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).

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 $supp(B_j) = [x_{j-d-1}, x_j]$,