

Journal of Computational Biology, 1996, v3, n4:573-576.

Principal Component Analysis and Large-Scale Correlations in Non-Coding Sequences of Human DNA.

Michael Teitelman, Frank H. Eeckman

Human Genome Center Informatics Group, Lawrence Berkeley National Laboratory,
MS 46A-1123, 1 Cyclotron Road, Berkeley, CA 94720

Email: teitel@genome.lbl.gov

Telephone: (510)486-6569

Fax: (510)486-4711

Keywords: long range correlations, DNA, principal component analysis.

Abstract

We have calculated a full set of second-order correlation functions of nucleotides in non-coding DNA. They are found to be independently invariant in regard to permutations of A and T, and also C and G. Considering correlation functions as a 4×4 matrix with a symmetrical basis we have found the principal components - objects with zero cross-correlations. These three principal components are present the base compositions: (A+T-C-G), (A-T), (C-G). The long range behavior of these principal components yield power-law dependencies with different critical exponents.

Recent studies of large-scale statistical properties of DNA sequences reveal the presence of non-trivial long-range correlations (Peng *et al*, 1992; Li and Kaneko, 1992; Voss 1992). These correlations are found in non-coding DNA and can be described by power-law dependencies.

This finding may be of universal importance, but it does not answer the question of what the specific structure is of the stochastic process that represents nucleotide sequences.

We have developed an approach based on the study of families of correlation functions to further study DNA statistics in great detail. We are looking for relationships between the different correlation functions and are attempting to find principal components.

In this correspondence we show some results and demonstrate the method using a full set of second-order correlation functions for nucleotides calculated from a representative set of introns in human DNA sequences, each having a length of more than 10,000 bases. We extracted this set from Genbank, release 89, 1995, using a procedure described in Kulp *et al*, 1996.

The second-order correlation functions we used describe the actual frequency of finding a pair of nucleotides in one strand of DNA separated by a given distance between the individual bases. We subtracted all uncorrelated contributions obtained by combining lower order functions to yield the true correlations of a nucleotide pair. Let us define a 4-component vector $\vec{v}(x)$ for each nucleotide site x , which is equal to $\vec{a} = (1, 0, 0, 0)$ for A , $\vec{c} = (0, 1, 0, 0)$ for C , $\vec{g} = (0, 0, 1, 0)$ for G and $\vec{t} = (0, 0, 0, 1)$ for T . Then the correlation function for A and C at given distance x is defined as follows

$$\langle A, C \rangle = \sum_{y=1}^{N-x} \frac{1}{N-x} (\vec{a} \cdot \vec{v}(y)) (\vec{c} \cdot \vec{v}(x+y)) - \left(\sum_{y=1}^{N-x} \frac{1}{N-x} \vec{a} \cdot \vec{v}(y) \right) \left(\sum_{z=1}^{N-x} \frac{1}{N-x} \vec{c} \cdot \vec{v}(z+x) \right),$$

where N is the length of nucleotide sequence. All other correlation functions are defined analogously.

We averaged over all introns and found 16 correlation functions $\langle a_1, a_2 \rangle$, where a_1 and a_2 may be A , C , G , or T , for a range of internucleotide distances of 1 to 1,000.

Basically, these functions display noisy behavior and to study them on long scales we have used smoothing on a scale of 30 bases. The smoothed correlation functions show very regular behavior with distance. The most obvious feature is the presence of two symmetries.

For any given distance, the 16 correlation functions may be considered as a 4×4 matrix with components $\langle a_i, a_j \rangle$. Then the matrix components obtained by interchanging A and T show a very similar dependence on distance. For example, the components

$\langle A, A \rangle$, $\langle T, T \rangle$ are always very close to each other, the same is true for the pair ($\langle A, C \rangle$, $\langle T, C \rangle$) and so on. See Figure 1. In other words, the correlation matrix is invariant under interchanges of A and T : $A \rightarrow T$, $T \rightarrow A$. A second symmetry is present when interchanging C 's and G 's: $C \rightarrow G$, $G \rightarrow C$.

