CROATICA CHEMICA ACTA
CCACAA **81** (4) 657–664 (2008)

ISSN-0011-1643
CCA-3287
*Original Scientific Paper*

# A New Similarity/Diversity Measure for the Characterization of DNA Sequences

**Roberto Todeschini,* Davide Ballabio, Viviana Consonni and Andrea Mauri**

** Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, 20126 Milano, Italy*

RECEIVED MARCH 24, 2006; REVISED JULY 30, 2008; ACCEPTED SEPTEMBER 22, 2008

In this paper, a new similarity/diversity measure is proposed as a new approach to the analysis of sequential data, where useful information can be also obtained by the ordering relationships between the sequence elements. This methodology has been applied to characterize DNA sequences, evaluating their similarity/diversity. 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. DNA sequences taken from the first exons of the beta-globins for eight different species have been analyzed. 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 literature is given.

*Keywords*
DNA
partial ordering
Hasse matrix
distances
similarity/diversity
rank correlation

## INTRODUCTION

In several fields sequential data are very common, *i.e.* data where some property is ordered by a ranking variable such as the intensity of signals obtained by mass spectrometry ordered by increasing masses, the intensity of IR/UV signals ordered by wave lengths, the intensity of 1D – NMR spectra ordered by chemical shifts, and, in general, all the spectra achieved along time. Analogously, data based on natural sequences, such as DNA or protein sequences, can be also considered as sequential data.

In particular, 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 letter (A, T, G, C), which denote the four nucleic acid bases. Even when sequences are not too long, the searching for their similarity/diversity is not usually easy as shown by several sequence comparisons considered in literature papers.

As previously proposed,[1–4] a possible strategy to compare DNA primary sequences is the representation of each sequence by a suitable matrix and then the extraction of the corresponding matrix invariants. Matrix invariants have been widely used in several QSAR studies and represent a short-cut to synthesize matrix properties.

In this paper, a new approach to obtain fingerprints of DNA primary sequences is proposed exploiting the partial ordering approach based on the so-called Hasse matrices. The similarity/diversity between two sequences is obtained by the definition of a distance between the corresponding Hasse matrices. These distances have some

* Author to whom correspondence should be addressed. (roberto.todeschini@unimib.it)

useful properties and seem to show a high sensitivity to changes in structure sequences. Partial ordering was already used with the aim of comparing proteomic maps,[5–7] even if the ranking was used in order to get embedded graphs, while Hasse diagrams were not evaluated for this aim. In the first part of the paper, the theory of the partial ordering is presented together with the proposed distance between Hasse matrices; then, some examples with a final comparison among eight DNA sequences of the first exon of beta-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.

### *Partial Ordering (PO)*

Partial Ordering is a ranking approach where the relationship of »incomparability« is added to the classical relationships of »greater than«, »less or equal than«, *etc*.[8–10]

Given a set Q of $n$ elements, each described by a vector $\boldsymbol{x}$ of $p$ variables (attributes), the two elements $s$ and $t$ belonging to Q are comparable if *for all* 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 $t \rhd s$, *i.e*. $t$ covers $s$ (or $s$ is *covered* by $t$). The request »for all« is very important and is called the *generality principle*:

$$t \rhd s \Leftrightarrow x_j(t) \geq x_j(s) \qquad \forall j \qquad (1)$$

The ordering relationships between all 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:

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

If the entry $s$-$t$ contains +1, the entry $t$-$s$ contains –1; if the entry $s$-$t$ contains 0, also the entry $t$-$s$ contains 0. Then, the Hasse matrix is a squared $n \times n$ antisymmetric matrix, whose elements take only the values 0 and ±1. Moreover, in presence of elements having the same variable values (for all the variables), in both 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 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.

In order 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 ($H_{ii} = P_i / P_{MAX}$).

### *Hasse Similarity/Diversity Measures*

Let be $\boldsymbol{H}^A$ and $\boldsymbol{H}^B$ two $n \times n$ Hasse matrices obtained by two different realizations of the variables defining $n$ elements, *i.e*. 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:

$$d_D(A,B) = \frac{\sum_{i=1}^{n} \left| H_{ii}^A - H_{ii}^B \right|}{n} \qquad d_H(A,B) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left| H_{ij}^A - H_{ij}^B \right|}{n \cdot (n-1) / 2} \qquad (3)$$

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 zero to one. This is obvious for the diagonal contribution using scaled values, but not for the off-diagonal contribution.

