CROATICA CHEMICA ACTA
CCACAA **81** (2) 253–261 (2008)

ISSN-0011-1643
CCA-3237
Original Scientific Paper

Analysis of Tagged* Sequences by Line Distance Matrices and Grid Paths**

Agnes Pisanski-Peterlin[a] and Tomaž Pisanski[b,*]**

[a]*Faculty of Arts, University of Ljubljana, Slovenia*

[b]*IMFM, University of Ljubljana, Ljubljana, and University of Primorska, Koper, Slovenia*

RECEIVED MARCH 26, 2007; ACCEPTED DECEMBER 10, 2007

Keywords
protein sequence
graphical representation
graph
metadiscourse
virtual genetic code
preview
review
eigenvalue
grid paths

Sequences with certain types of elements emphasized arising in various branches of science and humanities can be analyzed using similar mathematical tools. In this paper, we present a novel approach applicable to DNA and protein sequences, on the one hand, and text analysis *via* metadiscourse on the other.

INTRODUCTION AND MOTIVATION

It is not uncommon for studies of different phenomena in science to be prone to the same mathematical apparatus. Calculus, initially invented for use in mechanics (Newton) and geometry (Leibniz), turned out to be quite useful in diverse, unrelated branches of science from quantum physics and biology to economics.

We could give a number of examples of phenomena, models and theories to illustrate how disconnected topics from different disciplines show the same mathematical approach. However, we will mention only a few: the count of paths and the count of Kekule valence structures in the class of benzenoids representing a lattice. Another case relates to the count of rotamers (isomers of *n*-alkanes in 3-D space embedded on a diamond lattice) with their representation in ternary code.[1–4]

With the advance of combinatorial chemistry, molecular descriptors that were once invented for modeling of only limited classes of molecules,[5–7] such as paraffins, have gained on importance as a collective classification and identification tool for potentially useful new molecules with prescribed properties. One of the key issues of contemporary bioinformatics is sequence analysis. Since both DNA and written text come in the form of a finite sequence, it is not surprising that some methods first designed by computer scientists for text analysis have found applications in computational biology.

* The word tag has a new meaning and should not be confused with Expressed Sequences Tags or ESTs.

** Dedicated to Nenad Trinajstić on the occasion of his 70th birthday.

*** Author to whom correspondence should be addressed. (E-mail: tomaz.pisanski@fmf.uni-lj.si)

In this paper, we highlight three ideas that were used in the past in quite different contexts and show their universal applicability.

1. A tagged sequence is a sequence in which symbols are grouped into classes, called tags, and each tag is assigned a numerical value. A sequence can then be represented as a single number expressed in some base or as a polynomial. This approach is known in combinatorics as the generating function approach.

2. Binary sequences can be described as certain grid paths. In the case of certain metadiscourse elements, such grid paths give a visual representation of previews and reviews. Several parameters can be defined in order to study the similarities or dissimilarities of texts.

3. Several novel methods have been proposed in the study of DNA sequences. We single out the line distance matrix and colored map visualization.

As a synthesis, we exchange the approaches: we apply grid path techniques to DNA sequencing and the line distance matrix and colored map visualization to metadiscourse.

The concept of metadiscourse has received a great deal of attention in recent years within the context of applied linguistics. Research into metadiscourse has been carried out by a number of authors,[8–16] many of their studies focusing on intercultural rhetorical differences in academic discourse, in which successful intercultural discourse production/reception is essential. Hyland[15] defines metadiscourse as »the cover term for the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community«. Metadiscourse is used to organize a text and help the reader interpret and evaluate it.

For practical reasons, this analysis is limited to two types of metadiscourse used for signposting. Signposting is a type of metadiscourse which helps the reader understand how the propositional content of the text is organized. Two subtypes of signposting can be distinguished: retrospective signposting, or reviews**,** elements used to remind the reader of what has already been said in the text, and prospective signposting, or previews, elements used to announce what is about to be presented in the text. The sentences below are examples of a review and preview, respectively, both from mathematics research articles.

»We have noted in §1 that the Lebesgue integral is not powerful enough to integrate every derivative.«

»Having established our terminology, we now consider a central idea of this article, the Mandelbrot set, or Mandelset.«

Graphical representations of DNA sequences, initiated by Hamori[17] and subsequently expanded by Nandy and others,[18–22] illustrate some of the advantages of graphical representations of complex chemical, biochemical and biological systems. One of the advantages of graphical representations is that they allow visual inspection of similarities/dissimilarities between complex systems that need not be apparent from the raw experimental data. Later on, it was shown how graphical representations of DNA sequences lead to numerical characterizations of such sequences and a subsequent quantitative analysis of similarities/dissimilarities between the sequences based on mathematical invariants of the sequences,[21,23–25] and most recently how graphical representation of bio-sequences can lead to sequence alignment.[26]

Extension of such considerations to proteome maps[27–37] has offered the possibility of representing proteome maps in a digital format suitable for computer storing, search, and processing. These more recent extensions of graph theoretical methodologies to complex biochemical systems would not have been possible without continuing efforts within Chemical Graph Theory to characterize molecular systems, including applications to the quantitative structure-property/activity relationship, by mathematical invariants derived from the matrices representing such systems.[38,39]