Fickett *et al*, 1992 found that there is a significant tendency for any region of either genome to have a strand-symmetric base composition. However we can see even more symmetry is the sequence.

These two symmetries of the correlation matrix give us a clue about how to choose the best representation for the matrix. Let us consider a 4-dimensional vector space, where the 4 vector components correspond to the nucleotide letters A, C, G, T . Then we use an orthogonal set of vectors with invariant direction under transformations of correlation matrix symmetries. This set is as follows (normalization factors are omitted) $\vec{v}_1 = (1, 0, 0, -1)$, $\vec{v}_2 = (0, 1, -1, 0)$, $\vec{v}_3 = (1, -1, -1, 1)$, $\vec{v}_4 = (1, 1, 1, 1)$. For permutation of A and T the transformation of the basis is $\vec{v}_1 \rightarrow -\vec{v}_1$, $\vec{v}_2 \rightarrow \vec{v}_2$, $\vec{v}_3 \rightarrow \vec{v}_3$, $\vec{v}_4 \rightarrow \vec{v}_4$. When components C and G are interchanged then $\vec{v}_1 \rightarrow \vec{v}_1$, $\vec{v}_2 \rightarrow -\vec{v}_2$, $\vec{v}_3 \rightarrow \vec{v}_3$, $\vec{v}_4 \rightarrow \vec{v}_4$.

Using this vector basis we have found the principal components of the correlations. With this basis the correlation matrix has the following representation $p_{\alpha, \beta} = \sum_{i, j} v_{\alpha}^i \langle a_i, a_j \rangle v_{\beta}^j$, where $\alpha, \beta = 1, 2, 3$. Figure 2 shows how components $p_{\alpha, \beta}$ behave with an internucleotide distance.

The symmetrical basis is a basis of eigenvectors. Within this basis the non-diagonal components of the correlation matrix are not exactly zero but fluctuate around zero. These components are much smaller than the diagonal components. Remarkably, the eigenvector basis is the same over all distances.

The correlation matrix has three nontrivial eigenvalues for eigenvectors \vec{v}_1 , \vec{v}_2 and \vec{v}_3 . For the vector \vec{v}_4 the eigenvalue is exactly zero. This is simply a result of a condition that the probability of finding any of four nucleotides at any site is 1 and it follows from our definition of the correlation functions.

These eigenvectors correspond to three base compositions: (A+T-C-G), (A-T), (C-G) which are present three independent stochastic processes. That is in consistency with symmetry properties which require zero cross-correlations between compositions with different sets of eigenvalues: for a pair of transformations ($A \leftrightarrow T$, $C \leftrightarrow G$) the eigenvalues are (1,1) for (A+T-C-G), (-1,1) for (A-T) and (1,-1) for (C-G).

When we consider the distance dependence of the three nonzero principal components of the correlation matrix on a log scale we find that they behave according to a power-law dependency for scales from a few tens up to 1,000 bps, see Figure 3. The critical

exponents are different from each other. In fact two of them are equal and the third is different, and the average of all three is found to be in agreement with the observations of other authors (Buldyrev *et al*, 1995).

The observed symmetries for interchanging letters A, T and C, G mean that there is no difference in long-range statistics between two complementary strands of non-coding DNA and therefore no preferred reading direction.

We can conclude also that there is an additional symmetry in the correlations because two eigenvalues are found to be very close to one another and may be considered equal. Therefore, we can allow for a continuous rotational transformation of the correlation matrix in subspace of the A and T components.

The observed symmetries impose very important restrictions on large scale stochastic models of DNA sequence. Not only should the process be scaling invariant, but it should also obey the global symmetries described in this study.

Acknowledgements

We wish to thank Martin Reese and Jean Thierry-Mieg for useful discussion and help with selecting the data set.

This work was supported by the U.S. Department of Energy under Contract Number DE-AC03-76SF00098.

