BioSystems 100 (2010) 132–143

Contents lists available at ScienceDirect

BioSystems

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

Spectral density ratio based clustering methods for the binary segmentation of protein sequences: A comparative study

Alexis Ioannou [a], Konstantinos Fokianos [a], Vasilis J. Promponas [b,*]

[a] Department of Mathematics & Statistics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus
[b] Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus

ARTICLE INFO

Article history:
Received 23 July 2009
Received in revised form 12 February 2010
Accepted 23 February 2010

Keywords:
Distance measures
OMP topology prediction
Physicochemical parameters
Protein sequence segmentation
Spectral analysis
Periodogram
Time series

ABSTRACT

We compare several spectral domain based clustering methods for partitioning protein sequence data. The main instrument for this exercise is the spectral density ratio model, which specifies that the logarithmic ratio of two or more unknown spectral density functions has a parametric linear combination of cosines. Maximum likelihood inference is worked out in detail and it is shown that its output yields several distance measures among independent stationary time series. These similarity indices are suitable for clustering time series data based on their second order properties. Other spectral domain based distances are investigated as well; and we compare all methods and distances to the problem of producing segmentations of bacterial outer membrane proteins consistent with their transmembrane topology. Protein sequences are transformed to time series data by employing numerical scales of physicochemical parameters. We also present interesting results on the prediction of transmembrane β-strands, based on the clustering outcome, for a representative set of bacterial outer membrane proteins with given three-dimensional structure.

1. Introduction

It is rather common in practice that biological sequence analysis is applied for the detection of signals along nucleic acid or amino acid sequences. These signals encode different structural and/or functional features of the respective molecules. Hence, their detection is a delegate matter which is formulated as classification, prediction or segmentation problem in several cases, see Han (2005). It is pointed out though that solutions of such important problems depend upon a number of assumptions and the relevant biological process under investigation. Moreover, several classes of biological macromolecules exhibit periodical patterns in their primary structures. These types of patterns reflect their underlying properties, for instance they determine their coding potential (Yan et al., 1998), reading frame of DNA stretches (Shepherd, 1981) or the potential of formation of long helical coiled-coils (McLachlan and Stewart, 1976). An extensive list of examples is given by Pasquier et al. (1998) and references therein. Notice that the aforementioned patterns cover a broad range of underlying properties of the molecule. They describe either exact or inexact repetitive (sub)sequences (e.g. CGACGACGA or CGACGTCGACGA) but they also elucidate periodical (perfect or imperfect) appearances of similar physicochemical properties in the sequence, such as hydrophobicity, charge or residue size.

In several cases there exists adequate biological knowledge for a specific family of molecules under consideration. If there are available several experimentally determined three-dimensional structures or detailed functional data for a class of molecules, then the properties of the unknown biological signal are well defined. Therefore, successful predictors or classifiers can be built by exploiting the available prior knowledge. Any given information as such, can be converted into a meaningful model for the detection of the biological signal. Several examples have been recorded in the literature, see consensus sequences (Aitken, 2003), sequence logos (Schneider and Stephens, 1990), position specific scoring matrices (Gribskov et al., 1990), and their generalizations, generalized profiles (Bucher et al., 1996) and profile Hidden Markov Models (Eddy, 1998), among many others. However, there are cases where such information is not available. Therefore, exploratory data analysis techniques, such as clustering, can be employed as a first step towards identification of molecule specific informative features. These techniques allow for the development of a sound statistical model which incorporates all available information. It is crucial that application of such methods enables us to comprehend the underlying biological principles and mechanisms.

Clustering methods have been routinely applied to sequences of biological macromolecules for successful identification of biologically relevant – in a functional, structural or evolutionary sense – groups of data. When

* Corresponding author.
E-mail address: vprobon@ucy.ac.cy (V.J. Promponas).

0303-2647/$ – see front matter © 2010 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2010.02.008

considering though clustering of long sequences in the sequence domain-equivalently time domain – then the problem of high dimensionality – (number of sequences) × (number of sequence positions that constitute the sequence signal) – poses a major challenge. In addition, the biological sequences under comparison might have different lengths. So far, most of the clustering approaches rely on a single numerical distance computed for each pair of sequences. Such a summary is the result of a CPU intensive dynamic programming based pairwise sequence comparison, e.g. Smith and Waterman (1981), and clustering is obtained by either hierarchical or partitioning methods, or by more sophisticated methods, like stochastic graph clustering, see Enright et al. (2002).

In this article, we address the problem of clustering subsequences of biological macromolecules by time series methodology. In particular, we suggest the use of spectral domain methods to take into account the data dependence. The use of spectral domain methods effectively reduces the dimensionality of the sequences, while it preserves several of their periodical features. Therefore, meaningful distances between time series data can be calculated effectively. In doing so, we resort to the spectral density ratio model as introduced by Fokianos and Savvides (2008) and further studied by Savvides et al. (2008). The spectral density ratio model is a semiparametric model in the sense that the ratio of two or more unknown spectral density functions is modeled by a parametric function of cosines. The current work evaluates different clustering approaches based on the spectral density ratio model for addressing the problem of naïvely segmenting a single amino acid sequence into two classes of subsequences with biologically – either structurally or functionally – distinct characteristics.

