

# Mapping Equivalence for Symbolic Sequences: Theory and Applications

Liming Wang, *Student Member, IEEE*, and Dan Schonfeld, *Senior Member, IEEE*

**Abstract**—Processing of symbolic sequences represented by mapping of symbolic data into numerical signals is commonly used in various applications. It is a particularly popular approach in genomic and proteomic sequence analysis. Numerous mappings of symbolic sequences have been proposed for various applications. It is unclear however whether the processing of symbolic data provides an artifact of the numerical mapping or is an inherent property of the symbolic data. This issue has been long ignored in the engineering and scientific literature. It is possible that many of the results obtained in symbolic signal processing could be a byproduct of the mapping and might not shed any light on the underlying properties embedded in the data. Moreover, in many applications, conflicting conclusions may arise due to the choice of the mapping used for numerical representation of symbolic data. In this paper, we present a novel framework for the analysis of the equivalence of the mappings used for numerical representation of symbolic data. We present strong and weak equivalence properties and rely on signal correlation to characterize equivalent mappings. We derive theoretical results which establish conditions for consistency among numerical mappings of symbolic data. Furthermore, we introduce an abstract mapping model for symbolic sequences and extend the notion of equivalence to an algebraic framework. Finally, we illustrate our theoretical results by application to DNA sequence analysis.

**Index Terms**—DNA sequence analysis, symbolic signal processing, transform equivalence.

## I. INTRODUCTION

INFORMATION is provided in many forms. At times, information is conveyed numerically. More often, information is represented in the form of symbols such as characters, tags, etc. For example, the areas of genomic and proteomic signal processing focus on sequences of nucleotides and amino acids, respectively [1]. The aim of symbolic signal analysis is to process symbolic data elements in order to extract useful information.

In general, symbolic information is represented as a sequence of symbols (possibly of infinite length)  $\{a_i\}_{i=0}^{N-1}$ , where  $a_i \in \mathcal{A}$  and  $\mathcal{A}$  is a set of all possible symbols. For example,  $\mathcal{A}$  could be a collection of the 26-lowercase English letters, i.e.,  $\mathcal{A} = \{a, b, \dots, z\}$ , or the four nucleotides in a genomic sequence, i.e.,  $\mathcal{A} = \{A, T, G, C\}$ . In statistical literature, symbolic data is usually called categorical data [2]. The use of Markov

chain models and hidden Markov models has been examined for time-domain analysis of genomic and proteomic data [3]–[5]. However, we often seek to rely on frequency-domain analysis methods of symbolic signals. Unfortunately, symbolic sets do not generally possess an algebraic structure that allows us to define mathematical operations (e.g., group, ring, or field). In traditional signal processing, the set  $\mathcal{A}$  corresponds to real- or complex-valued numbers, i.e.,  $\mathcal{A} = \mathbb{R}$  or  $\mathbb{C}$ , which form an algebraic field. However, attempts to define mathematical operations such as addition and multiplication on symbolic data has raised many questions about the meaning of the results obtained using such methods.

Several techniques exist which incorporate numerical and symbolic processing in an effective way to develop symbolic analysis systems [6]. Software systems for symbolic computational algebra (e.g., Mathematica, Maple, etc.) represent a successful example of this approach. Such systems, however, are application-specific and difficult to realize for a broad class of symbolic signal processing applications. There are also various techniques for analyzing correlations, periodicities, etc. that do not require the aid of numerical symbol mappings. Among these techniques, the mutual information function (MIF) [7] is one of the most important. The main advantage of these methods is that numerical mappings are not required. Moreover, it can be shown that methods such as MIF can capture any type of statistical dependence. The main disadvantages of these techniques, however, are that they generally provide less specific information than correlation analysis and they often suffer from a systematic overestimation of mutual information for finite sequences. Nevertheless, in order to extract the mathematical and statistical information embedded in symbolic sequences, we wish to employ the powerful analysis tools developed in traditional signal processing, e.g., Fourier transform, correlation function, etc. We must therefore map the symbolic elements into numerical values. The resulting numerical sequence should preserve the information embedded in the symbolic data. Moreover, it should allow traditional signal processing techniques to extract the salient information about the symbolic sequences from the corresponding numerical signals. For instance, in DNA sequences, we have a finite alphabet associated with the four nucleotides in the genome, i.e.,  $\mathcal{A} = \{A, T, G, C\}$ . The mapping used for the representation of genomic data must preserve the inherent structure of DNA sequences. In particular, if we choose a mapping such as  $A \mapsto 1, T \mapsto 0, G \mapsto -1, C \mapsto 0$ , we would not preserve uniqueness since T and C are mapped to the same value.

Numerous mappings have been proposed for the numerical representation of DNA sequences. Buldyrev *et al.* [8] proposed

Manuscript received January 25, 2009; accepted May 26, 2009. First published June 30, 2009; current version published November 18, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Z. Jane Wang.

The authors are with the Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607-7053 USA (e-mail: lwang37@uic.edu; dans@uic.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TSP.2009.2026544

various mapping rules for the representation of nucleotide sequences into one-dimensional numerical sequences based on the purine–pyrimidine (RY) rule, hydrogen-bond energy rule, etc. Li and Kaneko [9] and Voss [10] used the indicator sequence method, which essentially maps the symbol to a standard basis of the 4-D Euclidean space  $\mathbb{R}^4$ . Berthelsen *et al.* [11] revised the method introduced in [8] by taking the molecular mass and hydrophobicity into account in representation of genomic data. Silverman and Linsker [12] relied on the simplex method, which maps the symbol to the vertices of a regular simplex. Cristea and Anastassiou [13], [14] proposed the tetrahedral mapping, which maps the nucleotides into corners of a tetrahedron. Stoffer *et al.* [15] introduced a mapping whose aim is to accentuate the periodic features embedded in genomic sequences for stationary symbolic sequence analysis. Wang and Johnson [16] extended the method proposed by Stoffer *et al.* [15] for non-stationary sequence analysis. Rushid and Tuqan [17] proposed the Z-curve mapping, which is a unique 3-D curve representation whose sequences are composed of binary values, i.e., 1 and  $-1$ . They also proposed a matrix-based framework to combine many widely used mapping strategies in genomic sequence analysis [18].

Each of the large number of numerical mappings used for the representation of genomic sequences can be justified for various applications. This raises several fundamental questions: What are the merits of each mapping used for the analysis of DNA sequences? How can we compare the results obtained from different numerical mappings? 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 [10], [19], [20]. Investigation using a large DNA sequence database did not resolve this dispute; in fact, the controversy grew even further [21]. Bouaynaya and Schonfeld [22], [23] shed light on this dilemma by demonstrating that a certain class of 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 [22] and [23] 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 [10], [19], [20].

To ensure a clear understanding of the implications of the different choices used for numerical representation of symbolic data, we must develop a fundamental new approach that can be used to characterize the fundamental properties of numerical mappings. Specifically, it is essential that we establish a mapping equivalence theory for symbolic data that can be used to guarantee consistency among a class of numerical representations. With the aid of a mapping equivalence theory we could determine whether different mappings should yield compatible results, i.e., whether the mappings used for the analysis of the

same data lead to consistent conclusions. Moreover, the theory can indicate when distinct mappings could lead to contradictory results and thus comparison of the corresponding conclusions is futile.

In this paper, we provide a mapping equivalence theory for the numerical representation of symbolic data undergoing transformation by an operator. We focus primarily on the mapping  $f : \mathcal{A} \rightarrow \mathbb{R}^n$  which maps the symbols to the  $n$ -dimensional Euclidean space. In Section II, we first propose a framework for the analysis of different numerical mappings undergoing transformation by an analytic operator using Taylor's expansion. Moreover, we emphasize the investigation of first- and second-order operators including the correlation function and Fourier transform. These operators are widely used in signal processing and analysis and thus play an important role in this presentation. In Section II-A, we provide an analysis of the correlation between different numerical mappings of a symbolic sequence. In particular, we derive conditions for strong equivalence captured by perfect correlation among distinct mappings. In Section II-B, we explore a relaxed similarity measure between distinct numerical mappings. Specifically, we provide conditions for weak equivalence which is characterized by preservation of the local extrema of the representation. In Section III, we introduce an abstract mapping model and extend the concept of equivalence to the generalized mapping model. In Section IV, we present experimental results which illustrate the significance of the proposed mapping equivalence theory in symbolic signal processing applications. In this presentation, the simulations are focused exclusively on analysis of genomic sequences. The results presented in this paper, however, are applicable for any symbolic signal modeled by a discrete alphabet with a finite cardinality and independent of particular statistical properties such as stationarity, etc. Finally, we provide a brief summary and discussion of our results in Section V.

## II. EUCLIDEAN MAPPING EQUIVALENCE FOR SYMBOLIC SEQUENCES

Given  $\{a_i\}_{i=0}^{N-1}$ , where  $a_i \in \mathcal{A}$  and  $|\mathcal{A}| < \infty$ , here  $|\cdot|$  means the cardinality of the set.  $f$  is a mapping from  $\mathcal{A}$  to  $\mathbb{R}^n$ , i.e.,  $f : a_i \mapsto x_i, x_i \in \mathbb{R}^n$ . After the mapping we obtain a vector sequence  $\{x_i\}_{i=0}^{N-1}$ .  $T : x_i \mapsto y_i$  is a transformation from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ .  $\Phi_l$  is an analytic operator on the numerical sequence and maps into  $\mathbb{R}$  parameterized by  $l \in \mathbb{R}$ . We also assume that  $\Phi_l \in L^2(l)$ . We classify the problems as in the following cases.

