Building Block and Building Rule: Dual Descriptor Method for Biological Sequence Analysis

Bin-Guang Ma[a, b, *]

[a] *Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Center for Advanced Study, Shandong University of Technology, Zibo 255049, P. R. China*

[b] *College of Chemistry and Chemical Engineering, Suzhou University, Suzhou 215006, P. R. China.*

*To whom correspondence should be addressed.

Telephone: ++86-533-2780271; Fax: ++86-533-2780271;

E-mail: bgMa@sdut.edu.cn

Abstract

The emergence of “Systems Biology” in recent years highlights the systematic viewpoint of bio-system modeling. Building on such a background, Dual Descriptor Method, a generic methodology for biological sequence analysis is proposed. From a systematic perspective, Dual Descriptor is defined as a two element set of Composition Weight Map and Position Weight Function which aim at reflecting the composition and permutation information of a sequence. An alternate training algorithm is provided to get an optimum description of the building patterns of the sequences. In this paper, dual descriptor method has been applied to the analysis of two typical problems of molecular biology: gene identification and the prediction of protein function. Satisfactory and insightful results are achieved. Owing to the generality of this methodology, dual descriptor method has wide application perspective for many problems of pattern recognition, especially those involved in “Systems Biology”.

Keywords: Biological sequence analysis; Gene identification; Protein function prediction; Pattern Recognition; Systems Biology

1. Background

After the completion of human genome sequencing project, we step into the post-genomic era (Comment, 2004; Lander, et al., 2001; Venter, et al., 2001). One ambitious goal of this era is to establish a "unified biology" which aims at unifying the biological disciplines from microscopic "molecular biology" to macroscopic "population biology" and even the "evolutionary biology" to achieve a profound comprehension of life (Nature Editorial, 2001). It has been acknowledged that the unification of biology lies in a synthesis of the human knowledge about biological systems (McDaniel and Weiss, 2005), which recently gives birth to a new field called "Systems Biology" (Aebersold, 2005; Church, 2005; Liu, 2005). The core spirit of systems biology is the systematical paradigm which is different from the traditional analytic paradigm where much attention is paid to the components of a system while little is known about the wholeness of the system as an ordered organization. The appearance of the subject "Systems Biology" milestoned a paradigm shift from analytic to systematic in the field of biological research.

From the systematical perspective, an abstract system *S* is modeled as a two element set $S = \{ E, R \}$, where *E* is the elements (building blocks) of the system and *R* is the mutual relations (building rules) between these elements. The building blocks of a system indicate "what are used" for the build of the system while the building rules of the system indicate "how to build" the system. This abstract system model is usually represented as a network (a graph from the perspective of Combinatorics) with *E* to be vertexes and *R* to be edges which has been exemplified by many

biological networks such as molecular metabolic networks (Fiehn and Weckwerth, 2003; Kell, 2004), gene regulation networks (Klemm and Bornholdt, 2005; Schlitt and Brazma, 2005) and protein interaction networks (Pellegrini, et al., 2004; Rousseau and Schymkowitz, 2005) *etc*.

Life is organized as a multi-leveled living system from the microscopic molecular components of a cell to the macroscopic population of individuals. Each level can be modeled as an abstract system, including the outcome of a genome sequencing project: biological sequence, which is the coarse-grained representation of polynucleotide or polypeptide at molecular level. If a biological sequence is viewed as a system, its building blocks are the nucleic acids or amino acids whose occurrence number carries the composition information of the sequence, and its building rules are the mutual relations between these molecules which reflect the permutation information of the sequence. Therefore the partition of the information of a character sequence into composition and permutation (Ma, 2007) is consistent with the systematical paradigm.

In the present work, we devise a novel generic methodology, called “dual descriptor method”, for biological sequence analysis by formulizing the ideas given in Ma (2007) from a systematical viewpoint. This methodology is used for the study of two typical problems of computational molecular biology: gene identification and the prediction of protein function. The average accuracy for the coding/non-coding recognition is more than 95% on a dataset of five bacterial genomes. And through the analysis by using dual descriptors, it is found that protein function prediction cannot be easily accomplished because there is high similarity between the building patterns

of the proteins of different functional groups. The results achieved are satisfactory and insightful, demonstrating the applicability of the present methodology to biological sequence analysis. Due to the generality of this methodology, it is also useful to many problems of pattern recognition, especially those involved in "Systems Biology".

2. Methodology

2.1. Quantization of character sequence

Suppose a character set $C=\{c_1,c_2,\cdots,c_i,\cdots,c_n\}$ $(n\geq 2)$. We use *C** to represent the set of the sequences composed of the characters in *C* and with finite lengths. Designate an order for the characters in *C* and then the order numbers constitute an integer set $N=\{1,2,\cdots,i,\cdots,n\}$. Construct the map $m:C\to N$. There are totally $n!$ kinds of such maps. Choosing any one of them, we can quantize the character sequence. For example, we can choose the following one:

$$m_1:C\ \to\ N\ =\ m_1:c_i\ \to\ i=\begin{matrix}\{\ c_1\ c_2\ \cdots\ c_i\ \cdots\ c_n\ \}\\ \updownarrow\ \updownarrow\ \cdots\ \updownarrow\ \cdots\ \updownarrow\\ \{\ 1\ \ 2\ \cdots\ \ i\ \cdots\ \ n\ \}\end{matrix} \tag{1}$$

Under this map, for each sequence $s\in C^*$, there is a corresponding number *N*. When $s=\varepsilon$ (null sequence), define $N=0$; when $s=c_{i_k}c_{i_{k-1}}\cdots c_{i_0}$,

$$N=i_k\cdot n^k+i_{k-1}\cdot n^{k-1}+\cdots+i_1\cdot n+i_0 \tag{2}$$

where $1\leq i_j\leq n,\ j=0,1,\cdots,k$. Suppose the length of the sequence *s* is *L*, then Eq. (2) can be rewritten as:

$$N=\sum_{k=1}^{L}n^{L-k}\cdot i_{c_k} \tag{3}$$

where i_{c_k} is the number *i* corresponding to the character c_k which appears at the

position k in the sequence s. Therefore, under the map m_1, for any sequence in C^*, there is a unique non-negative integer corresponding to it, i.e., via map m_1, we obtain a one-to-one map between C^* and Z^+, where Z^+ represents the non-negative integer set.

2.2. Definition of Dual Descriptor

On C^*, Dual Descriptor (DD) is defined as a two element set:

$$DD = \{M,\ P\} \tag{4}$$

where M is the Composition Weight Map (CWM) and P is the Position Weight Function (PWF) which are used to reflect the two aspects of information of a character sequence: composition and permutation.

M is a map from the character set C to a real number set X, i.e., $M: C \to X$, $X = \{x_1, x_2, \cdots, x_i, \cdots, x_n \mid (x_i \in R - \{0\})\}$; $M: C \to X$ is an extension of $m: C \to N$ for the image set from integer to real numbers. P is a real-valued function of position k in the character sequence s; $P(k)$ reflects the weight endued to the position k; $P(k)$ is an extension of the default position weight function of base n number system $I(k) = n^{L-k}$.

2.2.1. Pattern Description Function

For a sequence s composed of the characters in C and with length L, under the map $M: C \to X$, it can be converted into a real number sequence x, that is:

$$\begin{aligned} s &= [s[1], s[2], \cdots, s[k], \cdots, s[L]] \\ &\Downarrow \\ x &= [x[1], x[2], \cdots, x[k], \cdots, x[L]] \end{aligned} \tag{5}$$

where

$$x[k]=\begin{cases} x_1 & \text{if } (s[k]=c_1) \\ x_2 & \text{if } (s[k]=c_2) \\ \quad\vdots & \\ x_i & \text{if } (s[k]=c_i) \\ \quad\vdots & \\ x_n & \text{if } (s[k]=c_n) \end{cases} \qquad (k=1,2,\cdots,L;\ x_i\in X,\ c_i\in C)$$

For the character sequence *s*, its pattern description function is defined as:

$$N(k)=P(k)\times x[k] \quad (k=1,2,\cdots,L) \tag{6}$$

where the coefficient $P(k)$ before $x[k]$ is the position weight function.

