Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition 41 (2008) 2398–2412
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Clustering of biological time series by cepstral coefficients based distances

Alexios Savvides[a], Vasilis J. Promponas[b], Konstantinos Fokianos[a,*]

[a]*Department of Mathematics and 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*

Received 23 July 2007; received in revised form 19 November 2007; accepted 8 January 2008

Abstract

Clustering of stationary time series has become an important tool in many scientific applications, like medicine, finance, etc. Time series clustering methods are based on the calculation of suitable similarity measures which identify the distance between two or more time series. These measures are either computed in the time domain or in the spectral domain. Since the computation of time domain measures is rather cumbersome we resort to spectral domain methods. A new measure of distance is proposed and it is based on the so-called cepstral coefficients which carry information about the log spectrum of a stationary time series. These coefficients are estimated by means of a semiparametric model which assumes that the log-likelihood ratio of two or more unknown spectral densities has a linear parametric form. After estimation, the estimated cepstral distance measure is given as an input to a clustering method to produce the disjoint groups of data. Simulated examples show that the method yields good results, even when the processes are not necessarily linear. These cepstral-based clustering algorithms are applied to biological time series. In particular, the proposed methodology effectively identifies distinct and biologically relevant classes of amino acid sequences with the same physicochemical properties, such as hydrophobicity.

Keywords: Exponential model; Likelihood; Distance measures; Spectral analysis; Periodogram; Data mining; Protein sequence analysis

1. Introduction

Time series data arise in different disciplines including engineering, business, medicine among many others. Several experiments produce massive sets of dependent data and there is need for data reduction methods which allow classification or clustering. The purpose of clustering, which is the main interest to this work, is to obtain an assignment rule which divides the data set into homogeneous groups. Here the notion of homogeneity means that objects within a group are similar while objects between groups are not that similar. In the context of independent data these concepts have been studied extensively, see Johnson and Wichern [1] and Hastie et al. [2], for instance. However, for massive time series data sets, which appear very often in applications, there does not seem to be such an extensive literature. The aim of this work is to contribute towards this research by introducing similarity (or dissimilarity) measures based on the so-called cepstral coefficients. These quantities are simply the Fourier coefficients of the log spectrum of a stationary process. Hence, the methodology is developed in the frequency domain which simplifies calculations because of the asymptotic independence of periodogram ordinates. In addition, focusing on spectral domain yields clustering of time series with similar second order structure—that is a feature-based approach. This is one aspect of our work. The additional feature, which can be called as model-based approach, is the introduction of a model that links all the spectral densities of the processes under consideration. Accordingly, the model imposes that the log ratio of two spectral densities is linear in some parameters. Using this model, we estimate a distance between two or more time series and then apply a specific clustering algorithm. Some different approaches have been suggested in the literature but a detailed review of the topic is out of the scope of this article. We only mention here that the closest approach to ours is that taken by Kalpakis et al. [3] who have studied the euclidean distance between cepstral coefficients of two or more time

* Corresponding author. Tel.: +357 22 892 613.
E-mail address: fokianos@ucy.ac.cy (K. Fokianos).

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.01.002

series in the context of ARIMA modeling. However, there are fundamental differences between their approach and the method that is suggested in this work. More details and further references will be given later but the interested reader should see Liao [4] and Shumway and Stoffer [5] for a comprehensive review on the topic of clustering time series.

Motivation for studying this problem comes from the need of identifying similar physicochemical properties—such as hydrophobicity—of amino acid sequences, as proclaimed in the spectral domain. Identifying similar sequential and structural patterns is useful for successful discrimination among protein sequences which belong into distinct biologically relevant classes. Consider for instance, bacterial pathogenicity. It is well known that this physiological process, requires the export of certain proteins. These proteins are initially synthesized at the ribosomes which are located in the bacterial cytoplasm, and transported to the periplasmic space or the surrounding medium. This functionality is mediated by specialized transport molecules that help selected proteins to cross the otherwise impermeable to them lipid bilayer of the bacterial plasma membrane. Although several secretion pathways exist, it has been shown that—at least in the vast majority of known cases—all the information required for the initial export of bacterial proteins from the plasma membrane resides in short sequence segments. These segments are termed as signal peptides and are of variable length. They are located in the N-terminal region of the nascent polypeptide chain, see Emanuelsson et al. [6]. More specifically, these N-terminal signals are recognized by special protein molecules that assist their translocation through dedicated transmembrane protein channels located on the bacterial plasma membrane.

