A Striking Property of Genetic Code-like Transformations

Hillol Kargupta*
Department of Computer Science and Electrical Engineering,
University of Maryland Baltimore County,
Baltimore, MD 21250, USA

The gene expression process in nature plays a key role in evaluating the fitness of DNA through the production of different proteins in different cells. The production of proteins from DNA goes through different stages. Among others, the transcription stage produces the mRNA from the DNA and translation produces the amino acid sequence in proteins from the mRNA. The translation process is accomplished by mapping the mRNA sequence using a transformation called the *genetic code*. This code considers every consequent triplet (codon) of nucleic acids in the mRNA sequence and maps it to a corresponding amino acid. This paper shows that genetic code-like transformations (GCTs) introduce very interesting properties to the representation of a genetic fitness function. It presents a Fourier[1] analysis of GCTs. It points out that such transformations can convert some function representations of exponential description in Fourier basis to a description that is highly suitable for polynomial-complexity approximation. More precisely, such transformations can construct a Fourier representation with only a polynomial number of terms that are exponentially more significant than the rest. Polynomial-complexity approximation of functions from data is a fundamental problem in inductive learning, data mining, search, and optimization. Therefore the work has important implications in these areas. It is unlikely that such representations can be constructed for all functions. However, since such transformations appear to work well in nature, the class of such functions may not be trivial and should be explored further.

1. Introduction

Learning functions from data is important in many fields such as inductive learning, statistics, and data mining. It may also be important for nonenumerative optimization in the absence of sufficient domain knowledge; this is because such optimization may require inductively

*Electronic mail address: hillol@cs.umbc.edu.

[1]The analysis is identical to that using Walsh basis [4, 45]; however, the term Fourier is chosen because of its historical [18, 32] use in function approximation literature.

detecting the structure of the objective function for intelligent guessing about the desired solution.

Representation plays an important role in learning functions. For example, if the function has an exponentially large (in the number of variables defining the function) description in the chosen representation then its polynomial-time computation is not possible. On the other hand, a representation with polynomially bounded size is amenable to efficient computation. A function with an exponentially large description may be efficiently computed when it can be approximated using a function that has a polynomially bounded description size. This may be possible when the target function has an exponentially large representation with only a polynomial number of "significant" components. In that case, we may be able to neglect the "insignificant" components and still enjoy a high degree of accuracy. Therefore, constructing function representations with a "small" number of significant components is important for efficient function induction.

This paper considers Fourier basis representation of some well-known functions and shows that there exists a class of transformations that offers this property in the Fourier space under some practical conditions. The transformations are similar to the genetic code that transforms the genetic fitness function defined over the protein sequences to the mRNA representation in a living organism.

A living body starts its life from the DNA, the primary information carrier in genetics. Almost every critical activity of the organism is accomplished by proteins constructed from the DNA. The efficacy of the organism; that is, the genetic fitness, depends on the proteins. For some reason our body chooses different representations of the information stored in the proteins. It uses the mRNA and the DNA sequences to represent the proteins. It first transforms the DNA to the mRNA representation and subsequently to the protein before evaluating the fitness of the genome. This process of representation transformations is called *gene expression*. Representation transformations are often used in many fields like physics, engineering, machine learning, and mathematics for transforming difficult problems into suitable forms that are easier to solve. Therefore representation transformations in gene expression allude intriguing possibilities.

This paper investigates the possible role of the gene expression in making genetic search efficient. It considers one important part of gene expression, the translation, that transforms the mRNA sequence to protein. Translation is governed by the *genetic code*. This paper presents a Fourier analysis of genetic code-like transformations (GCTs) in the binary sequence space and demonstrates a quite interesting property of such representation transformations. It points out that there exist some GCTs that can convert some functions with an exponentially long description in Fourier basis to a representation where only a polynomial

number of terms are exponentially more significant than the rest when fitter proteins are given more copies through redundant and equivalent representation.

Section 2 describes the gene expression process in nature. Section 3 briefly reviews previous work on the computation in gene expression. Section 4 reviews the basics of Fourier representation. Section 5 analyzes the effect of GCTs on the representation of the genetic fitness function and proves the main results of this paper. Finally, section 6 concludes this paper.

2. Gene expression and the genetic code

The DNA is the primary carrier of the genetic information that is transmitted from one generation to another. DNA molecules consist of two long complementary chains held together by base pairs. DNA consists of four kinds of bases joined to a sugar-phosphate backbone. The four bases in DNA are *adenine* (A), *guanine* (G), *thymine* (T), and *cytosine* (C). Chromosomes are made of DNA *double helices*. Bases in DNA helices obey the *complementary base pairing rule*. T and G pair with A and C respectively. In other words, if the base at a particular position of a helix is T then the corresponding base in the other helix should be A. The information coded in the DNA is extracted during the process of gene expression.

Expression of genetic information coded in DNA requires construction of the mRNA sequence, followed by that of proteins. The main steps follow.

- *Transcription*. Formation of mRNA (messenger ribonucleic acid) from DNA.
- *Translation*. Formation of protein from mRNA.
- Protein folding.

In a particular cell, transcription produces the mRNA from a small portion of the DNA. The mRNA defines another level of representation of the genetic information. It consists of four types of bases joined to a ribose-sugar-phosphodiester backbone. The four bases are *adenine* (A), *uracil* (U), *guanine* (G), and *cytosine* (C). All the bases defining the mRNA are the same as those in DNA sequences, except that T is replaced by U. The mRNA is produced from the DNA by RNA Polymerase and the regulatory proteins following the complementary base-pairing rules similar to those in DNA. The RNA Polymerase initiates the transcription at a place of the DNA marked by the *promoter* region (*start site*). It splits the DNA double helix and continues generating the mRNA using one of the DNA strands as a template. The RNA Polymerase stops when it finds a termination signal sequence (*stop site*) in the DNA strand.

Protein feature	mRNA codons
Alanine	GCA GCC GCG GCU
Cysteine	UGC UGU
Aspartic acid	GAC GAU
Glutamic acid	GAA GAG
Phenylalanine	UUC UUU
Glycine	GGA GGC GGG GGU
Histidine	CAC CAU
Isoleucine	AUA AUC AUU
Lysine	AAA AAG
Leucine	UUA UUG CUA CUC CUG CUU
Methionine	AUG
Asparagine	AAC AAU
Proline	CCA CCC CCG CCU
Glutamine	CAA CAG
Arginine	AGA AGG CGA CGC CGG CGU
Serine	AGC AGU UCA UCC UCG UCU
Threonine	ACA ACC ACG ACU
Valine	GUA GUC GUG GUU
Tryptophan	UGG
Tyrosine	UAC UAU
STOP	UAA UAG UGA

Table 1. The universal genetic code.

Note that only a small portion of the DNA strand is transcribed and different cells may transcribe different regions of the DNA for producing proteins.

The mRNA acts as the template for protein synthesis. A protein is defined by a sequence of *amino acids*, joined by peptide bonds. The mRNA is transported to the cell cytoplasm for producing protein in the ribosome. There exists a set of rules that defines the correspondence between nucleotide triplets (known as *codons*) and the amino acids in proteins. This is known as the *genetic code*. Each codon is comprised of three adjacent nucleotides in a DNA chain and it produces a unique amino acid. With a few exceptions the genetic code for most eukaryotic and prokaryotic organisms is the same. An amino acid sequence defines a new representation of the information coded in mRNA.

The final level of representation of genetic information is defined by the three-dimensional structure of folded proteins. Although amino acid sequences fundamentally define proteins, formation of the three-dimensional structure of proteins involves a complex process, often called *protein folding*. This process involves interaction between multiple amino acid subsequences, resulting in the emergence of a folded structure from the sequence.

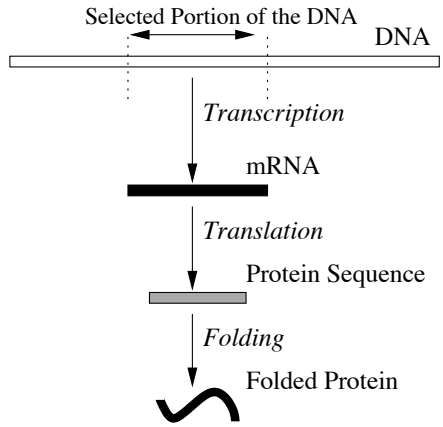

Figure 1. Different steps of gene expression.

Figure 1 shows the different steps of the gene expression process. Although proteins play a key role in determining genetic fitness, the purpose of the two additional layers of representation (mRNA and DNA) for representing the fitness function is not clear. Representation transformations are often used in physics, engineering, and machine learning for solving problems efficiently. Therefore, the role of gene expression in efficient genetic search is really intriguing. This paper investigates gene expression from this perspective. First, let us review the existing related literature.

3. Previous work

The importance of gene expression in genetic search was realized in the early days of the field of genetic algorithms (GAs). Holland [15] described the dominance operator as a possible way to model the effect of gene expression in diploid chromosomes. He also noted the importance of the process of protein synthesis from DNA in the computational model of genetics. Despite the fact that dominance maps are traditionally explained from the mendelian perspective, Holland made an interesting leap by connecting it to the synthesis of protein by gene signals, which today is universally recognized as gene expression. He realized the relation between the dominance operator with the "operon" model of the functioning of the chromosome in evolution [19] and pointed out the possible computational role of gene signaling in evolution [15].