2.2.2. Dual Formula

The sum of the first *l* items of *N*(*k*) is

$$S(l)=\sum_{k=1}^{l}N(k)=\sum_{k=1}^{l}P(k)x[k]=\sum_{x_i\in X}x_i\sum_{k_{x_i}}P(k_{x_i}) \tag{7}$$

where k_{x_i} represents the position *k* where c_i appears. $S(l)$ indicates some kind of dual relation and thus is called Dual Formula or Dual Variable. To know the dual relationship indicated by $S(l)$, see the following two special cases:

1) If the compositions are equally weighted, i.e., $x_i=\text{constant}=1 \quad (x_i\in X)$, then

$$S_{\bar{c}}(l)=\sum_{x_i\in X}1\sum_{k_{x_i}}P(k_{x_i}) \tag{8}$$

Let $P_{x_i}=\sum_{k_{x_i}}P(k_{x_i})$, then $S_{\bar{c}}=\sum_{x_i\in X}P_{x_i}$, i.e., P_{x_i} $(x_i\in X)$ represent the contribution of the character c_i to the sum from 1 to *l* of the value of the position weight function. Actually, P_{x_i} $(x_i\in X)$ is the position weighted frequencies when it is normalized by the sequence length *L*, which constitutes the "permutation part" of dual variable.

2) If the positions are equally weighted, i.e., $P(k)=\text{constant}=1 \quad (k=1,2,\cdots,L)$, then

$$S_{\bar{p}}(l) = \sum_{x_i \in X} x_i \sum_{k_{x_i}} 1 \tag{9}$$

Let $n_{x_i} = \sum_{k_{x_i}} 1$, and then n_{x_i} indicates the occurrence number of character c_i in the sequence *s*, namely, $\sum_{x_i \in X} n_{x_i} = l$. And then, $S_{\bar{p}} = \sum_{x_i \in X} x_i n_{x_i}$. Because n_{x_i} is just the occurrence number of the character c_i in the sequence *s*, x_i, multiplying on n_{x_i}, indicates the weight of n_{x_i} in the composition of the sequence *s*, and thus x_i $(x_i \in X)$ are called Composition Weight Factors (CWF), which constitute the "composition part" of dual variable.

The above two cases are two special cases where either the composition or the position is equally weighted. In general case, of course, neither of them (composition and position) is equally weighted.

2.2.3. Target Pattern and Standard Pattern

When both the composition and the position are equally weighted, i.e., $P(k) = \text{constant}$ and $x[k] = \text{constant}$, the pattern description function is also a constant: $N(k) = P(k)x[k] = \text{constant}$ $(k = 1, 2, \cdots, L)$, which is called Target Pattern. Without losing generality, suppose the constant = 1, then $N(k) = 1$ $(k = 1, 2, \cdots, L)$, which is called Standard Pattern.

2.3. Training DD to describe patterns of character sequence

Dual Descriptor can be trained on datasets. The training process of DD is the process of feature extraction from character sequence, which is implemented by minimizing the pattern deviation of a character sequence from a target pattern.

2.3.1. Pattern Deviation Function

To describe the pattern deviation of a sequence, we defined the Pattern Deviation

Function (PDF) as:

$$d = \frac{1}{L}\sum_{k=1}^{L}\left(N(k)-t\right)^2 \tag{10}$$

which represents the deviation of a sequence (whose pattern is described by $N(k)$) from a target pattern t. when $t=1$, d represents the deviation of the sequence from the standard pattern: $N(k)=1 \quad (k=1,2,\cdots,L)$.

2.3.2 Minimization of Pattern Deviation Function

The training of a DD is to minimize d. Substitute Eq. (6) into Eq. (10), we get

$$d = \frac{1}{L}\sum_{k=1}^{L}\left(P(k)x[k]-t\right)^2 . \tag{11}$$

$P(k)$ can be expanded on a set of basis functions $b_\gamma(k) \quad (\gamma=1,2,\cdots,m)$, i.e.,

$$P(k)=\sum_{\gamma} a_\gamma b_\gamma(k) \quad (\gamma=1,2,\cdots,m), \tag{12}$$

in which a_γ is independent of k and $b_\gamma(k) \quad (\gamma=1,2,\cdots,m)$ is the m items of the basis functions. The coefficients (a_γ) in the expanded form of the position weight function $P(k)$ are abbreviated as PWC (Position Weight Coefficients).

For a given CWM, $x_i \quad (x_i \in X_0)$ are constants. To minimize d, from $\frac{\partial d}{\partial a_\gamma}=0$, we get

$$\begin{aligned} u_{\alpha\beta} &= \sum_{k=1}^{L} b_\alpha(k)b_\beta(k)x[k]^2 \\ v_\alpha &= \sum_{k=1}^{L} b_\alpha(k)x[k] \end{aligned} \qquad (\alpha,\beta=1,2,\cdots,m) \tag{13}$$

where $b_\alpha(k)$ and $b_\beta(k)$ are the α-th and β-th basis function, respectively, and $x[k]\in X_0$ is the number at the k-th position in the real number sequence x. The coefficients of $P(k)$ can be written as a vector $\mathbf{a}$ which can be obtained from the

matrix $\mathbf{u}$ and the vector $\mathbf{v}$:

$$\mathbf{a} = \mathbf{u}^{-1}\mathbf{v} \tag{14}$$

where $\mathbf{a} = (a_1, a_2, \cdots, a_\gamma, \cdots, a_m)$ and the matrix $\mathbf{u}$ and the vector $\mathbf{v}$ are composed of the elements $u_{\alpha\beta}$ and v_α.

For a given PWF ($P_0(k)$), $a_\gamma(k)$ ($\gamma = 1, 2, \cdots, m$) are constants. To minimize *d*, from $\frac{\partial d}{\partial x_i} = 0$, we arrive at:

$$x_i = \frac{\sum_{k_{x_i}} P_0(k_{x_i})}{\sum_{k_{x_i}} P_0^2(k_{x_i})} \tag{15}$$

where k_{x_i} represents the position *k* in the real number sequence *x* where x_i appears.

2.3.3 Extracting common features of multiple sequences

For the situation of multiple sequences, to extract the common features of these sequences, the PDF for the ensemble of these sequences is defined as the mean value of the PDF values of these sequences:

$$D = \frac{1}{n}\sum_{j=1}^{n} d_j \tag{16}$$

where *n* is the number of the sequences, and $d_j = \frac{1}{L_j}\sum_{k=1}^{L_j}(N_j(k)-1)^2$ is the PDF value for the *j*-th sequence with the length L_j, and $N_j(k)$ is the pattern description function of the *j*-th sequence.

For a given CWM, from $\frac{\partial D}{\partial a_\gamma} = \sum_{j=1}^{n}\frac{\partial d_j}{\partial a_\gamma} = 0$, we arrive at:

$$\begin{aligned} U_{\alpha\beta} &= \sum_{j=1}^{n} u_{\alpha\beta}^{j} \\ V_{\alpha} &= \sum_{j=1}^{n} v_{\alpha}^{j} \end{aligned} \qquad (\alpha, \beta = 1, 2, \cdots, m). \tag{17}$$

At this time, the coefficient vector $\mathbf{a}$ is given by

$$\mathbf{a} = \mathbf{U}^{-1}\mathbf{V} \tag{18}$$

where the matrix $\mathbf{U}$ and the vector $\mathbf{V}$ are the sums of $\mathbf{u}$ and $\mathbf{v}$ (that are for single sequence), respectively.