- 1) Given  $T$ , determine the consistency between  $\Phi_l(\{x_i\}_{i=0}^{N-1})$  and  $\Phi_l(\{T(x_i)\}_{i=0}^{N-1})$ . We need also figure out the largest class of operators which shows consistent results for two mappings under the given  $T$ .
- 2) Given  $f$  and  $\Phi_l$ , 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. Also figure out the largest class of transformation  $T$  preserving the consistency for given mapping  $f$ .

The consistency here means we require the results under two different mappings to be similar in certain extent. In general  $\Phi_l$  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.  $\Phi_l : \mathbb{R}^{Nn \times 1} \rightarrow \mathbb{R}$ . Without using the common scalar form representation of Taylor's expansion [24], we shall present it in a concise form by using tensor product. First, we define the gradient operator  $\nabla$  as

$$\nabla = \left( \frac{\partial}{\partial x_1} \quad \frac{\partial}{\partial x_2} \quad \cdots \quad \frac{\partial}{\partial x_{Nn}} \right)^T. \quad (1)$$

Then the Taylor's expansion of  $\Phi_l$  at  $x_0$  can be represented as the following form:

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

where  $\times_i$  is the  $i$ th mode tensor product [25], and  $\nabla^i$  is the  $i$ th order gradient of  $\Phi_l$ , which is defined as

$$\nabla^i \Phi_l = \Phi_l \times_1 \nabla \times_2 \nabla \times_3 \cdots \times_i \nabla. \quad (3)$$

Furthermore,  $\nabla^0 \Phi_l$  is defined as  $\Phi_l$ . For first- and second-order terms, it is easy to check that it coincides with the well-known definition of *Gradient*  $\nabla \Phi_l(x)$  and *Hessian*  $\nabla^2 \Phi_l(x)$ . So we can rewrite the Taylor's expansion at  $x_0$  as

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

A metric or measure is needed for measuring the consistency. In general, there is no universal metric. Various operators may have different metrics for different purposes. In many cases, it is a reasonable principle to require the results of two different mappings to be similar in some extent. In light of this principle, we propose the following two kinds of metrics.

**A. Strong Equivalence: Perfect Correlation**

We will use the correlation coefficient to characterize the consistency. First we provide the definition of the correlation coefficient  $\rho$  used in this paper.

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

$$\begin{aligned} \rho &= \frac{\int_L [\Phi_l(\{x_i\}_{i=0}^{N-1}) - m(\Phi_l(\{x_i\}_{i=0}^{N-1}))]}{\sqrt{\int_L (\Phi_l(\{x_i\}_{i=0}^{N-1}) - m(\Phi_l(\{x_i\}_{i=0}^{N-1})))^2 d\mu}} \\ &\times \frac{[\Phi_l(\{T(x_i)\}_{i=0}^{N-1}) - m(\Phi_l(\{T(x_i)\}_{i=0}^{N-1}))]}{\sqrt{\int_L (\Phi_l(\{T(x_i)\}_{i=0}^{N-1}) - m(\Phi_l(\{T(x_i)\}_{i=0}^{N-1})))^2 d\mu}}. \end{aligned} \quad (5)$$

The use of abstract integration provides a unified framework for definition of the correlation coefficient. The measure  $\mu$  can be chosen to be any Borel measure such as the

Lebesgue–Stieltjes or counting measures depending on the properties of the operator. In practice, the measures we rely upon are mainly the counting measure and the Lebesgue measure.

It is well known that the correlation coefficient is between  $[-1, 1]$  [26]. The correlation coefficient can be 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  $\Phi_l(\{T(x_i)\}_{i=0}^{N-1})$  is a *strong equivalence* of  $\Phi_l(\{x_i\}_{i=0}^{N-1})$ . When the correlation coefficient is 1, it means the two mappings are the same up to a translation and scaling. This is the reason that it is called “strongly equivalent.” Unfortunately, there is no the universal equivalent transformation for arbitrary operator. However, because of the importance of second-order statistics, we shall emphasize on the second-order operators such as the correlation function. From now on we will focus on the transformation  $T$  from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ . For the case of mapping between Euclidean spaces with a different dimensions, we will present a detailed discussion in Section II-D.

We first consider the correlation function. The correlation function of a sequence is defined as

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

Then if  $\rho = 1$ , we have the following theorem on the strongly equivalent transformation  $T$ .

*Theorem 1:* For non-trivial operator and linear transformation  $T$ , 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  $\lambda \in \mathbb{R}$ .

*Proof:* If  $T(x_i) = \lambda \mathbf{R}x_i$ , and  $\mathbf{R}$  is orthogonal. Then

$$\begin{aligned} r_l(T(\{x_i\}_{i=0}^{N-1})) &= \frac{1}{N} \sum_{n=0}^{N-1} \lambda x_i^T \mathbf{R}^T \mathbf{R} x(n-l) \\ &= \frac{1}{N} \sum_{n=0}^{N-1} \lambda x_i^T x(n-l) \\ &= \lambda r_l(\{x_i\}_{i=0}^{N-1}). \end{aligned} \quad (7)$$

Conversely, if  $\rho = 1$ . and  $T(x_i) = \mathbf{A}x_i$ . Then

$$\begin{aligned} r_l(T(\{x_i\}_{i=0}^{N-1})) &= \frac{1}{N} \sum_{n=0}^{N-1} x_i \mathbf{A}^T \mathbf{A} x(n-l) \\ &= \lambda' r_l(\{x_i\}_{i=0}^{N-1}) + c \\ &= \lambda' \frac{1}{N} \sum_{n=0}^{N-1} x^T(n)x(n-l) + c \end{aligned} \quad (8)$$

where  $\lambda' \in \mathbb{R}$  and  $c \in \mathbb{R}$  is a constant. Since the equality holds for any sequence and any  $l$ . So  $\mathbf{A}^T \mathbf{A} = \lambda I$ , then  $\mathbf{A}$  is orthogonal. ■

Actually, this property not only holds for correlation operator, but also for a larger class of operators. Consider the Taylor's expansion of an operator  $\Phi_l$ . We would like first to introduce the definition of bounded linear operator and Riesz representation

theorem [27]. Then we will present a result for the first- and second-order bounded operators.

*Definition 2:* Let  $(X, \|\cdot\|)$  be a normed space. A operator  $f : X \rightarrow \mathbb{R}$  is a bounded operator if  $f$  is linear and there exists  $C > 0$ , such that  $|f| \leq C\|x\|$ .

The bounded operator can be thought as an analog of BIBO linear system in signal processing theory, which illustrates the good behaved operators. Furthermore if the space  $X$  is a Hilbert space, we have the following theorem to characterize any linear bounded operator.

*Theorem 2 (Riesz Representation Theorem for Hilbert Space):*  $X$  is a Hilbert space, then for any linear bounded operator  $\phi$ , there exists a unique  $y \in X$ , such that  $\phi(x) = \langle x, y \rangle$ .

Note that  $\mathbb{R}^n$  with the usual dot product is a Hilbert space. Therefore, Riesz Representation theorem for Hilbert space holds for  $\mathbb{R}^n$ . As before, we vectorize the vector sequence  $\{x_i\}_{i=0}^{N-1} x_i \in \mathbb{R}^n$  to a large vector  $x \in \mathbb{R}^{Nn \times 1}$ . Then any linear transformation  $T$  can be represented in the form

$$T = \begin{pmatrix} \mathbf{A}_{n \times n} & & & \\ & \mathbf{A}_{n \times n} & & \\ & & \ddots & \\ & & & \mathbf{A}_{n \times n} \end{pmatrix}_{Nn \times Nn} \quad (9)$$

i.e.,  $y = Tx$ .

Then we have the following theorem for equivalent transformation of first- and second-order operators.