Several other efforts have been made to model some aspects of gene expression. Diploidy and dominance have also been used elsewhere [1, 7, 16, 38, 40]. Most of these take their inspiration from the mendelian view of genetics. The under-specification and over-

specification decoding operator of messy GAs has been viewed as a mechanism similar to gene signaling [13]. The structured GA [8] also shares motivations from gene expression: it uses a structured hierarchical representation in which genes are collectively switched on and off. This provides the search algorithm with a richer representation and helps capture properties of the landscape better. An empirical study of genetic programming using artificial genetic code is presented in [31]. In [30] Kauffman offered an interesting perspective of the natural evolution that realizes the importance for gene expression. However, Kauffman's work does not explain the process in basic computational terms on analytical grounds and does not relate the issue to the complexity of search. The complex nature of the representation in the DNA itself created interest among the researchers. The eukaryotic DNA typically contains many segments that are not used in the gene expression process for producing proteins. An empirical investigation of the role of such "noncoding" segments (introns) in genetic search can be found in [46]. A survey of evolutionary algorithms with intron-based representations is presented in [47].

The neutral network theory [36, 39] also considers sequence-to-structure mapping from the perspective of random graph construction. This work approaches gene expression from the perspective of random graph construction and points out the existence of fitness invariant neutral networks. The translation process maps multiple mRNA sequences to the same protein sequence. As a result, it creates a genetic space that contains multiple genomes with the same fitness, known as *neutral networks*. This work provides interesting insights into the effect of such neutral networks in genetic search. However, its contribution towards polynomial-time representation construction of genetic fitness functions is not clear.

Another related effort to understand the properties of the fitness landscape defined by the mRNA can be found in [37]. This work presents a Fourier analysis of the landscapes derived from the RNAs using fast Fourier transformation (FFT). Although the time complexity of the FFT is better than the regular Fourier transformation, it still grows exponentially with respect to the number of feature variables defining the domain of the fitness function. This paper suggests that the genetic code that transforms the RNA to protein itself may help in designing a polynomial-time algorithm for the construction of the Fourier representation which the FFT cannot offer.

There also exists a body of literature that investigates evolution of the genetic code. An algebraic model for the evolution of the genetic code is presented in [17]. This work searches for symmetries in the genetic code and points out the existence of a unique approximate symmetry group compatible with the codon assignments. The main idea behind this work is to view the evolution of the genetic code as an iterative

process of representation decomposition. The genetic code is viewed as a 64-dimensional representation decomposed into several subrepresentations with respect to different subgroups. The number of amino acids correspond to the number of subrepresentations and the number of codons for any amino acid corresponds to the dimension of that subrepresentation. An extension of this work using Lie superalgebra is presented in [3]. Additional work on the different biological theories on the evolution of the genetic code can be found elsewhere [5, 10].

An alternate approach has been developed by Kargupta and his colleagues [2, 20–26, 28, 29]. This approach is mainly motivated by a perspective of the gene expression as a mechanism to make genetic search more efficient. This approach notes that the traditional model of evolutionary computation (based on selection, crossover, and mutation) [15] appears to have some serious scalability problems [42] for reasonably difficult problems. There are also few theoretical results available that prove guaranteed polynomial-time performance of existing evolutionary algorithms for reasonably difficult classes of problems. Since the existing models of evolutionary computation do not address the gene expression issue very well and gene expression changes the genetic representation, it may become a natural candidate for exploring the unknown mechanism that makes the genetic search in nature so efficient and scalable.

The early exploration of gene expression-like mechanisms for efficient inductive detection of function structure resulted in a class of heuristics-based techniques, known as the *gene expression messy GA* (GEMGA) [22]. In the recent past, more rigorous approaches using Fourier basis representations are suggested. Fourier representations expose the underlying function structure and are functionally complete. Therefore, if we can learn such representations quickly, the purpose of function induction is served. A randomized algorithm is presented in [27, 28] that can induce a representation in Fourier basis in polynomial time for problems with bounded variable interaction (BVI). The assumption of BVI makes sure that among ℓ features defining the search domain, only at most some k (a constant) number of variables can interact with each other. In other words, the overall fitness function can be decomposed into a collection of either overlapping or nonoverlapping subfunctions where each of the subfunctions can depend on at most k variables. This condition guarantees a polynomial-size description of the target function in Fourier representation. An alternate technique for estimating the Fourier representations is proposed elsewhere [18, 32]. An extension of this technique for detecting function structure in GAs is reported in [41].

Although many functions exist with a polynomial-size canonical representation, it is not clear why the natural genetic fitness function should have such a property. This paper suggests a possible direction to answer

this question. It shows that GCTs can construct a Fourier representation of at least some fitness functions where the contribution of Fourier coefficients involving some q features decreases exponentially with q. This may allow us to approximate the genetic fitness function with a Fourier representation that neglects the effect of Fourier coefficients associated with some k or higher features. In other words, the approximation will satisfy the BVI property. If that is the case, then we can induce such functions efficiently in polynomial time. The following section reviews the fundamentals of Fourier representation.

4. Fourier representation and function induction

The role of the genetic code in the evaluation of fitness can be understood in the context of an appropriately chosen set of basis functions. This paper uses the Fourier basis functions to do that. The representation is very similar to the Walsh basis [4, 45], frequently used by the GA community. This paper uses the Fourier representation because of its history in function induction literature [18, 32]. The following section presents a brief review of the Fourier basis and its relation with the problem of inducing functions from data.

4.1 A brief review of the Fourier basis

Fourier bases are orthogonal functions that can be used to represent any function. In this paper we consider functions of binary variables. Consider the function space over the set of all ℓ-bit strings. The Fourier basis set that spans this space is comprised of 2^ℓ functions. Each Fourier basis function is defined as follows:

$$\psi_{\mathbf{j}}(\mathbf{x}) = (-1)^{(\mathbf{x}\cdot\mathbf{j})}. \tag{1}$$

Where $\mathbf{j}$ and $\mathbf{x}$ are binary strings of length ℓ. In other words, $\mathbf{j} = j_1, j_2, \ldots j_\ell$, $\mathbf{x} = x_1, x_2, \ldots x_\ell$, and $\mathbf{j}, \mathbf{x} \in \{0,1\}^\ell$; $\mathbf{x}\cdot\mathbf{j}$ denotes the inner product of $\mathbf{x}$ and $\mathbf{j}$ which is nothing but $\sum_{i=1}^{\ell} x_i j_i$. $\psi_{\mathbf{j}}(\mathbf{x})$ can either be equal to 1 or -1. The string $\mathbf{j}$ is called a *partition*. The *order* of a partition $\mathbf{j}$ is the number of 1s in $\mathbf{j}$. A Fourier basis function depends on some x_i only when $j_i = 1$. Therefore a partition can also be viewed as a representation of a certain subset of x_is: every unique partition corresponds to a unique subset of x_is. If a partition $\mathbf{j}$ has exactly α number of 1s then we say the partition is of order α since the corresponding Fourier function depends on only those α number of variables corresponding to the 1s in the partition $\mathbf{j}$. Fourier bases are orthonormal. Therefore,

$$\frac{1}{2^\ell}\sum_{\mathbf{x}} \psi_{\mathbf{i}}(\mathbf{x})\psi_{\mathbf{j}}(\mathbf{x}) = 1 \quad \text{when} \quad \mathbf{i} = \mathbf{j}$$
$$= 0 \quad \text{when} \quad \mathbf{i} \neq \mathbf{j}.$$

A function $f : \mathbf{X}^\ell \to \Re$, that maps an ℓ-dimensional space of binary strings to a real-valued range, can be represented using the Fourier basis functions:

$$f(\mathbf{x}) = \sum_{\mathbf{j}} w_{\mathbf{j}} \psi_{\mathbf{j}}(\mathbf{x}) \tag{2}$$

where $w_{\mathbf{j}}$ is the Fourier coefficient (FC) corresponding to the partition $\mathbf{j}$

$$w_{\mathbf{j}} = \frac{1}{2^\ell} \sum_{\mathbf{x}} f(\mathbf{x}) \psi_{\mathbf{j}}(\mathbf{x}). \tag{3}$$

We note from equation (2) that a function can be expressed as a linear sum of the Fourier functions, each weighed by the corresponding FC. The FC $w_{\mathbf{j}}$ can be viewed as the relative contribution of the partition $\mathbf{j}$ to the function value of $f(\mathbf{x})$. Therefore, the absolute value of $w_{\mathbf{j}}$ can be used as the "significance" of the corresponding partition $\mathbf{j}$. If the magnitude of some $w_{\mathbf{j}}$ is very small compared to other coefficients then we may consider the $\mathbf{j}$th partition to be insignificant and neglect its contribution.

Fourier bases and their close relatives Walsh bases are frequently used to study the behavior of GAs. Walsh bases [4] were first used by Bethke [6] for analyzing GAs. Further investigation of this approach can be found elsewhere [9, 11, 12, 14, 33–35, 43, 44].

4.2 Function induction from data and Fourier basis

Function induction from data plays an important role in adaptation, machine learning, and nonenumerative black-box optimization. In function induction, the goal is to learn a function $\hat{f} : X^\ell \to Y$ from the data set $\Omega = \{(\mathbf{x}_{(1)}, \mathbf{y}_{(1)}), (\mathbf{x}_{(2)}, \mathbf{y}_{(2)}), \ldots (\mathbf{x}_{(k)}, \mathbf{y}_{(k)})\}$ generated by some underlying target function $f : X^n \to Y$, such that the $\hat{f}$ approximates f. Since Fourier basis is functionally complete, any function can be represented in Fourier basis. Therefore, learning a function $\hat{f}$ can be posed as the problem of approximating the Fourier representation of f. If we can accurately estimate the significant coefficients (coefficients with relatively large magnitude) of the Fourier representation of f then we can use those coefficients to define $\hat{f}$. The complexity of inducing a function in Fourier representation is directly proportional to the number of such coefficients.