Similarly, for a given PWF, from $\frac{\partial D}{\partial x_i} = \sum_{j=1}^{n} \frac{\partial d_j}{\partial x_i} = 0$, we arrive at:

$$x_i = \frac{\sum_j \sum_{k_{x_i}} P(k_{x_i})}{\sum_j \sum_{k_{x_i}} P^2(k_{x_i})}. \tag{19}$$

2.3.4. Choice of the basis functions

Basis functions in Eq. (12) are usually chosen as periodic functions, such as trigonometric function

$$P(k) = \sum_m a_m \cos\frac{2\pi k}{m} \qquad (m = 2, 3, \cdots) \tag{20}$$

or

$$P(k) = e^{k \bmod m} \qquad (m = 2, 3, \cdots), \tag{21}$$

because periodicity is a main difference between an ordered and an random sequence. Note that *m* starts from 2 rather than 1, because one – periodicity is the simple repeat of the same thing and thus meaningless here (just as in the case of the above character set *C* where *C* must have at least two elements because one element is unable to encode any information).

If the local information around some position in the sequence is concerned,

wavelet functions can also be considered, such as:

$$P(k) = \sum_{\alpha} \sum_{\beta} a_{\alpha,\beta} \psi(\alpha k - \beta) . \quad (22)$$

When $\alpha = 2^j$ and $\beta = n$, $P(k)$ is the usual discrete binary wavelet, and at that time, $a_{\alpha,\beta} =< P(k), \psi_{\alpha,\beta}(k) >$. The choice of wavelet function may extend the description ability of dual descriptor to a multileveled system owing to the multi-scaled analysis ability of wavelet as in the Mallat Algorithm (Mallat, 1999), which will be studied in the future.

2.3.5. Alternate training Process

The alternate training process for a DD on a sequence (or a set of sequences) consists of the following steps:

Step (1): Preparing a dataset composed of one or multiple sequences; set the maximum step number $N_{\max}$.

Step (2): Randomly construct a CWM, i.e, assign a random real number to each of the characters in *C* and these real numbers constitute the set of $x_i \quad (x_i \in X)$, and then use the minimization condition in Eq. (14) or Eq. (18) to obtain a corresponding PWF, namely, a set of coefficients $a_\gamma(k) \quad (\gamma = 1, 2, \cdots, m)$.

Step (3): With the set of coefficients $a_\gamma(k) \quad (\gamma = 1, 2, \cdots, m)$, use the minimization condition in Eq. (15) or Eq. (19) to obtain a CWM, namely, a set of $x_i \quad (x_i \in X)$; Generally speaking, the set of $x_i \quad (x_i \in X)$ obtained in this step is not the same as those are used in Step (1).

Step (4): Repeat Step (2) and Step (3) until the stop condition is satisfied.

Stop condition: the minimum *d* (or *D*) is achieved, or the maximum step number

N_{max} is reached.

In the training process, d (or D) becomes smaller and smaller. When d (or D) reaches its minimum value, the training process stops and an optimum DD is obtained on the dataset of the sequences used. The flowchart for the training of a DD on a dataset is illustrated in Fig. 1 (a).

Dual descriptor method can also be viewed as a kind of machine-learning approach. Different from other machine-learning approaches where local minimums are ubiquitous and cannot be tackled readily, the present method does not yield local minimum in principle because the PDF (Eq. (10)) is a quadric function and has only a unique global minimum. An object-oriented implementation of Dual Descriptor (the DD class written in Python language), which wraps the alternate training algorithm, is freely available from the author upon request.

2.4. Using DD as a sequence classifier

Because the result of the alternate training process is the acquirement of an optimum DD, the resultant DD carries the common features of the sequences that are used for the training, and then can be used as a classifier to identify sequence. For a sequence s to be identified, just use the resultant optimum DDp and DDn, which are obtained by training on the positive sample dataset and the negative sample dataset respectively, to calculate the PDF value d_p and d_n of this sequence, respectively. The classification of the sequence s is just to see which one of d_p and d_n is the smaller. If d_p is less than d_n, then the sequence s belongs to the Class P (positive sample); otherwise, the sequence s belongs to the Class N (negative sample). Fig. 1 (b) shows

the flowchart of using DD as a sequence classifier.

2.5. Evaluating the accuracy of classification

Three indicators are used to assess the classification ability of dual descriptor: Sensitivity, Specificity and Accuracy, which are commonly adopted in the evaluation of gene identification algorithms (Burset and Guigo, 1996). Sensitivity (Sn) represents the proportion of positive samples that have been correctly recognized as positive and Specificity (Sp) represents the proportion of negative samples that have been correctly recognized as negative. That is:

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \text{and} \qquad \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}, \tag{23}$$

where TP is the True Positive and FN is the False Negative and TN is the True Negative and FP is the False Positive. Accuracy (Ac) is simply defined as the average of Sn and Sp, i.e.,

$$\mathrm{Ac} = \frac{\mathrm{Sn} + \mathrm{Sp}}{2}. \tag{24}$$

For the details of the definition of these indicators, see Burset and Guigo (1996).

3. Materials

The annotated protein coding genes for five bacteria genomes *Wigglesworthia brevipalpis* (NC_004344), *Lactococcus lactis* (NC_002662), *Escherichia coli* K12 (NC_000913), *Bifidobacterium longum* (NC_004307) and *Streptomyces coelicolor* (NC_003888) were downloaded from ftp://ftp.ncbi.nih.gov/genomes/bacteria (corresponding to GenBank Release 153.0). For reliability, those genes with unknown or uncertain protein products (namely, those genes with “hypothetical”, “predicted”,

"probable" or "possible" in their protein product annotation) are removed from the dataset. For each species, a half of the genes are used as the training set and the other half as the test set in each round of the cross-validation tests.

The functional annotation for the proteins in the five bacteria genomes is according to COG. There are totally 25 functional groups in the updated version of COG database (Tatusov, et al., 2003; Tatusov, et al., 1997). For the purpose of demonstration, we focus on two functional groups: the functional group "translation, ribosomal structure and biogenesis" (symbolized by J) and the functional group "transcription" (symbolized by K). The protein sequences for the five bacteria genomes are extracted from their corresponding NC_xxxxxx.faa files that locate at the above directory. Some basic information of these genomes is listed in Table 1.

Table 1. Some basic information of the five bacteria genomes used in this study [a]

Species	Abbreviation	NC_number	GC%	Gene Number	Protein Number in J	Protein Number in K
Wigglesworthia brevipalpis	W. bre	NC_004344	22.5	126	107	21
Lactococcus lactis	L. lac	NC_002662	35.3	1466	114	119
Escherichia coli K12	E. col	NC_000913	50.8	2474	164	225
Bifidobacterium longum	B. lon	NC_004307	60.1	687	101	99
Streptomyces coelicolor	S. coe	NC_003888	72.1	5001	226	801

[a] The GC content (GC%) varies from the lowest 22.5 of *Wigglesworthia brevipalpis* to the highest 72.1 of *Streptomyces coelicolor* which covers the whole range of GC contents of the species currently available in the public database. The column "Gene Number" indicates the number of protein coding gene after removing those genes with uncertain protein products. The last two columns "Protein Number in J" and "Protein Number in K" indicate the protein numbers in the two functional groups J and K (according to the functional classification of COG database), respectively (including those proteins with uncertain annotation but the COG functional information can be inferred).

4. Results

4.1. Using DD to recognize protein coding genes

The first step for dealing with the genomic data that are acquired by sequencing projects is undoubtedly to find the protein coding genes. A variety of algorithms has been developed to fulfill this purpose and they are summarized in Ma (2007). Dual descriptor method, as a new methodology of sequence analysis proposed from the systematic viewpoint, can be added to the stock of these powerful tools. Owing to its "holograph" ability that not only the composition but also the permutation information of a sequence can be reflected, dual descriptor can be used for the task of gene identification.