Matrix representations of complex chemical and biological systems have made it possible to obtain useful numerical characterizations of DNA sequences,[21,23–25] proteome maps,[27–37] and the »degree of folding« of proteins.[40–42] The origin of this methodology can be traced to an earlier work on numerical characterization of the »degree of bending« of the molecular skeleton of smaller chain-like molecules, in which the notion of the distance/distance matrix, ***D*/*D***, was introduced.[43] The approach based on ***D*/*D*** matrices represents a special case of the numerical representation of complex systems with which a matrix can be associated. Such analysis can be also applied to systems of unknown geometry as long as one can associate with them suitable mathematical objects of definite geometry, which then allow matrix representation in the form of numerical matrices. Thus, when considering DNA, the geometry of DNA is not required (and is most often not known) because 2-D or 3-D geometry of a mathematical object representing DNA is used to construct a matrix that represents DNA. Moreover, one can construct a matrix for DNA even without any graphical representation just by manipulating sequential labels.[44] Hence, it may be possible to construct mathematical invariants even for structures that have no geometrical representation, actual or fictitious.

TAGGED SEQUENCES – A MATHEMATICAL MODEL

Let us assume we are given a finite alphabet Σ of symbols. As it is customary in the theory of formal languages we denote by Σ^* the set of all finite sequences form-ed from the symbols of Σ. Let $\Pi = \{\Pi_0, \Pi_1, ..., \Pi_{r-1}\}$ be an ordered set partition of alphabet Σ. This means that every symbol of Σ belongs to exactly one class of partition Π. Equivalently, we may define a mapping π: $\Sigma \rightarrow \{0, 1, 2, ...\}$ that assigns each symbol the index of its class: $\pi(a) = i$ if and only if $a \in \Pi_i$. We assume that $\pi(\Sigma)$ is an interval such that if $j \in \pi(\Sigma)$ and $0 \leq k < j$, then $k \in \pi(\Sigma)$. Note that $\pi(\Sigma)$ can be considered to be an alphabet, so each word from Σ^* can be mapped to a word from $\pi(\Sigma)^*$. The structure (Σ, Π, π) will be called a tagging system. For a given sequence w and a tagging system, the sequence w will be called a tagged sequence. A tagged sequence can be viewed as a natural number expressed in base r or as a polynomial. If w is a sequence, we will use w^R to denote its reverse, *i.e.*, the sequence obtained from w by reading it backwards. If a word w is tagged to $\pi(w)$ and the first symbol is mapped to a non-zero value, then we say that w is properly tagged. Sometimes only a subset $\Sigma' \subseteq \Sigma$ of symbols matters. In such a case, we may extend the mapping π: $\Sigma \rightarrow \{\varepsilon, 0, 1, ...\}$ where ε denotes the empty word and $\pi(a) = \varepsilon$ if and only if $a \in \Sigma \setminus \Sigma'$. The quadruple $(\Sigma, \Sigma', \Pi, \pi)$ is called a reducing tagging system.

Let us explain these concepts using two simple examples.

Example 1 – Let the alphabet be the standard Latin alphabet and let it be partitioned into five vowels A,E,I,O,U and consonants. If we map consonants to 0 and vowels to 1, the word DINOSAURUS is mapped to 0101011010.

Example 2 – Let the alphabet be the alphabet of the bases A,C,G,T. If we map A to 0, C to 1, G to 2 and T to 3, the DNA sequence ATGGTGCACCTGACTCCTGAG is mapped to a tagged sequence 032232101132013113202. Using a different tagging system, we may single out the occurrences of the letter A: Let $\Pi_0 = \{C, G, T\}$ and $\Pi_1 = \{A\}$. Then the same sequence is mapped to a binary tagged sequence 100000010000100000010.

METADISCOURSE MODELS AND SEQUENCES

In a study of metadiscourse models,[45,46] a series of scientific texts was first surveyed and prepared for analysis. The text that is being analyzed for its metadiscourse is divided into sections and sections are further subdivided into subsections. Appearances of previews and reviews are marked. Each preview is labeled by the letter P and a target section is indicated. In the same way, a review is labeled by the letter R.

EXOPHORIC TEXT I (P1 EXOPHORIC) (R1 I)(P2 EXOPHORIC)(P3 I) (R2 I) (R3 EXOPHORIC) (P4 I) (R4 EXOPHORIC) II (P5 II) (R5 I) (P6 II) (P7 II) III (R6 III) (R7 III) IV (R8 IV) (P8 VI) (P9 IV) V(P10 V) VI (P11 VI) (R9 IV)

All other symbols are irrelevant for our purposes and we keep only the symbols P and R and the reduced tagging sequence is obtained from:

$$w = \text{P R P P R R P R P R P P R R R P P P P R}$$

by replacing each (P) by 1 and (R) by 0. This defines a number written in binary that encodes the essence of the position of metadiscourse elements in a given text. In principle, other metadiscourse elements could have been taken and encoded in a similar way. By considering d distinct metadiscourse elements and using 0 to represent words, sections or other non-metadiscourse elements, it would be possible to represent the text under analysis as a number in base $(d+1)$ if no reductions are used.

