Journal of Fiber Bioengineering and Informatics 8:2 (2015) 347–355
doi:10.3993/jfbim00109

A B-spline Quasi-interpolation EMD Method for Similarity/Dissimilarity Analysis of DNA Sequences ⋆

Junsheng Zheng [a,b], Min Xu [c], Jihong Zhang [d,*], Qin Fang [e]

[a] *Department of Computer Science and Technology, Dalian Neusoft University of Information Dalian 116023, China*

[b] *School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China*

[c] *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China*

[d] *School of Science, Dalian Jiaotong University, Dalian 116028, China*

[e] *College of Information, Dalian University, Dalian 116622, China*

Abstract

A B-spline Quasi-interpolation Empirical Mode Decomposition (BSQI-EMD) method is presented and applied to similarity analysis of DNA sequences. The B-spline quasi-interpolation is used to approximate the extrema envelopes during the Intrinsic Mode Function (IMF) sifting process. The BSQI-EMD method is simple, easy to implement, and does not require solving any linear system of equations. Then we implement the classic EMD method and our method respectively. This work verifies our method's suitability and better performance for similarity/dissimilarity analysis among the coding sequences of the first exon of β-globin gene of ten different species.

Keywords: EMD; IMFs; B-spline Quasi-interpolation; Similarity/Dissimilarity Analysis, DNA Sequences

1 Introduction

With the huge amount of biological sequences discovered in the post-genomic era, it is very significant to develop fast and accurate methods for evaluating sequence similarity. Unlike the traditional methods such as the alignment methods, which are relatively precise, but sometimes they are very complex and time-consuming, especially when the data are great. The graphical representation methods can give us a visual representation.

Graphical representations of DNA sequences were initiated by Hamori and Ruskin [1] and expanded by Nandy, Randić and others [2–4]. The advantage of graphical representation of DNA

⋆Project supported by the Educational Commission of LiaoNing Province of China (No. L2012167) and the National Natural Science Foundation of China (Nos. 11301052, 11301045).

*Corresponding author.
Email address: iamzjh@126.com (Jihong Zhang).

1940–8676 / Copyright © 2015 Binary Information Press & Textile Bioengineering and Informatics Society
June 2015

sequences is that they allow visual inspection of data, helping in recognizing major differences, comparing various structures, and making the analysis of similarity among DNA sequences. It has been shown that some of the graphical representations lead to numerical characterizations of DNA sequences and quantitative measures of the degree of similarity/dissimilarity between the sequences [5–7].

One can convert a DNA sequence into a one-dimensional or multi-dimensional discrete complex sequences by the graphical representations. This method can also be called signalization of a DNA sequence. The DNA sequence is in one to one correspondence with its numerical signal sequence, so if we want to analyze and compare the features and similarity of DNA sequences, the only thing that we need to do is to compare the features and investigate the similarity of their signal sequences.

EMD is a nonlinear, non-stationary signal processing method proposed by Norden Huang [8, 9] et al. in 1998. With this method, a complicated data set can be decomposed into a small number of IMFs that admit well-behaved Hilbert transforms, with an additional residue. As the EMD method is one of the most popular tools in signal processing, we [10–12] have used the EMD method to make the similarity analysis among the long DNA sequences and the protein sequences [13].

However, in practice, the EMD has met several problems, such as boundary extension, curve fitting, stop criteria. In the classic EMD [8], one uses cubic spline functions to obtain the upper and lower envelopes of data. The cubic spline interpolating methods may produce large swings near the ends of data, which may make the decomposition of data inaccurate. Various methods are proposed to improve it. In [14], B-spline approach is proposed to fit the extremes of data, which improves the analytical performance. In [15], a rational spline EMD and flexible treatment of the end conditions are discussed. In [16], the TPS-RBF is made use of surface interpolation in bi-dimensional EMD.

An EMD method using the B-spline quasi-interpolation, abbreviated as the BSQI-EMD method, is first presented and used to compare the similarities among different species in this paper. Compared with the classic EMD method, our method is simple, easy to implement and showing better performance in similarity analysis of short DNA sequences. The paper is organized as follows: the B-spline quasi-interpolation is introduced and the BSQI-EMD method is proposed in Section 2, where the extrema envelopes are approximated by the B-spline quasi-interpolation during the IMF sifting process. A graphical representation method is introduced and similarity/dissimilarity analysis of DNA sequences by using the classic EMD method and our method is given in Section 3. We present the BSQI-EMD method and use it to carry out research on the similarities among the coding sequences of the first exon of β-globin gene of 10 different species respectively. Finally, we compare it with the EMD method.

