from the first exons of the β-globins for eight different species. Sensitivity analysis has been also performed, showing the high capability of this measure to take into account small modifications of the DNA sequences. Finally, a comparison with results obtained from the literature is given, together with a comparison with matrix invariants derived from the Hasse matrix.

INTRODUCTION

The concept of similarity and its dual concept of diversity play a fundamental role in several quantitative structure−activity relationship (QSAR) strategies, chemometrics and library searching methods, and virtual screening, as well as in relatively new fields such as genomics and proteomics. Several distance measures both for quantitative and for binary variables have been defined, such as, for example, Euclidean, Manhattan, Minlowski, and Camberra distances for quantitative variables and Hamming, Tanimoto, and Jaccard distances for binary variables. Distances are the quantitative measure of diversity between a pair of objects; thus, large distances indicate large diversity (small similarity), and small distances indicate small diversity (large similarity).

Studies of DNA primary sequencing have become a very important scientific goal, also considering the abundance of DNA sequence data for various species. DNA sequences can be represented as a sequence of four letters (A, T, G, and C), which denote four nucleic acid bases: adenine, thymine, guanine, and cytosine, respectively. Even when sequences are not too long, the searching for their similarity/diversity is a not trivial task, as shown by several sequence comparisons considered in the literature.

As previously proposed,[1−6] a possible strategy to compare DNA primary sequences is the representation of each sequence by a suitable graph-theoretical matrix and, then, the extraction of the corresponding matrix invariants. Matrix invariants have been widely used in several QSAR studies and represent a shortcut to synthesize matrix properties.

In this paper, a new approach to obtain fingerprints of DNA primary sequences is proposed exploiting the partial-ordering (PO) approach on the basis of the Hasse theory.

The similarity/diversity between two sequences is obtained by the definition of a distance between the corresponding Hasse matrices. These distances have some useful properties and seem to show a high sensitivity to changes in structure sequences. The theory of the partial ordering is presented together with the proposed distance between Hasse matrices; some examples with a final comparison among eight DNA sequences of the first exon of β-globin of different species are given.[1]

THEORY

The theory of the proposed approach to the similarity/diversity analysis of DNA sequences is presented, introducing some partial-ordering concepts, the Hasse matrix and the corresponding similarity/diversity measures, and the Hasse matrix invariants.

Partial Ordering (PO). Partial ordering is an approach to the ranking where the relationship of “incomparability” is added to the classical relationships of “greater than”, “less or equal than”, and so forth.[7−10]

Given a set Q of n elements, each described by a vector $\mathbf{x}$ of p variables (attributes), the two elements s and t

[*] Corresponding author telephone: +39-02-6448-2820; fax: +39-02-6448-2839; e-mail: roberto.todeschini@unimib.it.

[†] Department of Environmental Sciences, University of Milano.

[‡] Department of Food Science and Technologies, University of Milan.

10.1021/ci060099e CCC: $33.50
Published on Web 08/24/2006

belonging to Q are comparable if *for all* of the variables x_j either $x_j(t) \geq x_j(s)$ or $x_j(s) \geq x_j(t)$. If $x_j(t) \geq x_j(s)$ for all x_j ($j = 1, ..., p$), then $t \geq s$. The request "for all" is very important and is called the *generality principle*:

$$t \geq s \leftrightarrow x_j(t) \geq x_j(s) \qquad \forall\, j \in [1,p] \tag{1}$$

The ordering relationships between all of the pairs of elements are collected into the Hasse matrix; for each pair of elements s and t, the entry H_{st} of this matrix is

$$H_{st}\begin{cases} +1 & \text{if } x_j(s) \geq x_j(t), \quad \forall\, j \in [1,p] \\ -1 & \text{if } x_j(s) < x_j(t), \quad \forall\, j \in [1,p] \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

If the entry $s-t$ contains $+1$, the entry $t-s$ contains -1; if the entry $s-t$ contains 0, the entry $t-s$ also contains 0. Then, the Hasse matrix is a square $n \times n$ matrix whose elements take only the values 0 and ± 1; if pairs of equal elements are not present, it is also an antisymmetric matrix. In fact, in the presence of elements having the same values for all the variables, in both of the corresponding entries of the Hasse matrix ($s-t$ and $t-s$), a value equal to 1 is stored.

It is interesting to observe that the Hasse matrix contains a holistic view of all of the ordering relationships among the n elements belonging to the set Q. In other words, the Hasse matrix can be assumed as a fingerprint of the ordering relationships among the n elements.