THE STAIR DIAGRAM

For a given binary sequence w, we may define the stair diagram in a grid. We start at the bottom left corner in (0,0). If the next symbol is 0, we move one step to the right; if the symbol is 1, we move one step upwards. The result is a stair-like diagram from (0,0) to (a,b) where $a = 0(w)$ and $b = 1(w)$. The slope $k = b / a$ indicates the ratio between ones and zeros in w. If $k > 1$, there are

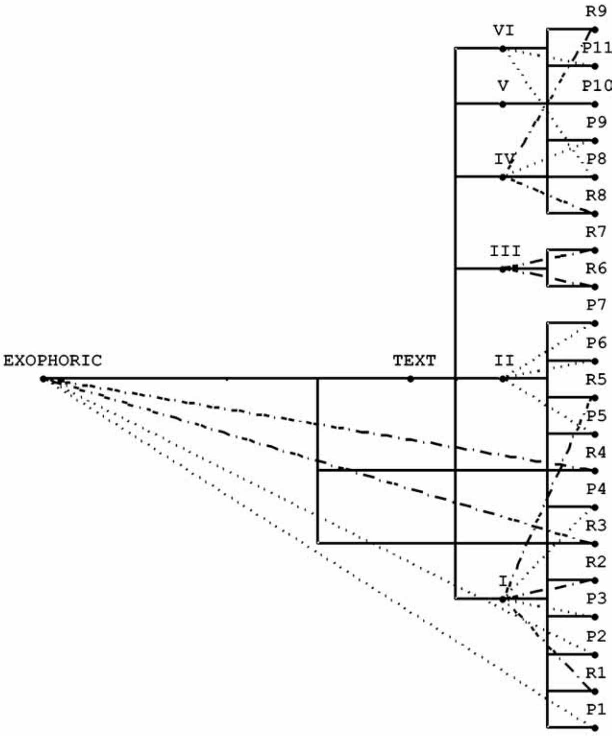

Figure 1. A tree composed from the original sequence[45] of metadiscourse data, showing not only the position of previews and reviews but also their scope in the tree-like structure of the document under analysis.[46] In this paper, only the linear structure of the document is considered.

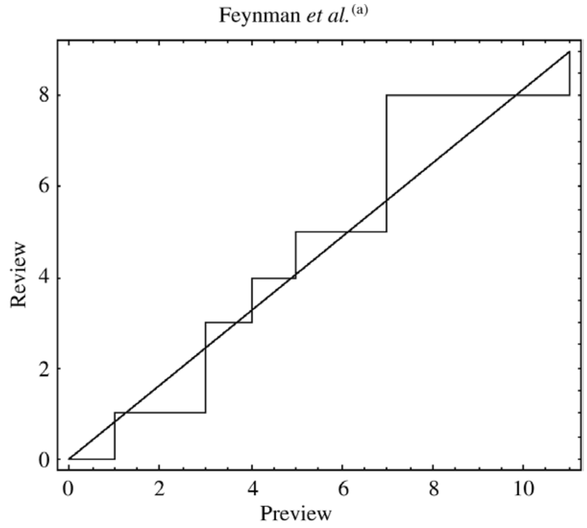

Figure 2. The stair diagram of the sequence of reviews and previews P R P P R R P R P R P P R R R P P P P R. The text starts in the bottom left and ends in the top right corner. It begins with a single preview (P) and a single review (R). Various parameters can be computed, such as the number of stairs or the ratio of the number of stairs 6 to the number of metadiscourse elements 20 equaling 3/10. The number of bends *b(w)* = 11. The slope of the stairs is defined by the quotient reviews(*w*)/previews(*w*) = 9/11.

(a) R. P. Feynman and R. B. Leighton, in: M. Sands (Ed.), The Feynman Lectures on Physics, Volume II. Mainly Electromagnetics and Matter, Addison Wesley, Reading, MA, USA, 1964. This work was the object of the lingvistic analysis in Refs. 45 and 46.

more ones than zeros in *w*. Clearly $|w| = a + b$ is the length of *w*. We may count the number of bends $b(w) := 01(w) + 10(w)$. Another interesting measure is how many of the stairs lie below and how many above the line passing through the points (0,0) and *(a,b)*. This line is called the critical line. Measuring the two areas is an interesting exercise in computer programming. The word *w* can be partitioned into sub-words that have the area alternating below and above the critical line.