In case that only two variables are considered in building the Hasse matrix and that 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 based on 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 trade-off between the ranking relationships and the property values. Therefore, the weighted standardized Hasse distance $d_W$ can be defined as:

$$d_w(A,B) = (1-w) \cdot d_H(A,B) + w \cdot d_D(A,B) \qquad 0 \leq d_w \leq 1 \quad (4)$$

where $w$ is a weighting term ranging between 0 and 1. Using a weight equal to zero, the distance is calculated taking into account only the ranking relationships, while a weight equal to one takes into account only the property values. 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 Manhattan distance calculated on the corresponding pairs of elements of two Hasse matrices, thus preserving all 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 $s = 1 - d_W$.

Only considering the $d_H$ term, a simple measure of rank correlation can be also derived as:

$$r_H = (1 - d_H) \cdot 2 - 1 \qquad -1 \leq r_H \leq +1 \qquad (5)$$

Unlikely the Spearman rank correlation, this correlation measure also takes into account the presence of incomparabilities.

### *Hasse Distance between Hasse Matrices of Different Size*

As previously explained, Hasse matrices are squared $n \times n$ antisymmetric matrices able to take into account the partial ordering of $n$ elements. When two sets of different element size are considered, *i.e*. the two sets are constituted by $n_1$ and $n_2$ elements, respectively, with $n_1 > n_2$, two Hasse matrices $\boldsymbol{H}1$ ($n_1 \times n_1$) and $\boldsymbol{H}2$ ($n_2 \times n_2$) of different size 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 smallest matrix ($n_2 \times n_2$) to the biggest one ($n_1 \times n_1$, the reference matrix), starting from the left-up corner and shifting the smallest matrix diagonally until the right-down 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 search the subset of ordered elements of the biggest matrix which is more similar to the $n_2$ ordered elements of the smallest matrix.

### *Application of the Hasse Theory to Sequential Data*

Data where an ordering variable is present can be considered as sequential data. For example, the intensity of signals obtained by mass spectrometry are ordered by increasing masses, the intensity of 1D – NMR spectra are ordered by chemical shifts, and, in general, all the spectra achieved along time are intrinsically ordered. Analogously, data based on natural sequences can be also considered as sequential data. In fact, a defined property of the elements of natural sequences, such as the letters of alphabetic sequences, the 4 nucleic bases of DNA sequences, the 20 aminoacids of protein sequences, the most relevant protein abundances of proteomic maps, can be used as ordering variable.

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 content is obtained by using only two variables. In fact, the incomparabilities between two cases $s$ and $t$ can be due to only one condition, *i.e*. when the two variables $X1$ and $X2$ show an opposite rank:

$X1(s) > X1(t)$ and $X2(s) < X2(t)$ or $X1(s) < X1(t)$ and $X2(s) > X2(t)$

For example, if three variables are taken into account, the incomparabilities 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 can not 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 literature.[1] These DNA sequences corresponding to the first exon of beta-globin of eight different species are collected in Table I. The calculations have been performed by a MATLAB[11] module developed by the authors.

## RESULTS AND DISCUSSION

In the proposed approach each nucleic acid is described by his relative molecular mass, $M_r$, as shown in Table II. This property gives the following rank of the four bases: C, T, A, G.

Since 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, *i.e*. scaling on the sequence of the ordered property, have been used to give higher importance to the differences in the sequence. 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 relative molecular mass of the nucleic acid sequence. For example, the Hasse matrix of a DNA sequence has been built by using the two variables (in bold in Table III) corresponding to the ID sequence and to the corresponding rank-scaled molecular mass.

The use of the first variable guarantees a non-trivial Hasse matrix simply based on a total ordering of the relative molecular mass. Other kinds of properties can be used instead of the molecular mass, then obtaining different orderings and different Hasse matrices. However,

TABLE I. The DNA sequences from the first exon of beta-globins for the eight considered species

**A** human beta-globin (92 bases)

ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGG
TGAGGCCCTGGGCAG

**B** goat alanine beta-globin (86 bases)

ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGG
CCCTGGGCAG

**C** opossum beta-hemoglobin beta M-gene (92 bases)

ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGG
TGAGGCCCTTGGCAG

**D** gallus gallus beta-globin (92 bases)

ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCT
GGCCAG

**E** lemur beta-globin (92 bases)

ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGG
CGAGGCCTTGGGCAG

**F** mouse beta-globin (93 bases)

ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGG
TGAGGCCCTGGGCAGG

**G** rabbit beta-globin (90 bases)

ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTG
GTGAGGCCCTGGGC

**H** rat beta-globin (92 bases)

ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATGTTGGCGC
TGAGGCCCTGGGCAG

when the matrix diagonal terms are not considered, any property producing the same order of C, T, A, G gives the same Hasse matrices and then the same distances.