To add more information to the Hasse matrix, the augmented Hasse matrix can be defined by adding to the main diagonal (zero in the original Hasse matrix) any property P of the elements. The property values of each set of n elements are scaled dividing each value by the maximum property value ($H_{ii} = P_i/P_{\mathrm{MAX}}$).

Hasse Similarity/Diversity Measures. Let H^A and H^B be two $n \times n$ Hasse matrices obtained by two different realizations of the variables defining n elements, representing two partial orderings A and B.

The distance between the two partial orderings can be obtained by summing up the differences between the corresponding matrix elements. The distance between A and B can be considered as the contribution of two terms:

$$d_{\mathrm{D}}(A,B) = \frac{\sum_{i=1}^{n} |H_{ii}^{A} - H_{ii}^{B}|}{n}$$

$$d_{\mathrm{H}}(A,B) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |H_{ij}^{A} - H_{ij}^{B}|}{n(n-1)/2} \tag{3}$$

where the first term d_{D} is the contribution to the distance due to the diagonal terms (the property values), while the second term d_{H} is the contribution to the distance due to the off-diagonal terms (the ranking relationships of the Hasse matrix). In both cases, the two distance terms d range from 0 to 1. This is obvious for the diagonal contribution using scaled values but not for the off-diagonal contribution.

In cases where only two variables are considered in building the Hasse matrix and where no discrepancy is observed between the ordering provided by the two variables, the corresponding Hasse matrix obtained contains only $+1$ and -1 values, meaning that a total ranking of the elements exists. If the Hasse matrix is obtained by using a second variable which provides an inverse ordering with respect to the first one, it will comprise only zero values, meaning that no ordering relationships exist among the elements on the basis of these variables. Then, it is noticeable that the maximum theoretical distance between these two matrices is $n \times (n - 1)$.

From the two contributions, a weighted standardized Hasse distance (WSHD) can be defined as a tradeoff between the ranking relationships and the property values. Therefore, the weighted standardized Hasse distance d_{W} can be defined as

$$d_{\mathrm{W}}(A,B) = (1 - w)d_{\mathrm{H}}(A,B) + wd_{\mathrm{D}}(A,B) \qquad 0 \leq d_W \leq 1 \tag{4}$$

where w is a weighting term ranging between 0 and 1. When a weight equal to zero is used, the distance is calculated taking into account only the ranking relationships, while a weight equal to 1 takes into account only the property values. In between, a weight equal to 0.5 takes equally into account both terms, resulting in a distance measure where both the ordering relationships among the elements and their property differences are equally considered.

Moreover, WSHD is a generalized Manhattan distance calculated on the corresponding pairs of elements of two Hasse matrices, thus preserving all of the metric properties of the Manhattan distance. This distance is straightforwardly interpretable as an absolute measure of distance (or as percentage $d \times 100$) or as an absolute measure of similarity after the transformation as $s = 1 - d_{\mathrm{W}}$ or as a correlation measure after the transformation:

$$r_{\mathrm{W}} = 2\cdot(1 - d_{\mathrm{W}}) - 1 \quad -1 \leq r_{\mathrm{W}} \leq +1 \tag{5}$$

The rank correlation r_{H} calculated for $w = 0$ (i.e., $d_{\mathrm{W}} = d_{\mathrm{H}}$) coincides with the Greiner–Kendall rank correlation index, defined as

$$\tau = \frac{4\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}^{+}}{n(n-1)} - 1 \qquad -1 \leq \tau \leq +1 \tag{6}$$

where d_{ij}^{+} is defined as

$$d_{ij}^{+} = \begin{cases} 1 & \text{if } i < j \text{ and } p_i < p_j \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

and p are the ranks of the samples.

Therefore, the Spearman rank correlation uses more information with respect to the Greiner–Kendall rank correlation: the Spearman index is more suitable if no information has to be discarded, while the Greiner–Kendall index is a more robust statistical index.

Hasse Distance between Hasse Matrices of Different Sizes. As explained above, Hasse matrices are square $n \times n$ antisymmetric matrices able to take into account the partial ordering of n elements. When two sets of different element sizes are considered, that is, the two sets are constituted by n_1 and n_2 elements, respectively, with $n_1 > n_2$, two Hasse

Table 1. DNA Sequences from the First Exon of β-Globins for the Eight Considered Species

A human β-globin (92 bases)
ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
B goat alanine β-globin (86 bases)
ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
C opossum β-hemoglobin β M-gene (92 bases)
ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
D *Gallus gallus* β-globin (92 bases)
ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
E lemur β-globin (92 bases)
ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
F mouse β-globin (93 bases)
ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
G rabbit β-globin (90 bases)
ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
H rat β-globin (92 bases)
ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG

matrices **H1** ($n_1 \times n_1$) and **H2** ($n_2 \times n_2$) of different sizes have to be compared. In this case, the WSHD distance is not univocally defined, and the algorithm has to be further developed.