Table 2. The sensitivity, specificity and accuracy for coding/non-coding sequence

recognition in self-tests and ten-fold cross-validation tests for the five bacterial genomes [a]

Species	Self-tests			Cross-validation tests		
	Sensitivity (%)	Specificity (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	Accuracy (%)
W. bre	97.62	99.21	98.41	98.57±1.75	97.94±1.07	98.25±1.23
L. lac	96.32	99.18	97.75	97.62±0.35	98.85±0.23	98.24±0.24
E. col	93.21	98.02	95.61	93.75±0.55	97.89±0.24	95.82±0.31
B. lon	96.80	99.42	98.11	96.40±0.58	98.28±0.77	97.34±0.55
S. coe	90.80	97.98	94.39	91.33±0.32	97.92±0.15	94.62±0.15

[a] In the Self-tests and the Cross-validation tests, the maximum training step number N_{max} (see Fig. 1) is set to be 5.

The results for the identification of protein coding genes from non-coding sequences are listed in Table 2. It can be found that the average accuracy is 96.85% in the self-tests and almost equally good in cross-validation tests, suggesting that the information encoded in the sequences has been captured by the dual descriptors in the alternate training process and that based on the specificity of this information for the coding genes, high accuracy of classification can be achieved. The accuracies for the five genomes are all more than 90% and insensitive to the GC contents of the genomes (GC% in the 4th column of Table 1), which shows the applicability of dual descriptor method to gene identification. It should be noticed that, after the training process, the accuracy for the identification of protein coding genes of *E. coli* K12 is improved from the original 91.64% (Ma, 2007) to the current 95.82% with an increase of near five percents, which demonstrates the validity of the alternate training process

to get an optimum description of the genes' common features. Fig. 2 gives an intuitive impression for the classification result using dual descriptor.

4.2. Using DD to describe the building patterns of protein coding genes

Nowadays, there are millions of protein coding genes that have been identified and stored in the public databases such as GenBank/EMBL/DDBJ (Benson, et al., 2005). With these already known genes, an equally important but less noticed task is to describe the features of genes to convert raw data into human understandable knowledge, which is a process of data-mining or knowledge-discovery. A proper methodology is necessary to achieve this goal. Dual descriptor is helpful to the implementation of this task owing to its "description" ability.

As shown above, there are two sets of parameters for a dual descriptor: Composition Weight Factors (CWF) and Position Weight Coefficients (PWC), corresponding to the two components of a dual descriptor, respectively, with the former to be the image set of Composition Weight Map (CWM) and the latter to be the coefficients of the expanded form of Position Weight Function (PWF). The intuitive meaning of the two sets of parameters (they correspond to building block usage and building rule usage, respectively) gives us the possibility to explore the building patterns of the sequences.

Table 3 shows the CWF and PWC of the obtained dual descriptors after the alternate training process, and a close-inspection of these parameters informs us of the building patterns of the protein coding genes. A CWF for a character carries the composition information of that character in the sequence. The larger its absolute

value is, the more important role the character plays in the building patterns of the sequence. From Table 3, it can be found that the first two largest (absolute value) CWF for the two AT-rich genomes (GC% < 50%) are x_G and x_T, while the first two largest CWF for the GC-rich genomes (GC% > 50%) are x_C and x_G, and for the AT & GC content-balanced genome (GC% = 50.8%), the first two largest CWF are x_T and x_G. Firstly, we notice that the common character in the top two CWF of the five genomes is G, which means guanine always plays important role in the building patterns of the protein coding genes no matter how rich its content is. Secondly, we can find that the pattern of T (with x_T to be the second largest CWF) in AT-rich genomes and the pattern of C (with x_C to be the second largest CWF) in CG-rich genomes are important ones for the building patterns of the sequences due to their abundance.

Table 3. The composition weight factors (CWF) and position weight coefficients (PWC) of the optimum dual descriptors for the five bacterial genomes

Species	CWF [a]				PWC (×100) [b]				
	x_A	x_C	x_G	x_T	a_2	a_3	a_4	a_5	a_6
W. bre	0.1614	2.956	-15.91	9.869	-0.0283	-2.6752	0.0436	0.0226	0.0190
L. lac	5.496	2.137	-44.09	38.70	-0.0106	-1.0885	0.0118	0.0164	0.0124
E. col	-3.069	0.9414	-16.12	28.56	-0.0142	-1.5072	0.0061	0.0111	0.0126
B. lon	-13.99	31.71	-29.52	26.24	-0.0077	-0.9506	0.0003	-0.0016	0.0122
S. coe	34.66	-60.82	51.89	-40.22	0.0042	0.4159	-0.0044	-0.0025	-0.0015

[a] x_N (N = A, C, G, T) are the composition weight factors whose absolute value indicates the contribution of the corresponding nucleotide to the building patterns of the protein coding genes. For all the five genomes, the initial values of CWF are the same, as: $x_A = 1$, $x_C = 2$, $x_G = 3$, $x_T = 4$.

[b] a_i (i = 2, 3, 4, 5, 6) are the coefficients of the expanded form of the corresponding position weight function whose absolute value indicates contribution of the periodicity – i in the building patterns of the sequences. For saving the print space, the PWC listed in this table have been amplified 100 times.

Likewise, PWC carry the permutation information: the larger the absolute value of a PWC is, the stronger the corresponding periodicity is in the building patterns of the sequence. From Table 3, it can be found that the three – periodicity (a_3) is the strongest signal and common in all the genomes used, which is acknowledged as the reflection of the triplet codes in coding sequences (Gutierrez, et al., 1994; Kobayashi, et al., 2003; Shepherd, 1981) and shows the unity of the building rules of protein coding genes. Besides, the periodicities 4 and 5 (a_4 and a_5) are relatively stronger than other periodicities in the building pattern of the coding sequences of the AT-rich genomes, while the periodicities 6 and 4 (a_6 = 0.0122 for *B. lon* and a_4 = -0.0044 for S. coe) are the stronger signals than other periodicities in the two GC-rich genomes, respectively. The genome with balanced AT & GC contents has the second strongest signal of two – periodicity (a_2 = -0.0142 for *E. col*). Additionally, the four – periodicity (a_4 = 0.0003 for *B. lon*) is particularly weak in the building pattern of the protein coding genes of the human gastrointestinal tract resident, *Bifidobacterium longum*. These differences, on the other hand, reflect the diversity of the building

rules of protein coding genes, which may embody the adaptation of the organisms to their living environments.

Owing to the alternate training process, a CWF also carries the permutation information by its sign. According to Eq. (15), we know that the positive sign of a CWF means that the corresponding character more frequently appears at the positions where position weight function get positive values while a negative sign of a CWF means the opposite situation. For example, according to Eq. (20) and the PWC listed in Table 3, the position weight function of the dual descriptor for the *E. coli* K12 genome can be written as:

$$\begin{aligned} P(k) &= \sum_{i=2}^{6} a_m \cos(\frac{2\pi k}{m}) \\ &= -0.0142\cos(\pi k) - 1.5072\cos(\frac{2\pi k}{3}) + 0.0061\cos(\frac{\pi k}{2}) \\ &+ 0.0111\cos(\frac{2\pi k}{5}) + 0.0126\cos(\frac{\pi k}{3}). \end{aligned} \tag{25}$$

Because $P(k)$ is approximated by trigonometric functions that are periodic, the approximation of $P(k)$ is also a periodic function whose periodicity is the least common multiple (LCM) of those of its expanded items. Since we take the first five items (from 2 to 6) of the series (20) as the approximation, the periodicity of $P(k)$ here is 60 which is the LCM of the periodicities 2, 3, 4, 5, 6. The values of $P(k)$ on the interval of [0, 59] (one period) are listed in Supplementary Table 1. A detailed examination of this list informs us that $P(k)$ gets negative values at the positions that can be divided exactly by 3 while gets positive values at other positions. Sifting the CWF for *E. col* in Table 3, we know that x_A and x_G take negative values, indicating that they appear more frequently at the locations of 0, 3, 6, ,…, and so on, meanwhile,

x_C and x_T take positive values, indicating that they have more preference for other positions. See Supplementary Fig. 1 for an intuitive impression of the building patterns reflecting such features in the *E. coli* K12 genome. Fig. 3 illustrates the meaning of CWF as the position weighted contents by a linear regression.