In order to illustrate the characteristics of the Hasse matrix and the corresponding Hasse diagram, a 20-length sequence constituted by 4 different elements has been arbitrarily defined:

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | T | G | G | T | G | C | A | C | C | T | G | A | C | T | C | C | T | G | A |

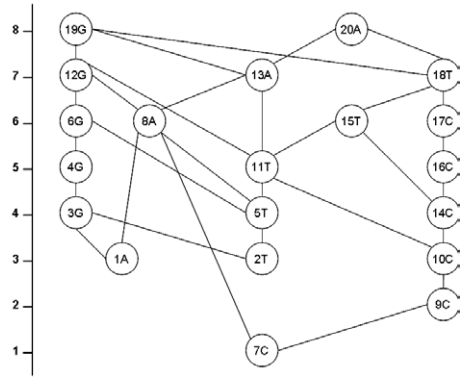


Figure 1. Hasse diagram obtained by the DNA sequence ATGGTGCACCTGACTCCTGA. For each element, the number corresponds to its absolute position in the sequence.

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 se-

TABLE II. Different representations of the DNA sequences

| Label | ID | $M_r$ | Scaled ID | Scaled $M_r$ |
|---|---|---|---|---|
| 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 |

TABLE III. The variables selected in this work for building the Hasse matrices (columns 1 and 4 in bold characters)

| ID | Base | $M_r$ | 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** |

quence of the elements, but also four linear extensions are highlighted, one for each different element (A, C, G, T). For example, the sequence of the element A is characterized by the path 1-8-13-20, while for the 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).

TABLE IV. Standardized Hasse distances (percentages) between human beta-globin sequences where only one base has been changed. The weight used is zero ($d_H$)

| | C | G | T | A |
|---|---|---|---|---|
| C | 0 | 0.681 | 0.215 | 0.466 |
| G | 0.681 | 0 | 0.466 | 0.215 |
| T | 0.215 | 0.466 | 0 | 0.251 |
| A | 0.466 | 0.215 | 0.251 | 0 |

*Sensitivity Analysis of the Weighted Hasse Distance*

In order to check the sensitivity of the proposed approach, three different cases have been studied.

In the first case, the human beta-globin has been considered as the reference sequence and three other sequences have been artificially produced changing only the position 10 (C in human beta-globin) with G, T and A, respectively (G, T and A). 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 IV.

As it can be easily observed and as it is expected, with respect to the human beta-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, G. The second most similar is the sequence Seq.A and the last one is the sequence Seq. G, producing the most remarkable change in the original ordering sequence. Since the Hasse distances can be interpreted as percentages, it can be also observed that all the distances are lower than 1 %, as expected for the minimal changes performed on the original sequence.

In the second case, the human beta-globin has been still considered as reference sequence (M0); other 6 sequences have been artificially generated with one modification in the 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 position 20, 30, 40, 50 and 60. In other words, the sequences M1, M2,..., M6 have 1, 2,..., 6 changes, respectively, with respect to the original human beta-globin; each of them preserves the changes of the previous modified sequences (Table V). As expected, the Hasse distances from the reference sequence M0 increase from sequence 1 to sequence 6 due to the increasing number of modifications. The three distance 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, being considered also the similarity due to the similar rankings of the bases.

The specific role of the off-diagonal elements of the Hasse matrix has been highlighted also comparing the human beta-globin and the opossum beta-globin (Table I). When only the off-diagonal terms are considered ($w = 0$), the distance between the two sequences 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 the sequences, substituting the base T with A for the human beta-globin and the base C with G for the opossum beta-globin. The contribution of the 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, considering only the diagonal terms ($w = 1$) 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.