The distance between the two matrices can be calculated by overlapping $n_1 - n_2 + 1$ times the smaller matrix ($n_2 \times n_2$) to the bigger one ($n_1 \times n_1$, the reference matrix), starting from the upper-left corner and shifting the smaller matrix diagonally to the lower-right corner. Each distance between the pair of matrices is calculated as explained above, and the smallest distance among the $n_1 - n_2 + 1$ distances is taken as the final distance. This procedure corresponds to the search for the subset of ordered elements of the bigger matrix which is more similar to the n_2 ordered elements of the smaller matrix.

Matrix Invariants. From a matrix representation of a molecular structure, several graph invariants can be calculated.[11,12] In this case, several mathematical invariants can be easily calculated by using the augmented Hasse matrix, obtained by the procedure outlined above. The most common matrix invariants are the matrix row sums and matrix eigenvalues/eigenvectors, calculated in this case on the absolute values of the matrix entries.

For example, from its eigenvalue vector $\{\lambda_1\lambda_2...\lambda_n\}$, a distance measure can be calculated by using as variables all of the eigenvalues (or a subset of these).

Application of the Hasse Theory to Sequential Data. Data including an ordering variable can be considered as sequential data. These can be characterized by an ordering variable (sequential integer numbers, variable X1) and a property variable (real numbers, variable X2).

Examples of sequential data are mass spectrometry signals, which are ordered by increasing masses, the intensity of signals being the property variable and their position in the spectrum being the ordering variable; IR/UV signals, the signal intensity being the property variable and the wavelength being the ordering variable; and 1D-NMR spectra, the signal intensity being the property variable and the chemical shifts being the ordering variable. In general, all of the spectra achieved along time are intrinsically ordered and can be analyzed as sequential data. Analogously, data based on natural sequences can be also considered as sequential data. In effect, a sequence of integer numbers representing the positions of the elements in the sequence is the ordering variable, while any property characterizing the elements of the sequence is the property variable.

In the case of DNA sequences, which are sequences of the four nucleic acids, the molecular weight (MW) can be chosen as the property characterizing the elements of the sequence, that is, the nucleic acids. For proteins, any physicochemical property of the 20 amino acids of protein sequences can be used as the property variable, while the most relevant protein abundances can be used in the case of proteomic maps.

This kind of data can be easily characterized by Hasse matrices and their similarity/diversity assessed by the previously defined Hasse distance. In this case, the maximum information about the sequence is obtained by using only two variables, that is, the ordering variable (X1) and the property variable (X2). In fact, in this case, the incomparabilities between two samples s and t can be due to only one condition, that is, when the two variables X1 and X2 show an opposite rank:

$$\mathrm{X1}(s) > \mathrm{X1}(t) \quad \text{and} \quad \mathrm{X2}(s) < \mathrm{X2}(t) \quad \text{or} \quad \mathrm{X1}(s) < \mathrm{X1}(t) \quad \text{and} \quad \mathrm{X2}(s) > \mathrm{X2}(t)$$

For example, if three variables are taken into account, the incomparabilities between two samples can be obtained by opposite ranks of X1−X2 or X1−X3 or X2−X3, with a loss of information. In fact, in this case, the presence of zero values in the Hasse matrix cannot be univocally related to a specific relationship.

DATA

With the goal of evaluating the performances of the proposed approach to the similarity/diversity analysis among sequences, eight DNA sequences have been taken from the