Note that the Fourier representation in the binary domain can potentially have 2^ℓ coefficients and estimating all of them will require exponential time. However, we may be able to get away with polynomial-time computation if there is a way to accurately approximate the function with an exponentially long description by a function with only a polynomially long description. For example, if the function has only a

polynomial number of significant FCs then we may be able to construct an approximation by considering only those significant coefficients and neglecting the rest. In that case we can write $\hat{f} = \sum_{\mathbf{j}} w''_{\mathbf{j}} \psi_{\mathbf{j}}(\mathbf{x})$, where $w''_{\mathbf{j}} = w_{\mathbf{j}}$ when $|w_{\mathbf{j}}| \geq \theta$ and $w''_{\mathbf{j}} = 0$ otherwise. $|w_{\mathbf{j}}|$ denotes the magnitude of $w_{\mathbf{j}}$ and θ represents the chosen threshold. If the number of FCs in $\hat{f}$ is polynomially bound then its Fourier representation can be computed in polynomial time [18, 27, 28, 32, 41].

Unfortunately, function representations may not always come with such a nice property. The following section points out that there exists some GCTs that can construct representations of functions with this very desirable property.

5. Exploring genetic code-like transformations

The genetic code transforms the mRNA sequence to the protein sequence by assigning one protein feature for every codon in the mRNA sequence. Although the cardinalities of the alphabet sets of the mRNAs and proteins are more than two, understanding the underlying computation may require abstraction. In this section we will do so by assuming that the protein and the mRNA sequences are binary strings. Our objective is to explore the effect of the genetic code-like representation transformations in the binary domain using Fourier analysis. In order to do that we first need to define what is meant by "genetic code-like transformations."

5.1 The notion of genetic code-like transformations

The genetic code defines the correspondence between an mRNA codon and a protein feature value. Although in nature the codons are defined by three mRNA feature values, the implication of the choice of number "three" is yet to be explained. Therefore, the current analysis will treat this as a parameter and the results of this paper can be specialized for any size of codons, including three. As noted earlier, the analysis considers the effect of such transformations in the binary space. Although strings are binary, we will continue to use the terms mRNA, protein, and genetic code accordingly for maintaining the link between biology and the current analysis.

Let us use $\mathbf{r}$ and $\mathbf{p}$ to represent the mRNA and the protein sequences respectively. Let ℓ_r and ℓ_p be their respective lengths. Just like the natural translation process, our artificial translation maps the mRNA sequence to the corresponding protein sequence using the genetic code. The mapping in translation will be denoted by η_c where the subscript c denotes the number of mRNA features that define a codon. If three features are used like natural codons, $c = 3$; η_c can be defined as η_c :

Protein feature	mRNA codon
1	100
1	000
1	001
1	010
0	111
0	101
0	110
0	011

Table 2. Code A: A GCT for binary representation. A single bit in the protein space maps to 3-bit codons in the mRNA space.

Protein feature	mRNA codon
0	100
0	000
0	001
0	010
0	111
0	101
0	110
1	011

Table 3. Code B: Another GCT for binary representation. Note that seven unique mRNA codons map to the protein feature value of zero.

$R^{\ell_r} \to P^{\ell_p}$. R^{ℓ_r} and P^{ℓ_p} denote the ℓ_r- and ℓ_p-dimensional space of all mRNAs and proteins respectively. Note that $\ell_r = c\,\ell_p$ and for binary representation $R = P = \{0, 1\}$.

Consider the GCTs presented in Tables 2 and 3. Note that the genetic code may be redundant. In other words, a unique protein feature value may be produced by several mRNA codons. This is also true for natural genetic code (Table 1). As a result, there exist many equivalent mRNA sequences that produce the same protein sequence. All these mRNA sequences have the same genetic fitness since they all map to the same protein sequence. So we can view the space of mRNAs grouped into different equivalence classes. We shall call this characteristic *translation introduced equivalence* (TIE) and these groups of equivalent mRNAs will be called the *TIE classes*. Let R_p be the TIE class for the protein sequence $\mathbf{p}$. We can also define R_p in the following manner: $R_p = \{\mathbf{r}_j | \mathbf{r}_j \overset{\eta_c}{\to} \mathbf{p}\}$. The cardinality of the set R_p depends on the genetic code and the protein sequence $\mathbf{p}$. Let a_0 and a_1 be the total number of codons

that map to a protein feature value of 0 and 1 respectively. Let $\ell_{p,0}$ and $\ell_{p,1}$ be the number of 0s and 1s in $\mathbf{p}$ respectively. Then the cardinality of the TIE class is $|R_p| = a_0^{\ell_{p,0}} a_1^{\ell_{p,1}}$.

Since one feature in the protein sequence maps to c mRNA features, partitions defined in the mRNA and the protein spaces can be associated with each other. Let $\mathbf{j}$ and $\mathbf{j}'$ be partitions in the mRNA and the protein spaces respectively. We will call $\mathbf{j}'$, the *reflection* of $\mathbf{j}$ in the protein space when $\mathbf{j}'_i = 1$ if and only if $\mathbf{j}$ takes a value of 1 at the location(s) corresponding to at least one of the mRNA features associated with $\mathbf{j}'_i$. If $\mathbf{j}$ has 0s at all the locations corresponding to $\mathbf{j}'_i$ then $\mathbf{j}'_i = 0$.

For example, the reflection of the partition $\mathbf{j} = 101000$ using a genetic code of codon size three is $\mathbf{j}' = 10$. The left three bits of $\mathbf{j}$ are associated with the leftmost bit of $\mathbf{j}'$. Since two of those three bits are set to 1, $\mathbf{j}'_0 = 1$. However, none of the rightmost three bits in $\mathbf{j}$ takes the value 1, so the corresponding $\mathbf{j}'_1 = 0$. Note that the reflection of 100000 is also 10 since $\mathbf{j}'_0 = 1$ as long as at least one of the leftmost three bits is set to 1. Similarly the reflection of 100110 under a genetic code of codon size three is 11.

Note that different mRNA partitions may have the same reflection in the protein space. If q is the number of ones in $\mathbf{j}'$ then it is the reflection of $(2^c - 1)^q$ different partitions in the mRNA space. The number of 1s in $\mathbf{j}'$ will be called the *absolute order of partition* $\mathbf{j}$.

Once the protein sequence is constructed from the mRNA sequence, the protein folds into a three-dimensional structure and its shape determines its fitness. Let us use $f : P^{\ell_p} \rightarrow \Re^+$ for denoting this fitness function that maps the protein sequence to a nonnegative real-valued range. Since the protein sequences are produced from the mRNA sequences, we can also define the fitness over the domain of mRNA sequences. Let $\phi : R^{\ell_r} \rightarrow \Re^+$ be this fitness function defined over the mRNA representation. Therefore, $\phi(\mathbf{r}) = f(\mathbf{p}) = f(\eta_c(\mathbf{r}))$ and $\phi(\mathbf{r})$ can be viewed as a different representation of the genetic fitness function $f(\mathbf{p})$.

In this section we study the representations of $f(\mathbf{p})$ and $\phi(\mathbf{r})$. We will be particularly interested in the effect of the representation transformation η_c on the complexity of inducing the function. In other words, we would like to know if $\phi(\mathbf{r})$ has a more efficient description compared to that of $f(\mathbf{p})$. For example, if the size of the new representation is smaller by a considerable factor then its learning will be computationally easier. So it will be desirable over the original representation.

The rest of this paper will use two toy functions to illustrate the analytical observations. These functions are defined in the following. Let $\mathbf{x}$ be a boolean string of length ℓ and ones($\mathbf{x}$) returns the number of ones in $\mathbf{x}$.

1. Needle-in-a-haystack (NH) function:

$$f(\mathbf{x}, \mathbf{x}_{\text{opt}}) = \ell \quad \text{if} \quad \mathbf{x} = \mathbf{x}_{\text{opt}},$$
$$= 0 \quad \text{otherwise.} \tag{4}$$

   Where $\mathbf{x}_{\text{opt}}$ is the domain member with the maximum function value. Different NH functions can be defined using different choices for $\mathbf{x}_{\text{opt}}$.

2. Trap function:

$$f(\mathbf{x}) = \ell \quad \text{if ones}(\mathbf{x}) = \ell$$
$$= \ell - \text{ones}(\mathbf{x}) - 1 \quad \text{otherwise.}$$

The following section explores the change in the properties of the FCs under the genetic code-like representation transformations.

5.2 Exponential decay of individual Fourier coefficients

The jth FC in the mRNA space can be defined as:

$$w_{\mathbf{j}} = \frac{1}{2^{\ell_r}} \sum_{\mathbf{r}} \phi(\mathbf{r}) \psi_{\mathbf{j}}(\mathbf{r})$$
$$= \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f(\mathbf{p}) \sum_{\mathbf{r_i} \in R_p} \psi_{\mathbf{j}}(\mathbf{r_i}). \tag{5}$$

The magnitude of the second summation in equation (5) may take a value between 0 and $a_0^{\ell_{p,0}} a_1^{\ell_{p,1}}$ (cardinality of R_p) depending upon the nature of the set R_p. This imposes a scaling factor to the contribution of every unique protein sequence to the jth FC. Let us explore the effect of such scaling on the magnitude of an FC.