2 The BSQI-EMD Method

2.1 Univariate B-spline Quasi-interpolation

Let $I = [a, b]$, $h = (b - a)/n$, and $X_n = \{x_i = a + ih,\ 0 \le i \le n\}$ be a uniform partition, we denote $S_d(I, X_n)$ to be the space of spline of degree d and class C^{d-1} on X_n. A basis of this space is $\{B_j,\ j \in J\}$, with $J = \{1, 2, \ldots, n + d\}$. The support of B_j is $supp(B_j) = [x_{j-d-1}, x_j]$,

and $\mathcal{N}_j = \{x_{j-d}, \ldots, x_{j-1}\}$ is denoted to be the collection of all the knots of X_n in the interior of $supp(B_j)$. As usual, we add multiple knots at the endpoints: $a = x_{-d} = x_{-d+1} = \cdots = x_0$ and $b = x_n = x_{n+1} = \cdots = x_{n+d}$.

A univariate B-spline quasi-interpolation of degree d can be defined as operators of the form

$$Q_d s = \sum_{j \in J} \mu_j(s) B_j, \tag{1}$$

where $\mu_j(s)$ is a linear combination of discrete values of s at some points in the neighborhood of $supp(B_j)$ in [17, 18]. Denote Π_d to be the collection of univariate polynomials of degree $\leq d$, and we impose that Q_d is exact on the space Π_d, i.e., $Q_d p = p$, for all $p \in \Pi_d$. Then the approximation order of Q_d is h^{d+1}.

For the determination of the functional $\mu_j(s)(j \in J)$, the exactness of Q_d on Π_d amounts to solve a system of linear equations. The main advantage of spline quasi-interpolation is that they have a direct construction without solving any system of linear equations. Moreover, they have local property, in other words, the value of $Q_d s(x)$ depends on values of s in a neighborhood of x. Also, they have a rather small infinity norm, so that they are nearly optimal approximants.

Since the cubic spline has been used to the EMD method, we use cubic B-spline quasi-interpolation for the comparison in this paper. The coefficient functionals [18] of the cubic spline quasi-interpolation

$$Q_3 s = \sum_{j=1}^{n+3} \mu_j(s) B_j, \tag{2}$$

are given as

$$\begin{aligned}
&\mu_1(s) = s_0, \\
&\mu_2(s) = \frac{1}{18}(7s_0 + 18s_1 - 9s_2 + 2s_3), \\
&\mu_j(s) = \frac{1}{6}(-s_{j-3} + 8s_{j-2} - s_{j-1}), \quad 3 \leq j \leq n+1, \\
&\mu_{n+2}(s) = \frac{1}{18}(2s_{n-3} - 9s_{n-2} + 18s_{n-1} + 7s_n), \\
&\mu_{n+3}(s) = s_n,
\end{aligned} \tag{3}$$

where $s_i = s(x_i)(0 \leq i \leq n)$.

For $s \in C^4$, we have the error estimate [18]

$$||s - Q_3 s||_\infty = O(h^4). \tag{4}$$

2.2 The BSQI-EMD Method

The EMD method is an algorithm for the analysis of nonlinear signals that runs by dividing them into a number of amplitude and frequency modulated zero mean signals, named Intrinsic Mode Functions (IMFs). By using this data-driven algorithm, a non-linear signal can be decomposed adaptively as follows:

$$s(x) = \sum_{i=1}^{n} c_i(x) + r_n(x)$$

where $c_i(x)$ is an IMF and $r_n(x)$ is the residue. Each IMF must satisfy the following two terms: (1) the number of extreme and the number of zero crossings must either be equal or differ by at most one in the whole data set; (2) at any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is close to zero.

Actually, the EMD method is a sifting process by extracting the IMFs from the signal until the residue being either the mean trend or a constant. At the same time, one can reconstruct the original signal just by adding all IMFs and the residue without any information loss.