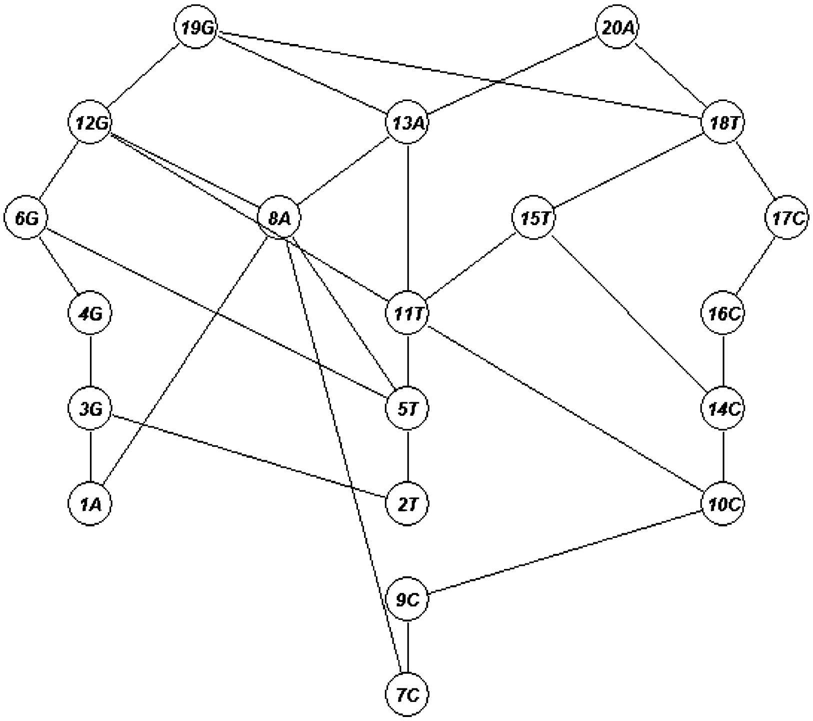

Figure 1. Hasse diagram obtained by the sequence in the text. For each element, the number corresponds to its absolute position in the sequence.

Table 2. Different Representations of the DNA Sequences

label	ID	MW	scaled ID	scaled MW
C	1	111.1	0.25	0.735
T	2	126.0	0.50	0.834
A	3	135.13	0.75	0.894
G	4	151.13	1.00	1.000

literature[1]. These DNA sequences, corresponding to the first exon of β-globin of eight different species, are collected in Table 1. Note that sequence G (rabbit) should be 92 bases and not 90 as reported in Table 1. However, for sake of comparison with previous papers, we have considered 90 bases as quoted by the authors.

The calculations have been performed by a MATLAB[13] module developed by the authors. The module is free and downloadable at www.disat.unimib.it/chm/. The Hasse diagram (Figure 1) has been produced by the software Hasse for Windows (v. 1.02)—GetSynapsed GmbH.

RESULTS AND DISCUSSION

In the proposed approach, each nucleic acid is described by its molecular weight, as shown in Table 2. This property gives the following rank of the four bases: C < T < A < G.

Because the scaled values are those used for the contribution of the diagonal term of the distance (d_D), the use of different scales gives different importance to this term. In this work, scaled ID values, that is, scaling on the sequence of the ordered property, have been used to give higher importance to the differences in the sequence instead of the scaled MW, which has a lower discriminant power.

Therefore, for each sequence, two variables are defined for building the corresponding Hasse matrix: the sequence ID number and the scaled ID on the ordered molecular weight of the nucleic acid sequence. Then, the partial ordering has been defined as

Table 3. Variables Selected in This Work for Building the Hasse Matrices (Columns 1 and 4 in Bold Characters)

ID	base	MW	scaled ID
1	A	135.13	**0.75**
2	T	126.0	**0.50**
3	G	151.13	**1.00**
4	G	151.13	**1.00**
5	T	126.0	**0.50**
...	...	...	**...**
90	C	111.1	**0.25**
91	A	135.13	**0.75**
92	G	151.13	**1.00**

$$x,y \in Q{:}x < y \leftrightarrow \mathrm{ID}(x) < \mathrm{ID}(y) \text{ and } \mathrm{Scaled\_ID}(x) < \mathrm{Scaled\_ID}(y) \quad (8)$$

where Q is the set of bases present in the DNA sequence.

For example, the Hasse matrix of a DNA sequence has been built by using the two variables (in bold in Table 3) corresponding to the ID sequence and to the corresponding rank-scaled molecular weight.

Other kinds of properties can be used instead of the molecular weight, then obtaining different orderings and different Hasse matrices. However, when the matrix diagonal terms are not considered, any property producing the same order of C, T, A, and G gives the same Hasse matrices and then the same distances.

To illustrate the characteristics of the Hasse matrix and the corresponding Hasse diagram, a 20-length sequence constituted by four different elements has been arbitrarily defined:

ATGGTGCACCTGACTCCTGA

Table 4. Standardized Hasse Distances between Human β-Globin Sequences Where Only One Base Has Been Changed[a]

	Seq. C	Seq. G	Seq. T	Seq. A
Seq. C	0	0.669	0.215	0.454
Seq. G	0.669	0	0.454	0.215
Seq. T	0.215	0.454	0	0.239
Seq. A	0.454	0.215	0.239	0

[a] The weight used is zero (d_H). Distances as percentages.