References

- Buldyrev,S.V., Goldberger,A.L., Havlin,S., Mantegna,R.N., Masta,M.E., Peng,C.-K., Simons,M., Stanley,H.E., 1995. Long-range correlation properties of coding and non-coding DNA sequences: GenBank analysis. *Physical Review E*, 51(5), 5084–5091.
- Fickett,J.W., Torney,D.C., Wolf,D.R., 1992. Base compositional structure of genomes. *Genomics*. 13(4), 1056–64.
- Kulp,D., Haussler,D., Reese,M.G., Eeckman,F.H., 1996. A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA. *Proceedings, 4th International Conference on Intelligent Systems for Molecular Biology*. St. Louis., in press.
- Li,W. and Kaneko, K., 1992. Long-range correlation and partial $1/f^\alpha$ spectrum in a noncoding DNA sequence. *Europhysics Letters*. 17(7), 655–660.
- Peng,C-K., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Sciortino,F., Simon,M. and Stanley,H.E., 1992. Long-range correlations in nucleotide sequences. *Nature*. 356, 168–170.
- Voss,R., 1992. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Physical Review Letters*. 68(25), 3805–3808.

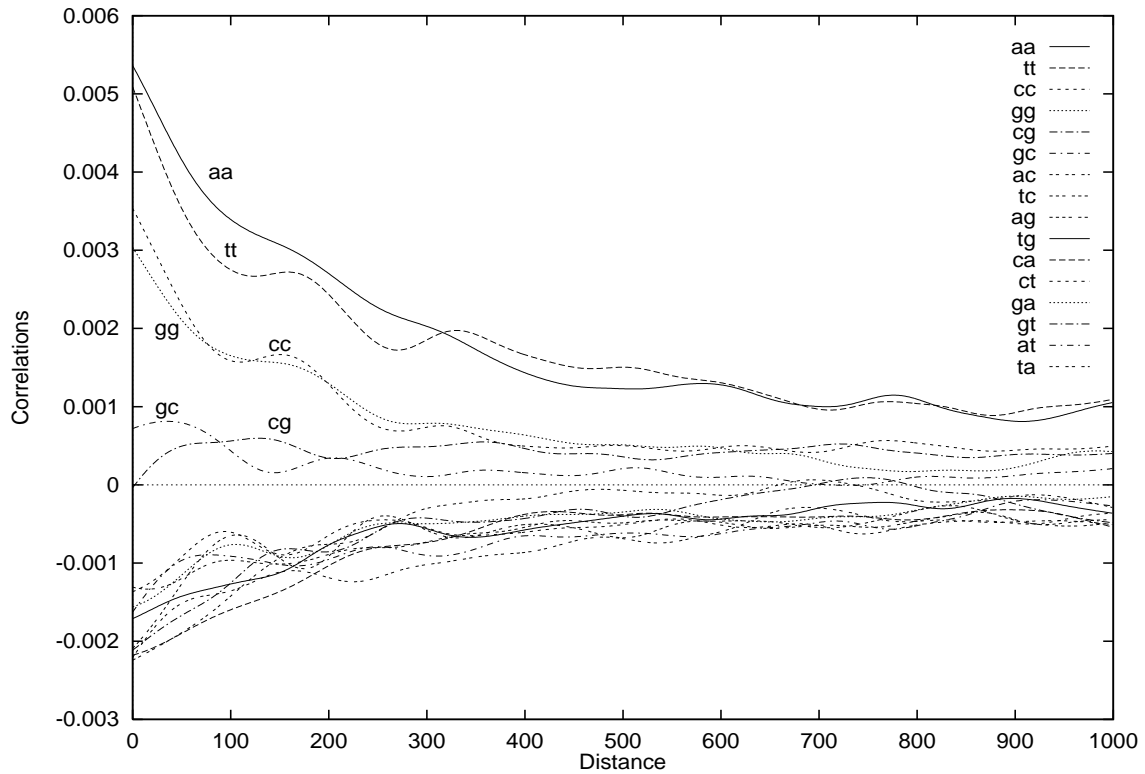


Figure 1. 16 correlation functions $\langle A(0)A(t) \rangle$, $\langle A(0)C(t) \rangle$, ... vs distance t , smoothed over 30 bp.