4.3. Using DD to analyze the problem of protein function prediction

One important task of molecular biology in post-genomic era is the annotation of protein function. The basic methods for this task are experimental techniques such as gene inactivation and microarray expression (Collins, et al., 2003; Kobayashi, et al., 2003) *etc*. When the experimental data accumulate to a certain degree, theoretical prediction of protein function is also possible by using a knowledge-based approach, which relies on the similarity between the newly sequenced and the function-known proteins. Dual descriptor can help answer the question of "to what degree can such a similarity-based approach be successful" owing to its analysis ability.

Table 4. The sensitivity, specificity and accuracy for the five genomes used for the analysis of protein function prediction

Species	Fun J vs. Fun K [a]			Fun J vs. Rand [b]			Fun K vs. Rand [c]			Rand J vs. Rand K [d]		
	Sn	Sp	Ac	Sn	Sp	Ac	Sn	Sp	Ac	Sn	Sp	Ac
W. bre	78.50	80.95	79.73	77.57	68.22	72.90	90.48	71.43	80.95	74.77	80.95	77.86
L. lac	60.53	66.39	63.46	75.44	57.02	66.23	79.83	63.87	71.85	76.32	73.11	74.71
E. col	59.76	66.67	63.21	78.05	62.20	70.12	75.11	65.78	70.44	77.44	63.56	70.50

B. lon	56.44	59.60	58.02	78.22	67.33	72.77	69.70	78.79	74.24	71.29	70.71	71.00
S. coe	59.29	59.93	59.61	79.65	62.39	71.02	81.02	61.67	71.35	82.74	64.04	73.39
Average	62.90	66.71	64.81	77.79	63.43	70.61	79.23	68.31	73.77	76.51	70.47	73.49

[a] The column "Fun J vs. Fun K" lists the evaluation indicators (Sensitivity, Specificity and Accuracy) for the identification between the protein sequences belonging to the two functional groups J and K of COG.

[b] The column "Fun J vs. Rand" lists those indicators for the identification between the protein sequences of the functional group J and their corresponding randomly shuffled sequences.

[c] The column "Fun K vs. Rand" lists the indicators for the identification between the protein sequences of the functional group K and their corresponding randomized sequences.

[d] The last column "Rand J vs. Rand K" lists the indicators for the identification between the two groups of randomized sequences which are generated by shuffling the corresponding protein sequences of the two functional groups, respectively.

The evaluation indicators for the protein function prediction from amino acid sequence are listed in Table 4. As shown, the average accuracy for the prediction of the two functional groups J and K is 64.81% and that for the identification of the functional sequences from their corresponding randomized sequences is 70.61% and 73.77%, respectively. That the average accuracy (64.81%) for the identification between the proteins of the two functional groups is less than those (70.61% and 73.77%) for the identification between the functional sequences and random sequences means that the protein sequences for the two functional groups are more similar to each other than their similarity to their corresponding randomized

sequences (with the same amino acid composition). Another point should be noticed is that, even the accuracy (73.49%) for the identification between the two sets of random sequences is higher than that (64.81%) for the identification between the two functional groups, which means at least two things: (1) the functional sequences have common features which differentiate them from random sequences; (2) the protein function prediction, at least for the two functional groups J and K used in the present study, cannot be easily achieved.

Table 5. The PWC for the building rules of the proteins of the two functional groups J and K in the five bacterial genomes

Species	PWC for Fun J (×100)					PWC for Fun K (×100)					p [a]
	a_2	a_3	a_4	a_5	a_6	a_2	a_3	a_4	a_5	a_6	
W. bre	4.88	2.33	8.52	7.30	-1.33	0.286	1.31	10.17	-6.61	6.77	0.62
L. lac	3.24	6.85	6.79	1.88	2.69	5.08	4.48	0.276	4.64	8.51	0.89
E. col	1.35	4.09	5.32	4.87	4.12	2.69	2.36	5.40	4.68	3.91	0.79
B. lon	4.28	-3.78	3.86	9.51	5.82	1.62	3.14	4.37	8.33	6.57	0.62
S. coe	2.01	6.62	2.55	6.89	2.74	1.26	2.39	3.58	3.48	2.37	0.19

[a] The last column "p" indicates the p – values in the paired t – test which is performed in a purpose of comparing the periodicities (a_2 - a_6) in the building patterns of the protein sequences of the two functional groups J and K.

To know the intrinsic difficulty of the task of protein function prediction, we can analyze the parameters of the obtained dual descriptor. The PWC for the five bacterial

genomes are listed in Table 5. Through a paired t – test, we find that the p – values for the five genomes are all larger than 0.1 which means that there is no significant difference between the building rules (revealed by periodicities) of the protein sequences of the two functional groups. That's why the ascertainment of protein function still mainly relies on experiments and why theoretical prediction are only applicable to those proteins that can be found high homolog to the function-known proteins by sequence alignments (Marti-Renom, et al., 2000; Wang, et al., 2000). There is still a long way to go to achieve an accurate prediction of a protein's function from its amino acid sequence.

5. Conclusion and further thinking

The above results demonstrate the application of dual descriptor method in two typical problems of computational molecular biology: gene identification and the prediction of protein function. In the former case, the building blocks are the nucleotides and the building rules of the protein coding genes are revealed as (short range) periodicities of these nucleotides in the building patterns of the sequences. In the latter case, the building blocks are the amino acids and the building rules of the protein sequences of the two functional groups J and K have no significant difference, which makes the protein function prediction a hard task to accomplish, at least for the two functional groups used in this study.

However, the protein function prediction is not impossible because there may be other ordered information hidden in the sequences which cannot be revealed by (short

range) periodicities but can differentiate the two functional groups of protein sequences. In fact, periodicity is only a kind of symmetry: translational symmetry. Other symmetries such as rotational symmetry or centrosymmetry can also be considered to reflect the ordered (non-random) information of a sequence (which will be studied later).

As a generic mathematical model, the systematic viewpoint of dual descriptor equips it with two sides of ability: owing to its “holograph” ability, dual descriptor can be used as a feature extractor and data classifier, while owing to it “description” ability, it can be used as a pattern finder and problem analyzer for any problems involving sequence analysis. In fact, many problems in “Systems Biology”, such as the identification of essential genes in the regulation networks or the prediction of “hub” proteins in protein interaction networks and so forth, can be analyzed by this methodology. Furthermore, many (if not any) systems can be represented by character sequences, especially in the simulation using computers where any objects are implemented in the form of binary sequences in the memory of the computing system. Therefore, the current methodology has a very wide scope of application.

Acknowledgement

This work was supported by the National Basic Research Program of China (2003CB114400) and the National Natural Science Foundation of China (30100035 and 30570383). The author is also supported by the fund of research project of Shandong University of Technology (2004KJM29).

Reference

Aebersold, R. 2005. Molecular Systems Biology: a new journal for a new biology? Mol Syst Biol 1, msb4100009-E4100001-msb4100009-E4100002.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. Nucl. Acids Res. 33**,** D34-38.

Burset, M., and Guigo, R. 1996. Evaluation of gene structure prediction programs. Genomics 34, 353-367.

Church, G.M. 2005. From systems biology to synthetic biology. Mol Syst Biol 1, 2005 0032.