We revise the EMD method by using the B-spline quasi-interpolation to approximate the extrema envelopes and name it the BSQI-EMD method. Compared with previous methods, which interpolate the envelopes, our method has the following advantages:

(1) It has an explicit expression (see Eq. (2) and Eq. (3)) and does not require solving any linear system of equations, so that we do not meet the question of the ill-condition of the matrix; (2) It has high accuracy (see Eq. (4)), and is easy to implement. Therefore, we can save computational time and decrease numerical errors.

3 Similarities/Dissimilarities among the Coding Sequences of the First Exon of β-gene of 10 Species

3.1 Graphical Representation for the Coding Sequences

In order to analyze the DNA sequence, we transform it into a one-dimensional or multi-dimensional discrete complex sequence by a graphical representation method. Thus, we can get a numerical signal sequence and a graph.

We assign A, T to $+x$-axis, and G and C to $-x$-axis, then we take [19, 20, 10] the map: $A \to x = +1$, $T \to x = +2$, $G \to x = -1$, $C \to x = -2$. Suppose that $S = s_1 s_2 \cdots$ may be an arbitrary DNA sequence, s_i stands for a unit, then we can get a recursive formula as follows:

$$s(x_i) = \frac{1}{d}(s(x_i) + s(x_{s_{i+1}})), \tag{5}$$

where $s(x_i)$ stands for a one-dimensional numerical sequence, d is a non-zero real number, and $s(x_0) = 0$.

We will characterize the coding sequences of the first exon of β-globin gene of 10 species, as shown in Table 1. For example, take $d = 2$, we can calculate the corresponding coordinates of the sequences of the 10 species, and the graphical representation is shown in Fig. 1 (where n is the number of the coding sequence). From the graph, we could see the numerical sequences are similar to each other, so it is difficult to distinguish them directly. Therefore, we will use the BSQI-EMD method to separate them from each other.

3.2 Similarities/Dissimilarities Analysis by the BSQI-EMD Method

We select the 10 species for analyzing features and comparing similarities. Firstly, according to the graphical presentation Eq. (5), we can convert the coding sequences of the first exon of β-gene of the 10 species into the numerical signal sequences. Secondly, we can use the classic EMD method

Table 1: The coding sequences of the first exon of β-globin gene of different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGG TCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGG GGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGG GGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTG GGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGG GGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGG GGAAAGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGG GGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAG GTGAAAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG

and our BSQI-EMD method described in Section 2 to the 10 signals to get their corresponding IMFs and residues. Finally, we compare the different species conveniently by comparing their residues for the residue being either the mean trend or a constant.

As described in [10], the classic EMD is suitable for long DNA sequences (see Fig. 2 for the residue comparison for long DNA sequences, we can compare the similarity very clearly), but it, being affected by endpoints, may not be fit to compare short DNA sequences. In order to find whether our method is suitable for the similarity analysis of short DNA sequences, we take the first exon of β-gene of the 10 species for examples to make the comparison. Then we will get 10 graphs of the 10 species by the EMD and the BSQI-EMD method respectively. We focus on the corresponding residues, so we just give the graph of the EMD comparison of the rat for observation. This is shown in Fig. 3, where the left plot is the result of the classic EMD method, the right one is the result of our method, and the symbol c_i denotes the IMF and r denotes the residue.

The BSQI-EMD method is not an interpolating method, so it will get a different decomposition from the classic EMD method. As shown in Fig. 3, the numerical sequences are decomposed into a number of different IMFs and the residue respectively by the two EMD methods.

The residue denotes the trend of a numerical sequence and is monotonic or a constant, so we can separate the sequences by comparing their residues easily. We give Fig. 4 and Fig. 5, where the residues by the two methods are printed in a single plot respectively. From Fig. 4 and Fig. 5,