More specifically, the biological motivation of this work originates from the problem of predicting the topology of bacterial outer membrane proteins (OMPs) which form transmembrane β-barrels – see Fig. 1. These proteins constitute a substantial fraction of the outer membrane of Gram-negative bacteria, Bagos et al. (2005). Moreover, they exhibit a remarkable and diverse functional repertoire because they provide molecular recognition sites (Morona et al., 1985; Vogt and Schulz, 1999) and mediate the passive (respectively, active) transport of small (respectively, large) molecules (Cowan et al., 1992; Zhou et al., 1995; Braun et al., 2000; Danelon et al., 2003). For a comprehensive review on the major functional and structural features of this eminent class of proteins, the interested reader is referred to the recent reviews by Schulz (2003) and Bagos et al. (2005).

The paper is organized as follows: The next section gives a brief introduction to the spectral density ratio model, which is the main instrument for defining spectral domain based metrics between time series for the purpose of clustering. Section 3 briefly reviews meaningful measures of spectral similarity between stationary time series. Section 4 illustrates an application of the proposed approach to both simulated and real data examples. In fact, the real data consist of a carefully compiled set of bacterial outer membrane β-barrel proteins with known structures. We also present interesting results on the prediction of transmembrane β-strands, based on the clustering outcome, for a representative set of bacterial outer membrane proteins with given three-dimensional structure. The article concludes with a discussion about the suggested methodology.

2. The Spectral Density Ratio Model

Consider now G biological sequences and suppose that these are viewed as independent stationary time series. We denote each one of them by $\{Y_{jt}, t = 1, 2, \ldots, N\}$ for $j = 1, 2, \ldots, G$. Assume that $\{Y_{jt}, t = 1, 2, \ldots, N\}$ possesses a spectral density function $\lambda_j(\omega)$,

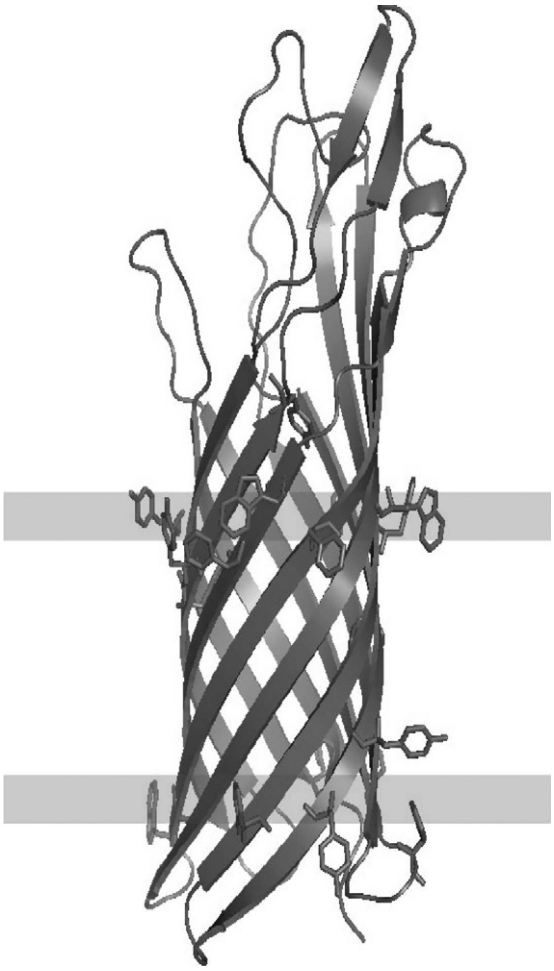

Fig. 1. Cartoon representation of the structure of the OpcA Outer Membrane Adhesin/Invasin from *Neisseria meningitides*, a typical β-barrel outer membrane protein (PDB ID: 1k24). The horizontal lines depict the position of the lipid bilayer based on the so-called aromatic belts.

$j = 1, 2, \ldots, G$. A useful example of stationary time series is that of a linear process which is formed by independent and identically distributed innovations and absolute summable autocovariance function, $\gamma_j(\cdot)$. Estimation of the unknown spectral density function is based on the periodogram ordinates

$$I_j(\omega_i) = \frac{1}{2\pi N}\left|\sum_{t=1}^{N} Y_{jt}\exp(-it\omega_i)\right|^2, \quad (1)$$

where $\omega_i = 2\pi i/N$, $i = 1, 2, \ldots, [(N-1)/2]$, are the so-called Fourier frequencies. An important result for linear processes is that the periodogram ordinates are asymptotically independent with distribution proportional to a chi-square random variable with two degrees of freedom. For a precise statement, see Brockwell and Davis (1991, Thm. 10.3.2). Furthermore, if $\log\lambda_j(\omega)$ is absolutely integrable on $(0, 1)$ then the Fourier coefficients of the expansion of $\log\lambda_j(\omega)$ are defined by

$$\theta_{jr} = \int_0^1 \log\lambda_j(\omega)\cos(2\pi r\omega)d\omega, \quad r = 0, 1, \ldots, \quad (2)$$

and are referred as cepstral correlations (or cepstral coefficients). A small number of cepstral coefficient suffices to discriminate between the second order characteristics of two or more time series.