Collins, J.E., Goward, M.E., Cole, C.G., Smink, L.J., Huckle, E.J., Knowles, S., Bye, J.M., Beare, D.M., and Dunham, I. 2003. Reevaluating Human Gene Annotation: A Second-Generation Analysis of Chromosome 22. Genome Res. 13, 27-36.

Comment, N. 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 931-945.

Editorial, N. 2001. Post-genomic cultures. Nature 409, 545.

Fiehn, O., and Weckwerth, W. 2003. Deciphering metabolic networks. Eur J Biochem 270, 579-588.

Gutierrez, G., Oliver, J.L., and Marin, A. 1994. On the origin of the periodicity of three in protein coding DNA sequences. J Theor Biol 167, 413-414.

Kell, D.B. 2004. Metabolomics and systems biology: making sense of the soup. Current Opinion in Microbiology 7, 296-307.

Klemm, K., and Bornholdt, S. 2005. Topology of biological networks and reliability of information processing. PNAS 102, 18414-18419.

Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., et al. 2003. Essential Bacillus subtilis genes. Proc Natl Acad Sci U S A 100, 4678-4683.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. Nature 409, 860-921.

Liu, E.T. 2005. Integrative biology and systems biology. Mol Syst Biol 1, msb4100008-E4100001-msb4100008-E4100001.

Ma, B.G. 2007. How to describe genes: enlightenment from the quaternary number system. Biosystems 90, 20-27.

Mallat, S. 1999. A Wavelet Tour of Signal Processing (Second Edition), Academic Press.

Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. 2000. Comparative protein structure modeling of genes and genomes. *Annu Rev* Biophys Biomol Struct 29, 291-325.

McDaniel, R., and Weiss, R. 2005. Advances in synthetic biology: on the path from prototypes to applications. Current Opinion in Biotechnology 16, 476-483.

Pellegrini, M., Haynor, D., and Johnson, J.M. 2004. Protein Interaction Networks. Expert Review of Proteomics 1, 239-249.

Rousseau, F., and Schymkowitz, J. 2005. A systems biology perspective on protein

structural dynamics and signal transduction. Current Opinion in Structural Biology 15, 23-30.

Schlitt, T., and Brazma, A. 2005. Modelling gene networks at different organisational levels. FEBS Letters 579, 1859-1866.

Shepherd, J.C. 1981. Periodic correlations in DNA sequences and evidence suggesting their evolutionary origin in a comma-less genetic code. J Mol Evol 17, 94-102.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tatusov, R.L., Koonin, E.V., and Lipman, D.J. 1997. A genomic perspective on protein families. Science 278, 631-637.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. Science 291, 1304-1351.

Wang, Y., Bryant, S., Tatusov, R., and Tatusova, T. 2000. Links from genome proteins to known 3-D structures. Genome Res. 10, 1643-1647.

Figure Legends

Fig. 1. Flowcharts for the alternate training process of DD and using DD as a classifier. (a) The alternate training process of DD on a dataset. “Randomly construct a CWM” means assigning a random number to each CWF. Usually, the assignment of the values for the CWF is according to the natural order of the characters in the alphabet, i.e., the map in Eq. (1) is used. (b) Optimum DD used as a classifier. “Optimum DDp” denotes the dual descriptor that is trained on the positive samples and obtained when the termination condition of the training process is reached, while “Optimum DDn” denotes that for the negative samples.

Fig. 2. Coding and non-coding sequences (of *E. coli* K12). The abscissa axis indicates the pattern deviation values of the sequences described by the optimum DDp and the ordinate axis indicates those values described by the optimum DDn. Red triangles represent the protein coding genes and black circles represent the non-coding sequences which are generated by shuffling the corresponding coding sequences. Because DDp is optimum for the building patterns of coding sequences, the coding sequences can get relatively less values than the non-coding sequences and thus the points representing the coding sequences locate at the left half of the figure while those representing the non-coding sequences locate at the right half of the figure. Similarly, because DDn is optimum for the building patterns of non-coding sequences, the non-coding sequences can get relatively less values than the coding-sequences and thus the points representing the non-coding sequences locate at the lower half of the

figure while those for the coding sequences locate at the upper half of the figure. The overlapped area between coding and non-coding constitutes the false positive and false negative of the recognition.

Fig. 3. The linear correlation between the sum of position weighted contents N*w* (on the interval of one period of the corresponding position weight function, here the period length is 60, see Supplementary Table 1 for details) and the CWF x_N of the Dual Descriptor for the protein coding genes of *E. coli* K12. The correlation coefficient is as high as 0.986, which indicates that the meaning of a CWF is the position weighted content of the corresponding character.