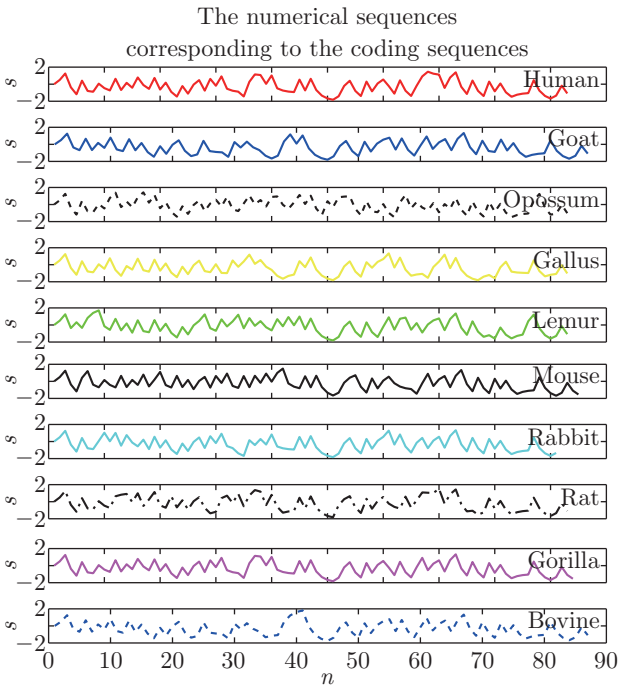


Fig. 1: The common chimp's nonlinear signal

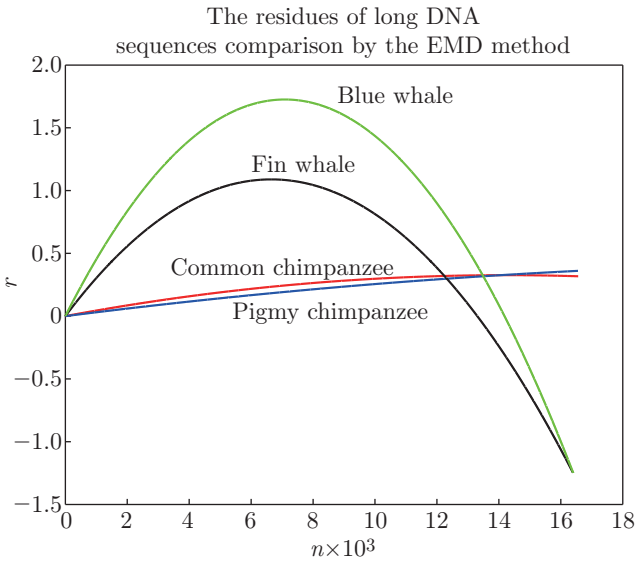


Fig. 2: The residue comparison by the classic EMD method for long DNA sequences

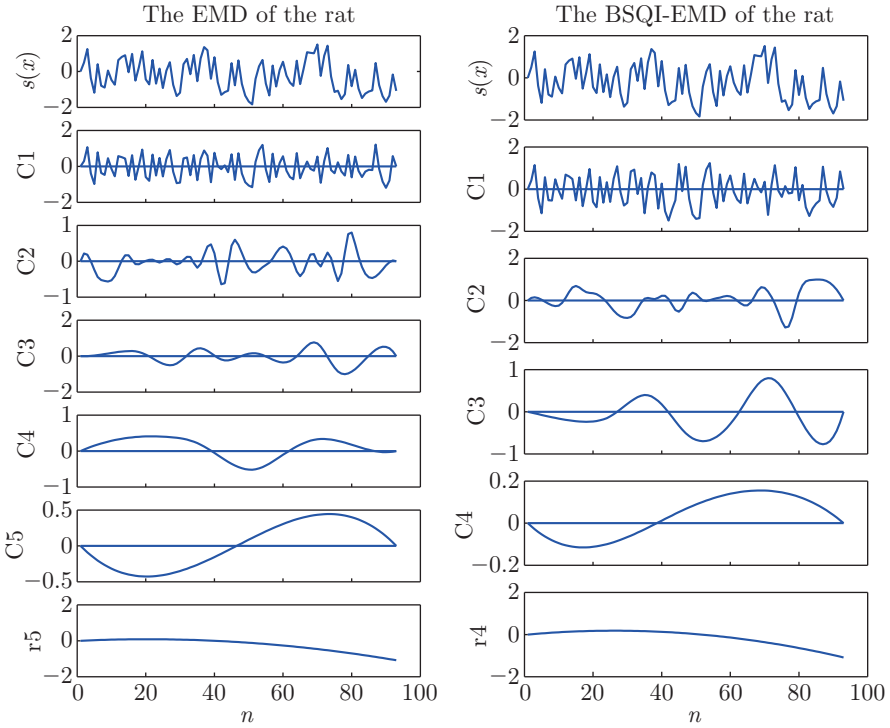


Fig. 3: The EMD and the BSQI-EMD comparison of the rat

we can find that the different species have the different residues. To make a comparison, we refer to the Randić's result [4] shown in Table 2. The closer the two residues are, the more similar the two species are. In Fig. 4, we note that the residues of the pairs (bovine, goat), (human,

gorilla) are closer with each other, and this is in accordance with Table 2. The residues of the pairs (human, opossum), (lemur, rat) are closer, it is not in accordance with Table 2. We guess that the reason may be related to the graphical presentation method or the endpoints effect of the EMD method.