Since one would expect to have more previews in the first part of discourse and more reviews in the last part of discourse, one would expect to have the stairs mostly below the critical line. With regard to Figure 2, we mention in passing that with suitable orientation the same diagram represents a path in a lattice and can be used for the count of paths in benzenoid hydrocarbons[47] and configurations in CI (configuration interaction)[48] calculations of quantum chemistry. There is also a relationship to Catalan numbers.[49] In graph theory, several matrices are associated to a graph. In the following, we use important properties of the distance matrix, an object well-known in graph theory.[50]

THE SEQUENCE AND A SYMBOL

Let us consider a DNA sequence (of four nucleotides A,T,G,C) and represent distances between occurrences of A (or distances between T, or G or C). For example, the first exon of human β-globin gene starts as: ATGGTGCACCTGACTCCTGAG... We will consider only the sequence composed of the first 21 nucleotides and focus on G only. Instead of taking G, we could have taken any other base and repeat the analysis four times.

$$w = \mathrm{ATGGTGCACCTGACTCCTGAG}$$

Positions of G in the sequence can be written in the vector

$$\boldsymbol{t} = (3,4,6,12,19,21).$$

If we are interested only in the appearance of G in *w*, we denote this as:

$$(w,\mathrm{G}) = \mathrm{ATGGTGCACCTGACTCCTGAG}$$

where any symbol following the rightmost occurrence of G may be deleted. Since other symbols are irrelevant, (*w*,G) in fact represents:

$$(w,\mathrm{G}) = \mathrm{NNGGNGNNNNNGNNNNNNGNG}$$

where N stands for any symbol different from G.

Reducing the Sequence to a Single Number

We may reduce the information stored in *(w,x)* to a single number, which we denote by *N(w,x)*. Note that the matrix ***D****(w,*G*)* does not change if some symbols different from G are added at the tail of the sequence *(w,*G*)* = NNGGNGNNNNNGNNNNNNGNG. Hence, we may encode this in binary:

$$(w,\mathrm{G}) = 001101000001000000101.$$

Since this sequence ends in 1, it represents proper tagging. By reversing the binary sequence, we obtain:

$$R(w,\mathrm{G}) = 101000000100000101100$$

We may read this as a natural number *N(w,*G*)* represented in binary.

$$N(w,\mathrm{G}) = 1010000000100000101100_2 = 1312812$$

Line Distance Matrix from a Number

The process can be reversed. This means that we may associate to each natural number *N* the distance matrix ***LD***(*N*) by first writing *N* in binary, reversing the sequence and then associating the matrix to the sequence. The number of rows and columns of ***LD*** is the number of ones in the binary representation of *N*. Furthermore, matrix ***LD*** is defined as follows:

$(\boldsymbol{LD})_{ij}$ represents the distance between the i-th and j-th occurrences of »1« in the binary representation of N.

This works in a more general setting. Let w be a sequence and let x be a symbol occurring in such a sequence. We denote such a pair by *(w,x)*.

Examples:

1. For example, w may be a DNA sequence and x any of the base codons.

2. Sequence w may represent a protein and x the position of a given amino acid.

3. w may represent a text, a word or a sentence and x may be any letter appearing in a given text.

4. Sequence w may represent a given text, the symbols are words and x is a given preview (or review).

We will model various sequences as special graphs, called paths. The edges of such a path can be weighted by the actual distance between two consecutive occurrences of a given symbol x. To each weighted path, we may associate a line distance matrix $\boldsymbol{LD}(w,x)$.

The distance matrix of a graph can be easily constructed using the adjacency matrix of the graph, $\boldsymbol{A}(w,x)$, and one of the standard algorithms for constructing the distance matrix of a graph.[51–55] The elements $[\boldsymbol{D}]_{ij}$ of $\boldsymbol{D}$ are defined as follows:

$[\boldsymbol{D}]_{ij} = d$ if the occurrences of symbols x_i and x_j are at a distance d (1).

In case w = ATGGTGCACCTGACTCCTGAG

(w,G) = NNGGNGNNNNNGNNNNNNGNG

where X stands for any symbol different from G and its corresponding line distance matrix is:

$$\boldsymbol{D}(w,\text{G}) = \begin{bmatrix} 0 & 1 & 3 & 9 & 16 & 18 \\ 1 & 0 & 2 & 8 & 15 & 17 \\ 3 & 2 & 0 & 6 & 13 & 15 \\ 9 & 8 & 6 & 0 & 7 & 9 \\ 16 & 15 & 13 & 7 & 0 & 2 \\ 18 & 17 & 15 & 9 & 2 & 0 \end{bmatrix}$$

In the paper by Jaklič, Pisanski, and Randić,[56] it is proven that line distance matrices of size n have one positive and $n–1$ negative eigenvalues. Visual representation of Cauchy's interlacing property for line distance matrices is considered. It is also shown that the line distance matrix is completely determined by its first row. In particular, if the sequence $\boldsymbol{t} = (t_1, t_2,..., t_n)$, $0 < t_1 < t_2 <... < t_n$, be a given vector. Then the $n \times n$ matrix:

$$(\boldsymbol{D})_{ij} = \begin{cases} t_i - t_j \,, j \le i \\ t_j - t_i \,, j > i \end{cases}$$

is the distance matrix for the interval with the subdivision given by $0 < t_1 < t_2 <... < t_n$.