Figures

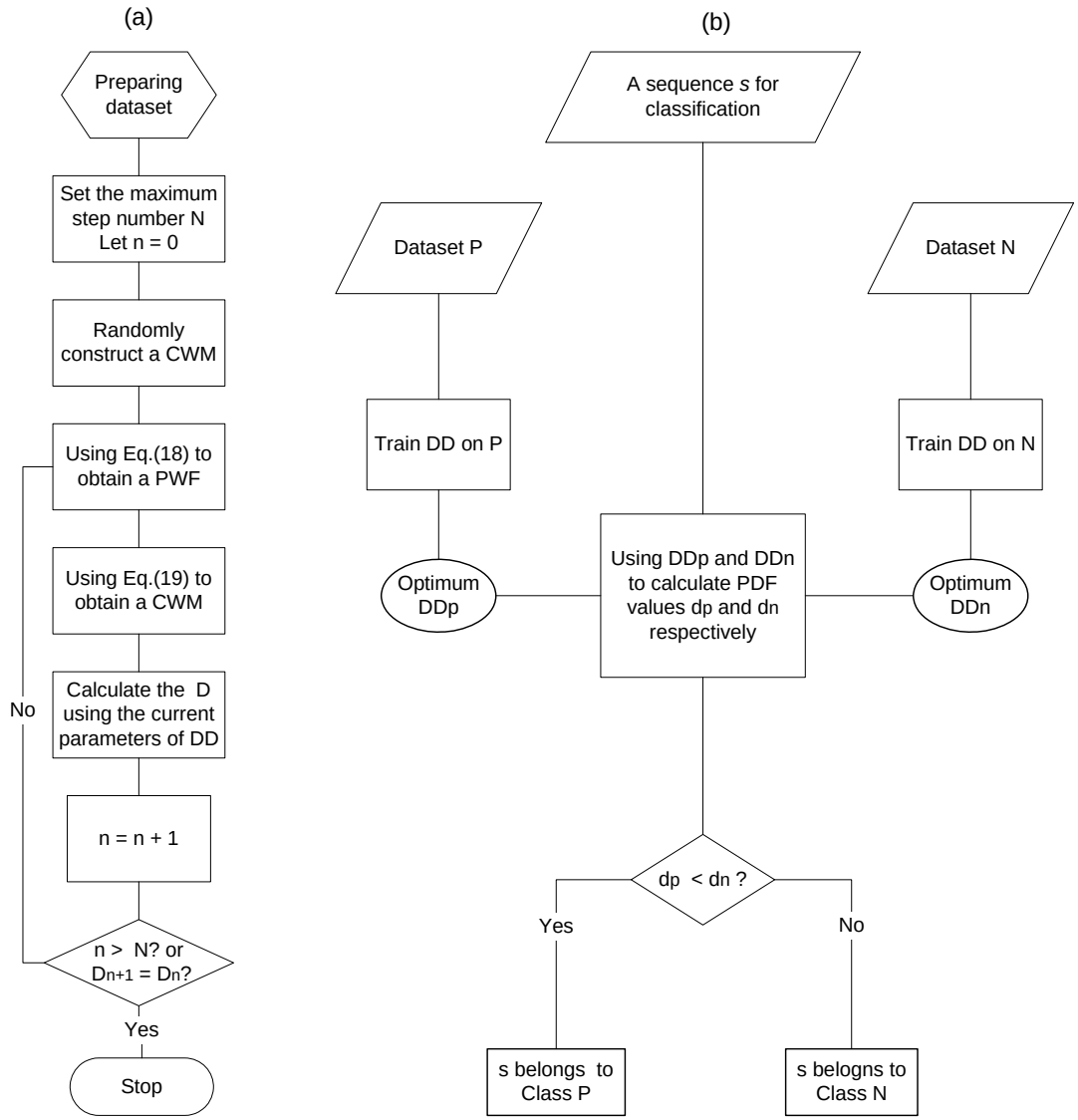


Fig. 1. Bin-Guang Ma

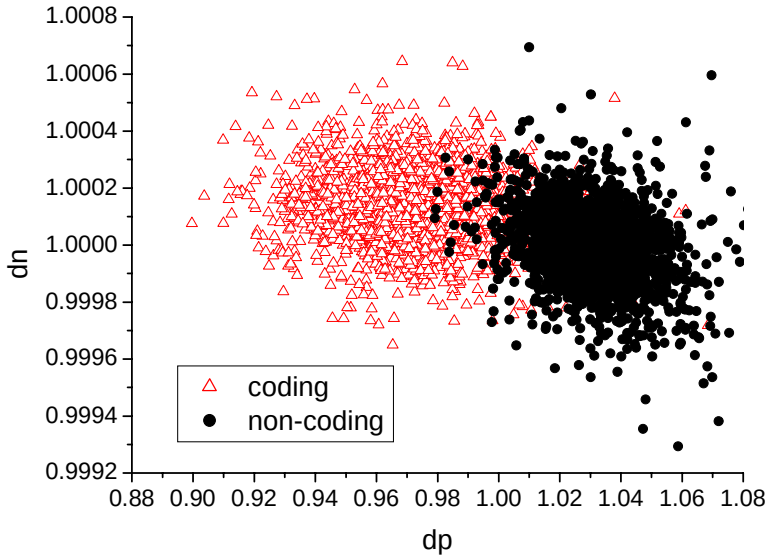


Fig. 2. Bin-Guang Ma

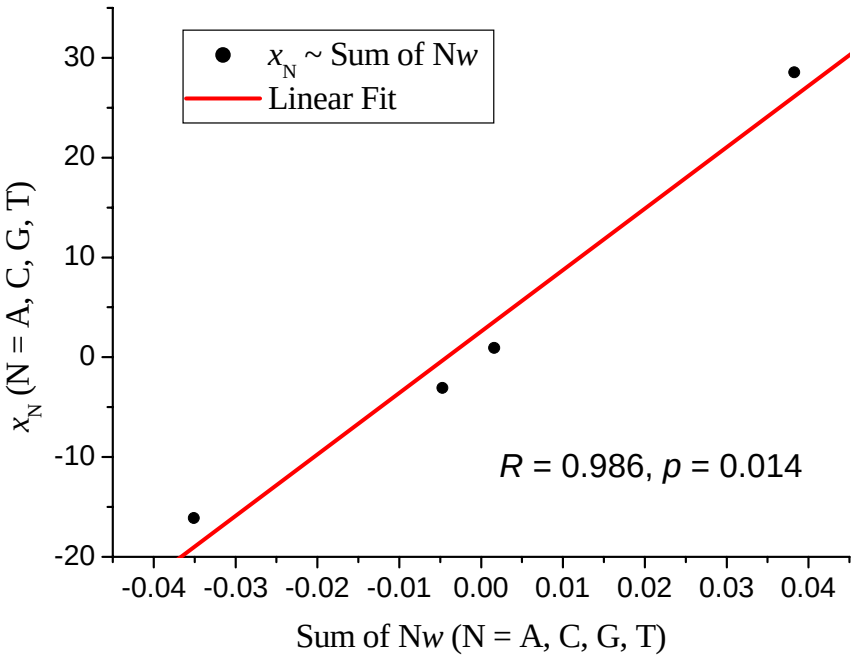
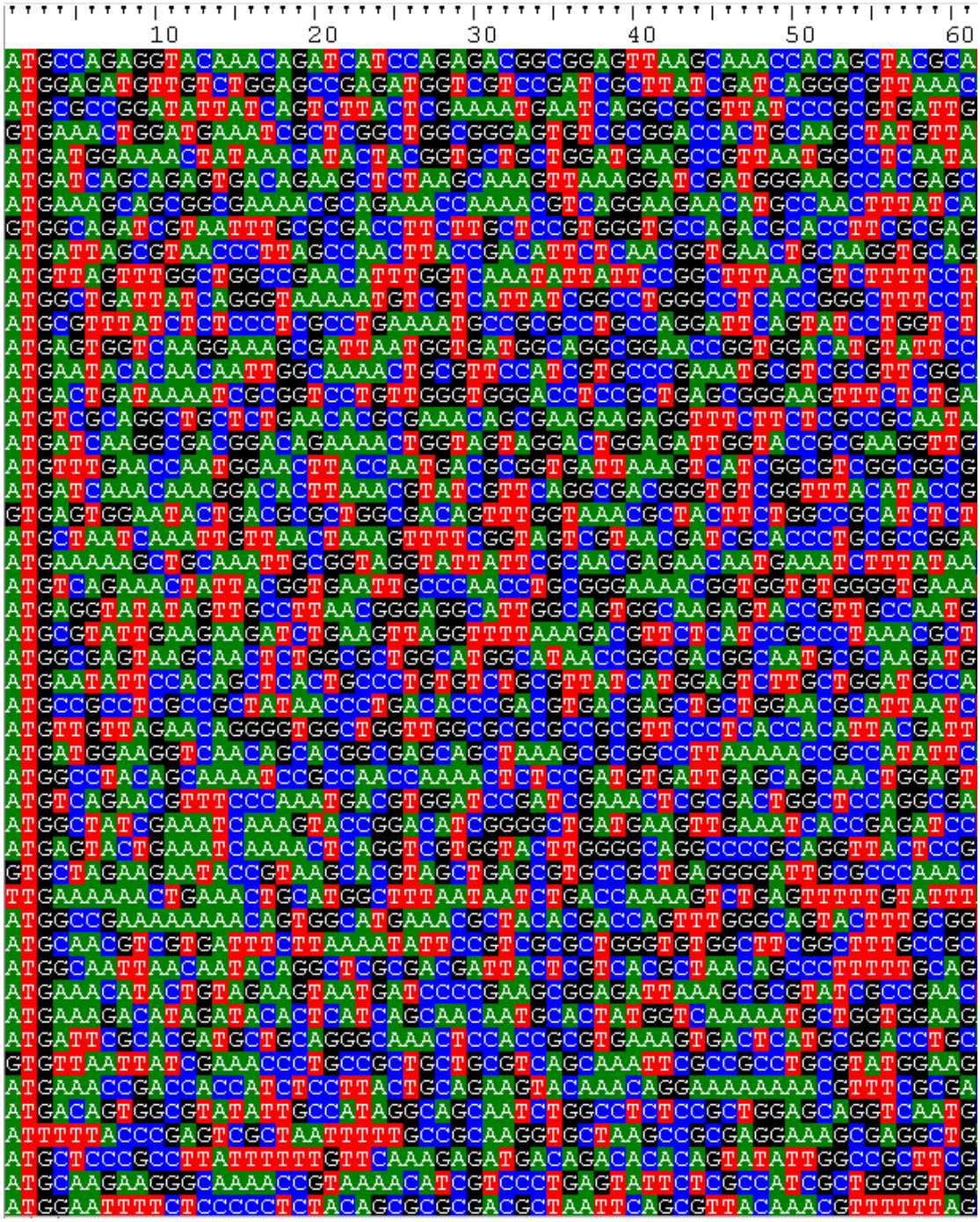


Fig. 3. Bin-Guang Ma

Supplementary Materials



Supplementary Fig. 1. Illustration of the building patterns of the protein coding genes of *E. coli* K12. Adenines locate at the positions with Green background color; Cytosines locate at the positions with Blue background color; Guanines locate at the positions with **Black** background color; Thymines locate at the positions with Red background color. The sequences are truncated from the beginning to the length of 60

that is one period of the corresponding position weight function. This figure is generated using BioEdit software (Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl. Acids. Symp. Ser. 41, 95-98).