Table 5. Euclidean Distances between Human β-Globin Sequences Where Only One Base Has Been Changed[a]

	Seq. C	Seq. G	Seq. T	Seq. A
Seq. C	0	0.938	0.397	0.678
Seq. G	0.938	0	0.766	0.388
Seq. T	0.397	0.766	0	0.461
Seq. A	0.678	0.388	0.461	0

[a] The first five eigenvalues derived from the Hasse matrix have been used for the Euclidean distance calculation.

In Figure 1, the Hasse diagram of this sequence is represented. As it can be easily noted, the information contained in the diagram not only considers the absolute sequence of the elements but also four chains are highlighted, one for each different element (A, C, G, and T). For example, the sequence of element A is characterized by the path 1−8−13−20, while for element C, the path is 7−9−10−14−16−17. The links between pairs of nodes represent ordering relationships between the elements, while elements on the same horizontal level are incomparable elements (not linked among them).

Sensitivity Analysis. To check the sensitivity of the proposed approach, three different cases have been studied.

In the first case, the human β-globin has been considered as the reference sequence (Seq. C) and three other sequences have been artificially produced, changing only position 10 (C in human β−globin) with G, T, and A, (Seq. G, Seq. T, and Seq. A, respectively). This means that only one base has been changed over a sequence of 92 bases.

For each sequence, the distances calculated from the Hasse matrix are collected in Table 4, and the distances calculated from the Hasse matrix invariants, considering the first five eigenvalues, are collected in Table 5.

As it can be easily observed and as it is expected, with respect to the human β-globin, the most similar modified sequence is the sequence Seq. T, which gives an ordering inversion of only one place with respect to the initial ordering C, T, A, and G. The second most similar is sequence Seq. A, and the last one is sequence Seq. G, producing the most remarkable change in the original ordering sequence. As the Hasse distances can be interpreted as percentages, it can be also observed that all of the distances are lower than 1%, as expected for the minimal changes performed on the original sequence.

Moreover, with different digits, also the Euclidean distances obtained from the Hasse matrix invariants reflect the expected similarity/diversity measures.

Table 6. Standardized Hasse Distances between Human β-Globin Sequences Where the Sequences from 1 to 6 Were Progressively Modified with Respect to the Reference Sequence mod. 0[a]

	mod. 0	mod. 1	mod. 2	mod. 3	mod. 4	mod. 5	mod. 6
mod. 0	0	0.669	0.884	1.087	1.338	2.043	2.246
mod. 1	0.669	0	0.215	0.418	0.669	1.374	1.577
mod. 2	0.884	0.215	0	0.227	0.478	1.183	1.386
mod. 3	1.087	0.418	0.227	0	0.251	0.956	1.159
mod. 4	1.338	0.669	0.478	0.251	0	0.705	0.908
mod. 5	2.043	1.374	1.183	0.956	0.705	0	0.203
mod. 6	2.246	1.577	1.386	1.159	0.908	0.203	0

[a] The weight used is zero (d_H). Distances as percentages.

Table 7. Euclidean Distances between Human β-Globin Sequences Where the Sequences from 1 to 6 Were Progressively Modified with Respect to the Reference Sequence mod. 0[a]

	mod. 0	mod. 1	mod. 2	mod. 3	mod. 4	mod. 5	mod. 6
mod. 0	0	0.973	0.728	0.839	0.739	1.109	0.817
mod. 1	0.973	0	0.412	0.437	0.571	0.587	0.492
mod. 2	0.728	0.412	0	0.272	0.351	0.567	0.511
mod. 3	0.839	0.437	0.272	0	0.555	0.712	0.697
mod. 4	0.739	0.571	0.351	0.555	0	0.384	0.403
mod. 5	1.109	0.587	0.567	0.712	0.384	0	0.485
mod. 6	0.817	0.492	0.511	0.697	0.403	0.485	0

[a] The first five eigenvalues derived from the Hasse matrix have been used for the Euclidean distance calculation.

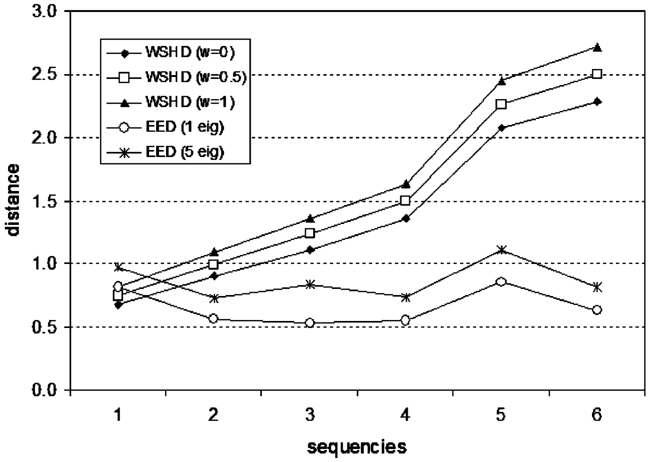

Figure 2. Comparison of the distances between the original human β-globin and six modified sequences. The distances have been calculated with the weighted standardized Hasse distance (WSHD) with different weights (0, 0.5, and 1) and by calculating the Euclidean distance with 1 and 5 eigenvalues.