Since line distance matrices are symmetric, their eigenvalues are real. It is interesting to study the distribution of eigenvalues. Jaklič, Pisanski, and Randić[56] obtained the following results.

Theorem. – (Jaklič, Pisanski, Randić):[56] Let $\boldsymbol{D} \in R^{n\times n}$ be a line distance matrix, defined by a vector $\boldsymbol{t}$ and let $\boldsymbol{D}^{(i)} := \boldsymbol{D}(1{:}i,1{:}i)$, i = 1, 2,...,n be its principal submatrices. Let $\lambda_i(\boldsymbol{D}^{(i)}) \le \lambda_{i-1}(\boldsymbol{D}^{(i)}) \le ... \le \lambda_1(\boldsymbol{D}^{(i)})$ be the eigenvalues of matrix $\boldsymbol{D}^{(i)}$. Then $\lambda_1(\boldsymbol{D}^{(i)}) > 0$, $\lambda_2(\boldsymbol{D}^{(i)}) < 0$ for $i > 1$ and $\lambda_i(\boldsymbol{D}^{(1)}) = 0$.

Since, $\boldsymbol{D} = \boldsymbol{D}^{(n)}$ the following corollary holds.

Corollary. – (Jaklič, Pisanski, Randić):[56] A line distance matrix has exactly one positive eigenvalue.

According to Cauchy's Interlacing Theorem,[57] the eigenvalues of principal submatrices of line distance matrices interlace. We may apply this analysis to the sequence of previews and reviews:

P	R	P	P	R	R	P	R	P	R	P	P	R	R	R	P	P	P	P	R
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

CASE I, SYMBOL P:
t_{P} = [1 3 4 7 9 11 12 16 17 18 19]

The line distance matrix can be computed directly from the definition of a distance matrix. However, in that case, the first row of the matrix determines the rest of the matrix. The matrix is symmetric. Hence, if we construct the upper triangle, the rest of the matrix is determined. Each subsequent row of the upper triangle is obtained by subtracting the first element of the previous row from the current element of the previous row:

0	2	3	6	8	10	11	15	16	17	18
2	0	1	4	6	8	9	13	14	15	16
3	1	0	3	5	7	8	12	13	14	15
6	4	3	0	2	4	5	9	10	11	12
8	6	5	2	0	2	3	7	8	9	10
10	8	7	4	2	0	1	5	6	7	8
11	9	8	5	3	1	0	4	5	6	7
15	13	12	9	7	5	4	0	1	2	3
16	14	13	10	8	6	5	1	0	1	2
17	15	14	11	9	7	6	2	1	0	1
18	16	15	12	10	8	7	3	2	1	0

Eigenvalues (P):

–49.9751, –14.5629, –4.2278, –2.9771, –2.4206, –1.4112, –0.9754, –0.8010, –0.7894, –0.5823, 78.7228.

Eigenvalues of the principal submatrices can be written in a triangular structure:

0										
–2.0000	2.0000									
–3.2019	–0.9112	4.1131								
–6.4069	–2.8158	–0.7967	10.0194							
–10.8861	–3.2883	–1.6257	–0.7952	16.5953						
–15.8284	–4.2554	–2.2911	–1.2817	–0.7941	24.4506					
–20.8542	–5.2138	–2.5326	–1.4534	–0.8391	–0.7939	31.6869				
–27.0232	–9.1552	–3.6648	–2.5111	–1.4055	–0.8033	–0.7905	45.3536			
–34.5758	–11.7979	–3.9527	–2.5253	–1.4144	–0.9566	–0.7975	–0.7878	56.8080		
–42.3298	–13.3568	–4.0783	–2.5584	–1.7864	–1.4008	–0.8045	–0.7914	–0.6560	67.7624	
–49.9751	–14.5629	–4.2278	–2.9771	–2.4206	–1.4112	–0.9754	–0.8010	–0.7894	–0.5823	78.7228

Their graphical representation is shown in Figure 3. Cauchy's interlacing property for the eigenvalues of the principal submatrices of the line distance matrix ***D*** is presented.

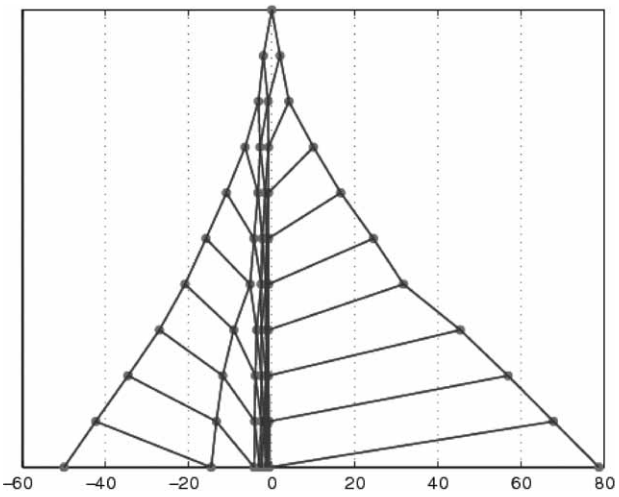

Figure 3. Cauchy's interlacing property for previews (P).