Supplementary Table 1. The nucleotide contents and their corresponding position weighted contents at the positions within the length of one period of the position weight function (on the dataset of protein coding genes of *E. coli* K12)

k	*P*(*k*)		Nucleotide Contents				Position Weighted Contents			
	sign	value	A	C	G	T	A*w*	C*w*	G*w*	T*w*
0	-	0.01491531	0.280	0.232	0.345	0.143	-0.004	-0.003	-0.005	-0.002
1	+	0.00777518	0.272	0.218	0.171	0.338	0.002	0.002	0.001	0.003
2	+	0.0071802	0.166	0.264	0.334	0.236	0.001	0.002	0.002	0.002
3	-	0.01514622	0.256	0.244	0.354	0.147	-0.004	-0.004	-0.005	-0.002
4	+	0.00742621	0.296	0.236	0.180	0.288	0.002	0.002	0.001	0.002
5	+	0.0078519	0.183	0.270	0.290	0.257	0.001	0.002	0.002	0.002
6	-	0.0151139	0.256	0.244	0.350	0.150	-0.004	-0.004	-0.005	-0.002
7	+	0.00765104	0.301	0.226	0.176	0.297	0.002	0.002	0.001	0.002
8	+	0.00730207	0.181	0.273	0.289	0.256	0.001	0.002	0.002	0.002
9	-	0.01502208	0.250	0.248	0.348	0.153	-0.004	-0.004	-0.005	-0.002
10	+	0.00738107	0.294	0.225	0.178	0.303	0.002	0.002	0.001	0.002
11	+	0.00777518	0.178	0.274	0.293	0.255	0.001	0.002	0.002	0.002
12	-	0.01511617	0.251	0.246	0.352	0.152	-0.004	-0.004	-0.005	-0.002
13	+	0.00765104	0.297	0.226	0.178	0.299	0.002	0.002	0.001	0.002
14	+	0.00730434	0.176	0.279	0.289	0.256	0.001	0.002	0.002	0.002
15	-	0.01494535	0.248	0.244	0.357	0.151	-0.004	-0.004	-0.005	-0.002
16	+	0.00742621	0.291	0.230	0.180	0.299	0.002	0.002	0.001	0.002

17	+	0.00765104	0.176	0.276	0.291	0.256	0.001	0.002	0.002	0.002
18	-	0.01523804	0.248	0.248	0.351	0.153	-0.004	-0.004	-0.005	-0.002
19	+	0.00777518	0.295	0.224	0.181	0.299	0.002	0.002	0.001	0.002
20	+	0.00750293	0.175	0.281	0.290	0.253	0.001	0.002	0.002	0.002
21	-	0.01502208	0.250	0.247	0.352	0.152	-0.004	-0.004	-0.005	-0.002
22	+	0.0071802	0.291	0.227	0.182	0.300	0.002	0.002	0.001	0.002
23	+	0.00765104	0.174	0.278	0.290	0.257	0.001	0.002	0.002	0.002
24	-	0.01499203	0.246	0.248	0.354	0.152	-0.004	-0.004	-0.005	-0.002
25	+	0.0078519	0.288	0.228	0.183	0.301	0.002	0.002	0.001	0.002
26	+	0.00730434	0.168	0.28	0.299	0.254	0.001	0.002	0.002	0.002
27	-	0.01514622	0.24	0.247	0.361	0.152	-0.004	-0.004	-0.005	-0.002
28	+	0.00730207	0.291	0.230	0.178	0.301	0.002	0.002	0.001	0.002
29	+	0.00777518	0.171	0.280	0.294	0.254	0.001	0.002	0.002	0.002
30	-	0.01503718	0.243	0.247	0.361	0.149	-0.004	-0.004	-0.005	-0.002
31	+	0.00777518	0.290	0.228	0.180	0.302	0.002	0.002	0.001	0.002
32	+	0.00730207	0.173	0.278	0.296	0.253	0.001	0.002	0.002	0.002
33	-	0.01514622	0.243	0.253	0.353	0.151	-0.004	-0.004	-0.005	-0.002
34	+	0.00730434	0.290	0.223	0.182	0.305	0.002	0.002	0.001	0.002
35	+	0.0078519	0.171	0.277	0.298	0.253	0.001	0.002	0.002	0.002
36	-	0.01499203	0.242	0.247	0.357	0.154	-0.004	-0.004	-0.005	-0.002
37	+	0.00765104	0.289	0.222	0.186	0.302	0.002	0.002	0.001	0.002
38	+	0.0071802	0.171	0.276	0.297	0.255	0.001	0.002	0.002	0.002
39	-	0.01502208	0.241	0.247	0.361	0.150	-0.004	-0.004	-0.005	-0.002
40	+	0.00750293	0.294	0.224	0.183	0.299	0.002	0.002	0.001	0.002
41	+	0.00777518	0.171	0.276	0.296	0.257	0.001	0.002	0.002	0.002
42	-	0.01523804	0.246	0.245	0.359	0.150	-0.004	-0.004	-0.005	-0.002
43	+	0.00765104	0.295	0.224	0.182	0.299	0.002	0.002	0.001	0.002
44	+	0.00742621	0.175	0.281	0.296	0.247	0.001	0.002	0.002	0.002
45	-	0.01494535	0.245	0.248	0.359	0.148	-0.004	-0.004	-0.005	-0.002

46	+	0.00730434	0.288	0.226	0.184	0.302	0.002	0.002	0.001	0.002
47	+	0.00765104	0.173	0.279	0.293	0.255	0.001	0.002	0.002	0.002
48	-	0.01511617	0.248	0.246	0.355	0.151	-0.004	-0.004	-0.005	-0.002
49	+	0.00777518	0.289	0.227	0.183	0.301	0.002	0.002	0.001	0.002
50	+	0.00738107	0.175	0.278	0.290	0.257	0.001	0.002	0.002	0.002
51	-	0.01502208	0.247	0.244	0.357	0.152	-0.004	-0.004	-0.005	-0.002
52	+	0.00730207	0.289	0.227	0.183	0.301	0.002	0.002	0.001	0.002
53	+	0.00765104	0.173	0.279	0.299	0.249	0.001	0.002	0.002	0.002
54	-	0.0151139	0.244	0.250	0.355	0.152	-0.004	-0.004	-0.005	-0.002
55	+	0.0078519	0.289	0.223	0.186	0.302	0.002	0.002	0.001	0.002
56	+	0.00742621	0.170	0.278	0.298	0.254	0.001	0.002	0.002	0.002
57	-	0.01514622	0.238	0.246	0.359	0.156	-0.004	-0.004	-0.005	-0.002
58	+	0.0071802	0.287	0.226	0.186	0.300	0.002	0.002	0.001	0.002
59	+	0.00777518	0.171	0.275	0.299	0.255	0.001	0.002	0.002	0.002

Note: k denotes the positions in the coding sequences; $P(k)$ denotes the values of the position weight function at these positions; The "Nucleotide Contents" are the nucleotide frequencies at the positions i in the sequences where i mod 60 = k. The "Position Weighted Contents" Nw (N = A, C, G, T) are the "Nucleotide Contents" multiplied by the value of $P(k)$ at each position, i.e., Nw = N*$P(k)$. According to Eq. (15), Nw is correlated with the CWF: x_N (N = A, C, G, T) listed in Table 3 of the manuscript. The linear correlation coefficient is shown on Fig. 3.