MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **58** (2007) 301-307

ISSN 0340 - 6253

On Description of Biological Sequences by Spectral Properties of Line Distance Matrices

Gašper Jaklič[*], Tomaž Pisanski[†], Milan Randić[‡]

(Received June 26, 2006)

Abstract

In the paper spectral properties of line distance matrices, associated with biological sequences, are studied. It is shown that a line distance matrix of size $n > 1$ has one positive and $n-1$ negative eigenvalues. Furthermore, a recently introduced conjecture that line distance matrices belong to a class of well known squared distance matrices, is confirmed. The interlacing property for line distance matrices is considered.

1 Introduction

One of the main areas of Bioinformatics is the study of biological sequences. It is well known that finding an optimal structural alignment between two protein sequences is an NP-hard problem [6], therefore only heuristics for comparison of sequences, based on computer techniques are known. Recently, alternative routes for quantitative measure of the degree of similarity of DNA sequences were considered [10, 11, 13], which have

[*]Institute of Mathematics, Physics and Mechanics, University of Ljubljana, Slovenia, gasper.jaklic@fmf.uni-lj.si *(Corresponding author)*

[†]Institute of Mathematics, Physics and Mechanics, University of Ljubljana, Slovenia and University of Primorska, Slovenia, tomaz.pisanski@fmf.uni-lj.si

[‡]Drake University, Emeritus & National Institute of Chemistry, Slovenia, Permanent address: 3225 Kingman Road, Ames, IA 50014, USA, mrandic@msn.com

also been extended to protein sequences [12, 14]. The novel methodology starts with a graphical representation of DNA, such as proposed by Nandy [7] and Jeffrey [4], which are subsequently numerically characterized by associating with the selected geometrical object that represents DNA, a matrix [8].

Another approach is to associate a matrix to a given sequence and study its properties instead. Such a representation, based on the sequential labels of each of the four nucleotides A, T, G, and C separately for construction of matrix elements is given in [9].

In this paper we study line distance matrices. A line distance matrix represents distances between points on the real line, therefore it gives a natural way of studying biological sequences. For example, one can associate four vectors, indicating the distances between the consecutive nucleotides of the same kind, with a given DNA sequence. The corresponding matrices for each of the four nucleotides A, T, G, and C give a representation of the DNA sequence. The study of spectral properties of the matrices and their principal submatrices gives us an insight into the DNA sequence.

The novelty of this particular approach is association of a matrix with partitioned lines, which allows construction of a set of invariants to characterize such lines. In this paper we study the eigenvalues of line distance matrices $D \in \mathbb{R}^{n\times n}$ and prove that their spectrum consists of only one positive and $n-1$ negative eigenvalues. Further, we prove that line distance matrices belong to a class of well known squared distance matrices [3, 15].

2 Spectral properties

Let $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$, $t_1 < t_2 < \cdots < t_n$, $t_i \in \mathbb{R}$, be a given position vector (i.e., a list of points on the real line). **A line distance matrix** $D \in \mathbb{R}^{n\times n}$, associated with $\boldsymbol{t}$ is defined as

$$d_{ij} = |t_i - t_j|.$$

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) in a vector $\boldsymbol{t}$. We associate a line distance matrix with the vector $\boldsymbol{t}$. A similar construction gives us matrices associated with nucleotides T, G and C. Those four distance matrices represent (part of) the given DNA sequence. A valuable insight into matrix properties can be obtained by the study of the distribution of its eigenvalues. Clearly, since line distance matrices are symmetric, their eigenvalues λ_i are real. Here is the main result of the paper.

Theorem 1 *Let $D \in \mathbb{R}^{n\times n}$ be a line distance matrix, associated with a vector $\boldsymbol{t}$ and let $D^{(i)} := D(1:i, 1:i)$, $i = 1, 2, \ldots, n$ be its principal submatrices. Let*

$$\lambda_i^{(i)} \le \lambda_{i-1}^{(i)} \le \cdots \le \lambda_2^{(i)} \le \lambda_1^{(i)}$$

be the eigenvalues of the matrix $D^{(i)}$. Then $\lambda_1^{(i)} > 0$, $\lambda_2^{(i)} < 0$ for $i > 1$ and $\lambda_1^{(1)} = 0$.