Another check of the sensitivity of the proposed approach was performed directly looking at the graphical representation of the 20-dimensional sequence shown in Figure 1. The basis A in position eight (8A) has a central role in the Hasse diagram and the greatest number of links

TABLE V. Standardized Hasse distances between human beta-globin sequences where the sequences from 1 to 6 were progressively modified with respect to the reference sequence M0. The weight used is zero ($d_H$)

| | M 0 | M 1 | M 2 | M 3 | M 4 | M 5 | M 6 |
|---|---|---|---|---|---|---|---|
| M 0 | 0 | 0.681 | 0.908 | 1.111 | 1.362 | 2.078 | 2.281 |
| M 1 | 0.681 | 0 | 0.227 | 0.430 | 0.681 | 1.398 | 1.601 |
| M 2 | 0.908 | 0.227 | 0 | 0.227 | 0.478 | 1.195 | 1.398 |
| M 3 | 1.111 | 0.430 | 0.227 | 0 | 0.251 | 0.968 | 1.171 |
| M 4 | 1.362 | 0.681 | 0.478 | 0.251 | 0 | 0.717 | 0.920 |
| M 5 | 2.078 | 1.398 | 1.195 | 0.968 | 0.717 | 0 | 0.203 |
| M 6 | 2.281 | 1.601 | 1.398 | 1.171 | 0.920 | 0.203 | 0 |

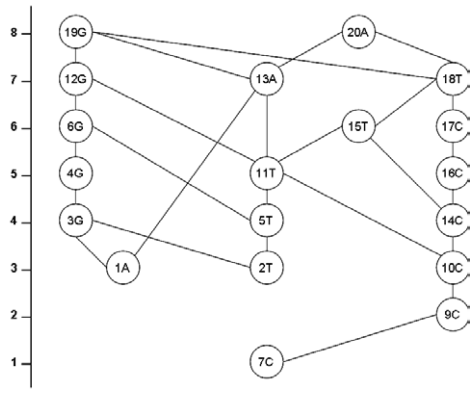


Figure 2. Hasse diagram obtained by removing the basis 8A.

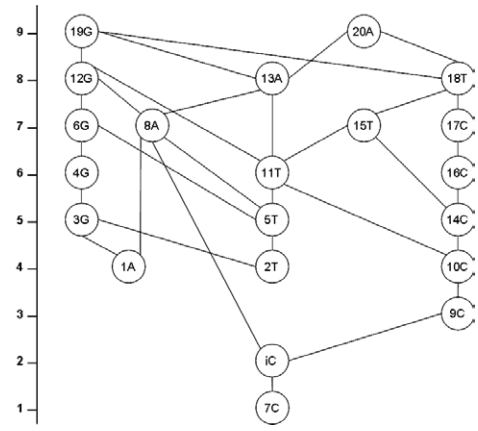


Figure 5. Hasse diagram obtained by inserting a new basis C (denoted as iC) before the basis 8A.

with the other bases (12G, 13A, 5T, 7C, 1A). Thus, removing, exchanging or shifting this basis could greatly influence the Hasse diagram. A simple study has been performed evaluating what happens (a) if the basis 8A is absent (Figure 2), (b) if an inversion of the positions in the sequence between basis 8A and 9C is performed (Figure 3), and (c) if it is shifted up and replaced by a new basis A (Figure 4), (d) by a new basis C (Figure 5), (e) by a new basis G (Figure 6), and (f) by a new basis T (Figure 7). These additional bases have been inserted in position 8 of the original sequence and denoted as iA, iC, iG, and iT, respectively; in all the new Hasse diagrams, the 20 bases have been denoted as in the original sequence. As it can be easily shown, the main structure of the original graphical representation remains the same in all the new graphs, while only local changes are observed.

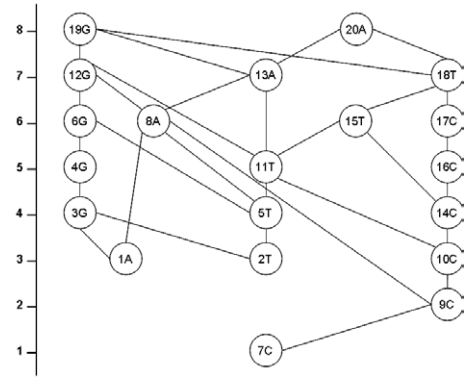


Figure 3. Hasse diagram obtained by inverting the positions of the bases 8A and 9C.

