Applied Mathematics Letters 24 (2011) 2052–2058

Contents lists available at ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

A Quasi-MQ EMD method for similarity analysis of DNA sequences☆

Jihong Zhang [a,b,*], Renhong Wang [a], Fenglan Bai [b], Junsheng Zheng [c]

[a] School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China

[b] School of Science, Dalian Jiaotong University, Dalian 116028, PR China

[c] Department of Computer Science and Technology, Neusoft Institute of Information, Dalian 116023, PR China

ARTICLE INFO

Article history:
Received 20 November 2010
Received in revised form 23 May 2011
Accepted 31 May 2011

Keywords:
EMD
IMF
MQ-RBF quasi-interpolation
DNA sequences
Similarity analysis

ABSTRACT

An empirical mode decomposition (EMD) method based on Multi-Quadrics radial basis function (MQ-RBF) quasi-interpolation (the Quasi-MQ EMD method) is presented and applied to similarity analysis of DNA sequences. The MQ-RBF quasi-interpolation is taken to approximate the extrema envelopes during the intrinsic mode function (IMF) sifting process. Our method is simple, easy to implement, and does not require solving any linear system of equations. Then we use the classic EMD method and our method to compare the local similarities among DNA sequences respectively. The work tests our method's suitability and better performance for local similarity analysis of DNA sequences by using the mitochondria of four different species.

1. Introduction

The research into biological sequences is a crucial and basic part of scientific study. It is valuable to use the similarity of the mitochondrial DNA sequences, which is called conserved sequences, to study relationships among different species. In order to analyze the DNA sequence, one converts it into one-dimensional or multi-dimensional discrete complex sequences [1–6]. It can be called signalization of a DNA sequence. The DNA sequence is in one to one correspondence with its signal sequence, so if we want to analyze and compare the features and similarity of DNA sequences, the only thing that we need do is compare the features and investigate the similarity of their signal sequences.

EMD is a nonlinear, non-stationary signal processing method proposed by Norden Huang et al. [7] in 1998. It is an adaptive and nonlinear signal decomposition approach. It can extract these intrinsic modes from the original signal, based on the local characteristic scale of data itself, and represent each intrinsic mode as an IMF, which meets the following two conditions: (1) in the whole data set, the number of extrema and the number of zero crossings must either equal or differ at most by one; (2) at any point, the mean value of the envelopes defined by the local maxima and the envelopes defined by the local minima is close to zero. The two conditions ensure that an IMF is a nearly periodic function and the mean is close to zero. 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 being either the mean trend or a constant. We use the EMD method to analyze the similarity of DNA sequences by comparing the corresponding residues in [8].

However, in practice, the EMD method has met several problems, such as boundary extension, curve fitting and stop criteria. In the classic EMD [7] method, one uses cubic spline functions to obtain the upper and lower envelopes of data. The

☆ Manuscript received October 13, 2010. This work was supported by the National Natural Science Foundation of China (Nos. U0935004, 11071031, 11001037, 10801024), the Fundamental Research Funds for the Central Universities (DUT10ZD112, DUT10JS02, DUT11LK34) and Educational Commission of Liaoning Province of China (Grant No. 2009A125).

* Corresponding author at: School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, PR China. Tel.: +86 411 86243032; fax: +86 411 86243032.

E-mail addresses: iamzjh@126.com (J. Zhang), renhong@dlut.edu.cn (R. Wang), bfl0219@163.com (F. Bai), zhengjunsheng@neusoft.edu.cn (J. Zheng).

0893-9659/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2011.05.041

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 [9], a B-spline approach is proposed to fit the extremes of data, which improves the analytical performance. In [10], a rational spline EMD and flexible treatment of the end conditions are discussed. In [11,12], the TPS-RBF is made use of surface interpolation in bi-dimensional EMD. See papers [13–15] for other works on the EMD methods.

RBF is a useful tool for fitting scattered data, because of its accuracy, spectral convergence, simplicity and ease of implementation. Among all RBFs currently in use, the MQ-RBF [16] is probably best understood, both theoretically and practically, and it usually ranks the best in accuracy. However, in order to obtain the MQ-RBF interpolation, one must resolve a linear system of equations. When the number of samples is large, the method shows the typical drawbacks of global methods, since the interpolation is influenced by all the data. Moreover, the condition number of the interpolation matrix heavily relies on the data density, which leads to unstable solutions or unacceptable computational costs [17]. There are different ways to overcome this ill-conditioning problem, and the MQ quasi-interpolation method [18,19] is one of them.