*Theorem 3:* Any non-trivial bounded linear operator does not have any non-trivial (scaled identity mapping) linear *strongly equivalent* transformation. If rotation is a *strongly equivalent* transformation for a bounded operator whose Taylor's expansion does not have the third or higher-order terms, then its Taylor's expansion can not have first-order term and the Hessian  $\nabla^2 \Phi_l(x)$  must have the form

$$\nabla^2 \Phi_l(x) = \begin{pmatrix} k_{11} I_{n \times n} & k_{12} I_{n \times n} & \cdots & k_{1N} I_{n \times n} \\ k_{21} I_{n \times n} & k_{22} I_{n \times n} & \cdots & k_{2N} I_{n \times n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{N1} I_{n \times n} & k_{N2} I_{n \times n} & \cdots & k_{NN} I_{n \times n} \end{pmatrix}, \quad (10)$$

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

For Fourier transform, in many situations, we focus exclusively on the modulus of the transform of symbolic data, i.e., we discard the phase information. Since the module of continuous-time Fourier transform is invariant under rotation, it is tempting to conclude that rotation is an equivalent transformation for the Fourier transform. However, the widely used form of the Fourier transform used in much of the literature devoted to DNA sequence analysis [16] is different from the classical multi-dimensional Fourier transform. Fortunately, we are able to show that rotation still yields an equivalent transformation. We first define the Fourier transform as

$$\hat{f}_m = \frac{1}{N^2} \|X \mathbf{L}_F\|_2^2 \quad (11)$$

where  $X$  is a  $n \times N$  matrix, whose  $i$ th column is  $x_i$ .  $\mathbf{L}_F$  is the frequency vector, i.e.,

$$\mathbf{L}_F = \left( e^{\frac{-2\pi j m 0}{N}} \quad e^{\frac{-2\pi j m 1}{N}} \quad e^{\frac{-2\pi j m 2}{N}} \quad \cdots \quad e^{\frac{-2\pi j m (N-1)}{N}} \right)^T. \quad (12)$$

If we vectorize  $X$  to  $x \in \mathbb{R}^{Nn \times 1}$  as before,  $\hat{f}_m$  can also be represented as  $\hat{f}_m = 1/N^2 \|Ax\|_2^2$ , where

$$A = \left( e^{\frac{-2\pi j m 0}{N}} I_{n \times n} \quad \cdots \quad e^{\frac{-2\pi j m (N-1)}{N}} I_{n \times n} \right)_{n \times N}. \quad (13)$$

Notice that  $\hat{f}_m = 1/N^2 (x^H A^H A x)$ , which is a second-order operator and  $A^H A$  is of the form

$$A^H A = \begin{pmatrix} \bar{w}_0 w_0 I_{n \times n} & \bar{w}_0 w_1 I_{n \times n} & \cdots & \bar{w}_0 w_N I_{n \times n} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{w}_N w_0 I_{n \times n} & \bar{w}_N w_1 I_{n \times n} & \cdots & \bar{w}_N w_N I_{n \times n} \end{pmatrix} \quad (14)$$

where  $w_i = e^{(-2\pi j m i)/N}$ . By Theorem 3, rotation is a strongly equivalent transformation for Fourier transform.

### B. Weak Equivalence: Preservation of Local Extrema

In the previous section, we employed the correlation coefficient as a metric to characterize the similarity for an operator under transformation. However, as we can see, the *strong equivalence* basically requires the result to be "exactly" the same. While in many situations, we do not focus on whether or not 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 of the result. For example, when we use the correlation function, in many cases, we only care where the peak and valley points are located and the changing trends, which are used to determine the periodicity structure of certain patterns. In these cases, what we really need is to preserve the local extremums and local trend under the a transformation. So we first give the definition of local minimum and maximum preserving similarity, or in this paper what we call *weakly equivalent*.

*Definition 3:* Given  $\{a_i\}_{i=0}^{N-1}$ , where  $a_i \in \mathcal{A}, |\mathcal{A}| < \infty$ .  $f : a_i \mapsto x_i, x \in \mathbb{R}^n, T : x_i \mapsto y_i$  is a transformation from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , and  $\Phi_l$  is an operator on the numerical sequence. We say  $T$  is *weakly equivalent*, if for every  $x$ , which is a local minimal or maximal for  $\Phi_x(\{x_i\}_{i=0}^{N-1})$ , then  $\Phi_x(\{T(x_i)\}_{i=0}^{N-1})$  is also a local minimal or maxima, l 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  $\Phi_l$  is twice differentiable with respect to  $l$ , then  $T$  is *weakly equivalent*, if for any  $l$ , where  $(\partial \Phi_l(\{x_i\}_{i=0}^{N-1})) / (\partial l) = 0$ , the following conditions hold:

$$\frac{\partial \Phi_l(\{T(x_i)\}_{i=0}^{N-1})}{\partial l} = 0 \quad (15)$$

and

$$\frac{\partial^2 \Phi_l(\{x_i\}_{i=0}^{N-1})}{\partial l^2} \cdot \frac{\partial^2 \Phi_l(\{T(x_i)\}_{i=0}^{N-1})}{\partial l^2} \geq 0. \quad (16)$$

*Proof:* If  $l$  is a local maximal or local minimal, then  $(\partial\Phi_l(\{x_i\}_{i=0}^{N-1}))/(\partial l) = 0$  and  $(\partial^2\Phi_l(\{x_i\}_{i=0}^{N-1}))/(\partial l^2) \leq 0$  or  $(\partial^2\Phi_l(\{x_i\}_{i=0}^{N-1}))/(\partial l^2) \geq 0$ . By the definition of weak equivalence, (15) and (16) follow. ■

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

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

$$(\Phi_l(\{x_i\}_{i=0}^{N-1}) - \Phi_{l-1}(\{x_i\}_{i=0}^{N-1})) \cdot (\Phi_l(\{T(x_i)\}_{i=0}^{N-1}) - \Phi_{l-1}(\{T(x_i)\}_{i=0}^{N-1})) \geq 0. \quad (17)$$

*Proof:* Without loss of generality, we assume  $l$  is a local maximal for  $\Phi_l(\{x_i\}_{i=0}^{N-1})$ . then  $\Phi_l(\{x_i\}_{i=0}^{N-1}) \geq \Phi_{l-1}(\{x_i\}_{i=0}^{N-1})$  and  $\Phi_l(\{x_i\}_{i=0}^{N-1}) \geq \Phi_{l+1}(\{x_i\}_{i=0}^{N-1})$ . If (17) holds, we have  $\Phi_l(\{T(x_i)\}_{i=0}^{N-1}) \geq \Phi_{l-1}(\{T(x_i)\}_{i=0}^{N-1})$  and  $\Phi_l(\{T(x_i)\}_{i=0}^{N-1}) \geq \Phi_{l+1}(\{T(x_i)\}_{i=0}^{N-1})$ . Thus,  $l$  is also a local maximal for  $\Phi_l(\{T(x_i)\}_{i=0}^{N-1})$ . ■

As the importance of second-order statistics, specially we would like to investigate the weakly equivalent transformation for the correlation function. We first introduce a lemma.

*Lemma 1:* If the transformation  $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is an inner-product preserving isometry, i.e.,  $x^T y = T(x)^T T(y), \forall x, y \in \mathbb{R}^n$ , then  $(x) = \mathbf{R}x$ , where  $\mathbf{R}$  is an orthogonal matrix. Hence,  $T$  is a bijective isometry.

*Proof:* First let  $x = y$ , we have  $\|x\|_2 = \|T(x)\|_2$ , i.e.,  $T$  preserves the Euclidean norm. Since  $T(x_0)$  is on the ball  $\|x\|_2 = \|x_0\|_2$ , we have  $T(x) = \mathbf{R}(x)x$ , where  $\mathbf{R}(x)$  is an orthogonal matrix function. Let

$$\mathbf{R}(x) = (u_1(x) \quad u_2(x) \quad \dots \quad u(x)) \quad (18)$$

where  $\{u_i(x)\}$  is orthonormal. Furthermore, let  $x = (1, 0, 0, \dots, 0)^T$  and  $y = (y_1, 0, 0, \dots, 0)^T$ . then we have

$$x^T y = T(x)^T T(y) \quad (19)$$

$$\begin{aligned} y_1 &= x^T \mathbf{R}(x)^T \mathbf{R}(y) y \\ &= y_1 u_1^T((1, 0, 0, \dots, 0)^T) u_1(y). \end{aligned} \quad (20)$$

Therefore  $u_1^T((1, 0, 0, \dots, 0)^T) u_1(y) = 1$ . By *Cauchy-Schwartz* inequality, we have

$$\begin{aligned} 1 &= u_1^T((1, 0, 0, \dots, 0)^T) u_1(y) \\ &\leq \|u_1((1, 0, 0, \dots, 0)^T)\|_2 \cdot \|u_1(y)\|_2 = 1. \end{aligned} \quad (21)$$

The equality holds if and only if  $u_1(y) = \lambda u_1((1, 0, 0, \dots, 0)^T)$ , but  $\|u_1(y)\|_2 = 1$ , thus  $\lambda = 1$ . Therefore  $u_1(y) = u_1((1, 0, 0, \dots, 0)^T), \forall y \in \mathbb{R}^n$ . By the same arguments we can show  $u_i(y) = u_i(e_i), \forall i = 1, \dots, n$ , where  $e_i$  is the standard basis of  $\mathbb{R}^n$ . So  $\mathbf{R}(x) = \mathbf{R}$  is a constant orthogonal matrix, thus  $T(x) = \mathbf{R}x$ . This also shows  $T$  is a bijective isometry. ■

For correlation function, we have the following theorem showing that generally speaking, rotation can be thought as the “only” weakly equivalent transformation.

*Theorem 4:* 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.