Data: 10 introns of Human DNA, each is bigger than 10,000bp, total length $\sim 200,000$ bp.

There are obviously **two independent symmetries**:

first, by permutation $A \leftrightarrow T$; and **second**, by permutation $C \leftrightarrow G$.

With these symmetry transformations the correlation behavior remains near the same, while correlation functions transform as follows:

1. $A \leftrightarrow T$: $\langle AA \rangle \rightarrow \langle TT \rangle$, $\langle AC \rangle \rightarrow \langle TC \rangle$, $\langle AG \rangle \rightarrow \langle TG \rangle$, ..

2. $C \leftrightarrow G$: $\langle AA \rangle \rightarrow \langle AA \rangle$, $\langle AC \rangle \rightarrow \langle AG \rangle$, $\langle AG \rangle \rightarrow \langle AC \rangle$, ..

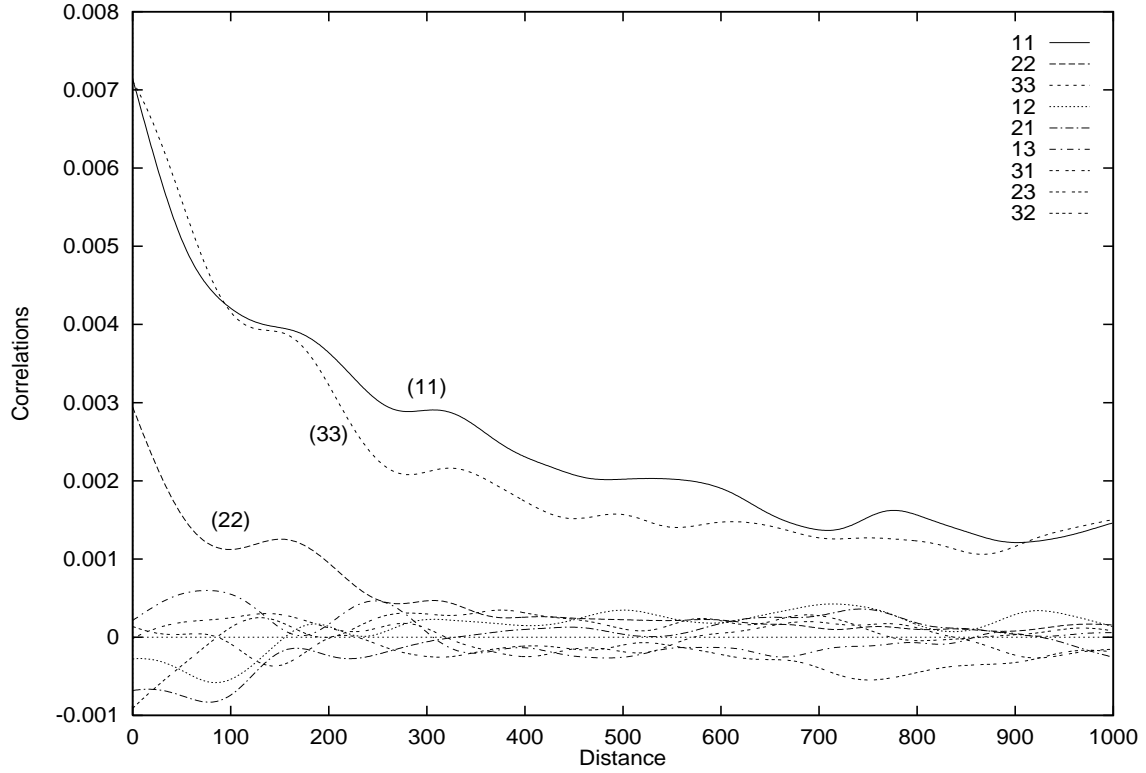


Figure 2. 3x3 correlation matrix vs distance within symmetrical basis :

$$X_1 \sim (A + T - C - G); \quad X_2 \sim (C - G); \quad X_3 \sim (A - T); \quad X_0 = (A + T + C + G) \equiv 1$$

Diagonal elements are always positive: $\langle X_i(0)X_i(t) \rangle \gg 0$.