The best characterized secretion pathway so far is the so-called Sec-dependent pathway. A general feature of signal peptides is a tripartite sequential structure consisting of a positively charged n-region, followed by a hydrophobic h-region and a polar uncharged c-region of variable respective lengths, see Emanuelsson et al. [6]. A novel pathway, namely the Tat (Twin arginine translocation) pathway, has been recently discovered in bacteria, Berks [7] and Berks et al. [8]. Although there seems to exist significant differences in the molecular mechanism of protein export via those two pathways, proteins entering the Tat pathway have signal peptides with a tripartite structure resembling the one of proteins exported by the Sec molecular machinery. Nevertheless, they often possess two (initially thought invariant) consecutive arginine amino acid residues at the border between the n- and h-regions [9]. Aiming to reveal hidden protein sequence features, we have performed an analysis of fixed-length N-terminal amino acid sequences (50 residues long) from different bacterial proteins with experimental evidence for the presence of a secretory N-terminal signal peptide, either of the classical or of the Tat form. To carry out the identification of similar hidden features in amino acid sequences of well-characterized proteins bearing either Tat- or Sec-machinery specific secretory signal peptides we consider standard spectral domain clustering techniques together with the cepstral distance which was discussed earlier. It is not aimed to develop a prediction method neither to benchmark different clustering techniques, but rather to illustrate the power of novel cepstral coefficient-based spectral analysis tools in biological sequence analysis.

As a general remark, spectral analysis tools have been extensively used in biological and biomedical research for more than two decades. In particular, there have been many successful applications in Computational Biology in diverge areas such as gene finding [10], periodicity analysis [11], proteomics [12], study of fibrous proteins [13] and protein functional classification [14]. Very recently, cepstral-based measures have been used in an application of protein classification, in a proof-of-principle demonstration [15].

The paper is organized as follows: The next section introduces notation and the main model that will be employed throughout the presentation. Section 3 introduces time domain distances—this topic is briefly discussed—and spectral domain distances. These distances will be compared with the cepstral coefficient-based metrics introduced in Section 4. In addition, it is shown how the cepstral measures can be estimated by means of the proposed model. Section 5 reports results from simulated and real data and the article closes with some remarks about the proposed method.

2. Model specification

In this section, we establish some notation and state some useful results and then we propose a model which yields the inferential output for cepstral coefficient-based clustering.

2.1. Notation and model assumptions

In what follows, we will focus on stationary processes which possess continuous spectrum even though in Section 5.1 we consider other examples to validate further the robustness of the proposed approach. In particular, we will be concerned with the general linear process of the form

$$X_t = \sum_{u=-\infty}^{\infty} g_u \varepsilon_{t-u}, \tag{1}$$

where $\{\varepsilon_t\}$ is a sequence of independent and identically distributed random variables with $\mathrm{E}[\varepsilon_t]=0$ and variance $\mathrm{Var}[\varepsilon_t]=\sigma_\varepsilon^2<\infty$. Note that we do not assume any normality in what follows. In addition, assume $\sum_{u=-\infty}^{\infty}|g_u|<\infty$. Then, it is well known that the autocovariance function $\gamma_x(.)$ of X_t satisfies $\sum_{h=-\infty}^{+\infty}|\gamma_x(h)|<\infty$. The last condition implies the existence of the so-called spectral density function of X_t which is given by

$$\lambda_x(\omega)=\frac{1}{2\pi}\sum_{h=-\infty}^{\infty}\gamma_x(h)\exp(-\mathrm{i}h\omega), \quad -\pi\leqslant\omega\leqslant\pi.$$

The spectral density function $\lambda_x(\omega)$ is a non-negative even function and therefore it suffices to consider $\omega>0$.