*Proof:* Consider the vector sequence  $\{x_i\}_{i=0}^{N-1}$ . The correlation function is  $r_l(\{x_i\}_{i=0}^{N-1}) = (1/N) \sum_{n=0}^{N-1} x^T(n)x(n-l)$ . Then we have

$$r_l - r_{l-1} = \frac{1}{N} \sum_{n=0}^{N-1} x^T(n)(x(n-l) - x(n-l+1)). \quad (22)$$

After the transformation, we have the correlation function  $r'_l = r_l(T(\{x_i\}_{i=0}^{N-1})) = (1/N) \sum_{n=0}^{N-1} T(x(n))^T T(x(n-l))$  and  $r'_l - r'_{l-1} = (1/N) \sum_{n=0}^{N-1} T(x(n))^T (T(x(n-l)) - T(x(n-l+1)))$ . By Proposition 2.  $T$  is weakly equivalent if  $r'_l - r'_{l-1}$  has the same sign as  $r_l - r_{l-1}$ .

Consider the alphabet  $\mathcal{A}' = \mathcal{A} \cup \{a-b | a, b \in \mathcal{A}\}$ , which means we consider the symbol “ $a-b$ ” as a new symbol and we extend the mapping  $f$  on the newly added symbols as  $f : (a-b) \mapsto (f(a) - f(b))$ , which also extends the transformation  $T$  for  $f(a-b)$  be  $T(f(a)) - T(f(b))$ . Thus, finding the weakly equivalent transformation is same to find the  $T$  which preserve the sign at each  $l$  of cross correlation function  $R_l$  for the sequence  $\{x_i\}_{i=0}^{N-1}$  and  $\{x_i - x_{i+1}\}_{i=0}^{N-1}$ .

Let  $\mathcal{F} = (a_1 \quad a_2 \quad \dots \quad a_A)^T$ , where  $a_i = f(x)^T f(y)$ ,  $(x, y)$  or  $(y, x) \in \mathcal{A} \times \mathcal{A}'$  and  $A = |\mathcal{A} \times \mathcal{A}'|$ . Also let  $N_l = (C_1 \quad C_2 \quad \dots \quad C_A)^T$ , where  $C_i$  is the counting number for the pair  $(x, y)$  corresponding to  $a_i$  which appears in the cross correlation function  $R_l$ . Therefore, we have

$$R_l = \frac{1}{N} \mathcal{F}^T N_l. \quad (23)$$

Define  $T(\mathcal{F}) = (b_1 \quad b_2 \quad \dots \quad b_A)^T$ , where  $b_i = T(f(x))^T T(f(y))$ , where  $(x, y)$  corresponds to  $(x, y)$  in the  $a_i$  in  $\mathcal{F}$ . Notice that  $N_l$  will not change since it is determined by the given sequence. After the transformation, the cross correlation functions becomes  $R'_l = (1/N) T(\mathcal{F})^T N_l$ . We need  $R_l$  and  $R'_l$  have the same sign for all  $l$ . Notice that  $\sum_{i=1}^A C_i = N$ . So every  $N_l$  for a given sequence corresponds to a point on the hyperplane  $\sum_{i=1}^A x_i = N$  in  $\mathbb{R}^A$ . If  $T$  preserves signs for all  $l$ , then for each  $N_l, T(\mathcal{F})$  should reside in the same half plane of  $\mathcal{F}$ . Because  $T(\mathcal{F})^T N_l$  should have the same sign of  $\mathcal{F}^T N_l$ . In general, the sign will not all be positive or negative, since that will means the correlation function is monotonic which in general is not valid for all sequences. Consider all possible symbol sequence of length  $N$  and  $l \in \mathbb{Z}$ . Then

$$T(\mathcal{F}) \in \bigcap_{\text{all sequences of length } N} \{x | (x^T N_l) \cdot (\mathcal{F}^T N_l) \geq 0\}. \quad (24)$$

This implies  $T$  should reside in the intersection of all half plane determined by all sequences of length  $N$ . Each half plane is a convex cone, therefore the intersection is still a convex cone as illustrated in Fig. 1. Since we have finitely many point on the hyperplane  $\sum_{i=1}^A x_i = N$ . All  $N_l$  reside in the first quadrant. We can always construct two hyperplanes whose intersection is in one quadrant. Thus, the intersection will be a small convex cone in a quadrant. If  $T(\mathcal{F})$  is in that convex cone, then the sign is all preserved, which means  $T$  is a weakly equivalent transformation. This proves the first claim.

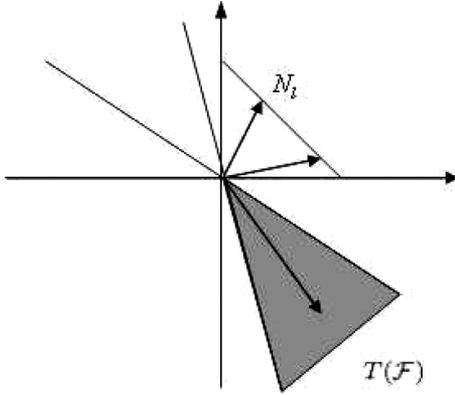


Fig. 1. Illustration for  $N_l$  and  $T(\mathcal{F})$ .  $T(\mathcal{F})$  should reside in the convex cone as the shaded area in the figure.

However, if we let  $N$  go to infinity, first notice that the intersection will not be empty set, since  $\mathcal{F}$  is always in the intersection. But the points  $N_l$  become dense in the first quadrant. We can have a sequence of  $N_l$  such that  $N_l^T \mathcal{F} \rightarrow 0^+$  and  $N_l'$  with  $N_l' \mathcal{F} \rightarrow 0^-$ . Therefore, the intersection will be squeezed to the line  $\lambda \mathcal{F}$ ,  $\lambda \geq 0$ . i.e., The  $T(\mathcal{F}) = \lambda \mathcal{F}$ . If the scaling  $\lambda = 1$ , then  $T(f(x))^T T(f(y)) = f(x)^T f(y)$ ,  $\forall f$ . Since the mapping is arbitrary, which means it should hold for any  $x$  and  $y$ . By Lemma 1,  $T$  is a rotation or a scaled rotation.

Conversely, by Theorem 1, we have rotation or scaled rotation is a strongly equivalent transformation. Thus, it follows it is also a weakly equivalent transformation. ■

From this theorem, we can see that rotation is essentially the only weakly equivalent transformation for correlation function. We can expand the class of operator having this property by combining Theorem 4 and Theorem 3 with some technical conditions, then we have the following corollary.

*Corollary 1:* If a second order operator whose Hessian is of the form as (10) and all  $k_i$  have the same sign and  $0 < \sum k_i < \infty$ , then rotation is essentially the only transformation which is both strongly and weakly equivalent.

*Proof:* If the Hessian has the form above, then

$$\Phi_l = \frac{1}{2} \text{sgn}(k_1) \sum_{n=0}^{N-1} |k_n| x(n)^T x(n-l) \quad (25)$$

where

$$\text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0. \end{cases}$$

Notice that the proof for correlation function follows here except we shall use  $N_l' = (\zeta_1 C_1 \quad \zeta_2 C_2 \quad \dots \quad \zeta_A C_A)^T$  instead of  $N_l$ , where  $\zeta_i$  are all positive and

$$\zeta_i = \sum_{j|\text{all terms having same pattern}} k_j.$$

If the length goes to infinity,  $N_l$  becomes dense, since  $0 < \sum k_i < \infty$ , then  $N_l'$  is in some non-degenerated bounded set, which is still dense and resides in the first quadrant. Then the argument above still valid. Thus, rotation is essentially the only weakly equivalent for this kind of operator. By Theorem 1, rotation is also the strongly equivalent transformation.

### C. Mapping Between 1-D Euclidean Spaces

The case in which the transformation  $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is limited to  $n = 1$  is particularly common in the literature. Moreover, this case stands out and deserves special attention since the only possible rotation on  $\mathbb{R}$  is obtained by scaling and interchanging the mapping values. From Theorems 1 and 4, we observe that if the mappings cannot be obtained by scaling and interchanging the mapping values, the correlation and Fourier analysis results obtained using these mappings are neither strongly nor weakly equivalent. Therefore, in the 1-D case, the equivalent mapping class under a given operator becomes fairly limited. In general, distinct mappings will usually lead to inconsistent correlation and Fourier analysis results. Another interesting fact about 1-D mappings that can be derived from our previous results is that if the mapping is binary (i.e., the range of the mapping can only takes two distinct values), then we observe that the correlation and Fourier analysis under any two such binary mappings are always consistent since we can always obtain one of the mappings by scaling and interchanging the mapping values of the other mapping.

### D. Mapping Between Euclidean Spaces of Different Dimensions

In the previous sections, we focused primarily on the transformation  $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , where  $n = m$ . In this section, we will present a brief discussion of the case where  $m \neq n$ . If  $m > n$ , which means  $T$  will transform the vector into a larger dimensional Euclidean space. However, since there is a natural embedding for  $\mathbb{R}^n$  into  $\mathbb{R}^m$ , we can always think the transformation as  $T' : \mathbb{R}^m \rightarrow \mathbb{R}^m$ . For second-order operators which are shown equivalent under rotation, we still have the same results in this situation, except the rotation matrix here means a matrix have orthonormal columns.

