APPLICATION OF VECTOR SPACE TECHNIQUES TO DNA

T. Bíró,[a] A. Czirók,[a] Á. Major,[b] and T. Vicsek[a]

[a] Department of Biological Physics, Eötvös University, Budapest, Puskin u 5-7, 1088 Hungary

[b] Department of Genetics, Eötvös University, Budapest, Múzeum krt 4/a, 1088 Hungary

Abstract

We propose the application of a recent vector space technique for the analysis of DNA sequences. In this approach the similarity of two series of symbols is measured by the dot products of vectors corresponding to the given sequences. In this paper we demonstrate that this measure, depending on the short-range properties of the sequences, may be used for identifying different elements of the genome.

1. Introduction

The information necessary for the correct protein synthesis of organisms is stored in the form of a sequence which can be considered as a long series of four "letters" `A`, `C`, `G` and `T` corresponding to the four bases in DNA molecules. This sequence is very long and extremely complex. In recent years sequence data have been collected through the joint effort of many research laboratories and the amount of related information is expected to increase by several orders of magnitude in the near future. Various methods have been developed to uncover the underlying structure of the sequences, but many of the fundamental questions concerning the ways the genetic information is coded are still open.

Several approaches have been used as models for DNA sequences. Markov processes were proposed in the 1980s[1]. However, since the early 1990s it has been demonstrated by methods taken from statistical physics[2] or statistical linguistics[3,4] that Markov models cannot give an adequate description of DNA, since they can-

not explain long-range correlations and Zipf-law-like behaviour of DNA-sequences. Mantegna et al.[4] showed that "higher order Markovian processes can mimic with increasing accuracy the observed statistical properties", but "the noncoding sequences cannot be described by a Markovian stochastic process."

Some nontrivial statistical properties of DNA sequences have been discovered by the investigation of n-tuple Zipf-plots, i. e. by plotting the frequencies of n-tuples in function of their range of occurrence[3-6]. This idea was borrowed from G. K. Zipf's early works on universal distribution of words in written texts[7,8].

One of the promising approaches to the interpretation of the organization of information in DNA has been the investigation of the so called "DNA walk"[2,9-13]. In this recent method DNA is mapped onto a process which can be regarded as a walk: each of the four "letters" of a DNA sequence is identified with a step in a given direction. Then, the specific features of this walk can be analyzed using methods borrowed from statistical physics.

The idea is to look for correlations. Two series of data (X and Y) are correlated if the corresponding elements of the series are not independent. In other words: if the fact that X_i (the $i-th$ element of X) is larger than the average value of X is typically accompanied by the fact that Y_i is also larger than the average of Y, then the two sets of data are positively correlated, and the corresponding correlation coefficient is a number larger than zero. The lack of corelation results in a coefficient equal to zero. Correlations within a single sequence of data can be defined similarly. In this case one can ask how the value X_i is related to the value X_{i+j}. By comparing the two values with the average of X one can get information about the question whether two values in the data set separated by j elements are correlated. An ordinary random walk has no long range correlations, thus, an analysis concentrating on the nature of correlations may detect relevant differences from random sequences.

In an alternative approach, the symbol sequence corresponding to a DNA molecule can be regarded as a written text composed by using four letters. Naturally, we do not know the "language" of the text, thus, when we are trying to get information about the sequence as a whole we are led to apply methods developed for analysing written (natural) texts of unknown origin.

Recently Damashek has applied a vector space technique[14] to a number of

texts written in different languages. Using his approach he has been able to find correlations, for example, between texts of different origin, but written in the same language. The correlations of texts of similar content could also be manifested. Here two sequences are correlated if the scalar product of the two vectors associated with them has a value different from that it would have for two uncorrelated sequences. In the first approximation his method is not sensitive to a class of encodings and can be used to identify the original language of encoded texts.

When applying the vector space technique to DNA sequences we look at DNA as an encoded text written in an unknown laguage, still, we expect to locate correlations between parts of the sequences due to similarities in their underlying structure.

2. The Vector-space Technique

Let us move an n character long "window" along our document, symbol by symbol. An *n-gram* or *n-tuple* means a sequence (i. e. a string) of n consecutive characters, that is what can be "seen in the window" at a particular position of the window. We denote each possible n-tuple with an index i, and we count the number m_i of the occurrences of the $i-th$ n-tuple in our text, for each possible i. The document can be characterized then by a "document" vector $\mathbf{x}$, whose components are the normalized occurrences as

$$x_i = \frac{m_i}{\sum_j m_j} \tag{1}$$

The similarity of two texts with document vectors $\mathbf{x}$ and $\mathbf{y}$ can be measured by the dot product

$$S = \cos\theta = \frac{\sum_i x_i y_i}{(\sum_i x_i^2 \sum_i y_i^2)^{1/2}} \tag{2}$$

The maximum of this measure of similarity is 1, in the case of identical vectors, which occurs in practice only if the two documents are identical. The minimum of the dot product is zero, in the case of orthogonal vectors, i. e. if there

is no n-tuple occurring in both documents. This measure is obviously symmetric, but $1 - S$ is not a distance in its mathematical sense, as it does not satisfy the triangle-inequality.