As noted earlier, the value of $\psi_{\mathbf{j}}(\mathbf{r})$ depends only on those features of $\mathbf{r}$ corresponding to the 1s in the partition $\mathbf{j}$. The mRNA features corresponding to the positions with 1s in the partition $\mathbf{j}$ may belong to (1) the same mRNA codon, (2) different codons, and (3) a combination of both. In other words, they originate from (1) the same protein feature (since one feature in the protein sequence maps to c features in the mRNA sequence), (2) different protein features, or (3) a combination of both respectively. Next, we are going to represent $\mathbf{j}$ using a collection of partitions $\{\mathbf{j}_0, \mathbf{j}_1, \ldots \mathbf{j}_q\}$ where $\mathbf{j}_0$ represents the null partition with all 0s and every $\mathbf{j}_{i \neq 0}$ represents a subpartition of the 1-contributing positions of $\mathbf{j}$ that contains only those features that belong to the same protein feature. Note that the reflection of any $\mathbf{j}_{i \neq 0}$ in the protein space has only one 1. The null partition always contributes a value of 1 and is introduced only for taking care of the case when the partition $\mathbf{j}$ is a sequence of all 0s. For example, consider a two-bit protein space that maps to a six-bit mRNA space. The partition 110001 in the mRNA

space can be represented in terms of the subpartitions 000000, 110000, and 000001. Note that $\psi_{110001}(\mathbf{r}) = \psi_{000000}(\mathbf{r})\psi_{110000}(\mathbf{r})\psi_{000001}(\mathbf{r})$. We can write $\psi_{\mathbf{j}}(\mathbf{r}) = \prod_{\alpha=0,1,\ldots q} \psi_{\mathbf{j}_\alpha}(\mathbf{r})$. Therefore, we can rewrite equation (5) as follows:

$$w_{\mathbf{j}} = \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f(\mathbf{p}) \sum_{\mathbf{r_i} \in R_p} \prod_{\alpha=0,1,\ldots q} \psi_{\mathbf{j}_\alpha}(\mathbf{r_i}). \tag{6}$$

All the defining bits (with partition value of 1) of some $\mathbf{j}_\alpha$ belong to only one protein feature by definition. Therefore, the value of $\prod_{\alpha=0,1,\ldots q} \psi_{\mathbf{j}_\alpha}(\mathbf{r_i})$ depends only on the portion of r_i defined by those q protein features. For any given combination of q protein feature values, we can define a subspace of mRNA subsequences.

If p_α is the protein feature value in a given $\mathbf{p}$ corresponding to the reflection of the α-partition in the mRNA space, then let R_{p_α} be the set of all mRNA codons that maps to p_α. Let $R_{j',p}$ be the cartesian product of the R_{p_α}s for $\alpha = 1, 2, \ldots q$. Every member of $R_{j',p}$ has cq mRNA features. For example, consider $\mathbf{p} = 110$ and $\mathbf{j} = 110000010$. So $\mathbf{j}_0 = 000000000$, $\mathbf{j}_1 = 110000000$, and $\mathbf{j}_2 = 000000010$; $\mathbf{j}_1' = 100$ and $\mathbf{j}_2' = 001$; $p_1 = 1$, $p_2 = 0$, and $\mathbf{j}' = 101$. In the case of code A, $R_{p_1} = \{100, 000, 001, 010\}$ and $R_{p_2} = \{111, 101, 110, 011\}$. Therefore $R_{101,110} = R_{p_1} \times R_{p_2}$.

Note that the basis function $\psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p})$ is well defined for any $\mathbf{r}_{j,p} \in R_{j',p}$ for any α since the feature values of $\mathbf{r}_{j,p}$ are defined for every defining location of the partition $\mathbf{j}_\alpha$. This is indeed a slight abuse of the symbols since the lengths of $\mathbf{j}_\alpha$ and $\mathbf{r}_{j,p}$ are not the same. However, we take that liberty since even if we pad $\mathbf{r}_{j,p}$ with 1s and 0s in order to make the lengths the same, the outcome will be identical. This is because the corresponding values in $\mathbf{j}_\alpha$ are 0 by definition.

Let $q_{p,j',0}$ and $q_{p,j',1}$ be the number of 0s and 1s in $\mathbf{p}$ that are covered by the fixed bits of $\mathbf{j}'$, the reflection of $\mathbf{j}$ in the protein space; $q_{p,j',0} + q_{p,j',1} = q$. In other words, q is the total number of 1s in $\mathbf{j}'$. Now note that for every $\mathbf{r}_{j,p} \in R_{j',p}$ there are $a_0^{\ell_{p,0} - q_{p,j',0}} a_1^{\ell_{p,1} - q_{p,j',1}}$ strings in the corresponding $R_{j',p}$. So we can write from equation (6),

$$\begin{aligned} w_{\mathbf{j}} &= \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f(\mathbf{p}) \; a_0^{\ell_{p,0} - q_{p,j',0}} a_1^{\ell_{p,1} - q_{p,j',1}} \sum_{\mathbf{r}_{j,p} \in R_{j',p}} \prod_{\alpha=0,1,\ldots q} \psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p}) \\ &= \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f(\mathbf{p}) \; a_0^{\ell_{p,0} - q_{p,j',0}} a_1^{\ell_{p,1} - q_{p,j',1}} \prod_{\alpha=0,1,\ldots q} \sum_{\mathbf{r}_{j,p} \in R_{p_\alpha}} \psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p}). \end{aligned} \tag{7}$$

Let $e_{j_\alpha,p}$ and $o_{j_\alpha,p}$ be the number of members in R_{p_α} that have an even and odd number of ones respectively over the partition $\mathbf{j}_\alpha$. For example, if $\mathbf{j}_\alpha = 110000$ and $p = 10$ then $e_{110000,10} = 2$ and $o_{110000,10} = 2$ for code

A shown in Table 2. Now using equation (7) we can write,

$$w_{\mathbf{j}} = \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f(\mathbf{p})\ a_0^{\ell_{p,0}-q_{p,j',0}} a_1^{\ell_{p,1}-q_{p,j',1}} \kappa \prod_{\alpha=0,1,\ldots q} |e_{j_\alpha,p} - o_{j_\alpha,p}|, \tag{8}$$

where $\kappa \in \{-1, 1\}$ and $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ denotes the magnitude of $(e_{j_\alpha,p} - o_{j_\alpha,p})$ for all $\alpha \neq 0$. The value of $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ can be determined directly from the genetic code. By definition, for the null partition $\alpha = 0$, we set $|e_{j_0,p} - o_{j_0,p}| = 1$. As before, this is done to take care of the case where $\mathbf{j}$ is comprised of only 0s resulting in $q = 0$.

Now let us specialize this equation for code A. For this code $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ is either 2 or 0 for all the partitions (except the partition with all 0s). If $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ is equal to zero for any given α and the corresponding protein feature value in $\mathbf{p}$ then the overall contribution of $\mathbf{p}$ to $w_{\mathbf{j}}$ is zero. Also note that since $\mathbf{p}$ is a binary string, any feature in $\mathbf{p}$ can take only two values: 0 and 1. Therefore, if $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ is 0 for a certain feature entry in $\mathbf{p}$ (corresponding to j_α in the mRNA space) $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ must be 0 for the complementary feature value of $\mathbf{p}$ at the same location. As a result, the corresponding coefficient $w_{\mathbf{j}}$ will be zero. Therefore for all nonzero $w_{\mathbf{j}}$s, except the coefficient w_0, the value of $|e_{j_\alpha,p} - o_{j_\alpha,p}|$ must be equal to 2; w_0 is the FC for the partition with all entries set to zero.

So for all nonzero coefficients except w_0 in the representation using the code A we can write,

$$\begin{aligned} w_{\mathbf{j}} &= \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} \kappa\ f(\mathbf{p})\ 4^{\ell_p - q}\ 2^q \\ &= \frac{1}{2^{\ell_p+q}} \sum_{\mathbf{p}} \kappa\ f(\mathbf{p}) \\ &\leq \frac{1}{2^{\ell_p+q}} \sum_{\mathbf{p}} f(\mathbf{p}) \leq \frac{w_0}{2^q}. \end{aligned} \tag{9}$$

Note that $w_0 = 1/2^{\ell_p} \sum_{\mathbf{p}} f(\mathbf{p})$. This equation shows an exponential decay in the magnitude of the coefficients as the partition index of the coefficients involve more and more defining bits. As we increase the value of q (the number of 1s in $\mathbf{j}'$, the reflection of the partition $\mathbf{j}$) the upper bound on the magnitude of the coefficient $w_{\mathbf{j}}$ decreases exponentially.

We can also specialize equation (8) for the genetic code B. Note that $|e_{j_\alpha,p} - o_{j_\alpha,p}| = 1$ for all α and $\mathbf{p}$. Also $a_1 = 1$ and $a_0 = 7$. Therefore,

$$w_{\mathbf{j}} = \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} \kappa f(\mathbf{p})\ 7^{\ell_{p,0}-q_{p,j',0}}.$$

Figure 2 (top) shows the effect of codes A and B on the Fourier representation of function NH. The figure also shows the magnitude

of the coefficients of the original Fourier representation without using the representation transformation where all the coefficients have the same magnitude. However, the magnitude decays exponentially with respect to the absolute order of the mRNA partitions for representations generated using codes A and B. Note that the magnitude of the nonzero coefficients corresponding to partitions with the same absolute order are the same. Figure 2 (bottom) shows similar results for the trap function.

Magnitudes of the individual coefficients do not tell the complete story. Construction of an efficient representation requires considering properties of all the coefficients together. This is particularly important for the current case since these transformations expand the domain and introduce many new partitions. Even if the magnitudes of individual coefficients decrease, the increased number of coefficients (recall that the mRNA representation uses more features) may result in no benefit towards reducing the description size of the overall function representation. In other words, a large number of small coefficients together may contribute significantly to the output of the function. The following section explores this issue and points out that although code A offers little benefit from this perspective, properties of code B are quite encouraging.