In the second case, the human β-globin has still been considered as the reference sequence (mod. 0); six other sequences have been artificially generated with one modification in sequence 1 (at position 10), with respect to the reference sequence. Then, iteratively, other modifications with respect to the reference sequence have been performed at positions 20, 30, 40, 50, and 60. In other words, the sequences mod. 1, mod. 2, ..., mod. 6 have one, two, ..., six changes, respectively, with respect to the original human β-globin; each of them preserves the changes of the previous modified sequences (Table 6). As expected, the Hasse distances from the reference sequence mod. 0 increase from sequence 1 to sequence 6 because of the increasing number of modifications.

For the same seven sequences, the Euclidean distances considering the first five eigenvalues obtained by the augmented Hasse matrix have also been calculated (Table 7). In this case, the Euclidean distances do not reflect the expected similarity/diversity among the sequences but appear quite confused. Similar results have been obtained using the leading eigenvalues or other numbers of eigenvalues.

Table 8. WSHD Distances among the Eight β-Globins Calculated for $w = 0$ (Upper Matrix) and $w = 0.5$ (Lower Matrix, in Italics) Are Shown[a]

	A − 92	B − 86	C − 92	D − 92	E − 92	F − 93	G − 90	H − 92
A − 92	0.000	**5.677**	9.281	10.201	**8.302**	13.426	**3.009**	**7.800**
B − 86	***6.472***	0.000	11.272	9.535	**8.167**	13.584	20.260	**8.358**
C − 92	*10.211*	*12.613*	0.000	13.055	12.279	15.444	11.199	13.784
D − 92	*11.214*	*10.581*	*14.136*	0.000	13.629	17.033	11.748	13.247
E − 92	***9.314***	***8.880***	*13.884*	*15.102*	0.000	12.649	**8.477**	11.801
F − 93	*15.816*	*16.385*	*17.777*	*20.065*	*15.156*	0.000	14.132	12.028
G − 90	***3.449***	*23.938*	*12.405*	*12.957*	***9.516***	*16.650*	0.000	**8.727**
H − 92	***8.520***	***9.557***	*15.180*	*14.640*	*13.238*	*14.030*	***9.641***	0.000

[a] The first eight smallest distances are in bold characters for both cases. Distances as percentages.

Table 9. First 20 Eigenvalues Obtained from the Augmented Hasse Matrix for the Eight β-Globins

Eigenvalue number	A − 92	B − 86	C − 92	D − 92	E − 92	F − 93	G − 90	H − 92
1	66.8370	60.8560	63.3260	62.0230	66.4250	66.3680	64.5810	63.7890
2	21.6360	21.5110	22.0490	24.0340	21.3860	21.8060	21.5380	21.9860
3	11.8560	11.5760	13.5270	13.9600	12.4270	13.5620	11.9500	13.4100
4	9.9913	9.2024	11.3640	9.6716	10.4940	10.3320	9.9139	11.0950
5	6.5643	6.4851	7.2851	6.6164	6.6291	6.7461	6.0333	6.5272
6	5.9406	4.7005	5.6812	5.7742	5.7750	6.0124	5.3401	5.9820
7	3.4956	3.0157	3.4825	3.6093	3.0988	3.8328	3.0186	3.7724
8	2.9880	2.7858	3.3710	3.4943	2.8527	2.7537	2.8190	3.4635
9	2.7144	2.5574	2.6200	2.6703	2.5372	2.5804	2.6844	2.7063
10	2.2204	2.2439	2.1711	2.3208	2.3610	2.3662	2.2797	2.5437
11	1.9809	1.8243	1.8283	1.9971	1.9867	1.9524	1.9305	1.9649
12	1.7078	1.6688	1.7884	1.9020	1.6054	1.6620	1.6558	1.6687
13	1.6258	1.5017	1.6903	1.6517	1.5155	1.5435	1.5615	1.4660
14	1.4542	1.3434	1.3500	1.4520	1.4508	1.4633	1.4181	1.3895
15	1.3258	1.2460	1.2593	1.1384	1.2103	1.3545	1.1513	1.2398
16	1.0986	1.0689	1.0400	0.9915	1.0703	1.1249	0.9381	1.0510
17	1.0160	0.8142	1.0018	0.9781	0.9991	1.0339	0.9087	0.9478
18	0.8839	0.7396	0.8754	0.9584	0.9453	0.9476	0.7773	0.9043
19	0.6902	0.6527	0.8181	0.7602	0.6940	0.8393	0.6931	0.7827
20	0.6672	0.6191	0.7936	0.6983	0.6606	0.7844	0.6664	0.7030

