CONSISTENCY IN REPRESENTATION AND TRANSFORMATION OF GENOMIC SEQUENCES

Liming Wang and Dan Schonfeld

Department of Electrical and Computer Engineering
University of Illinois at Chicago, Chicago, IL 60607
lwang37@uic.edu, dans@uic.edu

ABSTRACT

Many previous results in genomic sequence analysis have been derived based on the representation of genomic structures as numerical sequences. Various mapping strategies have been proposed for the representation of genomic and proteomic sequences. However, little is understood about the effect of specific choices of numerical mappings on the final analysis results. In fact, inconsistent numerical mappings could have led to contradictory results in genomic sequence analysis. In this paper, we propose a mathematical framework for analysis of the consistency in representation and transformation of numerical mappings of genomic sequences. We introduce strong and weak correlation metrics to characterize consistency measures among distinct numerical mappings. We derive sufficient conditions to ensure consistency among different numerical mappings. We present an important class of equivalent transforms under the proposed consistency conditions. We also derive a class of operators which is shown to be equivalent under rotation of numerical mappings. Finally, we conduct computer simulation experiments on DNA sequences which demonstrate the theoretical results.

1. INTRODUCTION

Processing of genomic sequences as represented by mapping of symbolic data into numerical signals is a commonly used technique. First it is required to map the genomic symbols into the numerical domain. Many kinds of mapping methods have been proposed for different areas. For example, in DNA sequence analysis, there are many mapping methods like mapping the original nucleotide sequence into one-dimensional numerical sequence [1]; indicator sequences method [2]; simplex method [3]; method emphasizing the periodic features for stationary symbolic sequence [4]; Method for non-stationary sequences [5], etc.

Each of the large number of numerical mappings used for the representation of genomic sequences can be justified for various applications. Indeed, it is impossible to determine which mapping is preferable. Furthermore, it is conceivable that distinct mappings could lead to contradictory conclusions. In fact, several contradictory results have arisen in the field of genomic sequence analysis. Most notably, the study of long-range correlations in coding and non-coding DNA sequences has been contested by several contradictory results [2, 6]. Investigation using a large DNA sequence database did not resolve this dispute; in fact, the controversy grew even further [7]. Bouaynaya and Schonfeld [8] shed light on this dilemma by demonstrating that genomic sequences are inherently non-stationary and thus one of the reasons for the contradictory conclusions stems from the use of stationary time-series analysis tools. Moreover, they determined experimentally that the results obtained remained invariant over a large class of numerical mappings used for the representation of DNA sequences. Nonetheless, the experimental study conducted by Bouaynaya and Schonfeld in [8] cannot be used to ascertain with certainty whether the different numerical mappings used for representation of genomic sequences contributed to the contradictory findings reported in the literature [2, 6].

Therefore it is important to investigate the connections between different mapping methods. Since if two different mapping methods are shown to be incompatible, i.e. for the same genomic sequence, it gives inconsistent analysis results. Then there is no reason to compare these two analysis results. Moreover some seemingly different methods may lead to similar analysis results.

In this paper, we provide a systematic method for analyzing the analysis results between different mappings. In Section 2, we first propose a framework for analyzing different mapping methods under any analytic operator using Taylor's expansion. In Section 3, we provide an analysis of the correlation between different mappings of a genomic sequence. In particular, we derive conditions for strong equivalence captured by perfect correlation among the distinct mappings. In Section 4, we explore a more relaxed similarity between different mappings. Specifically, we provide conditions for weak equivalence which is characterized by preservation of the local extrema of the representation. In Section 5, we present experimental results which illustrate the significance of the proposed equivalent mapping theory in genomic signal processing. Finally, we provide a brief summary and discussion of our results in Section 6.