Since $D = D^{(n)}$, we obtain the following corollary.

Corollary 1 *The spectrum of a line distance matrix D of size $n > 1$ consists of one positive and $n-1$ negative eigenvalues.*

The definition of line distance matrices contains the assumption $t_j \neq t_{j+1}, \forall j$ for a position vector $\boldsymbol{t}$. Obviously, the relation $t_j = t_{j+1}$ implies linearly dependent rows and columns of the line distance matrix D. If this assumption is neglected, we obtain the following result.

Corollary 2 *Let $D \in \mathbb{R}^{n\times n}$ be a line distance matrix, associated with a position vector $\boldsymbol{t}$ and $n > 2$. Then $\lambda_2^{(i)} = 0,\ i > 1$ iff $t_j = t_{j+1}$ for some $j < i$.*

Proof of Theorem 1: Let $p_i(x) := \det(D^{(i)} - xI)$ denote the characteristic polynomial of the matrix $D^{(i)}$. Clearly, $\lambda_1^{(1)} = 0$. By using ideas of Krattenthaler [5] it is easy to prove

$$\det D^{(i)} = (-1)^{i+1}2^{i-2}(t_i - t_1)\prod_{j=1}^{i-1}(t_{j+1} - t_j). \tag{1}$$

Since $\operatorname{trace} D^{(i)} = \sum_{j=1}^{i}\lambda_j^{(i)} = 0$ and $\det D^{(i)} = \prod_{j=1}^{i}\lambda_j^{(i)} \neq 0$,

$$\lambda_1^{(i)} > 0, \quad \lambda_i^{(i)} < 0, \quad i > 1,$$

in particular $\lambda_2^{(2)} < 0$. Cauchy's interlacing theorem [16] implies

$$\lambda_3^{(3)} \leq \lambda_2^{(2)} \leq \lambda_2^{(3)} \leq \lambda_1^{(2)} \leq \lambda_1^{(3)}.$$

Since $p_3(0) > 0$ by (1), $\lambda_2^{(2)} < 0$, $\lambda_1^{(2)} > 0$ and $p_3(x) = (\lambda_1^{(3)} - x)(\lambda_2^{(3)} - x)(\lambda_3^{(3)} - x)$, therefore $\lambda_2^{(3)} < 0$.
Now let by inductive supposition $\lambda_2^{(i-1)} < 0$. Cauchy's interlacing theorem implies

$$\lambda_3^{(i-1)} \leq \lambda_3^{(i)} \leq \lambda_2^{(i-1)} \leq \lambda_2^{(i)} \leq \lambda_1^{(i-1)} \leq \lambda_1^{(i)}.$$

Recall $\lambda_1^{(i)} > 0$ and $\lambda_j^{(i)} < 0,\ j \geq 3$. Since $p_i(x) = \prod_{j=1}^{i}(\lambda_j^{(i)} - x)$ and by (1) $\operatorname{sign}(p_i(0)) = (-1)^{i+1}$,

$$\lambda_2^{(i)} < 0.$$

This concludes the proof.

■

It is interesting to consider a relation between line distance matrices and well-known squared distance matrices. Recall that a matrix $S \in \mathbb{R}^{n\times n}$ is **a squared distance matrix**, if there are vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^r$ $(r \leq n)$, such that $s_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ for all $i, j = 1, 2, \ldots, n$ ([3, 15]). By using Theorem 1 and a characterization of squared distance matrices ([3]), we can prove the following claim.

Theorem 2 *Line distance matrices are squared distance matrices.*

Proof: Let D be a line distance matrix, associated with a position vector $\boldsymbol{t}$ and let us define $\boldsymbol{e} := [1, 1, \ldots, 1]^T \in \mathbb{R}^n$. Since by Corollary 1 D has exactly one positive eigenvalue, it is by characterization of squared distance matrices [3], Thm. 2.2, enough to prove, that there exists $\boldsymbol{w} \in \mathbb{R}^n$, such that $D\boldsymbol{w} = \boldsymbol{e}$ and $\boldsymbol{w}^T\boldsymbol{e} \geq 0$. Let us define $v := \frac{1}{t_n - t_1}$ and let $\boldsymbol{w} := [v, 0, \ldots, 0, v]^T$. Clearly $D\boldsymbol{w} = \boldsymbol{e}$ and $\boldsymbol{w}^T\boldsymbol{e} = 2v > 0$. This confirms that D is a squared distance matrix.