Table 10. Euclidean Distances between the Eight β-Globins[a]

	A − 92	B − 86	C − 92	D − 92	E − 92	F − 93	G − 90	H − 92
A − 92	0.000	5.981	3.510	4.814	**0.412**	**0.468**	2.256	3.047
B − 86	*6.191*	0.000	2.471	**1.168**	5.570	5.513	3.726	2.934
C − 92	*4.233*	*4.129*	0.000	**1.303**	3.099	3.042	**1.255**	**0.463**
D − 92	*5.810*	*3.959*	*3.034*	0.000	4.402	4.345	2.558	1.766
E − 92	***1.034***	*5.885*	*3.593*	*5.487*	0.000	**0.057**	1.844	2.636
F − 93	***1.877***	*6.179*	*3.365*	*5.012*	***1.454***	0.000	1.787	2.579
G − 90	***2.453***	*3.894*	*2.940*	*4.267*	***2.137***	***2.773***	0.000	**0.792**
H − 92	*3.668*	*4.309*	***1.099***	*3.126*	*3.095*	*2.809*	***2.463***	0.000

[a] In the upper matrix are those distances obtained from the leading eigenvalues, and in the lower matrix (in italics) are those obtained from the first 10 eigenvalues. The first eight smallest distances are shown in bold characters for both cases.

The profiles of the distances from the original human β-globin, obtained by the different methods, are shown in Figure 2. The three profiles obtained by setting the weights $w = 0$, $w = 0.5$, and $w = 1$ are similar, even if the similarities obtained by $w = 0$ are the highest ones, the similarity also being considered because of the similar rankings of the bases. The profiles obtained by calculating the Euclidean distance using only the leading eigenvalue (EED1) and the first five eigenvalues (EED5) do not seem to reflect the induced increasing diversity between the original human β-globin and the modified sequences.

In the third case, the specific role of the off-diagonal elements of the Hasse matrix has been highlighted comparing the human β-globin (A) and the opossum β-globin (C).

When only the off-diagonal terms are considered ($w = 0$), the distance between the sequences A and C is 8.815, while considering only the diagonal terms ($w = 1$), the distance is 10.598.

Then, a change of one base is performed in position 30 for both sequences, substituting the base T with A for the human β-globin and the base C with G for the opossum β-globin. The contribution of position 30 to the diagonal term is $|T - C| = 0.25$ for the two original sequences and $|A - G| = 0.25$ for the two modified sequences. As expected, when only the diagonal terms ($w = 1$) are considered, the distance is again 10.598; however, the distance calculated considering only the off-diagonal terms is 8.140. This difference is due to the different ordering relationships induced by the change in the two sequences, then taking into account the global change of the two sequences.

Comparison of β-Globins. The same approach presented above has been used for evaluating the similarity/diversity among the eight β-globins of Table 1.

Because the first exon of the eight β-globins is constituted by a different number of bases, the calculation of the Hasse distances between pairs of matrices of different sizes is

Table 11. Eight Most Similar Pairs of β-Globins from Different Approaches[a]

	rank							
method	R1	R2	R3	R4	R5	R6	R7	R8
WSHD(0)	***AG***	***AB***	***AH***	***BE***	AE	BH	EG	GH
WSHD(0.25)	***AG***	***AB***	***AH***	***BE***	AE	BH	EG	GH
WSHD(0.50)	***AG***	***AB***	***AH***	***BE***	AE	EG	BH	GH
WSHD(1)	***AG***	***AB***	***AH***	***BE***	AE	EH	GH	AC
EED-1	EF	AE	**AF**	CH	***FH***	***BD***	CG	CD
EED-5	CH	AE	EF	**AF**	**AG**	EG	GH	FG
EED-10	AE	CH	EF	***AF***	***AG***	EG	GH	FG
ref 1 − 1	BG	DE	***FH***	***AB***	EF	AE	DF	EH
ref 1 − 2	DE	***FH***	BG	AD	***AG***	AE	EF	DF
ref 1 − 3	DE	BG	***FH***	AD	AE	***AG***	EF	DG
ref 1 − 4	DE	BG	***FH***	AD	AE	***AG***	DG	EF
ref 4	**AF**	**AG**	**FH**	**AH**	**BE**	**AB**	**BF**	**BD**

[a] In bold italic characters are the pairs present also among the most similar pairs found in ref 4 (all in bold).

performed using the sequential algorithm explained above.