For the case  $m < n$ , we can also think as  $\mathbb{R}^m$  is embedded inside  $\mathbb{R}^n$  by the transform

$$y = \begin{pmatrix} I_{m \times m} \\ 0_{(n-m) \times m} \end{pmatrix}_{n \times m} x \quad (26)$$

where  $x \in \mathbb{R}^m$ ,  $y \in \mathbb{R}^n$ . Then we only need to research on the new transformation  $T' : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . However, in this case, we can see that we actually project the higher dimensional subspace into a lower dimensional space, rotation here in general is not an equivalent transformation anymore. Intuitively, because of the projection, we lose information projected on  $(n - m)$  dimensions. Therefore, rotation is no longer an equivalent transformation.

## III. ABSTRACT MAPPING EQUIVALENCE FOR SYMBOLIC SEQUENCES

### A. Abstract Mapping Model and Examples

In previous sections, we mainly focused on properties of mappings, which map symbols into vector space. However, it is not necessary to restrict to the vector space. Many classical concepts in numerical signal processing can be extended to various algebraic structures. For example, the Fourier transform

and Wavelet transform can be defined on group, ring and finite-field [28]–[30]. In this section, we introduce the generalized mapping to arbitrary semi-ring, ring or algebra structure. We shall also extend the notion of equivalence defined in previous section.

For given finite alphabet  $\mathcal{A}$ , we define  $F(\mathcal{A})$  as collection of all the symbolic sequences and define the binary operation as concatenating two symbolic sequences. It can be shown  $F(\mathcal{A})$  is a free semi-group [31].  $R$  is any semi-ring. Let  $\mathcal{R}$  to be collection of all maps from  $F(\mathcal{A})$  to  $R$ . For any  $f \in \mathcal{R}$ , we denote it as the formal series

$$f = \sum_{u \in F(\mathcal{A})} f(u)u \tag{27}$$

we define two operation  $+$  and  $\cdot$  on  $\mathcal{R}$  as

$$f + g = \sum_{u \in F(\mathcal{A})} (f(u) + g(u))u \tag{28}$$

$$(f \cdot g)(s) = \sum_{uv=s} (f(u)g(v))s. \tag{29}$$

With these two binary operations, we have the following proposition to show that we construct a new algebraic structure on  $\mathcal{R}$ .

*Proposition 3:*  $R$  is a semi-ring or ring, then  $(\mathcal{R}, +, \cdot)$  forms a semi-ring or ring, respectively.

*Proof:* By the definition of addition, we can see that if  $(R, +)$  is a commutative monoid or abelian group, then  $(\mathcal{R}, +)$  has the same property correspondingly. Therefore, it's enough to show to show  $(\mathcal{R}, \cdot)$  is a semi-group, i.e., we need to show the multiplication is associative.  $\forall f, g, h \in \mathcal{R}$

$$\begin{aligned} (fg)h(s) &= \sum_{xw=s} (fg(x))h(w) \\ &= \sum_{xw=s} \sum_{uv=x} (f(u)g(v))h(w) \\ &= \sum_{u(vw)=s} f(u)(g(v)h(w)) = f(gh)(s). \end{aligned} \tag{30}$$

Therefore the multiplication is associative. We proved the proposition.  $\blacksquare$

The ring  $\mathcal{R}$  is called as the semi-group ring of  $F(\mathcal{A})$  with coefficients in  $R$ . Furthermore, if  $R$  is a left- $R'$  module for some ring  $R'$ . We can define for any  $r \in R'$

$$rf = \sum_{u \in F(\mathcal{A})} rf(u)u \tag{31}$$

then  $\mathcal{R}$  is the left- $R'$  algebra. The  $\mathcal{R}$  may be interpreted as the generalized filter space, while the  $F(\mathcal{A})$  is as the signal space. The multiplication can be thought as the extension of discrete convolution. If we let  $R$  and  $R'$  to be  $\mathbb{R}$ ,  $\mathcal{A} = \{0, 1, 2, \dots\}$ , the multiplication degenerates to classical convolution. The symbol sequence is mapped into a numerical sequence.

Another example of the abstract mapping model is the probability model. Consider all the outcomes of the words in  $F(\mathcal{A})$ . Denote the outcome space as  $\Omega$ .  $\omega$  is a  $\sigma$ -algebra on  $\Omega$  and  $\mathbb{P}$  is a probability measure on  $\sigma$ . Notice that two set-operations on  $\sigma, \cap$

and  $\cup$  is analog of  $\cdot$  and  $+$ .  $\emptyset$  and  $\Omega$  can be seen as 0 and 1 respectively. Therefore,  $R = (\omega, \cup, \cap)$  forms a semi-ring. The probability measure mapping  $P$  here is interpreted as a semi-ring mapping from  $\mathcal{R}$  to the semi-group ring of  $F(\mathcal{A})$  with coefficients in  $\mathbb{R}$ , which is defined as

$$P(f) = \sum_{u \in F(\mathcal{A})} \mathbb{P}(f(u))u. \tag{32}$$

The probability operations then can be realized by algebraic operations on  $\mathcal{R}$  and the corresponding probability measure values are obtained after the mapping  $P$ .

### B. Abstract Mapping Equivalence

For generalized mapping, the equivalence problem is still worth for investigating. However, in the situation, it becomes much more difficult than in a  $\mathbb{R}$ -vector space. The  $R$  in general does not possess any meaningful ordering. Therefore, the definition of equivalence turns out to be limited for specific application. Nevertheless, as we mentioned before, in most cases, it is reasonable to require the result to be similar in certain extent. From now on, we assume  $R$  is a integral domain with unity 1. We introduce the following definition for abstract equivalence of a generalized mapping.

*Definition 4:* For any  $f, g \in \mathcal{R}$ ,  $f$  and  $g$  are abstractly equivalent, if the ideals they generated are the same, i.e.,  $(f) = (g)$ .

The next proposition shows the intuition and legitimacy of this definition.

*Proposition 4:*  $(f) = (g)$  if and only if  $f = ug$ , where  $u$  has multiplicative inverse.

*Proof:* If  $(f) = (g)$ , then  $f = u_1g$  and  $g = u_2f$  for some  $u_1, u_2 \in R$ . We have

$$\begin{aligned} f &= u_1u_2f \\ (1 - u_1u_2)f &= 0. \end{aligned} \tag{33}$$

Since  $R$  is integral domain, we have  $u_1u_2 = 1$ .  $u_1$  and  $u_2$  are units.

Conversely, if  $f = ug$ , then  $g = u^{-1}f$ . We have  $(f) \subseteq (g)$  and  $(g) \subseteq (f)$ , therefore  $(f) = (g)$ .  $\blacksquare$

A loose interpretation of Proposition 4 implies that abstractly equivalent mappings only differ by a ‘‘scale’’ and that the scale change can be ‘‘reversed.’’ Let us first consider the case where the semi-ring  $R$  is  $\mathbb{R}$  or  $\mathbb{C}$ . In this case,  $R$  forms a field and thus any non-zero element is a unit. It is easy to show that in this case strong equivalence implies abstract equivalence. To see that strong equivalence is a special case of abstract equivalence, let us consider mappings  $f$  and  $g$  to be defined at the origin ‘‘0’’ of the field (i.e., we ignore the translation between the mappings). If non-trivial mappings  $f$  and  $g$  are strongly equivalent, then  $f = cg$ , where  $c$  is a non-zero real or complex number. We observe that  $c$  is a unit and therefore its inverse  $c^{-1}$  exists. Finally, we note that the strongly equivalent mappings  $f$  and  $g$  are abstractly equivalent.

We now extend the discussion to semi-ring  $R$  given by  $\mathbb{R}^n$  or  $\mathbb{C}^n$ . We note that  $R$  forms a vector space over  $\mathbb{R}$  or  $\mathbb{C}$ . We recall that the orthogonal linear operator is a necessary and sufficient condition for strong equivalence under the correlation function. Moreover, we note that the set of orthogonal linear

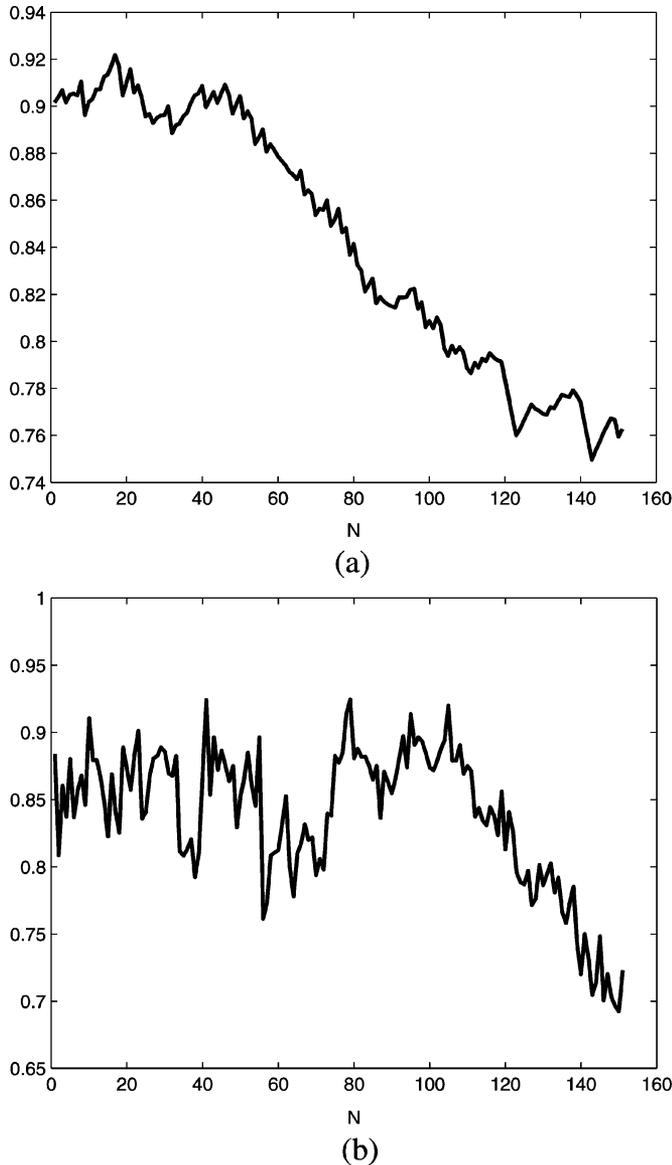


Fig. 2. Two consistency measurements for correlation results using two mapping methods change with the growth of sequence length  $N$  for AD169 DNA sequence. (a) Correlation coefficients for strong equivalence and (b) percentage of points preserving local extremes for weak equivalence.