As an exemple for languages, let us mention the 26 letters of the English alphabeth, space, dot and comma. Sequences of spaces should be substituted by a single space. As discussed by Damashek[14], texts of the same language and topic result in a noticeably higher similarity than documents of different languages. Same language but different topics result in a dot product which is still significantly higher than the case of different languages, but usually smaller than the case of texts with related topics.

The procedure can be improved by introducing centroid vectors, which are the averages of vectors taken from a given set of documents (e. g.: the set of the documents in a given language). The centroid vectors contain the common average features of the document set (e. g., the grammatical words in a language). By subtracting the centroid vector from the document vectors, we can increase the sensitivity of the similarity measure. This method gives an effective technique called *Acquaintance* for sorting by language, topic and sub-topic.

The effectivity of the procedure can be explained by three factors. The first one is the effect of the frequent words. For example, in the case of an English text, the most frequent 3-tuples are "$\sqcup th$" and "the" (where $\sqcup$ denotes the space) due to the many grammatical words beginning with "th", especially with "the" ('the', 'they', 'their', 'them', 'these', 'there', etc.). The effectivity of the method in sorting documents by topic is due to the nouns, adjectives and verbs characteristic to the topic. The second factor is the phonotactical features of the language considered, that are the phonological rules licensing or forbidding sequences of phonemes to appear at the beginning, middle or end of a word. These rules introduce definite short-range correlations into the text, hence influence the frequencies of the n-tuples. The third reason is not linguistical (neither semantical and syntactical as the first one was, phonological as the second one) but orthographical: one should think to the fact that the same sound is represented in English texts as "sh", in German documents as "sch", while in the French tradition we find "ch". Sequences of "th" are characteristic to English documents, "aux" to French ones, "sch" to German ones, while "qu" and "tio" to every topic using many words of Latin origin.

Let us suppose we have a multilingual document, such as the concatenation of English, Hungarian and French texts, as in the case of Fig. 1. We can move a box of length m ($m \geq 100$) along our document, step-by-step. We consider the content of this box as a "text" at each step, and plot the dot product of the vector given by the content of the box with a constant vector. Fig. 1 shows our results when we moved a box of length $m = 100$ along a multilingual document. The constant multiplier vector was made from a different English text, and we used $n = 3$-tuples. The similarity measure drastically drops when the end of the box is penetrating into a non-English part of the concatenated texts, i. e. m characters before it, as the position of the *first* character of the box is used as the position of the box itself. This characteristic "valley" lasts until the beginning (left end) of the box leaves this non-English sequence. So we can find a significant "foreign language valley", which is m characters longer than the length of the non-English text, and it is shifted by $^m/_2$ characters to the left.

This method turned out to be useful in differentiating between different parts of a document. In the following we will apply this method on DNA-sequences.

3. Applying the Vector Space Technique on DNA Sequences

In the case of DNA-sequences, our alphabet consists of four symbols, each representing one of the four bases. One of the most exciting questions concerning DNA sequences is the characteristic behaviour of coding and noncoding regions. Latter ones have turned out to behave like written texts, showing long-range correlations and power law-like Zipf-plot, while previous ones showed no long-range correlations and exponential Zipf-plot (similar to the distribution of single letters of the alphabet in written texts, and not like the distribution of words or n-tuples)[3,4].

We chose five primary transcripts from the human hemoglobin sequence HUMHBB, NCBI-GenBank, Release 92.0 (Dec. 1995). Clay et al.[15] state that – according to their results – coding regions typically have higher GC levels than introns of the same gene in human genome. (The GC level is the molar fraction of guanine + cytosine.) This statement predicts a difference in n-tuple frequencies between coding and non-coding sequences, similar to the effect of phonological rules and orthographical traditions on document vectors.

We produced two vectors from each primary transcripts ($n = 3$): the first one representing the concatenation of the primary transcript's exons (E1 - E5), the second one the concatenation of the head, the introns and the tail (I1 - I5). *Table 1* shows their dot products. The difference between the two groups is obvious. The similarity of two E-vectors is 0.94 ± 0.016, two I-vectors give 0.92 ± 0.03, while an E- and an I-vector produce 0.75 ± 0.08. We obtained the same result using the RATCRYG-sequence.