The distances for w equal to 0 (upper matrix) and w equal to 0.5 (lower matrix, in italics) are shown in Table 8, highlighting the eight smallest distances in bold characters.

In both cases, the pair of the most similar exons is constituted by the human (A) and rabbit β-globins (G).

The first 20 eigenvalues of the augmented Hasse matrix are shown in Table 9; the Euclidean distances calculated using the leading eigenvalues, and the first five eigenvalues of the augmented Hasse matrix are shown in Table 10.

Finally, a comparison between the most similar pairs of β-globins has been performed with the different approaches proposed in this work and in some literature papers.[1−4] In Table 11, the first eight most similar pairs of β-globins (R1−R8) are collected for different weights of WSHD; for Euclidean distances calculated with 1 (EED1), 5 (EED5), and 10 (EED10) eigenvalues obtained from the augmented Hasse matrix; and for two literature comparisons.[1,4] In bold, italic characters are shown the β-globin pairs that have been evaluated as the most similar by the counting of consecutive similar pairs of bases.[4]

As it can be noted, the five most similar pairs found by the Hasse approach are independent from the distance weight, and the first four are present in the ranking obtained by Randic.[4] In particular, AG corresponds to the human−rabbit, AB to the human−goat, AH to the human−rat, and BE to the goat−lemur β-globins. By considering the eigenvalue approaches, only some of them are also present in the similarity analysis proposed by Randic:[4] in particular, AF (human−mouse β-globin) and the previous AG pair; FH and BD are also found by Randic.[4] By considering the results achieved by Randic and Vracko,[1] only the pairs FH and AG coincide with the most similar pairs given by Randic,[4] as it has been noticed also by the authors, while some spurious similarities are probably found.

CONCLUSIONS

The proposed similarity/diversity measure appears as a new approach to sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements. In particular, the weighted Hasse distance shows some advantages: (a) the Hasse matrices and the corresponding distances are calculated by a straightforward algorithm; (b) the distance is naturally standardized, allowing a natural interpretation of the obtained values; (c) the distances are able to take into account the whole structure of the ranking relationships of the sequences; (d) the distances can be obtained by a flexible strategy (the weights) depending on the specific similarity/diversity study; (e) a simple rank correlation measure is derived, also taking into account incomparabilities among sequence elements.

Studies of the application of this new distance measure to whole coding DNA sequences and other kind of sequences (such as spectra) are in progress.

REFERENCES AND NOTES

(1) Randic, M.; Vracko, M. On the Similarity of DNA Primary Sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 599−606.

(2) Nandy, A. A. A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Applications to Globin Genes. *Curr. Sci.* **1994**, *66*, 309−313.

(3) Randic, M. Condensed Representation of DNA Sequences. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 50−56.

(4) Randic, M. Condensed Representation of DNA Sequences by Condensed Matrix. *Chem. Phys. Lett.* **2000**, *317*, 29−34.

(5) He, P.-a.; Wang, J. Characteristic Sequences for DNA Primary Sequence. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1080−1085.

(6) Roberts, F. S. Applications of Combinatorics and Graph Theory to the Biological and Social Sciences: Seven Fundamental Ideas. In *Applications of Combinatorics and Graph Theory to the Biological and Social* Sciences; Roberts, F., Ed.; Springer-Verlag: New York, 1989; pp 1−37.

(7) E. Halfon. On Ranking Chemicals for Environmental Hazard. *Environ. Sci. Technol.* **1986**, *20*, 1173−1179.

(8) Brüggemann, R.; Bartel, H.-G. A Theoretical Concept to Rank Environmentally Significant Chemicals. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 211−217.

(9) Pavan, M.; Todeschini, R. New Indices for Analysing Partial Ranking Diagrams. *Anal. Chim. Acta.* **2004**, *515*, 167−181.

(10) Brüggemann, R.; Franck, H.; Kerber, A. Proceedings of the Conference "Partial Orders in Environmental Sciences and Chemistry". *MATCH* **2004**, *54*, 485−689.

(11) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(12) Devillers, J.; Balaban, A. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon & Breach: Amsterdam, The Netherlands, 2000.

(13) *MATLAB*, v. 6.5; The MathWorks Inc.: Natick, MA, 2002.

CI060099E