5.3 Energy of the Fourier spectrum

The energy of the Fourier spectrum can be defined as

$$E = \sum_{\mathbf{j}} w_{\mathbf{j}}^2. \tag{10}$$

Let us now study the change in the overall energy of the spectrum due to the genetic code-like representation transformations. Using equation (5) and noting that $\psi_{\mathbf{j}}(\mathbf{x}) = \psi_{\mathbf{x}}(\mathbf{j})$ we can write

$$w_{\mathbf{j}}^2 = \frac{1}{2^{2c\ell_p}} \sum_{\mathbf{p},\mathbf{s}} f(\mathbf{p}) f(\mathbf{s}) \sum_{\mathbf{r}_i \in R_p, \mathbf{r}_k \in R_s} \psi_{\mathbf{j}}(\mathbf{r}_i) \psi_{\mathbf{j}}(\mathbf{r}_k)$$

$$\sum_{\mathbf{j}} w_{\mathbf{j}}^2 = \frac{1}{2^{2c\ell_p}} \sum_{\mathbf{p},\mathbf{s}} f(\mathbf{p}) f(\mathbf{s}) \sum_{\mathbf{r}_i \in R_p, \mathbf{r}_k \in R_s} \sum_{\mathbf{j}} \psi_{\mathbf{r}_i}(\mathbf{j}) \psi_{\mathbf{r}_k}(\mathbf{j}).$$

Exploiting the orthonormality condition we can write

$$\sum_{\mathbf{j}} w_{\mathbf{j}}^2 = \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f^2(\mathbf{p}) a_0^{\ell_{p,0}} a_1^{\ell_{p,1}}. \tag{11}$$

Let us now specialize this result for code A. For this code, $a_0 = a_1 = 4$ and $c = 3$. Substituting these values in equation (11) we get

$$E_R = \frac{1}{2^{\ell_p}} \sum_{\mathbf{p}} f^2(\mathbf{p}) = E_P,$$

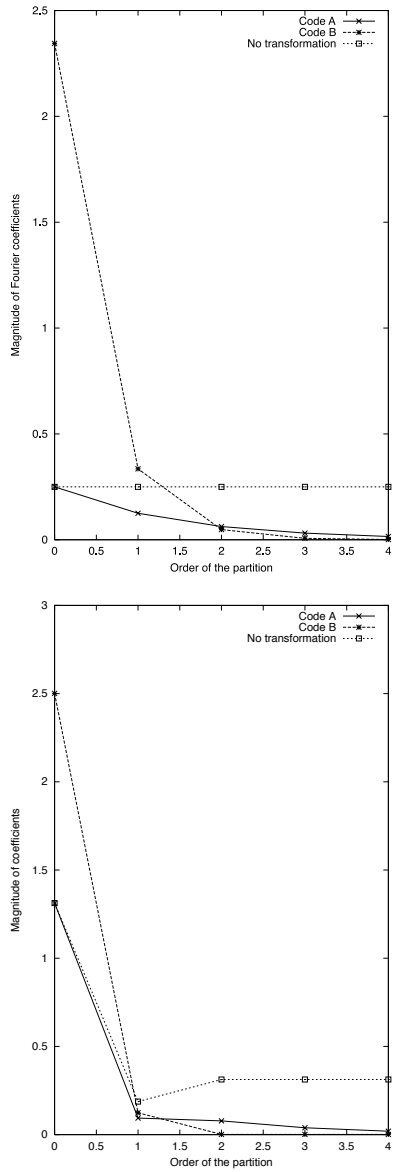

Figure 2. (Top) Variation of the magnitude of the Fourier coefficients with respect to the order (q) of the partitions in the original and transformed representations of the NH function. It shows the result using code A, code B, and no transformation. Note that the magnitude is invariant in the representation with no transformation. On the other hand it decays exponentially when the transformations are applied. (Bottom) Similar result for the trap function. Note that all coefficients of the same order have the same magnitude for both NH and trap functions.

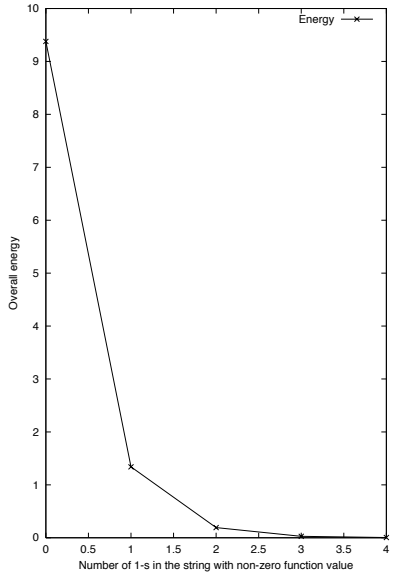

Figure 3. Variation of the overall energy with increasing number of 1s in the string ($\mathbf{x}_{\mathrm{opt}}$) with nonzero function value for the function NH. Code B is used.

where E_R and E_P are the energies of the mRNA and the protein spaces. The overall energy remains invariant under the transformation code A.

Code B however changes the overall energy. Figure 3 shows the different values of the energy for different choices of the string with a nonzero function value ($\mathbf{x}_{\mathrm{opt}}$) in the function NH using code B. The overall energy of the Fourier representation of the trap function with four variables using code B is approximately $2.2778E_P$.

Although the overall energy is an interesting property to observe, the most critical properties are the number of coefficients that significantly contribute to the overall energy of the representation and the location of those significant coefficients. If the number is small and the contribution from the rest is negligible, then we know that the function can be approximated using a small number of coefficients. If we also know the partitions that are associated with those significant coefficients then we should be able to efficiently compute the representation. The following section shows that both of these requirements can be satisfied by a class of GCTs.

5.4 Distribution of the energy in partitions of different order

The distribution of energy among the coefficients of different order is a very interesting property of a representation. For example, if we know that a representation has a small number of significant coefficients and

they are associated with a certain order of partitions then it will be easier to compute such a representation. In this section we shall study such properties of the Fourier representation produced by the GCTs.

Recall that the order of a partition $\mathbf{j}$ is the number of ones in $\mathbf{j}$; in other words, it is the number of features that define the corresponding basis function $\psi_{\mathbf{j}}(\mathbf{x})$. Let us define the *order-k energy*, $E^{(k)} = \sum_{\mathbf{j}|\text{ones}(\mathbf{j})=k} w_{\mathbf{j}}^2$. We can compute this for both the protein and the mRNA space. Note that an order-k partition in the mRNA space may correspond (through reflection) to a lower order partition in the protein space since multiple mRNA features are associated with the same protein feature. A careful study of the effect of the representation transformations on the order-k energy in the mRNA space may require understanding the properties of the coefficients in the mRNA space that correspond to exactly k features in the protein space. We are going to use the term *absolute order-k energy*, defined as $\mathcal{E}^{(k)} = \sum_{\mathbf{j}|\text{ones}(\mathbf{j}')=k} w_{\mathbf{j}}^2$; as defined earlier, $\mathbf{j}'$ is the reflection of $\mathbf{j}$ in the protein space. Just like the association between the partitions in the protein and the mRNA spaces through the concept of reflection, the distribution of energies in these two representations can be linked through the concept of absolute order-k energy.

Using equation (7) we can write

$$\begin{aligned}
w_{\mathbf{j}}^2 &= \frac{1}{2^{2c\ell_p}} \sum_{\mathbf{p},\mathbf{s}} f(\mathbf{p}) f(\mathbf{s}) \;\; a_0^{\ell_{p,0}-q_{p,j',0}+\ell_{s,0}-q_{s,j',0}} a_1^{\ell_{p,1}-q_{p,j',1}+\ell_{s,1}-q_{s,j',1}} \\
&\quad \cdot \prod_{\alpha=0,\cdots q} \sum_{\mathbf{r}_{j,p}\in R_{p_\alpha}, \mathbf{t}_{j,s}\in R_{t_\alpha}} \psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p}) \psi_{\mathbf{j}_\alpha}(\mathbf{t}_{j,s}) \\
&= \frac{1}{2^{2c\ell_p}} \sum_{\mathbf{p}} f^2(\mathbf{p}) \;\; a_0^{2(\ell_{p,0}-q_{p,j',0})} a_1^{2(\ell_{p,1}-q_{p,j',1})} \\
&\quad \cdot \prod_{\alpha=0,1,\cdots q} \sum_{\mathbf{r}_{j,p}\in R_{p_\alpha}, \mathbf{t}_{j,p}\in R_{p_\alpha}} \psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p}) \psi_{\mathbf{j}_\alpha}(\mathbf{t}_{j,p}) \\
&\quad + \frac{1}{2^{2c\ell_p}} \sum_{\mathbf{p}\neq\mathbf{s}} f(\mathbf{p}) f(\mathbf{s}) \sum_{\mathbf{r}_i\in R_p, \mathbf{r}_k\in R_s} \psi_{\mathbf{r}_i}(\mathbf{j}) \psi_{\mathbf{r}_k}(\mathbf{j}). \qquad (12)
\end{aligned}$$

Now summing both sides of equation (12) over all partitions and noting that the second term of the right-hand side disappears because of the orthonormality property, we can write

$$\begin{aligned}
\sum_{\mathbf{j}} w_{\mathbf{j}}^2 &= \frac{1}{2^{2c\ell_p}} \sum_{\mathbf{p}} f^2(\mathbf{p}) \;\; \sum_{\mathbf{j}} a_0^{2(\ell_{p,0}-q_{p,j',0})} a_1^{2(\ell_{p,1}-q_{p,j',1})} \\
&\quad \cdot \prod_{\alpha=0,1,\cdots q} \sum_{\mathbf{r}_{j,p}\in R_{p_\alpha}, \mathbf{t}_{j,p}\in R_{p_\alpha}} \psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p}) \psi_{\mathbf{j}_\alpha}(\mathbf{t}_{j,p}).
\end{aligned}$$

Comparing this with equation (11) we note that

$$\frac{1}{2^{c\ell_p}} \sum_{\mathbf{j}} a_0^{\ell_{p,0}-2q_{p,j',0}} a_1^{\ell_{p,1}-2q_{p,j',1}} \cdot \prod_{\alpha=0,1,\ldots q} \sum_{\mathbf{r}_{j,p}\in R_{p_\alpha}, \mathbf{t}_{j,p}\in R_{p_\alpha}} \psi_{\mathbf{j}_\alpha}(\mathbf{r}_{j,p})\psi_{\mathbf{j}_\alpha}(\mathbf{t}_{j,p}) = 1. \tag{13}$$

Now let us explore the rate of convergence of expression in the left-hand side of equation (13). If it approaches 1 very quickly with respect to increasing order of the coefficients ($q = q_{p,j',0} + q_{p,j',1}$) then we know that only a small number of low order coefficients mainly contribute to the overall energy of the Fourier representation.