The difference between the two sets of vectors can be demonstrated in another way as well. We prepared an $\mathbf{x}^{(naive)}$ "naive" (correlationless) vector from the RATCRYG-sequence (*Rattus norvegicus*, NCBI-GenBank, Release 92.0), which means that its ith component representing the n-tuple $\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}$ (where $\sigma_1...\sigma_4$ are the symbols `A`, `C`, `G` and `T`) is given by the product of its constituents' relative frequencies:

$$x_i^{(naive)} := P^{(naive)}(\sigma_{k_1}\sigma_{k_2}...\sigma_{k_n}) := \prod_{j=1}^{n} \nu_{k_j}$$

where ν_j denotes the relative frequency of the base σ_j in the whole RATCRYG-sequence. *Table 2* shows the dot products of this naive vector with the vectors representing the concatenation of the coding regions, as well as the concatenation of the noncoding regions of different primary transcripts of RATCRYG and of HUMHBB ($n = 4$). (The gene concentration pattern of the human genome is basically present in all vertebrates.[15]) In the case of the coding parts the product is 0.66 ± 0.05, while in the case of the noncoding parts it is significantly higher: 0.83 ± 0.03. This result means that the naive vector (0th order Markov-chain) gives a much better approximation of noncoding sequences' vectors than for that of the coding sequences.

This finding does not contradict the previous results proving the existence of *long* range correlations in intron-containing genes and their absence in intronless genes and cDNAs, as our results demonstrate stronger *short* range correlations in exons than in introns.

The results gave us the idea to use the "moving box technique"[10] for DNA-sequences, too. Fig. 2 shows a primary transcript sequence of HUMHBB. The naive vector is calculated from the relative frequencies of the bases in the whole HUMHBB

sequence ($n = 4$). A box of length $m = 100$ is moved along the primary transcript that contains a head, three exons, two introns and a tail. The pattern showed by the figure is characteristic of the HUMHBB-sequence: each exon is represented by a "valley" in the figure, shifted with $\frac{m}{2}$ positions to the left (as explained in the case of natural languages), the first two of the three valleys overlap, and there is a fourth valley, at about the two third of the primary transcript, before the last exon. This valley shows the existence of a sequence inside the second intron with statistical properties similar to those of the exons (or at least: dissimilar from the statistical properties of other intron sequences).

4. Conclusions

The idea of using Damashek's vector space technique in defining a similarity measure of symbol sequences as the dot product of vectors of n-tuples' frequencies turned out to be efficient in analyzing DNA-sequences. We found that exons and non-exonic regions of primary transcripts are represented in our multi-dimensional vector space as two distinct bundles of vectors, showing significantly higher similarity values among each other than between the bundles. The correlationless "naive" vector of the whole gene turned out to be more similar to the non-exonic regions than to the exons. The reason and significance of these observations remains to be clarified. But these results suggest that the relatively easy vector space technique is able to identify parts of DNA-sequences which are likely to be exons.

Acknowledgments This work was supported by OTKA F019299 and FKFP 0203/1997.

References

[1] Bruce S. Weir: Genetic Data Analysis, Methods for Discrete Population Genetic Data, 1990, Sinauer Associates, Inc. Publishers, Sunder-

land, Massachusetts, pp. 237-240.

[2] C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H. E. Stanley: Long-range correlations in nucleotide sequences, Nature, **356** (1992), pp. 168-170.

[3] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley: Linguistic Features of Noncoding Sequences, Physical Review Letters, **73**, 23 (1994), pp. 3169-3172.

[4] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H. E. Stanley: Systematic analysis of coding and noncoding DNA sequences using methods of statistical linguistics, Phys. Rev. E, **52**, 3 (1995), pp. 2939-2950

[5] A. Czirók, R. N. Mantegna, S. Havlin, H. E. Stanley: Correlations in binary sequences and a generalized Zipf analysis, Physical Review E, **52**, 1 (1995), pp. 446-452.

[6] A. Czirók, H. E. Stanley, T. Vicsek: Possible origin of power-law behavior in n-tuple Zipf analysis, Physical Review E **53**, 6371 (1996)

[7] G. K. Zipf: The Psychobiology of Language, Houghton Mifflin, Boston, 1935.

[8] G. K. Zipf: Human Behavior and the Principle of Least Effort, Addison-Wesley Press, 1949.

[9] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, J. M. Hausdorff, S. Havlin, J. Mietus, C.-K. Peng, F. Sciortino, M. Simons:Fractal landscapes in biological systems: Long-range correlations in DNA and interbeat heart intervals, Physica A, 191 (1992), 1-12.

[10] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, S. Havlin, S. M. Ossadnik, C.-K. Peng, M. Simons: Fractal Landscapes in Biological Systems, Fractals, **1**, 3 (1993), pp. 283-301.