Table 2: The similarity/dissimilarity matrix by Randić for the coding sequences of Table 1

Species	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine
Human	0	0.061	0.148	0.109	0.087	0.083	0.042	0.043	0.021	0.084
Goat		0	0.155	0.084	0.097	0.090	0.080	0.079	0.061	0.072
Opossum			0	0.129	0.093	0.130	0.149	0.143	0.160	0.207
Gallus				0	0.115	0.127	0.138	0.109	0.120	0.133
Lemur					0	0.050	0.081	0.078	0.099	0.155
Mouse						0	0.070	0.085	0.089	0.147
Rabbit							0	0.069	0.038	0.105
Rat								0	0.061	0.108
Gorilla									0	0.073
Bovine										0

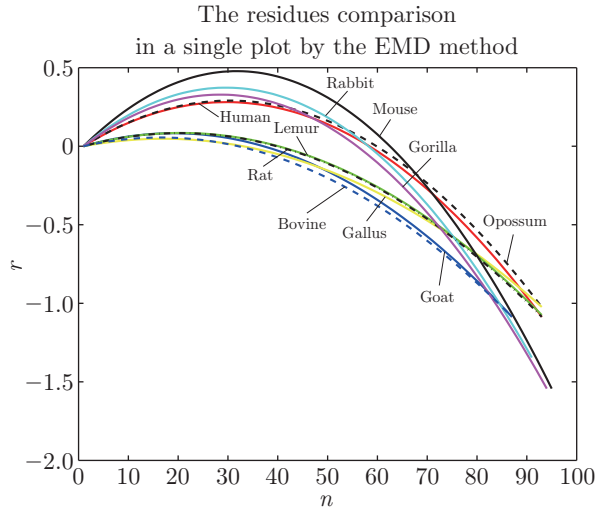


Fig. 4: The residues comparison in a single plot by the EMD method

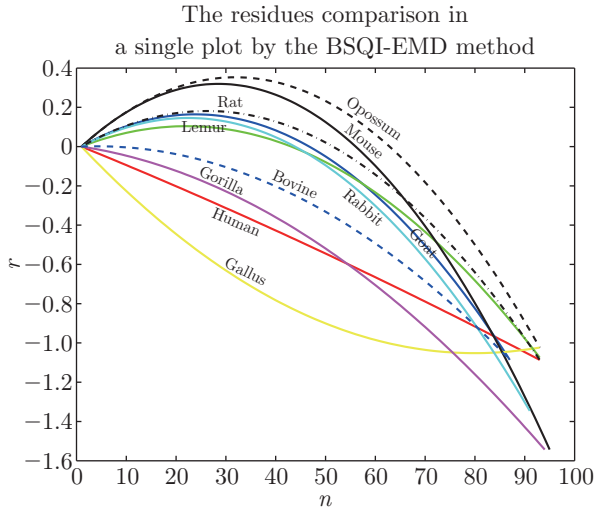


Fig. 5: The residues comparison in a single plot by the BSQI-EMD method

In Fig. 5, we see the residues of the pairs (human, gorilla), (rat, mouse) are closer, and it is in agreement with Table 2. The more important thing is that the residue of the opossum (the most remote species from the remaining mammals) is convex and at the top of the graph, moreover, the one of the gallus (the only non-mammalian representative) is concave and at the bottom of the graph, so it is in good agreement with Table 2. We can see from the above residues' graphs that the non-mammals and the mammals could be separated very clearly and even the ones who belong to the mammals can also be compared clearly.

On the basis of these findings we conclude that the BSQI-EMD method is more suitable for the similarities/dissimilarities among the coding sequences of the first exon of β-gene of 10 species.