■

Recently, Alfakih in [1] computed the characteristic polynomial of a squared distance matrix and studied properties of its nullspace and rangespace.

3 Interlacing property

Cauchy's interlacing theorem [2, 16], applied in the proof of Theorem 1, together with the results of Theorem 1, give us insight into the position of the eigenvalues of principal submatrices of line distance matrices. Let us consider a visualization of this interesting property.

For example, take the first exon of rabbit's β-globin gene, which starts as: ATGGTGCATCTGTCCAGTGAG... Positions of T in the sequence can be written in a vector

$$\boldsymbol{t} = [2, 5, 9, 11, 13, 18] \tag{2}$$

and its corresponding line distance matrix is

$$D = \begin{bmatrix} 0 & 3 & 7 & 9 & 11 & 16 \\ 3 & 0 & 4 & 6 & 8 & 13 \\ 7 & 4 & 0 & 2 & 4 & 9 \\ 9 & 6 & 2 & 0 & 2 & 7 \\ 11 & 8 & 4 & 2 & 0 & 5 \\ 16 & 13 & 9 & 7 & 5 & 0 \end{bmatrix}. \tag{3}$$

The positions of the nucleotide T in the DNA sequence are shown in Fig. 1. Note that the entries on the superdiagonal $3, 4, 2, 2, 5$ are the differences between adjacent components of the vector (2) (Fig. 1).

The eigenvalues of the principal submatrices of the matrix (3) can be written in a pyramidal structure:

$$\begin{array}{ccccccccccc} & & & & & 0 & & & & & \\ & & & & -3 & & -2.48 & & 3 & & \\ & & & -7.09 & & -2.91 & & -1.66 & & 9.57 & \\ & & -11.28 & & -4.13 & & -2.22 & & -1.29 & & 15.86 \\ & -15.52 & & -9.35 & & -3.03 & & -2.16 & & -1.25 & & 23.14 \\ -21.54 & & & & & & & & & & & & 37.33 \end{array} \tag{4}$$

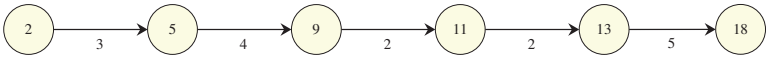

Figure 1: A path defines the line distance matrix (3).

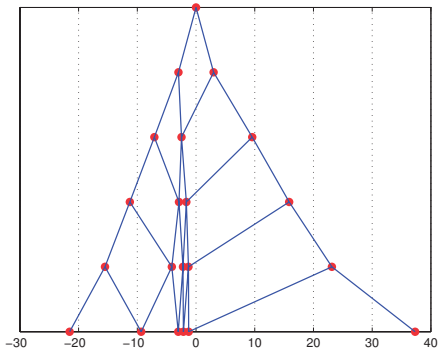

Figure 2: Cauchy's interlacing property for the line distance matrix (3).

Their graphical representation is shown in Fig. 2. Another characterization may be obtained by studying sequences of partial sums of eigenvalues $\left(\sum_{i=1}^{j} \lambda_i^{(k)}\right)_j$:

$$\begin{array}{ccccccccccccccc} & & & & & & 0 & & & & & & & & \\ & & & & & 3 & & 0 & & & & & & & \\ & & & & 9.57 & & 7.09 & & 0 & & & & & & \\ & & & 15.86 & & 14.20 & & 11.28 & & 0 & & & & & \\ & & 23.14 & & 21.86 & & 19.64 & & 15.52 & & 0 & & & & \\ 37.32 & & & 36.08 & & 33.92 & & 30.88 & & 21.54 & & 0 & & & \end{array}$$

4 Remarks

We have considered line distance matrices, introduced in Bioinformatics in [9] and in particular their spectral properties. Note that line distance matrices can be associated to any (noninteger) position vector with non-decreasing components (or associated to a path in graph theory). Line distance matrices have a structure similar to that of the well-known Toeplitz and Hankel matrices, since they are determined by their first row only. But here the rows are not simply shifts of the first row, but they represent distances between different elements in the position vector. At first it seems that no new information is obtained in this way, but the study of principal submatrices gives us novel information

on the subsequences of the given biological sequence.