operators forms an orthogonal group  $O(n)$  given by  $O(n) = \{M \in \mathbb{C}^{n \times n} : M^H M = I\}$ , where  $I$  denotes the identity operator. The orthogonal group contains the special orthogonal group  $SO(n)$  which represents usual rotations and is given by  $SO(n) = \{M \in \mathbb{C}^{n \times n} : M^H M = I \text{ and } \det(M) = 1\}$ . Finally, we observe that if  $f$  and  $g$  are strongly equivalent under the correlation function, then  $f = Mg$ , where  $M \in O(n)$ . Since  $M$  is in the orthogonal group  $O(n)$ , we note that it is a unit (i.e.,  $M^{-1} = M^H \in O(n)$ ). Therefore, we once again conclude that strong equivalence implies abstract equivalence.

#### IV. APPLICATIONS AND EXAMPLES IN GENOMIC SIGNAL PROCESSING

As we discussed in previous sections, approach of mapping the symbolic sequence to  $\mathbb{R}^n$  is a widely adopted method for

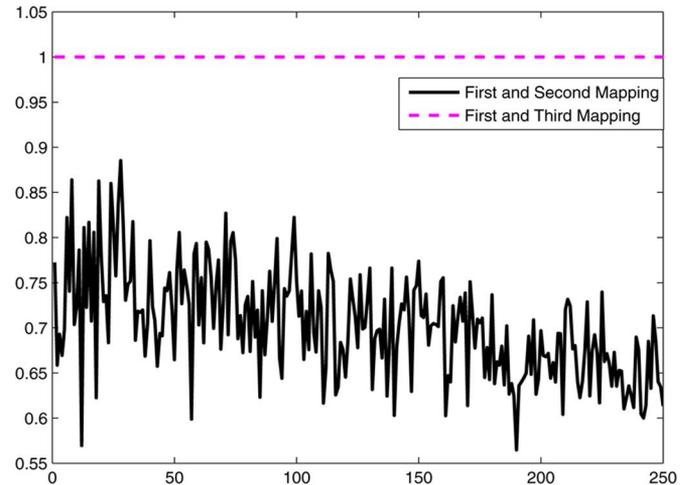


Fig. 3. Percentage of points preserving local extremes for Fourier transform using three different maps changes with growth of sequence length  $N$  for human gene AD169 sequences.

symbolic signal process. Therefore, the consistency problem for results using different mappings always arises. In this section, we will apply our theory to genomic signal processing.

We conduct experiments on Human gene AD169 sequence (GenBank accession no. X17403). We calculate the correlation function as in (6) using two different mappings. The first one maps the  $\mathcal{A} = \{A, T, G, C\}$  to the standard basis of  $\mathbb{R}^4$  correspondingly. 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 [10], [11]. In Fig. 2(a), we show the changing of correlation coefficient between the two correlation results with growth of DNA sequence length  $N$ , and in Fig. 2(b) we show how the percentage of the points having same local extremum property in two results grows with  $N$ . The second mapping is not obtained by rotation of the first mapping. As a result, all these two metrics have a decreasing trend with the grown of length  $N$ .

The example shows the same trends for two metrics between two mappings. The similarity between the two results become less and less, which finally may lead to an inconsistent analysis results due to the fact that two chosen mapping methods are not equivalent for the correlation function. Thus, it does not make sense to make comparison between the analysis result for a given gene sequence under these two mapping methods.

In Fig. 4, we show the consistency measurements between the power spectrum under the previous two mapping methods. Although the equivalent transform we analyzed before does not mainly focus on power spectrum, we can still find that the power spectrum results using these two different mappings have the trend to be inconsistent. Since the correlation and power spectrum are widely used and pervasive in statistic analysis, it suggests the consistency problem should not be neglected when comparing analysis results.

Research on statistical properties of coding and non-coding regions in nucleotide sequences is an important topic in genomic signal processing [22], [23]. We shall also conduct experiments on coding and non-coding regions of Human gene NOC2L using the two mapping methods introduced earlier. As

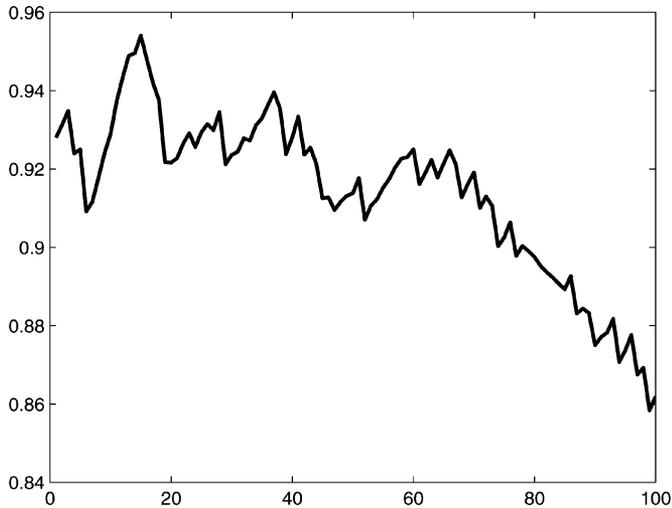


Fig. 4. Correlation coefficient of the power spectrum changes with growth of sequence length  $N$  for Human gene AD169 sequences under the two mapping methods.

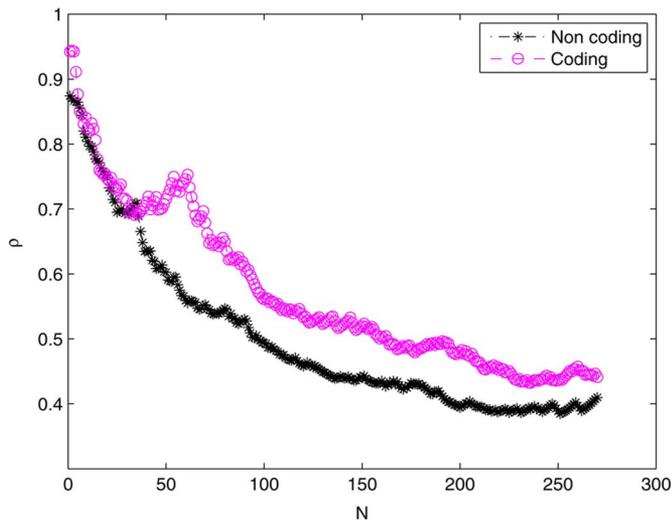


Fig. 5. Correlation coefficients for strong consistency measure of correlation functions changes with growth of  $N$  for coding and non-coding regions of Human gene NOC2L.

shown in Fig. 5, the consistency measure between the correlation functions decays as the length  $N$  increases. As a result of our analysis, we note that the correlation results under the two mappings are inconsistent in the long run. Furthermore, any comparison between the analysis results obtained by relying on these correlation functions becomes increasingly unreliable. Another interesting result can be observed in Fig. 5, where the decay rate of the non-coding region is faster than the coding region. This phenomenon could be attributed to the fact that the coding regions can be viewed as more random than the non-coding regions. Nevertheless, the main conclusion that we draw our attention to is that the consistency of the correlation between non-equivalent mappings decays as the sequence length increases for both coding and non-coding regions.