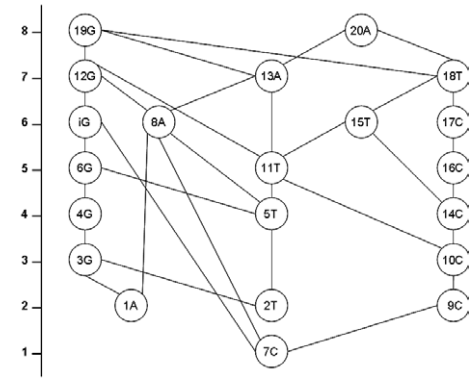


Figure 6. Hasse diagram obtained by inserting a new basis G (denoted as iG) before the basis 8A.

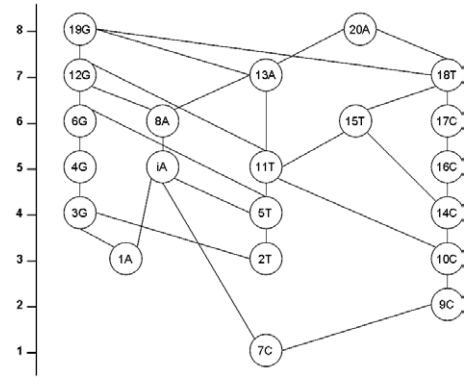


Figure 4. Hasse diagram obtained by inserting a new basis A (denoted as iA) before the basis 8A.

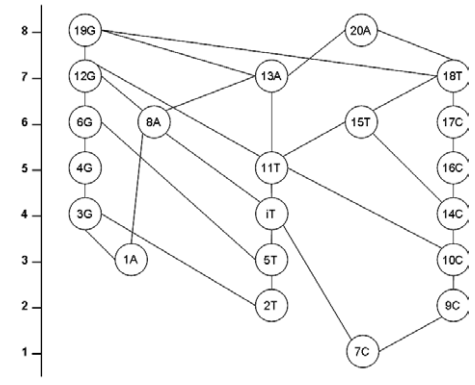


Figure 7. Hasse diagram obtained by inserting a new basis T (denoted as iT) before the basis 8A.

### *Comparison of Eight Beta-globins*

The same approach presented above has been used for evaluating the similarity/diversity among the eight beta--globins of Table I. Since the first exon of the eight beta--globins is constituted by different numbers of bases, the calculation of the Hasse distances between pairs of matrices of different size is performed using the sequential algorithm previously explained.

The distances of the eight beta-globins have been calculated using three different weights: $w = 0$ refers to distances only based on the off-diagonal terms of the Hasse matrix; $w = 0.25$ refers to distances based on the off-diagonal terms of the Hasse matrix (75 %) and on the distance calculated considering the diagonal terms (25 %); $w = 0.5$ refers to distances where both contributions are equally considered (50 %). The distances for $w$ equal to 0 (upper matrix) and $w$ equal to 0.5 (lower matrix, in italics) are shown in Table VI, 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 beta-globins (G).

Finally, a comparison between the most similar pairs of beta-globins has been performed. In Table VII, the first eight most similar pairs of beta-globins (R1-R8) are colected for different distance weights, together with some literature results.[4] In bold italic characters are shown the beta-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 Randić.[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 beta-globins. By considering the results achieved by Randić and Vracko,[1] only the pairs FH and AG coincide with the most similar pairs given in Randić,[4] as it has been noticed also by the authors, while some spurious similarities are probably found.

## CONCLUSIONS

Some interesting results are achieved by applying the proposed weighted Hasse distance for the similarity/diversity analysis of DNA sequences. In particular, the weighted Hasse distance shows some advantages: (a) the distance is naturally standardized, allowing a natural inter-

TABLE VI. The WSHD distances among the 8 beta-globins calculated for $w = 0$ (upper matrix) and $w = 0.5$ (lower matrix, in italics) are shown. The first eight smallest distances are in bold characters for both cases

| | **A – 92** | **B – 86** | **C – 92** | **D – 92** | **E – 92** | **F – 93** | **G – 90** | **H – 92** |
|---|---|---|---|---|---|---|---|---|
| **A – 92** | 0 | **6.238** | 8.815 | 10.690 | **8.313** | 13.963 | **3.558** | **7.836** |
| **B – 86** | ***4.474*** | 0 | 11.272 | 9.535 | **8.167** | 13.584 | 20.260 | **8.358** |
| **C – 92** | *6.296* | *8.104* | 0 | 13.055 | 12.279 | 15.444 | 11.199 | 13.784 |
| **D – 92** | *7.612* | *6.778* | *9.168* | 0 | 13.629 | 17.033 | 11.748 | 13.247 |
| **E – 92** | ***5.991*** | ***5.777*** | *8.891* | *9.738* | 0 | 12.649 | **8.477** | 11.801 |
| **F – 93** | *10.276* | *10.211* | *11.243* | *12.588* | *9.529* | 0 | 14.132 | 12.028 |
| **G – 90** | ***2.540*** | *14.986* | *7.962* | *8.371* | ***6.126*** | *10.387* | 0 | **8.727** |
| **H – 92** | ***5.509*** | ***6.050*** | *9.796* | *9.437* | *8.455* | *8.819* | ***6.222*** | 0 |