[11] G. Dietler, Y.-C. Zhang: Crossover from White Noise to Long Range Correlated Noise in DNA Sequences and Writings, Fractals, **2** (1994), 4, pp. 473-479.

[12] W. Ebeling, A. Neiman: Long-range correlations between letters and sentences in texts, Physica A 215 (1995), pp. 233-241.

[13] H. Herzel, I. Große: Correlations in DNA sequences: The role of protein coding segments, Physical Review E, **55**, 1 (1997), pp. 800-810.

[14] M. Damashek: Gauging Similarity with n-tuples: Language-Independent Categorization of Text, Science, **267** (1995), pp. 843-848.

[15] O. Clay, S. Cacciò, S. Zoubak, D. Mouchiroud and G. Bernardi: Human Coding and Noncoding DNA: Compositional Correlations, Mol. Phylogen. Evol., Vol. 5., No. 1. (1996), pp. 2-12.

Figure captions

Figure 1. The similarity between the content of a box of length $m = 100$ moving along the concatenation of texts in different languages, and an arbitrary English text. The concatenation consists of an about 8000 character long English text (E1), an 2640 character long Hungarian text (H), another English text (E2) that is 6450 character long, and a 2860 character long French text (Fr). The multiplying vector was constructed from a different English text. A deep "valley" can be observed starting at the point where the end of the box is penetrating into the non-English texts, i.e. m characters before the non-English texts. The x axis represents the place of the beginning of the box, and the valley lasts until the box contains any part of this text. These valleys differ significantly enough to split the symbol sequence to sub-sequences according to similarity, meaning the same language in our case. ($n = 3$ for all vectors.)

Figure 2. The region of a primary transcript in the HUMHBB sequence (positions 34496...36087), that contains a head, three exons (E), two introns (I), and a tail. A box of lenght $m = 100$ was moved along the sequence, and at each step we plotted the dot product of the vector prepared from the actual content of the box, to the naive (correlationless) vector calculated from the base frequencies in the entire HUMHBB sequence. ($n = 4$ for both vectors.) Characteristic valleys can be seen where the box contains exon sequences, and another valley appears at about the two third of the primary transcript.

Table captions

Table 1. Similarity among different regions of primary transcripts. E1-E5 denotes the concatenations of the exons of the five CDSs (coding regions) of HUMHBB, while I1-I5 denotes the concatenation of the corresponding non-coding sequences, i. e. heads, introns and tails. Our similarity measure, the dot product of the vectors representing the "texts" to be compared (n = 3-tuples were used), is symmetrical, and its range is between 0 and 1, where 0 means the lack of similarity, while 1 denotes maximal similarity. The similarity measure is significantly higher in the case of two coding or two non-coding sequence (0.94 ± 0.016 for exons, 0.92 ± 0.03 for non-coding "texts") than in the case of a coding and a non-coding sequence (0.75 ± 0.08).

Table 2. Similarity of coding and non-coding regions of mammalian CDSs to the naive (correlationless) vector made from the entire RATCRYG-sequence. We prepared the concatenation of the coding and the one of the non-coding parts for each of the five CDSs of the HUMHBB and for the the five CDSs of the RATCRYG sequence. We also prepared a correlationless vector using the base frequencies in the whole RATCRYG sequence. The similarity measures, i. e. the dot products shows that the non-coding sequences resemble more the naive vector (0.83 ± 0.03) than the coding sequences of either species (0.69 ± 0.013 for Rattus norvegicus, 0.62 ± 0.04 for Homo sapiens). ($n = 4$)

Figure 1

Figure 2

	E1	E2	E3	E4	E5	I1	I2	I3	I4	I5
E1	1.00	0.92	0.92	0.95	0.93	0.83	0.77	0.74	0.83	0.90
E2		1.00	0.97	0.93	0.95	0.73	0.66	0.62	0.73	0.85
E3			1.00	0.94	0.94	0.73	0.65	0.61	0.71	0.83
E4				1.00	0.95	0.78	0.71	0.67	0.79	0.87
E5					1.00	0.76	0.69	0.64	0.77	0.86
I1						1.00	0.92	0.90	0.94	0.93
I2							1.00	0.98	0.91	0.88
I3								1.00	0.90	0.86
I4									1.00	0.94
I5										1.00

Table 1

	coding	non-coding
RATCRYG 1	0.68	0.79
RATCRYG 2	0.71	0.87
RATCRYG 3	0.70	0.84
RATCRYG 4	0.70	0.86
RATCRYG 5	0.68	0.80
HUMHBB 1	0.69	0.84
HUMHBB 2	0.59	0.83
HUMHBB 3	0.59	0.80
HUMHBB 4	0.63	0.85
HUMHBB 5	0.62	0.81

Table 2