We calculate the Fourier transform as defined in (11) on Human gene AD169 sequence. The first mapping is chosen as before, which maps the  $\mathcal{A} = \{A, T, G, C\}$  to the standard basis of  $\mathbb{R}^4$ , respectively. Then we use second

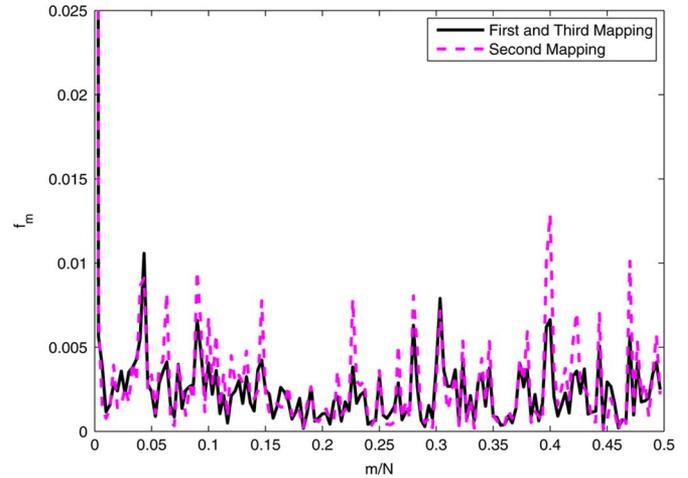


Fig. 6. Fourier transform using three mapping methods.

mapping strategy, which maps A to  $(0.9912, 0.1322, 0, 0)$ , T to  $(0.8367, -0.239, 0.1195, 0.4781)$ , G to  $(-0.7505, -0.5361, -0.2144, 0.3216)$  and C to  $(0.7804, -0.5103, -0.2401, -0.2701)$ . The third strategy maps A to  $((1/\sqrt{2}), 0, (1/\sqrt{2}), 0)$ , T to  $(0, (1/\sqrt{2}), 0, (1/\sqrt{2}))$ , G to  $(-(1/\sqrt{2}), 0, (1/\sqrt{2}), 0)$  and C to  $(0, -(1/\sqrt{2}), 0, (1/\sqrt{2}))$ . We have normalized the mappings so that it will not change the energy of the result. The second mapping is not obtained by rotating the first mapping. While the third mapping is obtained by rotating the first mapping. In Fig. 3, we show the weak equivalence metrics for Fourier analysis results using these three different mappings. The consistency between results using the first and second mapping becomes less and less. While the results using the first and third mapping are completely consistent. In Fig. 6, we show the analysis result of the three mappings. As we showed before, the results using the first and third mapping are exactly same, since rotation is a strongly equivalent transformation. We can find many differences between the results using the first and second mapping, especially at the peaks. We also calculate the correlation coefficient between them, which is 0.82. The peak here means the repeat pattern of some periodic sequences, however, since we have shown that the mapping is not equivalent here, it makes no reason to debate on possible conflicting analysis results for this gene sequence.

In all of the experiments conducted we observe that rotation serves as the unique equivalent transformation for the correlation function. Rotation also provides a strongly equivalent transformation for Fourier and spectrum analysis. Mappings which are not equivalent lead to inconsistent results as the sequence length  $N$  increases. However, we must point out that the opposite may not be true: specifically, for a fixed-length sequence, the consistency between mappings does not necessarily decay as the difference between mappings increases, measured in the sense of rotation equivalence, i.e., the similarity between the first mapping and any mapping obtained by rotating the second mapping.

## V. CONCLUSION

In this paper, we presented a novel framework for analysis of the equivalence of distinct numerical mappings of symbolic

sequences undergoing a transformation by an operator. We introduced a strong equivalence property that demands perfect correlation between the transformations of distinct numerical representations. We also characterized the weak equivalence property which requires the preservation of the extrema in the transformation of the numerical representations. We studied the mapping equivalence theory for general operators by using Taylor's approximation. Moreover, we focused on first- and second-order operators such as the correlation function and Fourier transform. Furthermore, we derived the largest class of equivalent mappings which lead to consistent results when undergoing transformation by a class of operators. We demonstrated that rotation plays an important role in characterization of equivalence between distinct mappings. We subsequently derived a class of operators which is equivalent under rotations. We also introduced an abstract mapping model and extended the notion of equivalence to a more general algebraic structure. We presented simulations of the mathematical and statistical properties of genomic sequences in order to demonstrate the implications of the proposed mapping equivalence theory. Our results suggest that one of the reasons for inconsistency in the analysis of genomic data reported in the theoretical biology literature as well as many other related areas can be attributed to incompatibility of the numerical representation of symbolic data.

#### APPENDIX

*Proof of Theorem 3:* Notice that  $\mathbb{R}^n$  with inner product  $\langle x, y \rangle = x^T y$  is a Hilbert space. So for any linear bounded operator,  $\exists y \in \mathbb{R}^n$ , such that  $\Phi_l(x) = x^T y$ . So  $\Phi_l(T(x)) = T^T x^T y$ .  $T$  is strongly equivalent, therefore  $\Phi_l(T(x)) = \lambda \Phi_l(x) + c$  for some  $\lambda$  and  $c$ . Then we have  $T = \lambda I_{Nn \times Nn}$ , i.e.,  $T$  is a trivial scaled identity transform. This finishes the proof of the first claim. We claim if a non-trivial operator  $\Phi_l(x)$  whose Taylor's expansion has no terms of order higher than or equal to three has a non-trivial linear *strongly equivalent* transformation, then it must only have the second order term and the constant term. We can always scale or add constant for the transformation to get a strongly equivalent result after transformation. So without loss of generality, we assume the result after the transformation is exactly the same as the previous one, i.e., If  $\Phi_l(x) = (1/2)\mathbf{x}^T \nabla^2 \Phi_l(0) \mathbf{x}$  and  $\Phi_l(T(x)) = \Phi_l(x)$ , then we have

$$\frac{1}{2} \mathbf{x}^T \nabla^2 \Phi_l(0) \mathbf{x} = \frac{1}{2} \mathbf{x}^T T^T (\nabla^2 \Phi_l(0)) T \mathbf{x} \quad (34)$$

this equality holds for any  $\mathbf{x} \in \mathbb{R}^{Nn \times 1}$ . Therefore, we have  $T^T (\nabla^2 \Phi_l(0)) T = \nabla^2 \Phi_l(0)$ .  $T$  is a rotation, i.e.,

$$T = \text{diag}(\mathbf{R}_{n \times n}, \dots, \mathbf{R}_{n \times n}). \quad (35)$$

Since  $\mathbf{R}^T \mathbf{R} = I$ , we have

$$(\nabla^2 \Phi_l(0)) T = T \nabla^2 \Phi_l(0). \quad (36)$$

Because  $\mathbf{R}^T \mathbf{R} = \mathbf{R} \mathbf{R}^T = I$ ,  $T^T T = T T^T$ , therefore  $T$  is normal,  $T$  is unitarily diagonalizable [32].  $\mathbf{R}$  is also normal. Therefore,  $\exists V$  unitary, such that  $\mathbf{R} = V^H \Lambda' V$ , where  $\Lambda'$  is a diagonal matrix. Since  $\mathbf{R}$  is real orthogonal, the eigenvalues

of  $\mathbf{R}$  are on the unit sphere  $S^1$ . Without loss of generality, we assume  $\mathbf{R}$  has two eigenvalues, 1 and  $\mu$ . Let the algebraic multiplicity of 1 be  $i$ , then the algebraic multiplicity of  $\mu$  is  $n - i$ . So we have

$$\Lambda' = \text{diag}(I_{i \times i}, \mu I_{(n-i) \times (n-i)}). \quad (37)$$

Let  $U = \text{diag}(V, \dots, V)$ ,  $\Lambda = \text{diag}(\Lambda', \dots, \Lambda')$ . From (36), we have  $(\nabla^2 \Phi_l(0)) U^H \Lambda U = U^H \Lambda U \nabla^2 \Phi_l(0)$ . Therefore, we have

$$U (\nabla^2 \Phi_l(0)) U^H \Lambda = \Lambda U \nabla^2 \Phi_l(0) U^H. \quad (38)$$

Let  $\tilde{X} = U (\nabla^2 \Phi_l(0)) U^H$ . We have  $\tilde{X} \Lambda = \Lambda \tilde{X}$ . By using the Jordan canonical form [33, Ch. VIII], we have that all  $\tilde{X}$  which commutes with  $\Lambda$  must have the form as follows:

$$\tilde{X} = \begin{pmatrix} A_{11} & 0 & A_{12} & 0 & \cdots & A_{1M} & 0 \\ 0 & B_{21} & 0 & B_{22} & \cdots & 0 & B_{2M} \\ A'_{12} & 0 & A_{31} & 0 & \cdots & A_{3M} & 0 \\ 0 & B'_{22} & 0 & B_{41} & \cdots & 0 & B_{4M} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ A'_{1M} & 0 & A'_{3M} & 0 & \vdots & A_{(2N-1)1} & 0 \\ 0 & B'_{2M} & 0 & B'_{4M} & \cdots & 0 & B_{(2N)1} \end{pmatrix}. \quad (39)$$

Every non-zero submatrix in  $\tilde{X}$  is an arbitrary upper triangular submatrix which has identical diagonal entries. All submatrices  $A_{kl}$  and  $A'_{kl}$  have size  $i \times i$ , all  $B_{kl}$  and  $B'_{kl}$  have size  $(n-i) \times (n-i)$ . Thus, all  $\nabla^2 \Phi_l(0)$  satisfies (36) are of the form  $\nabla^2 \Phi_l(0) = U^H \tilde{X} U$ . However, since the  $\Phi_l(x)$  is analytic, the Hessian must be symmetric. Therefore, all the submatrices of  $\tilde{X}$ ,  $A_{kl}$  and  $B_{kl}$  have the form  $k_{kl} I$  and  $A'_{kl}$  and  $B'_{kl}$  have also the form  $k_{kl} I$ . Notice that this is for a given rotation. Since (36) holds for any given rotation, we have that

$$\nabla^2 \Phi_l(0) \in Y = \bigcap_{U \text{ Unitary}} \{U^H \tilde{X} U\}. \quad (40)$$