TABLE VII. The eight most similar pairs of beta-globins from different approaches. In bold italic characters the pairs present also among the most similar found in Ref. 4 (all in bold)

| | 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 | BH | EG | GH |
| WSHD(1) | ***AG*** | ***AB*** | ***AH*** | ***BE*** | AE | EH | GH | AC |
| 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*** |

pretation of the obtained values; (b) the distances are able to take into account the whole structure of the ranking relationships of the sequences; (c) the distances can be obtained by a flexible strategy (the weights) depending on the specific similarity/diversity study; (d) a simple rank correlation measure is derived, also taking into account incomparabilities among sequence elements, (e) the Hasse matrices and the corresponding distances are calculated by a straightforward algorithm.

This new representation of the DNA sequence by using the Hasse matrices seems to be very promising due to the capability of the Hasse matrix to encode local modifications in the sequences and to preserve all the remaining order relationships. The Hasse matrix can also be used as a graph-theoretical matrix to derive several descriptors, such as, for example, the different functions of the eigenvalues[12] and to use these descriptors for similarity/diversity analysis by using different measures of distance. In this case, the problem due to the different size of the Hasse matrix, being its dimension dependent of the size of the sequence, is overcome by definition.

## REFERENCES


1. M. Randić and M. Vracko, *J. Chem. Inf. Comput. Sci*. **40** (2000) 599–606.
2. A. A. Nandy, *Curr. Sci*. **66** (1994) 309–313.
3. M. Randić, *J. Chem. Inf. Comput. Sci*. **40** (2000) 50–56.
4. M. Randić, *Chem. Phys. Lett.* **317** (2000) 29–34.
5. M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci* **42** (2002) 983–992.
6. M. Randić, *Int. J. Quant. Chem.* **90** (2002) 848–858.
7. M. Randić, J. Zupan, M. Novič, B. D. Gute, and S. C. Basak, *SAR QSAR Environ. Res.* **13** (2002) 689–703.
8. R. Brüggemann and H.-G. Bartel, *J. Chem. Inf. Comput. Sci*. **39** (1999) 211–217.
9. M. Pavan and R. Todeschini, *Anal. Chim. Acta*. **515** (2004) 167–181.
10. R. Brüggemann, H. Franck, and A. Kerber, *MATCH-Commun. Math. Comput. Chem*. **54** (2004) 485–689.
11. MATLAB (ver. 6.5); The MathWorks Inc., Natick (MA), USA.
12. V. Consonni and R. Todeschini, *MATCH-Commun. Math. Comput. Chem*. **60** (2008) 3–14.


## SAŽETAK

### Novo mjerilo za karakterizaciju sličnosti/različitosti nukleotidnih sljedova DNA

**Roberto Todeschini, Davide Ballabio, Viviana Consonni i Andrea Mauri**

U ovom je radu predloženo novo mjerilo sličnosti/različitosti, kao novi pristup u analazi podataka sekvenciranja, koji također pruža korisne informacije dobivene uređivanjem odnosa među sekvencnim elementima. Ta je metodologija primijenjena na karakerizaciju sekvenci DNA, kojom se vrednuje njihova sličnost/različitost. Nova predložena udaljenost (ponderirana standardizirana Hasseova udaljenost) vrednovana je između parova Hasseovih matrica izvedenih iz klasičnih parcijalnih pravila uređivanja. Ona se može prirodno standardizirati, omogućujući stoga interpretaciju tih udaljenosti kao apsolutnih vrijednosti (npr. postotka) i izvođenje jednostavnih indeksa sličnosti i korelacije. Analizirani su nukleotidni sljedovi DNA, koji pripadaju prvom eksonu beta-globinskih gena iz osam različitih organizama. Provedena je također analiza osjetljivosti, koja je pokazala sposobnost tih novih mjerila da uzmu u obzir male preinake u sekvencama DNA. Na kraju je dana i usporedba s literaturnim podatcima.