2. SEQUENCE REPRESENTATION AND TRANSFORMATION

Given $\{a_i\}_{i=0}^{N-1}$, where $a_i \in \mathcal{A}$. The set $\mathcal{A}$ could be collection of nucleotides, amino acids, etc. f is a mapping from $\mathcal{A}$ to $\mathbb{R}^n$, i.e. $f : a_i \mapsto x_i,\ x \in \mathbb{R}^n$. After the mapping we have a numerical sequence $\{x_i\}_{i=0}^{N-1}$. $T : x_i \mapsto y_i$ is a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$. $\mathbf{\Phi}_l$ is an analytic operator on the numerical sequence and maps into $\mathbb{R}$ parameterized by $l \in \mathbb{R}$. For example, $\mathbf{\Phi}_l$ could be genomic correlation function or Fourier transform, etc. We also assume that $\mathbf{\Phi}_l \in L^2$. We classify the problems as the following cases.

1. Given T, find out the consistency between $\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1})$ and $\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1})$. Also find the largest class of operator which is consistent under the given T.
2. Given f and $\mathbf{\Phi}_l$, if f and $T \circ f$ are consistent for any symbolic sequence $\{a_i\}_{i=0}^{N-1}$. Find out the largest class of such transformation T which preserves the consistency.

The consistency means we require the results under two different mappings to be similar in some extent. In general $\mathbf{\Phi}_l$ may not be linear. We will use Taylor's expansion to expand the operator. We vectorize the vector sequence $\{x_i\}_{i=0}^{N-1}\ x_i \in \mathbb{R}^n$ to a large vector $x \in \mathbb{R}^{Nn\times 1}$. Consider the Taylor's expansion of the analytic operator. $\mathbf{\Phi}_l : \mathbb{R}^{Nn\times 1} \to \mathbb{R}$. Unlike the common scalar form representation of Taylor's expansion. We shall present it in a concise form involving tensor product. First we define the gradient operator ∇ as

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} & \frac{\partial}{\partial x_2} & \cdots & \frac{\partial}{\partial x_{Nn}} \end{pmatrix}^T \tag{1}$$

Then the Taylor's expansion of $\mathbf{\Phi}_l$ at x_0 can be represented as the following form,

$$\mathbf{\Phi}_l = \sum_{i=0}^{\infty} \frac{1}{i!} (\nabla^i \mathbf{\Phi}_l)(x_0) \times_i (X - x_0) \times_{i-1} \cdots \times_1 (X - x_0) \tag{2}$$

Where $\times_i$ is the i^{th}-mode tensor product [9], and ∇^i is the i^{th}-order gradient of $\mathbf{\Phi}_l$, which is defined as,

$$\nabla^i \mathbf{\Phi}_l = \mathbf{\Phi}_l \times_1 \nabla \times_2 \nabla \times_3 \cdots \times_i \nabla \tag{3}$$

Furthermore, $\nabla^0 \mathbf{\Phi}_l$ is defined as $\mathbf{\Phi}_l$. For one and terms, the i^{th}-order gradient coincides with the traditional definition of *Gradient* $\nabla \mathbf{\Phi}_l(x)$ and *Hessian* $\nabla^2 \mathbf{\Phi}_l(x)$. So we can rewrite the Taylor's expansion as,

$$\begin{aligned} \mathbf{\Phi}_l = {} & \mathbf{\Phi}_l(x_0) + \nabla \mathbf{\Phi}_l(x_0)^T (x - x_0) \\ & + \frac{1}{2} (x - x_0)^T \nabla^2 \mathbf{\Phi}_l(x_0)(x - x_0) \\ & + \sum_{i=3}^{\infty} \frac{1}{i!} (\nabla^i \mathbf{\Phi}_l)(x_0) \times_i (x - x_0) \times_{i-1} \cdots \times_1 (x - x_0) \end{aligned} \tag{4}$$

In order to characterize the consistency, we need to have a metric to measure the consistency. In general, there is no universal metrics. Various operators may have different metrics for different purposes. However, in most situations, it is reasonable to require the results to be similar in certain extent. Thus we propose the following two kinds of metrics.

3. STRONG EQUIVALENCE: PERFECT CORRELATION

We will use the correlation coefficient to characterize the consistency. First we give the definition of the correlation coefficient ρ.