4 Conclusion

This work has proposed the BSQI-EMD method and applied it to divide the nonlinear numerical sequence, converted from a DNA sequence, into a set of IMFs and a residue. The residue represents the trend of the numerical sequence and is monotonic or a constant, so we can conveniently and intuitively establish the similarity of different DNA sequences by comparing the corresponding residues. In the end, we get a good method which could separate DNA sequences of the non-mammals from that of the mammals very clearly and even separate the ones who belong to the mammals from each other clearly by the suitability for the coding sequences of the first exon of β-gene of the 10 species.

Compared with the EMD method, our method is simple, easy to implement for its explicit expression, and more suitable for the similarity analysis of short DNA sequences. Similarity analysis of biological sequences are still in development, and we hope that this method can be used widely and be a complementary tool for the biological sequences analysis.

References

[1] E. Hamori, J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, J. Biol. Chem., 258 (1983), 1318-1327

[2] M. A. Gates, A simple way to look at DNA, J. Theor. Biol., 119 (1986), 319-328

[3] A. Nandy, P. Nandy, Graphical analysis of DNA sequence structure - II: Relative abundances of nucleotides in DNAs, gene evolution and duplication, Curr. Sci., 68 (1995), 75-85

[4] M. Randić, M. Vracko, N. Lers, D. Plavsic, Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chem. Phys. Lett., 371 (2003), 202-207

[5] M. Randić, A. T. Balaban, On a four-dimensional representation of DNA primary sequences, J. Chem. Inform. Comput. Sci., 43 (2003), 532-539

[6] B. Liao, T. M. Wang, Analysis of similarity of DNA sequences based on triplets, J. Chem. Inf. Comp. Sci., 44 (2004), 1666-1670

[7] F. L. Bai, Y. Z. Liu, T. M. Wang, A representation of DNA primary sequences by random walk, Math. Biosci., 209 (2007), 282-291

[8] N. E. Huang, Z. Shen, S. R. Long et al., The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, P. Roy. Soc. Lond. A, 454 (1998), 903-995

[9] N. E. Huang, M. L. Wu, W. D. Qu, Applications of Hilbert-Huang transform to non-stationary financial time series analysis, Appl. Stoch. Model. Bus. Ind., 19 (2003), 245-268

[10] F. L. Bai, J. H. Zhang, J. S. Zheng, Similarity analysis of DNA sequences based on EMD method, Appl. Math. Lett., 24 (2011), 232-237

[11] J. H. Zhang, R. H. Wang, F. L. Bai, J. S. Zheng, A Quasi-MQ EMD method for similarity analysis of DNA sequences, Appl. Math. Lett., 24 (2011), 2052-2058

[12] J. H. Zhang, R. H. Wang, F. L. Bai, J. S. Zheng, A RK-EMD method for similarity analysis of DNA sequences, Journal of Information and Com., 9 (2012), 565-573

[13] J. H. Zhang, J. S. Zheng, F. L. Bai, L. W. Liu. Similarity analysis of protein sequences based on the EMD method, Journal of Fiber Bioengineering and Informatics, 7, 2014, 387-395

[14] Q. H. Chen, N. E. Huang, S. Riemenschneider and Y. S. Xu, A B-spline approach for empirical mode decompositions, Adv. Comput. Math., 24 (2006), 171-195

[15] G. G. S. Pegram, M. C. Peel, T. A. McMahon, Empirical mode decomposition using rational splines: An application to rainfall time series, Proc. R. Soc. A, 464 (2008), 1483-1501

[16] J. C. Nunes, Y. Bouaoune, É. Deléchelle, O. Niang, P. Bunel, Image analysis by bidimensional empirical mode decomposition, Image. Vision. Comput., 21 (2003), 1019-1026

[17] P. Sablonnière, Recent progress on univariate and multivariate polynomial and spline quasi-interpolants, Int. Ser. Numer. Math., 151 (2005), 229-245

[18] P. Sablonnière, Univariate spline quasi-interpolants and applications to numerical analysis, Rend. Sem. Mat. Univ. Pol. Torino., 63 (2005), 211-222

[19] F. L. Bai, D. C. Li, T. M. Wang, A new mapping rule for RNA secondary structures with its applications, J. Math. Chem., 43 (2008), 932-943

[20] J. Zupan, M. Randić, Algorithm for coding DNA sequences into "Spectrum-Like" and "Zigzag" representations, J. Chem. Inf. Model., 45 (2005), 309-313