An EMD method using MQ-RBF quasi-interpolation, named the Quasi-MQ EMD method, is 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 local similarity analysis of DNA sequences. The paper is organized as follows: the MQ-RBF quasi-interpolation is introduced in Section 2. In Section 3, we propose the Quasi-MQ EMD method, where the extrema envelopes are approximated by the MQ-RBF quasi-interpolation during the IMF sifting process. Similarity analysis of DNA sequences by using the classic EMD and our method is given in Section 4. We select mitochondrial DNA sequences of four species—the common chimpanzee (D38116), pygmy chimpanzee (D38113), fin whale (X61145), and blue whale (X72204)—as our research objects. Finally, we make use of the EMD method and our method to carry out research on the similarity, respectively.

2. MQ-RBF quasi-interpolation

The MQ-RBF $\varphi(r) = \sqrt{r^2 + c^2}$, was proposed by Hardy [20] in 1971, where $r = \|x\|$, $c > 0$ is called the shape parameter. A review by Franke [16] showed that the MQ outperformed 29 methods in terms of accuracy and efficiency. Although the MQ interpolation is always solvable when the data points are distinct, the resulting matrix quickly becomes ill-conditioned as the number of points increases. The quasi-interpolation method is a good choice to overcome this problem.

Wu and Schaback [19] proposed the univariate MQ quasi-interpolation scheme L_D with the MQ function $\varphi_j(x) = \sqrt{(x - x_j)^2 + c^2}$, $j = 0, 1, \ldots, n$ and proved that the scheme is shape preserving and produces linear polynomials. Given points $\{(x_j, f_j)\}_{j=0}^n$, where $x_0 < x_1 < \cdots < x_n$, and $f_j = f(x_j)$, the scheme L_D is defined as follows:

$$(L_D f)(x) = f_0\alpha_0(x) + f_1\alpha_1(x) + \sum_{j=2}^{n-2} f_j\psi_j(x) + f_{n-1}\alpha_{n-1}(x) + f_n\alpha_n(x), \tag{2.1}$$

where

$$\alpha_0(x) = \frac{1}{2} + \frac{\varphi_1(x) - (x - x_0)}{2(x_1 - x_0)}, \qquad \alpha_1(x) = \frac{\varphi_2(x) - \varphi_1(x)}{2(x_2 - x_1)} - \frac{\varphi_1(x) - (x - x_0)}{2(x_1 - x_0)},$$
$$\alpha_n(x) = \frac{1}{2} + \frac{\varphi_{n-1}(x) - (x_n - x)}{2(x_n - x_{n-1})}$$
$$\alpha_{n-1}(x) = \frac{(x_n - x) - \varphi_{n-1}(x)}{2(x_n - x_{n-1})} - \frac{\varphi_{n-1}(x) - \varphi_{n-2}(x)}{2(x_{n-1} - x_{n-2})},$$
$$\psi_j(x) = \frac{\varphi_{j+1}(x) - \varphi_j(x)}{2(x_{j+1} - x_j)} - \frac{\varphi_j(x) - \varphi_{j-1}(x)}{2(x_j - x_{j-1})}, \quad j = 2, \ldots, n-2.$$

Theorem 2.1 (*[19]*). *For $f \in C^2[a, b]$, the quasi-interpolation $L_D f$ defined by Eq. (2.1) on the points $x_0 < x_1 < \cdots < x_n$ satisfies the error estimate: $\|f - L_D f\|_\infty \le k_1 h^2 + k_2 ch + k_3 c^2 |\log h|$, where $h = \max_{1\le j\le n}\{x_j - x_{j-1}\}$, and k_1, k_2, k_3 are positive constants independent of h and c. $L_D f(x)$ has an error $O(h^2)$ only if $c^2|\log c| = O(h^2)$.*

Theorem 2.2 (*[21]*). *The quasi-interpolation scheme $L_D f$ defined by Eq. (2.1) is variation-diminishing.*

3. The quasi-MQ EMD method

We use the MQ-RBF quasi-interpolation to approximate the extrema envelopes. Compared with previous methods, which interpolate the envelopes, our method has the following advantages.

(1) It has a simple expression (see Eq. (2.1)) and does not require solving any linear system of equations;
(2) MQ-RBF has high accuracy, spectral convergence, and is easy to implement.