CASE II: If we take (R) instead of (P) in the same sequence, we obtain the following results.

$$t_{\mathrm{R}} = [2\ 5\ 6\ 8\ 10\ 13\ 14\ 15\ 20]$$

Line distance matrix:

0	3	4	6	8	11	12	13	18
3	0	1	3	5	8	9	10	15
4	1	0	2	4	7	8	9	14
6	3	2	0	2	5	6	7	12
8	5	4	2	0	3	4	5	10
11	8	7	5	3	0	1	2	7
12	9	8	6	4	1	0	1	6
13	10	9	7	5	2	1	0	5
18	15	14	12	10	7	6	5	0

Eigenvalues (R):

–33.0633 –12.0259 –5.4694 –3.0583 –1.6223
–1.3094 –0.7868 –0.6435 57.9789

and the interlacing eigenvalues of the principal submatrices are presented in Figure 4.

0								
–3.0000	3.0000							
–4.5529	–0.9568	5.5096						
–6.4069	–2.8158	–0.7967	10.0194					
–9.5062	–4.2182	–1.5162	–0.7881	16.0287				
–14.5817	–6.0459	–2.6183	–1.3732	–0.7868	25.4059			
–20.4478	–6.8576	–2.9045	–1.4270	–0.9071	–0.7867	33.3308		
–26.0871	–7.4434	–3.0907	–1.7722	–1.3317	–0.7870	–0.6528	41.1649	
–33.0633	–12.0259	–5.4694	–3.0583	–1.6223	–1.3094	–0.7868	–0.6435	57.9789

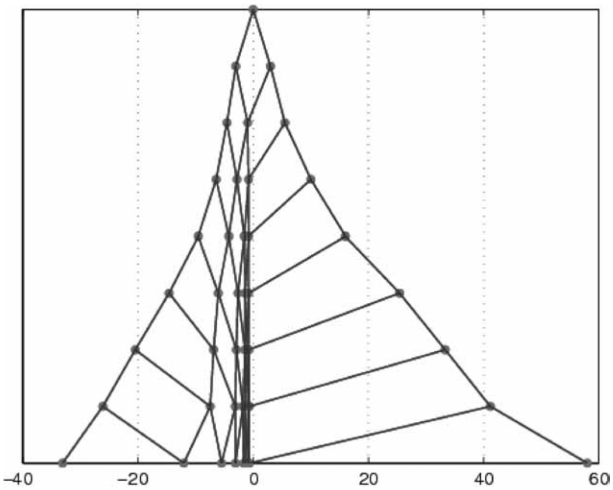

Figure 4. Cauchy's interlacing property for reviews (R).

Let us Visualize this Interesting Property

There is one vertex in the top layer, two vertices in the next layer, *etc*. Each layer has one more vertex than the previous layer. The horizontal location of each vertex of the k-th layer is determined by the corresponding eigenvalue of the k by k principal submatrix. The i-th vertex in the k-th layer is joined to the i-th and (i+1)-st vertex in the (k+1)-st layer. Due to the interlacing property, the grid is planar and there is no intersection of the edges.

SEQUENCES AND COLORED MAPS

Milan Randić[22] associated a planar map to a DNA sequence in such a way that each of the four bases was encoded with a color. This approach was later generalized.[58,59] Here, we briefly touch upon the subject. An interested reader can find further details in our references.

The same sequence can be regarded as a path in various meshes. In the examples presented here we take the sequence:

$$w = \mathrm{P\ R\ P\ P\ R\ R\ P\ R\ P\ R\ P\ P\ R\ R\ R\ P\ P\ P\ P\ R}$$

If we put it as a spiral into a triangular mesh, we obtain a two-colored map containing 4 »R« regions and three »P« regions, see Figures 5 and 6. The same sequence can be put into the square mesh, see Figure 7 or into a hexagonal mesh, see Figure 8.

Maps can be used for visual representation of a complicated sequence. On the other hand, from each map one can generate an auxiliary graph whose vertices are colored regions and edges are inserted if two regions are adjacent. In case only two colors are used, the auxiliary graph is necessarily bipartite. For instance, the auxiliary graph of Figure 8 is a path P_4 on 4 vertices.

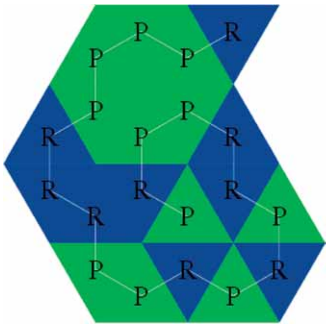

Figure 5. Triangular mesh.

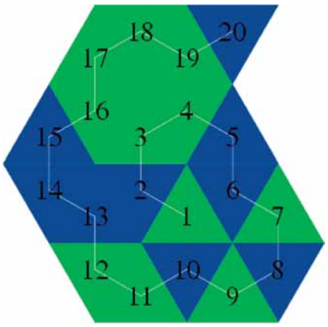

Figure 6. The order of symbols in sequence spirals in the triangular mesh.