Definition 1. *Given $\{a_i\}_{i=0}^{N-1}$, where $a_i \in \mathcal{A}$. $f : a_i \mapsto x_i,\ x \in \mathbb{R}^n$, $T : x_i \mapsto y_i$ is a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, $\mathbf{\Phi}_l$ is an operator on the numerical sequence. $m(\mathbf{\Phi}_l)$ is the mean value of the $\mathbf{\Phi}_l$ in the space of parameter l. The correlation coefficient is defined as*

$$\rho = \frac{\int_l [\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1}) - m(\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1}))]}{\sqrt{\int_l (\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1}) - m(\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1})))^2 dl}} \frac{[\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1}) - m(\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1}))] dl}{\sqrt{\int_l (\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1}) - m(\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1})))^2 dl}} \tag{5}$$

It is well known that the correlation coefficient is between $[-1, 1]$. The correlation coefficient can used as a measure to characterize the similarity of two different mappings. For a given T, if $\rho = 1$, then we say the transformation T is a *strongly equivalent* transformation of the map f for an operator and $\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1})$ is a *strong equivalence* of $\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1})$. When correlation coefficient is 1, it means the results under two mappings are the same only up to a translation and scaling. That is the reason why it is called "strongly equivalent". Unfortunately, there is no the universal equivalent transformation for any operator. However, because of the importance of second-order statistics, we shall emphasize on the second-order operators. We consider the genomic correlation function, which plays a vital role in genomic signal processing The genomic correlation function of a sequence is defined as

$$r_l = \frac{1}{N} \sum_{n=0}^{N-1} x^T(n) x(n - l) \tag{6}$$

If $\rho = 1$, we have the following theorem on the transformation T for correlation function.

Theorem 1. *If the transformation T is linear, then the correlation coefficient $\rho = 1$ if and only if the transformation T can be represented as $T(x_i) = \lambda \mathbf{R} x_i$, $\mathbf{R}$ is an orthogonal matrix and $\lambda \in \mathbb{R}$.*

Actually, this property not only holds for genomic correlation function, but also for a large class of operators. In order to show this class of operator, we first introduce the definition of bounded linear operator.

Definition 2. *Let $(X, \|\cdot\|)$ be a normed space. An operator f is a bounded functional if f is linear and there exists $C > 0$, such that $|f| \le C\|x\|$.*

The bounded operator can be thought as the BIBO linear system in signal processing theory, which illustrates the good-behaved operator. Then we have the following theorem.

Theorem 2. *Any bounded linear operator can only have a trivial (scaled identity transformation) linear strongly equivalent transformation. Moreover, given a bounded operator whose Taylor's expansion order is less than or equal to two, then rotation is a linear strongly equivalent transformation if and only if the operator does not have the first-order component and the Hessian $\nabla^2\mathbf{\Phi}_l(x)$ has the form*

$$\nabla^2\mathbf{\Phi}_l(x) = \begin{pmatrix} k_{11}I_{n\times n} & k_{12}I_{n\times n} & \cdots & k_{1N}I_{n\times n} \\ k_{21}I_{n\times n} & k_{22}I_{n\times n} & \cdots & k_{2N}I_{n\times n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N1}I_{n\times n} & k_{N2}I_{n\times n} & \cdots & k_{NN}I_{n\times n} \end{pmatrix} \tag{7}$$

, where $k_{ij} \in \mathbb{R}$ and $k_{ij} = k_{ji}$, $\forall i \neq j$.

Furthermore, using Theorem 2 we can also show that rotation is a strongly equivalent transformation for genomic Fourier analysis.

4. WEAK EQUIVALENCE: PRESERVATION OF LOCAL EXTREMA

In previous section, we introduce a metric to measure the similarity between two mappings for an operator. However, as we can see, the *strong equivalence* needs the result to be "exactly" the same. While in many situations, we do not care too much whether the result under two mapping strategies are exactly the same, i.e. the true numerical value of the result, but the relative relation or the relative trend in the result. For example, when we use correlation function, in many cases, we only care where is the peak point and valley point. Because this may suggest certain pattern appears more frequently than any other one. In these cases, what we really want is to preserve the local maximal and minimal for different mapping. So we first give the definition of local minimum and maximum preserving similarity or for which in this paper what we call *weakly equivalent*.