Using equations (8) and (13) we can write

$$\frac{1}{2^{c\ell_p}} \sum_{\mathbf{j}} a_0^{\ell_{p,0}-2q_{p,j,0}} a_1^{\ell_{p,1}-2q_{p,j,1}} \prod_{\alpha=0,1,\ldots q} (|e_{j_\alpha,p} - o_{j_\alpha,p}|)^2 = 1. \tag{14}$$

Although $(|e_{j_\alpha,p} - o_{j_\alpha,p}|)^2 = (e_{j_\alpha,p} - o_{j_\alpha,p})^2$ we have left the |.| symbol in place since earlier we defined that $|e_{j_\alpha,p} - o_{j_\alpha,p}| = 1$ for the partition with all 0s (i.e., $q = 0$). Equation (14) essentially controls the distribution of the eneregy with respect to the order of the partitions. GCTs that can provide an exponential convergence of the left-hand side of this equation to 1 with respect to increasing order, will also offer an exponential decay in the energy.

Let us now specialize equation (14) for code A. Since $a_0 = a_1 = 4$ and $c = 3$ for code A,

$$\frac{1}{2^{\ell_p}} \sum_{\mathbf{j}} 2^{-4q} \prod_{\alpha=0,1,\ldots q} (|e_{j_\alpha,p} - o_{j_\alpha,p}|)^2 = 1. \tag{15}$$

This can be further simplified by counting the number of partitions of absolute order q that are associated with nonzero coefficients and noting that $|e_{j_\alpha,p} - o_{j_\alpha,p}| = 2$ for all of them. There are $\binom{\ell_p}{q}$ order-q partitions and by studying the genetic code A we observe that any protein feature corresponds to four choices (note that the null partition is not a choice) in the mRNA partition-space that can have nonzero coefficients. In other words, there are only four choices of partitions over a particular codon that can have nonzero coefficients. In a partition of absolute order equal to q there are 4^q such partitions. Therefore,

$$\frac{1}{2^{\ell_p}} \sum_{q=0}^{\ell_p} \binom{\ell_p}{q} 2^{-4q} 4^q 2^{2q} = 1$$

$$\frac{1}{2^{\ell_p}} \sum_{q=0}^{\ell_p} \binom{\ell_p}{q} = 1. \tag{16}$$

Note that the case for a null partition ($q = 0$) is taken care of since $4^0 = 1$. So the absolute order-k energy for the code A is

$$\mathcal{E}_A^{(k)} = \frac{1}{2^{\ell_p}} \binom{\ell_p}{k} E_P. \tag{17}$$

This equation clearly shows that the distribution of the energy among different orders is controlled by only the total number of partitions in the protein space with the same order. This is identical to the distribution of order-k energy in the protein space. In other words, code A does not really change the distribution of the energy among different orders. The magnitudes of the individual coefficients decay only because the order-k energy is distributed among an increased number of partitions. This also means that code A does not necessarily offer a better representation that is easier to approximate using a smaller number of coefficients.

The theoretical observations are supported by the experimentally computed values of the FCs. Figure 4 (top) shows the distribution of absolute order-k energy for the function NH using code A. Figure 4 (bottom) shows the distribution of the energy in the protein space for the same function NH. As we see, both distributions are identical.

Let us now specialize equation (14) for code B. First note that $|e_{j_\alpha,p} - o_{j_\alpha,p}| = 1$ for code B. Therefore,

$$\frac{1}{2^{c\ell_p}} \sum_{\mathbf{j}} a_0^{\ell_{p,0}-2q_{p,j',0}} a_1^{\ell_{p,1}-2q_{p,j',1}} = 1.$$

This can be further simplified by counting the number of partitions in the mRNA space for each absolute order value of ones($\mathbf{j}'$),

$$\frac{1}{2^{c\,\ell_p}} \sum_{q=0}^{\ell_p} \sum_{q_{p,j,0}=\max(0,q-l_p+l_{p,0})}^{\min(q,\ell_{p,0})} \cdot \binom{\ell_{p,0}}{q_{p,j',0}} \binom{\ell_p-\ell_{p,0}}{q-q_{p,j',0}} (2^c-1)^q a_0^{\ell_{p,0}-2q_{p,j',0}} = 1, \tag{18}$$

where q is essentially ones($\mathbf{j}'$). The term with a_1 disappeared since $a_1 = 1$ for code B. Note that the left-hand side of equation (18) contains a summation over different values of q from zero through ℓ_p. We are interested in the convergence of the left-hand side to 1 as we continue to add the contributions for different values of q. In order to study that let us define,

$$g(\ell_p, \ell_{p,0}, k) = \frac{1}{2^{c\,\ell_p}} \sum_{q=0}^{k} \sum_{q_{p,j',0}=\max(0,q-l_p+l_{p,0})}^{\min(q,\ell_{p,0})} \cdot \binom{\ell_{p,0}}{q_{p,j',0}} \binom{\ell_p-\ell_{p,0}}{q-q_{p,j',0}} (2^c-1)^q a_0^{\ell_{p,0}-2q_{p,j',0}}$$

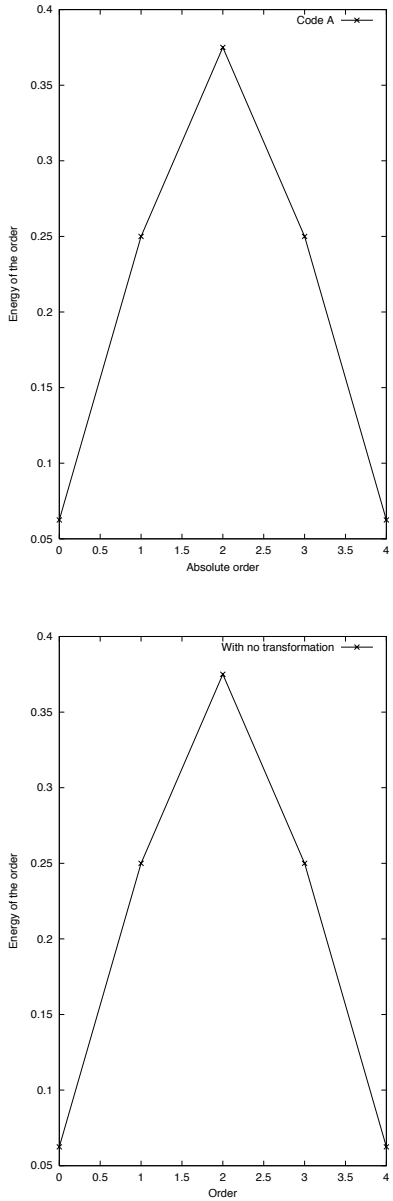

Figure 4. (Top) Distribution of the absolute order energy using code A and the function NH with $\mathbf{x}_{\mathrm{opt}} = 0000$. (Bottom) Distribution of the energy in the protein space (i.e., with no representation transformation).

where $0 \le k \le \ell_p$. We would like to study the convergence of $g(\ell_p, \ell_{p,0}, k)$ to 1 as k increases from 0 through ℓ_p. Note that $g(\ell_p, \ell_{p,0}, k)$ is also a function of $\ell_{p,0}$, the number of 0s in a sequence $\mathbf{p}$. Since we are dealing with boolean sequences, $\ell_{p,0}$ is sufficient to define any particular $\mathbf{p}$.

Figure 5 shows the variation of $g(4, \ell_{p,0}, k)$ with respect to increasing k and different $\ell_{p,0}$s. Since the convergence characteristic depends only on the number of 0s in $\mathbf{p}$, not their exact locations in the string, the variations are shown for the four different types (note that $\ell_p = 4$) of $\mathbf{p}$s. Figure 6 presents the variation of $g(300, \ell_{p,0}, k)$ for two boundary cases $\ell_{p,0} = 0$, $\ell_{p,0} = 300$, and the intermediate case $\ell_{p,0} = 150$.

Both Figures 5 and 6 convey an important message. Note that in both cases, $g(\ell_p, \ell_{p,0}, k)$ approaches 1 faster (with respect to k) when $\ell_{p,0}$ is large. This is simply because code B assigns seven mRNA codons for the protein feature 0. Now let us write the overall energy of the representation constructed using code B,

$$E_R = \sum_{\mathbf{j}} w_{\mathbf{j}}^2 = \frac{1}{2^{c\ell_p}} \sum_{\mathbf{p}} f^2(\mathbf{p}) a_0^{\ell_{p,0}} g(\ell_p, \ell_{p,0}, \ell_p). \tag{19}$$

Note that the protein sequences with high genetic fitness contribute significantly to the overall energy E_R since $f^2(\mathbf{p})$ will be large for them. Moreover, the effect of $f^2(\mathbf{p})$ on the energy gets scaled up by the factor $a_0^{\ell_{p,0}}$. This essentially means that fitter protein sequences with a large number of 0s will mainly contribute to E_R. Now note that for proteins with a large number of 0s (i.e., relatively large $\ell_{p,0}$) the function $g(\ell_p, \ell_{p,0}, k)$ approaches 1 very fast. In other words, the main portion of the overall energy comes from the highly fit proteins that have a greater number of equivalent mRNA representations (implied by large value of $\ell_{p,0}$ and bias of code B towards the protein feature 0).