Off-diagonal elements are close to zero: $\langle X_i(0)X_j(t) \rangle \sim 0, i \neq j; i, j = 1, 2, 3$.

Because $\langle X_1(0)X_1(t) \rangle \sim \langle X_3(0)X_3(t) \rangle$ there is a continuous symmetry by rotation in the plane of X_1 and X_3 components.

Note: those symmetries related to a long-range behavior. There are no such symmetries for short-range correlations.

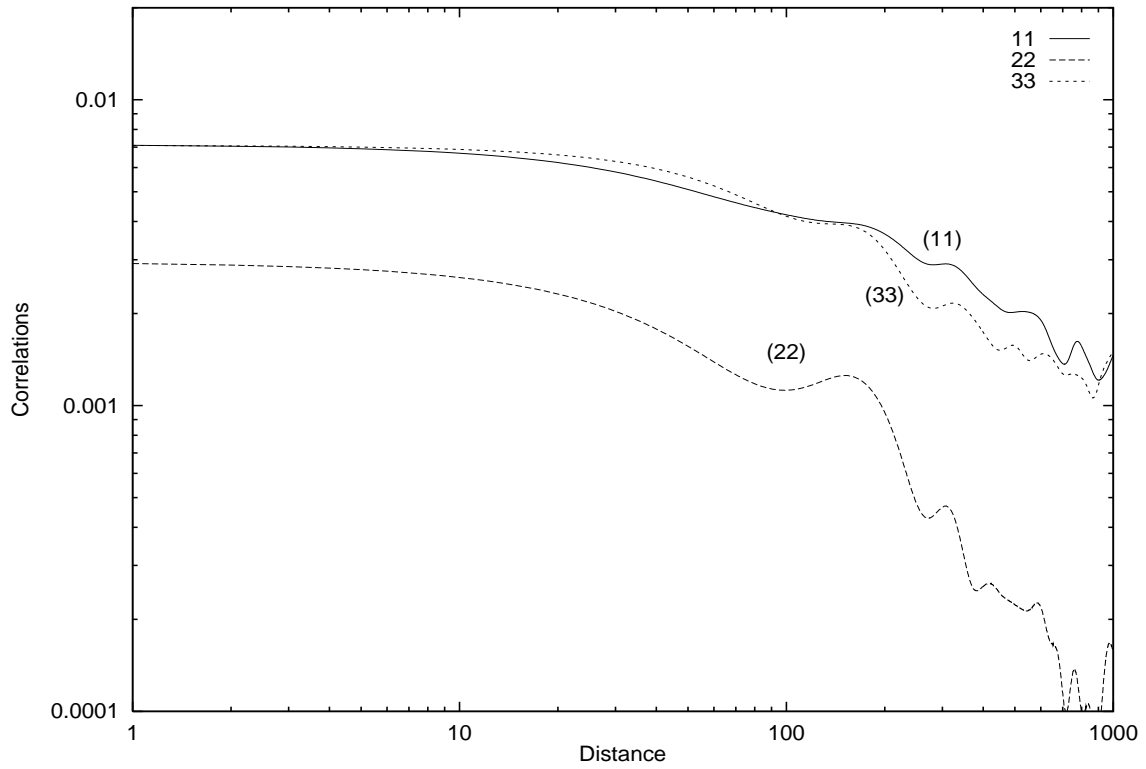


Figure 3. Correlations in log scales.

Power law behavior is present for distances greater than 100 bp, with, generally, different exponents for different components.