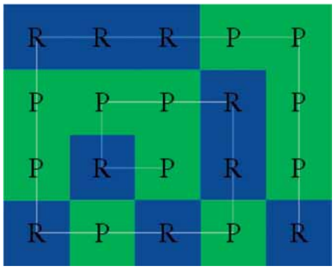

Figure 7. Square mesh.

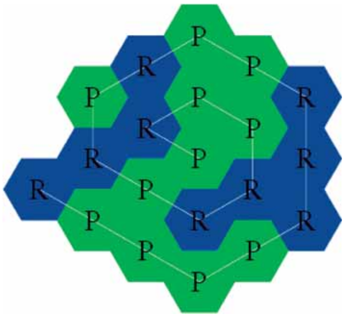

Figure 8. Hexagonal mesh.

CONCLUSIONS

In this paper, we have shown that several questions related to sequences arising from a variety of problems from unrelated scientific disciplines can be studied with the same mathematical model, which we call tagged sequences. We have not shown specific applications of this approach. This will be the goal of another study. However, several examples have been presented and some interesting observations follow if we transfer the newly discovered techniques from linguistics into bioinformatics and *vice versa*.

Acknowledgments. – This work was supported in part by the Slovenian Research Agency under Research Grants P1-0294, J1-6062, and L1-7230. The authors thank Gašper Jaklič for his help with computations and for careful reading of the manuscript and to Milan Randić whose broad and deep knowledge of the presented material helped in shaping the manuscript. Special thanks go to the anonymous referees whose suggestions improved the quality of presentation.