Equation (18) can also be further specialized for the function NH. If the string with all 0s is the optimal solution then it is the only member of the domain that contributes to the FCs. Therefore we can eliminate the summation over all ps by only the string with all 0s. Noting that $\ell_{p,0} = \ell_p$, $q_{p,j,0} = q$ we can write

$$\frac{1}{2^{c\ell_p}} \sum_{q=0}^{\ell_p} \binom{\ell_p}{q} (2^c - 1)^q a_0^{\ell_p - 2q} = 1. \tag{20}$$

These theoretical observations are also supported by experimentally computed values of the FCs. Figure 7 shows the distribution of absolute order-k energy for the function NH using code B. As we see, the contribution to the overall energy from the coefficients of a certain order diminishes as the order increases when the optimal solution contains all 0s. The NH function is an extreme case where everyone but one domain

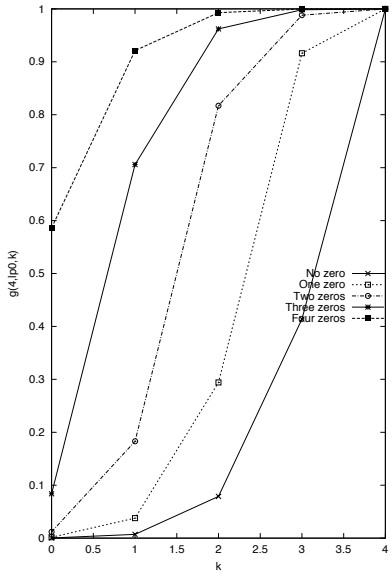

Figure 5. Variation of the $g(4, \ell_{p,0}, k)$ with respect to increasing k. The variation is shown for different types of **p**s. $\ell_p = 4$ and the code B is used.

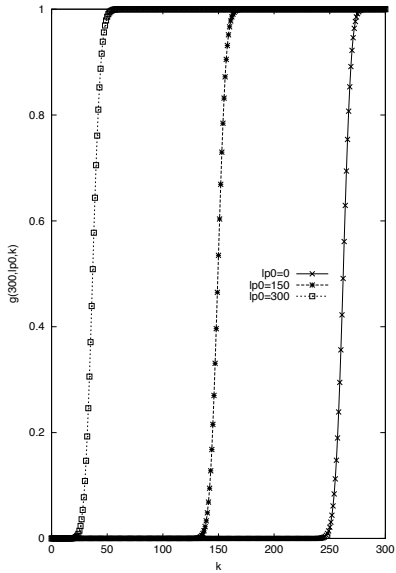

Figure 6. $g(300, \ell_{p,0}, k)$ with respect to increasing k for two boundary cases $\ell_{p,0} = 0$, $\ell_{p,0} = 300$, and the intermediate case $\ell_{p,0} = 150$. This shows the convergence rate for a larger number of protein features using code B.

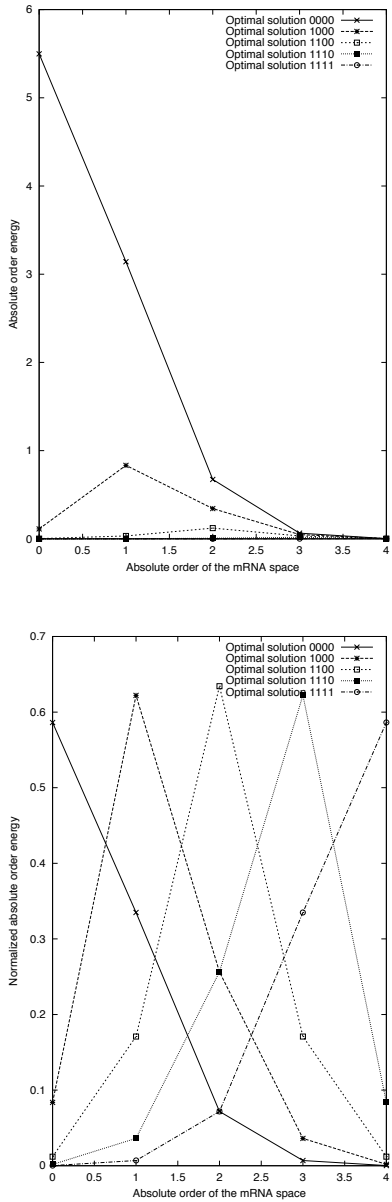

Figure 7. (Top) Distribution of the absolute order energy using code B for function NH. (Bottom) Distribution of the normalized value of the absolute order energy using code B for function NH.

member has a zero function value. As a result the distribution of the absolute order-k energy depends solely on the property of the sequence $\mathbf{x}_{\text{opt}}$. If we set $\mathbf{x}_{\text{opt}}$ to sequences with a smaller number of 0s the decay is preceded by an increase in energy.

Now let us consider a different example using the trap functions. Note that in this case although the sequence with all 1s has the highest function value, there are other sequences that have a nonzero function value. The sequences with a larger number of 0s have relatively high fitness values. This also matches with the bias of the genetic code B. Therefore we should expect a good approximation using the low order coefficients.

Figure 8 (top) shows the distribution of absolute order-k energy using the code A, code B, and no transformation for a trap function with $\ell_p = 4$. Note that the distribution of the energy using no transformation and that using code A are identical as noted in equation (17). Also note that the absolute order-k energy decreases exponentially for code B. Figure 8 (bottom) shows the order-k energy using both codes A and B for the trap function. Note that this is the order-k energy of the mRNA representation, not the absolute order-k energy and 4-bit protein sequences map to 12-bit mRNA sequences.

It is important to realize that the match between the bias of the genetic code and the representation of the fitter proteins may not be difficult to achieve. Fitter proteins will have larger values of $f^2(\mathbf{p})$. If we assign a greater number of codons to the most frequent feature value (either 1 or 0 in the case of binary strings) used in the fitter proteins, then the corresponding scaling factor ($a_0^{\ell_{p,0}}$ in the case of code B) will also be large. For these proteins $g(\ell_p, \ell_{p,0}, k)$ also approaches 1 very fast with respect to k. In the case of code B, the larger the value of $\ell_{p,0}$ in a protein, the higher the rate of convergence and the larger the scaling factor. On the other hand, the proteins with frequent feature values that have a smaller number of codons assigned (1 in the case of code B) will have a slower convergence rate for $g(\ell_p, \ell_{p,0}, k)$ and a smaller scaling factor. In the case of code B it will be protein sequences with smaller values for $\ell_{p,0}$ (i.e., strings with relatively more 1s). Although the convergence rate will be slow, if the fitnesses of these proteins are relatively low, their contribution to the overall energy will be low since the scaling factor will be small for them. Note that if the fitness value is relatively small compared to the scaling factor, the latter will play a more significant role. For binary representation, the issue is assigning a codon distribution among two possible protein features 0 and 1. For representations with higher cardinality, the code introduces richer transformations and we need to further explore the implications.

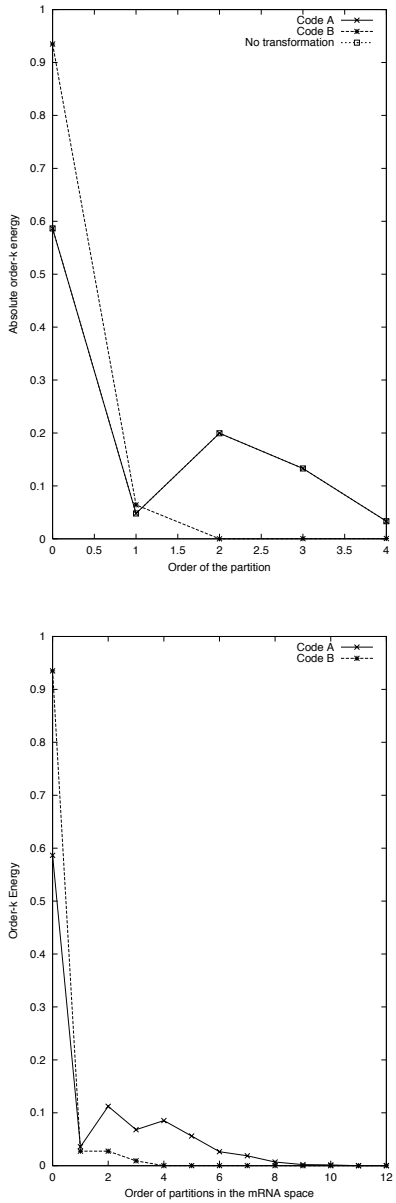

Figure 8. (Top) Distribution of the absolute order energy using codes A and B for the trap function. (Bottom) Distribution of the order-k energy using codes A and B for the trap function. Note that 4-bit protein space maps to 12-bit mRNA representation since the codon size is three.

6. Conclusions

This paper offered some intriguing properties of genetic code-like transformations (GCTs) that may be extremely useful for inducing a function from observed data. It showed that there exist some GCTs that can construct a Fourier representation of some fitness functions where the low order coefficients are exponentially more significant than the higher order coefficients. This is a very critical property that allows a polynomial-complexity approximation of an exponentially long function representation.

The paper demonstrated this by considering two GCTs and a pair of functions known to have exponentially long Fourier representations. It first showed that the magnitude of individual Fourier coefficients (FCs) decay at an exponential rate as the order (the number of features associated with the coefficients) increases. However, such decay in individual coefficients does not guarantee efficient representation. This is because a large number of small coefficients together may contribute an insignificant amount to the overall function value. So we needed to explore the variation of the energy (sum of the square of the coefficients) of the spectrum with respect to increasing order. We noted that one of the transformations (code B) generated an exponentially decaying energy distribution. This guarantees that a low order approximation of the function will be accurate since the cumulative contribution from the higher order terms is negligible.

Although the results are presented in the context of specific GCTs, this paper makes an effort to characterize the class of GCTs that offer such useful properties. Equation (14) essentially controls this property. GCTs that can provide an exponential convergence of the left-hand side of this equation to 1 will also offer an exponential decay in the energy with respect to increasing order. Code B does that; however, code A does not. This paper also outlines a physical conjecture for constructing such transformations. It suggests that one possible way to construct such GCTs may be to assign more equivalent copies to protein sequences with higher genetic fitness values by introducing redundancy in the genetic code.

The implication of this paper on the field of evolutionary computation is important. A technique for efficient and scalable induction of function representation will be useful in almost every application of evolutionary algorithms. Examples include evolving programs, learning classifiers, detecting patterns from data, and optimization. This work also suggests that we should rethink our existing models of evolutionary computation. We need to further explore the computational role of gene expression. That may ultimately lead us toward unveiling the true power of genetic search.