The leading eigenvalues of the principal submatrices of line distance matrices (shown as the most right hand side entries of the pyramidal structure (4)), can serve as additional line descriptors. From the leading eigenvalues we can construct a new position vector $\boldsymbol{t}' = (\lambda_1^{(1)}, \lambda_1^{(2)}, \lambda_1^{(3)}, \dots)$ and construct a line distance matrix of the second order, and continue the process.

Similarly, we can construct the Laplace matrix, associated with the given position vector or generalize the outlined approach from DNA sequences to other biological and non-biological sequences. A natural generalization is a study of protein sequences. Here we can consider line distance matrices based on the distances between the 20 natural amino acids in some prescribed order.

5 Software

Mathematica and Matlab programs for studying line distance matrices and their spectral properties for various biological sequences are available at the website:
`http://www.fmf.uni-lj.si/~jaklicg/ldmatrix.html`

Acknowledgement

Research was supported in part by grants P1-0294 and Z1-7330-0101 from Ministrstvo za visoko šolstvo, znanost in tehnologijo Republike Slovenije.

References

[1] Alfakih, A.Y. 2006. On the nullspace, the rangespace and the characteristic polynomial of Euclidean distance matrices. *Linear Algebra and its Applications* 416, 348–354.

[2] Demmel, J.W. 1997. *Applied numerical linear algebra*. SIAM, Philadelphia.

[3] Hayden, T.L., Reams, R., Wells, J. 1999. Methods for constructing distance matrices and the inverse eigenvalue problem. *Linear Algebra Appl.* 295, 97–112.

[4] Jeffrey, H.I. 1990. Chaos game representation of gene structure. *Nucleic Acid Res.* 18, 2163–2170.

[5] Krattenthaler, C. 1999. Advanced determinant calculus. *Séminaire Lotharingien Combin.* 42, (The Andrews Festschrift).

[6] Lathrop, R.H. 1994. The protein threading problem with sequence amino acid interaction preferences is NP-complete. *Protein Eng.* 7, 1059–1068.

[7] Nandy, A. 1994. A new graphical representation and analysis of DNA sequence structure, I. Methodology and application to globin gene. *Curr. Sci.* 66, 309–313.

[8] Randić, M., Vračko, M., Nandy, A., Basak, S.C. 2000. On 3-D Representation of DNA Primary Sequences. *J. Chem. Inf. Comput. Sci.* 40, 1235–1244.

[9] Randić, M., Basak, S.C. 2001. Characterization of DNA primary sequences based on the average distances between bases. *J. Chem. Inf. Comput. Sci.* 41, 561–568.

[10] Randić, M., Lerš, N., Plavšić, D. 2003. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.* 371, 202–207.

[11] Randić, M., Vračko, M., Zupan, J., Novič, M. 2003. Compact 2-D graphical representation of DNA. *Chem. Phys. Lett.* 373, 558–562.

[12] Randić, M. 2004. 2-D Graphical representation of proteins based on virtual genetic code. *SAR & QSAR in Environ. Res.* 15, 147–157.

[13] Randić, M. 2004. Graphical representation of DNA as a 2-D map. *Chem. Phys. Lett.* 386, 468–471.

[14] Randić, M., Zupan, J. 2004. Highly compact 2-D graphical representation of DNA sequences. *SAR & QSAR in Environ. Res.* 15, 191–205.

[15] Schoenberg, I.J. 1935. Remarks to Maurice Frechet's article Sur la definition axiomatique d'une classe d'espace distancies vectoriellement applicable sur l'espace de Hilbert, *Ann. Math.* 36 (3), 724–732.

[16] Trefethen, L.N., Bau, D. 1997. *Numerical linear algebra*. SIAM, Philadelphia.