REFERENCES

1. M. Randić, *Int. J. Quantum Chem: Quantum Chem. Symp*. **7** (1980) 187–197.
2. M. Randić and M. Razinger, *On Characterization of 3D Molecular Structure*, in: *From Chemical Topology to Three-dimensional Geometry*, Plenum Press, New York, 1977, pp. 159–236.
3. A. T. Balaban, *Rev. Roum. Chim.* **2** (1976) 1049–1071.
4. A. T. Balaban, *MATCH Commun. Math. Comput. Chem.* **2** (1976) 51–61.
5. H. Wiener, *J. Am. Chem. Soc.* **69** (1947) 17–20.
6. H. Hosoya, *Bull. Chem. Soc. Jpn.* **44** (1971) 2332–2339.
7. M. Randić, *J. Am. Chem. Soc*. **97** (1975) 6609–6615.
8. W. J. Kopple, *Coll. Compos. Commun.* **36** (1985) 82–94.
9. A. Mauranen, *Engl. Specif. Purp.* **12** (1993) 3–22.
10. P. Intaraprawat and M. S. Steffensen, *J. Second Lang. Writ.* **4** (1995) 253–272.
11. D. Bunton, *Engl. Specif. Purp.* **18** (1999) S41–S56.
12. P. A. Fuertes-Olivera, M. Velasco-Sacristán, A. Arribas-Baño, and E. Samaniego-Fernández, *J. Pragmatics* **33** (2001) 1291–1307.
13. T. Dahl, *J. Pragmatics* **36** (2004) 1807–1825.
14. K. Hyland, *J. Second Lang. Writ.* **13** (2004) 133–151.
15. K. Hyland, *Metadiscourse. Exploring Interaction in Writing*, London and New York, Continuum, 2005.
16. E. Ifantidou, *J. Pragmatics* **37** (2005) 1925–1953.
17. E. Hamori, *BioTechniques* **7** (1989) 710–720.
18. A. Nandy, *Curr. Sci.* **66** (1994) 309–314.
19. A. Nandy, *Curr. Sci.* **70** (1996) 661–668.
20. P. M. Leong and S. Morgenthaler, *Comput. Appl. Biosci.* **11** (1995) 503–507.
21. M. Randić, M. Vračko, N. Lerš, and D. Plavšić, *Chem. Phys. Lett*. **368** (2003) 1–6.
22. M. Randić, *Chem. Phys. Lett*. **386** (2004) 468–471.
23. M. Randić, M. Vračko, N. Lerš, and D. Plavšić, *Chem. Phys. Lett*. **371** (2003) 202–207.
24. M. Randić, M. Vračko, J. Zupan, and M. Novič, *Chem. Phys. Lett.* **373** (2003) 558–562.
25. A. T. Balaban, D. Plavšić, and M. Randić, *Chem. Phys. Lett.* **379** (2003) 147–154.
26. M. Randić, *J. Chem. Inf. Comput. Sci*. **41** (2001) 1330–1338.
27. M. Randić, J. Zupan, and M. Novič, *J. Chem. Inf. Comput. Sci*. **41** (2001) 1339–1344.
28. M. Randić, F. Witzmann, M. Vračko, and S. C. Basak, *Med. Chem. Res*. **10** (2001) 456–479.
29. M. Randić, *Quantitative Characterization of Proteomics Maps by Matrix Invariants*, in: P. M. Conn (Ed.), *Handbook of Proteomic Methods,* Humana Press Inc., Totowa, NJ, 2003, pp. 429–450.
30. M. Randić, M. Novič, and M. Vračko, *J. Proteome Res*. **1** (2002) 217–226.
31. M. Randić and S. C. Basak, *J. Chem. Inf. Comput. Sci.* **42** (2002) 983–992
32. M. Randić, J. Zupan, M. Novič, B. D. Gute, and S. C. Basak, *SAR QSAR Environ. Res.* **13** (2002) 689–703.
33. Ž. Bajzer, M. Randić, D. Plavšić, and S. C. Basak, *J. Mol. Graphics Modell*. **22** (2003) 1–9.
34. M. Randić, N. Lerš, D. Plavšić, and S. C. Basak, *Croat. Chem. Acta* **77** (2004) 345–351.
35. M. Randić, N. Lerš, D. Plavšić, and S. C. Basak, *J. Proteome Res.* **3** (2004) 778–785.
36. Ž. Bajzer, S. C. Basak, M. Vračko, M. Grobelšek, and M. Randić, *Use of Proteomics Based Biodescriptors in the Characterization of Chemical Toxicity,* in: M. J. Cunningham (Ed.), *Genomics and Proteomic Application in Toxicity Testing*, Humana Press, Tatowa, NJ, 2004.
37. M. Randić, N. Lerš, D. Vukičević, D. Plavšić, B. D. Gute, and S. C. Basak, *J. Proteome Res*. **4** (2005) 1347–1352.
38. J. Devillers and A. T. Balaban (Eds.), *Topological Indices and Related Descriptors, in: QSAR and QSPR*, Gordon and Breach, Amsterdam, 1999.
39. R. Todeschini and V. Consonni, *Handbook of Molecular Descriptors, Methods and Principles in Medicinal Chemistry,* Vol. 11, Wiley-VCH, Weinheim, 2000.
40. M. Randić and G. Krilov, *Chem. Phys. Lett*. **272** (1997) 115–119.
41. M. Randić and G. Krilov, *Int. J. Quantum Chem.* **65** (1997) 1065–1076.
42. M. Randić and G. Krilov, *New J. Chem*. **28** (2004) 1608–1614.
43. M. Randić, A. F. Kleiner, and L. M. De Alba, *J. Chem. Inf. Comput. Sci.* **34** (1994) 277–286.
44. M. Randić and A. T. Balaban, *J. Chem. Inf. Comput. Sci*. **43** (2003) 532–539.
45. A. Pisanski-Peterlin, *Konvencije rabe metabesedilnih elementov*, PhD Thesis, Ljubljana, 2005.
46. A. Pisanski-Peterlin, *Engl. Specif. Purp.* **25** (2005) 307–319.
47. W. J. He and W. C. He, *Theor. Chim. Acta* **75** (1989) 389–400.
48. I. Shavitt, *The Method of Configuration Interaction*, in: H. F. Schaefer (Ed.)*, Methods of Electronic Structure Theory*, New York, Plenum Press, 1977, pp. 189–275.
49. A. Tucker, *Applied Combinatorics,* 4th Ed., New York, Wiley, 2002.
50. F. Buckley and F. Harary, *Distances in Graphs*, Addison-Wesley, Redwood City, CA, 1990.
51. M. Barysz, D. Plavšić, and N. Trinajstić, *MATCH Commun. Math. Comput. Chem.* **19** (1986) 89–116.
52. B. Mohar and T. Pisanski, *J. Math. Chem.* **2** (1988) 267–277.
53. W. R. Müller, K. Szymanski, J. V. Knop, and N. Trinajstić, *J. Comput. Chem.* **8** (1987) 170–173.
54. Z. Mihalić, D. Veljan, D. Amić, S. Nikolić, D. Plavšić, and N. Trinajstić, *J. Math. Chem.* **11** (1992) 223–258.

55. H. P. Schultz, *J. Chem. Inf. Comput. Sci*. **40** (2000) 1158–1159.
56. G. Jaklič, T. Pisanski, and M. Randić, *J. Comput. Biology* **13** (2006) 1558–1564.
57. J. W. Demmel, *Applied Numerical Linear Algebra,* Philadelphia, SIAM, 1997.
58. M. Randić, N. Lerš, D. Plavšić, S. C. Basak, and A. T. Balaban, *Chem. Phys. Lett. Inf. Model.* **407** (2005) 205–208.
59. M. Randić, A. T. Balaban, M. Novič, A. Založnik, and T. Pisanski, *Period. Biol.* **107** (2005) 403–414.

SAŽETAK

Analiza nizova svrstanih elemenata pomoću matrice udaljenosti na pravcu i staza na rešetci

Agnes Pisanski-Peterlin i Tomaž Pisanski

Nizovi čiji elementi pripadaju određenim vrstama, javljaju se u raznim područjima prirodnih i humanističkih znanosti, i mogu se analizirati sličnim matematičkim sredstvima. U radu je prikazan novi pristup, primjenljiv za analizu DNA i proteinskih nizova kao i za analizu strukture teksta.