Definition 3. *Given $\{a_i\}_{i=0}^{N-1}$, where $a_i \in \mathcal{A}$. $f : a_i \mapsto x_i,\ x \in \mathbb{R}^n$, $T : x_i \mapsto y_i$ is a transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$, $\mathbf{\Phi}_l$ is an operator on the numerical sequence. We say T is* weakly equivalent*, if for any l, any local minima or maxima for $\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1})$ we have $\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1})$ has the same local minima or maxima respectively.*

A few easy observations and results follow. By definition strong equivalence implies weak equivalence. Moreover, we have the following propositions to determine weak equivalence.

Proposition 1. *If $\mathbf{\Phi}_l$ is twice differentiable with respect to l, the T is* weakly equivalent*, if for any l, where $\frac{\partial\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1})}{\partial l} = 0$, the following conditions hold*

$$\frac{\partial\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1})}{\partial l} = 0 \tag{8}$$

and

$$\frac{\partial^2\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1})}{\partial l^2} \cdot \frac{\partial^2\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1})}{\partial l^2} \geq 0 \tag{9}$$

If $l \in \mathbb{Z}$, Then we have the following criterion to determine weak equivalence.

Proposition 2. *T is* weakly equivalent*for an operator $\mathbf{\Phi}_l$ where $l \in \mathbb{Z}$, if for any l, the following condition holds*

$$\begin{aligned}&(\mathbf{\Phi}_l(\{x_i\}_{i=0}^{N-1}) - \mathbf{\Phi}_{l-1}(\{x_i\}_{i=0}^{N-1})) \\ &\cdot (\mathbf{\Phi}_l(\{T(x_i)\}_{i=0}^{N-1}) - \mathbf{\Phi}_{l-1}(\{T(x_i)\}_{i=0}^{N-1})) \geq 0\end{aligned} \tag{10}$$

Again, as the importance of genomic correlation function, especially we'd like to investigate the weak equivalent transformation for the genomic correlation function. Then we have the following theorem showing that rotation can be thought essentially as the only weakly equivalent transformation for genomic correlation function.

Theorem 3. *For a fix length sequence, any transformation which only brings small enough changes to the inner product value under previous mapping will be a weakly equivalent transformation for correlation function. However, if the length goes to infinity, then rotation (or scaled rotation) is the only weakly equivalent transformation for correlation function.*

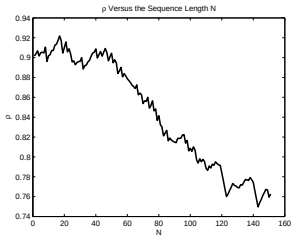
5. GENOMIC SEQUENCE ANALYSIS

We conduct the experiments on human gene AD169 sequences (GenBank accession no. X17403) and rhodopsin gene sequence (GenBank accession no. U49742). We calculate the correlation function as in (6) using the mapping, which maps the $\mathcal{A} = \{\mathtt{A},\ \mathtt{T},\ \mathtt{G},\ \mathtt{C}\}$ to the standard basis of $\mathbb{R}^4$. Then we use another mapping strategy, which maps `A` to $(-1, 0, 0, 0)$, `T` to $(1, 0, 0, 0)$, `G` to $(0, 1, 0, 0)$ and `C` to $(0, -1, 0, 0)$. These are two widely used mapping methods. In Fig. 1 (a), (c), we show the changing of correlation coefficient between the two correlation results with growth of DNA sequence length N and in (b), (d) we show how the percentage of the points having same local extremum property in two results grows with N. As our theories point out, all these two metrics have a decreasing trend with the grown of length N. The similarity between the two results become less and less, which finally will lead to a inconsistent analysis results.

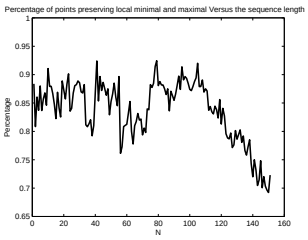
In Fig. 2, we show the strong equivalence measurements between the genomic power spectrum under the previous two mapping methods. We can find that the power spectrum results using these two different mappings have the trends to be inconsistent. Since the genomic correlation function and power spectrum are widely used and pervasive in genomic sequences analysis, it suggests the consistency problem should not never be neglected when comparing genomic sequence analysis results.