The rotation is arbitrary, (37) should hold for any  $i = 0, \dots, n$ . Claim that the  $N \times N$  principal matrice in  $\tilde{X}$  must satisfy

$$\text{diag}(A_{(i)j}, B_{(i+1)j}) = k_{ij} I_{n \times n} \quad (41)$$

$$\text{diag}(A'_{(i)j}, B'_{(i+1)j}) = k'_{ij} I_{n \times n} \quad (42)$$

where  $i = 1, 3, 5, \dots, (2N - 1)$  and  $j = 1, \dots, M$ . Because we know that

$$\text{diag}(A_{(i)j}, B_{(i+1)j}) = \text{diag}(k_{ij} I_{i \times i}, k_{(i+1)j} I_{(n-i) \times (n-i)}) \quad (43)$$

but if  $k_{ij} \neq k_{(i+1)j}$ , then this implies  $\{U^H \tilde{X} U\} \cap \{U^H \tilde{X}' U\} = \emptyset$ , where (37) for  $\tilde{X}'$  is of the form

$$\Lambda' = \text{diag}(I_{(i+1) \times (i+1)}, \mu I_{(n-i-1) \times (n-i-1)}). \quad (44)$$

Since we know  $Y$  is not empty, we get a contradiction here. Therefore, (41) and (42) hold. If we choose  $U = I$ , we have  $Y \subset \tilde{X}$ , where for  $\tilde{X}$ , (41) and (42) hold. It's straightforward

to check that such  $\tilde{X}$  is commutable with  $T$ . Therefore,  $Y = \tilde{X}$ . Finally we show that  $\nabla^2 \Phi_l(0)$  must have the form as in (10). If we expand the operator at any other point  $x_0$ , the only second order term is  $(1/2)x^T \nabla^2 \Phi_l(x_0)x$ . Repeat the previous argument, we will have (10). ■

#### ACKNOWLEDGMENT

The authors would like to thank Prof. A. Willsky from the Massachusetts Institute of Technology for discussions that served as the impetus for the work presented in this paper.

#### REFERENCES

- [1] P. Vaidyanathan, "Genomics and proteomics: A signal processor's tour," *IEEE Circuits Syst. Mag.*, vol. 4, no. 4, pp. 6–29, 2004.
- [2] A. Agresti, *Categorical Data Analysis*. New York: Wiley, 1990.
- [3] A. Krogh, I. Mian, and D. Haussler, "A hidden Markov model that finds genes in E. coli DNA," *Nucleic Acids Res.*, vol. 22, no. 22, pp. 4768–4778, 1994.
- [4] S. Salzberg, A. Delcher, S. Kasif, and O. White, "Microbial gene identification using interpolated Markov models," *Nucleic Acids Res.*, vol. 26, no. 2, pp. 544–548, 1998.
- [5] S. Salzberg, D. Searls, and S. Kasif, *Computational Methods in Molecular Biology*. Amsterdam, The Netherlands: Elsevier Science & Technology, 1998.
- [6] A. Oppenheim and S. Nawab, *Symbolic and Knowledge-Based Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
- [7] H. Herzel and I. Große, "Measuring correlations in symbol sequences," *Physica A*, vol. 216, no. 4, pp. 518–542, 1995.
- [8] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, "Long-range correlation properties of coding and noncoding DNA sequences: Genbank analysis," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 51, pp. 5084–5091, 1995.
- [9] W. Li and K. Kaneko, "Long-range correlation and partial 1/f spectrum in a noncoding DNA sequence," *Europhys. Lett.*, vol. 17, p. 655, Feb. 1992.
- [10] 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.
- [11] C. Berthelsen, J. Glazier, and M. Skolnick, "Global fractal dimension of human DNA sequences treated as pseudorandom walks," *Phys. Rev. A, Gen. Phys.*, vol. 45, no. 12, pp. 8902–8913, 1992.
- [12] B. D. Silverman and R. Linsker, "A measure of DNA periodicity," *J. Theoretic. Biol.*, vol. 118, pp. 295–300, 1986.
- [13] P. Cristea, "Conversion of nucleotides sequences into genomic signals," *J. Cellular Molecul. Med.*, vol. 6, no. 2, pp. 279–303, 2002.
- [14] D. Anastassiou, "Genomic signal processing," *IEEE Signal Process. Mag.*, vol. 18, no. 4, pp. 8–20, Jul. 2001.
- [15] 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.
- [16] W. Wang and D. H. Johnson, "Computing linear transforms of symbolic signals," *IEEE Trans. Signal Process.*, vol. 50, no. 3, pp. 628–635, Mar. 2002.
- [17] A. Rushdi and J. Tuqan, "Gene identification using the Z-curve representation," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 2006, vol. 2, pp. II–II.
- [18] A. Rushdi and J. Tuqan, "The role of the symbolic-to-numerical mapping in the detection of DNA periodicities," in *Proc. IEEE Int. Workshop Genomic Signal Processing Statistics*, 2008, pp. 1–4.
- [19] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, "Long-range correlations in nucleotide sequences," *Nature*, vol. 356, no. 6365, pp. 168–170, Mar. 1992.
- [20] A. Arneodo, E. Bacry, P. V. Graves, and J. F. Muzy, "Characterizing long-range correlations in DNA sequences from wavelet analysis," *Phys. Rev. Lett.*, vol. 74, no. 16, pp. 3293–3296, Apr. 1995.

- [21] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg, and J. L. Oliver, "Identifying characteristic scales in the human genome," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 75, p. 032903, 2007.
- [22] N. Bouaynaya and D. Schonfeld, "Non-stationary analysis of DNA sequences," in *Proc. IEEE/SP 14th Workshop Statistical Signal Processing*, 2007, pp. 200–204.
- [23] N. Bouaynaya and D. Schonfeld, "Nonstationary analysis of coding and noncoding regions in nucleotide sequences," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 3, pp. 357–364, Jun. 2008.
- [24] L. J. Corwin and R. Szczaiba, *Calculus in Vector Spaces*, 2nd ed. New York: Macel Dekker, 1994.
- [25] L. D. Lathauwer, "Signal processing based on multilinear algebra," Ph.D. dissertation, Electr. Eng. Dept.—ESAT, K.U. Leuven, Leuven, Belgium, 1997.
- [26] R. Larsen and M. Marx, *An Introduction to Mathematical Statistics and its Applications*. Englewood Cliffs, NJ: Prentice-Hall, 2005.
- [27] Y. Eidelman, V. Milman, and A. Tsolomitis, *Functional Analysis*. Providence, RI: Amer. Math. Soc., 2004.
- [28] G. Caire, R. Grossman, and H. Poor, "Wavelet transforms associated with finite cyclic groups," *IEEE Trans. Inf. Theory*, vol. 39, no. 4, pp. 1157–1166, Jul. 1993.
- [29] M. Swanson and A. Tewfik, "A binary wavelet decomposition of binary images," *IEEE Trans. Image Process.*, vol. 5, no. 12, pp. 1637–1650, Dec. 1996.
- [30] M. Poeschel and J. Moura, "Algebraic signal processing theory," *Arxiv Preprint Cs/0612077*, 2006.
- [31] D. Dummit and R. Foote, *Abstract Algebra*. New York: Wiley, 2004.
- [32] R. A. Horn and C. R. Johnson, *Matrix Analysis*. New York: Cambridge Univ. Press, 1988.
- [33] F. R. Gantmacher, *The Theory of Matrices*. New York: Chelsea, 1959, vol. I and II, Reprinted by Amer. Math. Soc..



**Liming Wang** (S'08) was born in Hefei, China, in July 1984. He received the B.S. degree in electronic engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2006. He is currently working towards the M.S. degree in mathematics and the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Illinois at Chicago.

His research interests include communication and signal processing, information theory, statistical modeling and genomic signal processing.



**Dan Schonfeld** (M'90–SM'05) received the B.S. degree in electrical engineering and computer science from the University of California at Berkeley, and the M.S. and Ph.D. degrees in electrical and computer engineering from The Johns Hopkins University, Baltimore, MD, in 1986, 1988, and 1990, respectively.

In 1990, he joined the University of Illinois at Chicago, where he is currently a Professor in the Department of Electrical and Computer Engineering. He has authored over 170 technical papers in various journals and conferences. His current research

interests are in multi-dimensional signal processing; image and video analysis; computer vision; and genomic signal processing.

Dr. Schonfeld is currently also serving as Associate Editor on Image and Video Storage, Retrieval and Analysis of the IEEE TRANSACTIONS ON IMAGE PROCESSING and Associate Editor on Video Analysis of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He has served as an Associate Editor on Multidimensional Signal Processing and Multimedia Signal Processing of the IEEE TRANSACTIONS ON SIGNAL PROCESSING as well as an Associate Editor on Nonlinear Filtering of the IEEE TRANSACTIONS ON IMAGE PROCESSING.