Acknowledgments

This work was supported by the United States National Science Foundation Grants IIS-9803660 and IIS-0083946. The author would also like to thank Alden Wright, Robert Heckendorn, and Dirk Thierens for very useful comments on this paper.

References

[1] J. D. Bagley, "The Behavior of Adaptive Systems which Employ Genetic and Correlation Algorithms," *Dissertation Abstracts International*, **28**(12):5106B, 1967 (University Microfilms Number 68-7556).

[2] S. Bandyopadhyay, H. Kargupta, and G. Wang, "Revisiting the GEMGA: Scalable Evolutionary Optimization through Linkage Learning," in *Proceedings of the IEEE International Conference on Evolutionary Computation* (IEEE Press, 1998).

[3] J. Bashford, I. Tsohantjis, and P. Jarvis, "A Supersymmetric Model for the Evolution of the Genetic Code," to be published in *National Academy of Science USA*, **95**(3), 1998.

[4] K. G. Beauchamp, *Applications of Walsh and Related Functions* (Academic Press, USA, 1984).

[5] P. Beland and T. Allen, "The Origin and Evolution of the Genetic Code," *Journal of Theoretical Biology*, **170** (1994) 359–365.

[6] A. D. Bethke, "Comparison of Genetic Algorithms and Gradient-based Optimizers on Parallel Processors: Efficiency of Use of Processing Capacity," *Technical Report Number 197* (University of Michigan, Logic of Computers Group, Ann Arbor, 1976).

[7] A. Brindle, *Genetic Algorithms for Function Optimization*, unpublished doctoral dissertation (University of Alberta, Edmonton, Canada, 1981).

[8] D. Dasgupta and D. R. McGregor, "Designing Neural Networks using the Structured Genetic Algorithm," *Artifical Neural Networks*, **2** (1992) 263–268.

[9] S. Forrest and M. Mitchell, "The Performance of Genetic Algorithms on Walsh Polynomials: Some Anomalous Results and their Explanation," in *Proceedings of the Fourth International Conference on Genetic Algorithms*, edited by R. K. Belew and L. B. Booker (Morgan Kaufmann, San Mateo, CA, 1991).

[10] S. Fukuchi, T. Okayama, and J. Otsuka, "Evolution of Genetic Information Flow from the Viewpoint of Protein Sequence Similarity," *Journal of Theoretical Biology*, **171** (1994) 179–195.

[11] D. E. Goldberg, "Genetic Algorithms and Walsh Functions: Part I, A Gentle Introduction," *Complex Systems*, **3**(2) (1989) 129–152, (also TCGA Report 88006).

[12] D. E. Goldberg, "Genetic Algorithms and Walsh Functions: Part II, Deception and its Analysis," *Complex Systems*, **3**(2) (1989) 153–171, (also TCGA Report 89001).

[13] D. E. Goldberg, B. Korb, and K. Deb, "Messy Genetic Algorithms: Motivation, Analysis, and First Results," *Complex Systems*, **3**(5) (1989) 493–530, (also TCGA Report 89003).

[14] R. Heckendorn and D. Whitley, "Predicting Epistasis from Mathematical Models," *Journal Of Evolutionary Computation*, **7**(1) (1999) 69–101.

[15] J. H. Holland, *Adaptation in Natural and Artificial Systems* (University of Michigan Press, Ann Arbor, 1975).

[16] R. B. Hollstien, "Artificial Genetic Adaptation in Computer Control Systems," *Dissertation Abstracts International*, **32**(3):1510B, 1971 (University Microfilms Number 71-23,773).

[17] J. Hornos and Y. Hornos, "Algebraic Model for the Evolution of the Genetic Code," *Physical Review Letters*, **71**(26) (1993) 4401–4404.

[18] J. Jackson, *The Harmonic Sieve: A Novel Application of Fourier Analysis to Machine Learning Theory and Practice*, PhD thesis (School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1995).

[19] F. Jacob and J. Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," *Molecular Biology*, **3** (1961) 318–356.

[20] H. Kargupta, "The Gene Expression Messy Genetic Algorithm," in *Proceedings of the IEEE International Conference on Evolutionary Computation* (IEEE Press, 1996).

[21] H. Kargupta, "Gene Expression: The Missing Link of Evolutionary Computation," in *Genetic Algorithms in Engineering and Computer Science*, edited by C. Poloni, D. Quagliarella, J. Periaux, and G. Winter (John Wiley & Sons Ltd., 1997).

[22] H. Kargupta, "SEARCH, Computational Processes in Evolution, and Preliminary Development of the Gene Expression Messy Genetic Algorithm," *Complex Systems*, **11**(4) (1997) 233–287.

[23] H. Kargupta, "Gene Expression and Large Scale Evolutionary Optimization," in *Computational Aerosciences in the 21st Century* (Kluwer Academic Publishers, 1998).

[24] H. Kargupta and S. Bandyopadhyay, "Further Experimentations on the Scalability of the GEMGA," in *Lecture Notes in Computer Science: Parallel Problem Solving from Nature* (Springer-Verlag, 1998).

[25] H. Kargupta and S. Bandyopadhyay, "A Perspective on the Foundation and Evolution of the Linkage Learning Genetic Algorithms," *Computer Methods in Applied Mechanics and Engineering*, **186** (2000) 269–294. Special Issue on Genetic Algorithms, Guest Editors: Goldberg, D. E. and Deb, K.

[26] H. Kargupta, D. E. Goldberg, and L. W. Wang, "Extending the Class of Order-*k* Delineable Problems for the Gene Expression Messy Genetic Algorithm," in *Proceedings of the Second Annual Conference on Genetic Programming* (Morgan Kaufmann Publishers, 1997).

[27] H. Kargupta and H. Park, "Fast Construction of Distributed and Decomposed Evolutionary Representation," in *Late Breaking Papers of the Genetic and Evolutionary Computation Conference*, pages 139–148, 1999. Extended version is in communication.

[28] H. Kargupta and K. Sarkar, "Function Induction, Gene Expression, and Evolutionary Representation Construction," in *Proceedings of the Genetic and Evolutionary Computation Conference, Orlando, FL, AAAI Press* (Morgan Kaufmann, 1999).

[29] H. Kargupta and B. Stafford, "From DNA to Protein: Transformations and their Possible Role in Linkage Learning," *Proceedings of the Seventh International Conference on Genetic Algorithms*, July 1997.

[30] S. Kauffman, *The Origins of Order* (Oxford University Press, New York, 1993).

[31] R. Keller and W. Banzhaf, "The Evolution of Genetic Code in Genetic Programming," in *Proceedings of the Genetic and Evolutionary Computation Conference* (Morgan Kaufmann Publishers, 1999).

[32] S. Kushilevitz and Y. Mansour, "Learning Decision Trees using Fourier Spectrum," in *Proceedings of the 23rd Annual ACM Symposium on the Theory of Computing*, 1991.

[33] G. E. Liepins and M. D. Vose, "Polynomials, Basic Sets, and Deceptiveness in Genetic Algorithms," *Complex Systems*, **5**(1) (1991) 45–61.

[34] C. K. Oei, "Walsh Function Analysis of Genetic Algorithms of Nonbinary Strings," unpublished master's thesis (University of Illinois at Urbana-Champaign, Department of Computer Science, 1992).

[35] S. Rana, R. B. Heckendorn, and D. Whitley, "A Tractable Walsh Analysis of SAT and its Implications for Genetic Algorithms," in *Proceedings of the AAAI-98* (AAAI Press, 1998).

[36] C. Reidys and S. Fraser, "Evolution of Random Structures," *Technical Report 96-11-082* (Santa Fe Institute, Santa Fe, 1996).

[37] D. Rockmore, P. Kostelec, W. Hordijk, and P. Stadler, "Fast Fourier Transform for Fitness Landscapes," *Technical Report 99-10-068* (Santa Fe Institute, Santa Fe, 1999).

[38] R. S. Rosenberg, "Simulation of Genetic Populations with Biochemical Properties," *Dissertation Abstracts International*, **28**(7):2732B, 1967 (University Microfilms Number 67-17,836).

[39] P. Schuster, "The Role of Neutral Mutations in the Evolution of RNA Molecules," in *Theoretical and Computational Methods in Genome Research*, edited by S. Suhai (Plenum Press, New York, 1997).

[40] R. E. Smith, "An Investigation of Diploid Genetic Algorithms for Adaptive Search of Nonstationary Functions," *TCGA Report Number 88001* (University of Alabama, The Clearinghouse for Genetic Algorithms, Tuscaloosa, 1988).

[41] D. Thierens, "Estimating the Significant Non-linearities in the Genome Problem-coding," in *Proceedings of the Genetic and Evolutionary Computation Conference* (Morgan Kaufmann Publishers, 1999).

[42] D. Thierens, "Scalability Problems of Simple Genetic Algorithms," *Evolutionary Computation*, **7**(4) (1999) 331–352.

[43] M. Vose and A. Wright, "The Simple Genetic Algorithm and the Walsh Transform: Part I, Theory," *Journal of Evolutionary Computation*, **6**(3) (1998) 253–274.

[44] M. Vose and A. Wright, "The Simple Genetic Algorithm and the Walsh Transform: Part II, The Inverse," *Journal of Evolutionary Computation*, **6**(3) (1998) 253–274.

[45] J. L. Walsh, "A Closed Set of Orthogonal Functions," *Ann. Journ. Math.*, **55**, 1923.

[46] A. Wu and R. Lindsay, "Empirical Studies of the Genetic Algorithm with Non-coding Segments," *Journal of Evolutionary Computation*, **3**(2) (1995) 121–147.

[47] A. Wu and R. Lindsay, "A Survey of Intron Research in Genetics," in *Parallel Problem Solving from Nature–PPSN IV*, edited by H. Voigt, W. Ebeling, I. Rechenberg, and H. Schwefel (Springer-Verlag, Berlin, 1996).