6. CONCLUSION

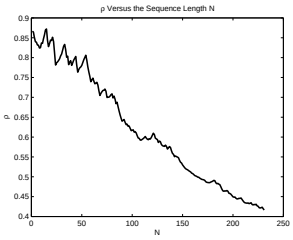
In this paper, we presented a mathematical framework for analysis of the consistency in representation and transformation of numerical mappings of genomic sequences. We



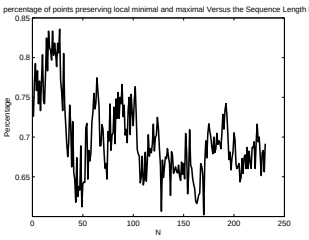
(a) Correlation coefficients for strong equivalence



(b) Percentage of points preserving local extremes for weak equivalence

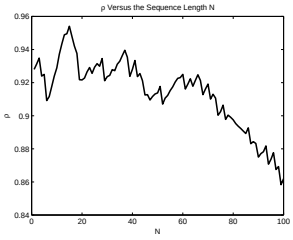


(c) Correlation coefficients for strong equivalence

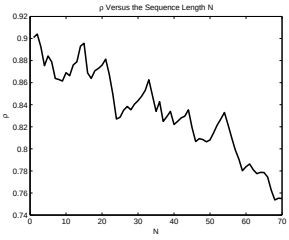


(d) Percentage of points preserving local extremes for weak equivalence

Figure 1. (a), (b) shows the correlation coefficient and percentage of points preserving local extremes of the power spectrum change with growth of sequence length N for human gene AD169 sequences respectively using the first mapping. (c) and (d) shows the strong and weak equivalence measurements respectively change with growth of sequence length N for rhodopsin gene sequences using the second map.



(a) Strong equivalence measurement for power spectrum of AD169 sequences



(b) Strong equivalence measurement for power spectrum of rhodopsin gene sequences

Figure 2. (a), (b) shows the correlation coefficient between the results of power spectrum using two different mappings changing with growth of sequence length N for Human gene AD169 sequences and rhodopsin gene sequences respectively.

introduced strong and weak correlation metrics for characterization of consistency measures among distinct numerical mappings. We derived sufficient conditions to ensure consistency among different numerical mappings. We presented an important class of equivalent transforms under the proposed consistency conditions. We also derived a class of operators which is shown to be equivalent under rotation of numerical mappings. Finally, we conducted computer simulation experiments on DNA sequences which demonstrate the theoretical results. Our results suggest a possible reason for inconsistent and often contradictory results obtained in long-range correlation analysis of DNA sequences and other areas in genomic analysis. The proposed approach for analysis of numerical mappings can be extended to symbolic signal processing for numerous applications beyond genomic and proteomic sequences.

7. REFERENCES

[1] S. V. Buldyrev et al., "Long-range correlation properties of coding and noncoding dna sequences: Genbank analysis," *Phys. Rev. E*, vol. 51, pp. 5084–5091, 1995.

[2] R. F. Voss, "Evolution of long-range fractal correlations and 1/f noise in dna base sequences," *Phys. Rev. Lett.*, vol. 68, pp. 3805–3808, 1992.

[3] B. D. Silverman and R. Linsker, "A measure of dna periodicity," *Journal of Theoretical Biology*, vol. 118, pp. 295–300, 1986.

[4] D. S. Stoffer, D. E. Tyler, and A. J. McDougall, "Spectral analysis for categorical time series: Scaling and the spectral envelope," *Biometrika*, vol. 80, pp. 611–622, 1993.

[5] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. on Signal Processing*, vol. 50, no. 3, pp. 628–635, March 2002.

[6] C. K. Peng et al., "Long-range correlations in nucleotide sequences," *Nature*, vol. 356, no. 6365, pp. 168–170, March 1992.

[7] P. Carpena et al., "Identifying chracteristic scales in the human genome," *Phys. Rev. E*, vol. 75, 2007.

[8] N. Bouaynaya and D. Schonfeld, "Nonstationary Analysis of Coding and Noncoding Regions in Nucleotide Sequences," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 3, pp. 357–364, 2008.

[9] L. De Lathauwer, *Signal Processing Based on Multilinear Algebra*, Ph.D. thesis, K.U. Leuven E.E. Dept.-ESAT, Belgium, 1997.