Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 28 (2007) 2133–2142
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Error probabilities for local extrema in gene expression data

Perry Groot [a,*], Christian Gilissen [b], Michael Egmont-Petersen [b]

[a] Institute for Computing and Information Sciences, Radboud University Nijmegen, 6525 ED Nijmegen, The Netherlands
[b] Department of Human Genetics, UMC St. Radboud Nijmegen, The Netherlands

Received 15 February 2007
Available online 19 July 2007

Communicated by M. Singh

Abstract

Current approaches for the prediction of functional relations from gene expression data often do not have a clear methodology for extracting features and are not accompanied by a clear characterisation of their performance in terms of the inherent noise present in such data sets. Without such a characterisation it is unclear how to focus on the most probable functional relations present. In this article, we start from the fundamental theory of scale-space for obtaining features (i.e., local extrema) from gene expression profiles. We show that under the assumption of Gaussian distributed noise, repeatedly measuring a local extrema behaves like a bivariate Gaussian distribution. Furthermore, the error of not re-observing local extrema is phrased in terms of the integral over the tails of this bivariate Gaussian distribution. Using integration techniques developed in the 1950s, we demonstrate how to compute these error probabilities exactly.

Keywords: Statistics; Scale-space theory; Bivariate Gaussian integration; Gene expression

1. Introduction

In the last few years, various international genome projects have yielded the near complete molecular sequences of a large number of species, including human. Novel high-throughput methodologies such as microarray-based gene expression profiling are now being used to generate genome-wide transcriptomic data sets at an ever-increasing rate to analyse and monitor the effects of intrinsic and exogenous variables on living cells, tissues, and organs. In general, the timing of mRNA expression for a given gene has been found to correlate well with the function of the resultant protein (Bähler, 2005; Bozdech et al., 2003).

Identification of functional relations from gene expression data has remained difficult because of the inherent noise in such data sets. Methods that have been developed to determine such relations include clustering algorithms like hierarchical clustering, *K*-means clustering, and self-organising maps (Eisen et al., 1998; Datta and Datta, 2003). The simplest approach to clustering is to select a gene and determine the nearest neighbouring genes according to a distance measure between gene expression profiles. This approach, called hierarchical clustering, allows the clustering of groups of genes that are co-regulated. As yet, however, it is unclear how well certain distance functions can deal with noise which is inherent to gene expression data sets. Model-based approaches like dynamic Bayesian networks offer more flexible techniques that can, in principle, deal with the inherent noise of gene expression data sets (Friedman et al., 2000; Husmeier, 2003). However, such model-based approaches preferably use discretised expression data mapping expression levels

* Corresponding author. Tel.: +31 24 3652075; fax: +31 24 3653366.
E-mail addresses: perry@cs.ru.nl (P. Groot), C.Gilissen@antrg.umcn.nl (C. Gilissen), M.EgmontPetersen@antrg.umcn.nl (M. Egmont-Petersen).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.06.017

to some discrete representation, which raises methodological questions. The interpretation of time series gene expression data sets is complex and is still regarded to be an open problem (Storey et al., 2005).

In this article, we will focus on the representation proposed by Egmont-Petersen et al. (2004), which only registers the local extrema of the time course gene expression. This representation focusses on the most likely time points a gene changes from up regulated to becoming down regulated or vice versa and allows the prediction of functional relations without regarding the amplitude of the signal (cf. Section 2). We will start the analysis of gene expression profiles by using a fundamental approach developed in the computer vision community, called scale-space theory (Koenderink, 1984), for analysing images and signals at multiple scales (Section 3). This allows us to formulate criteria for detecting local extrema in noisy signals. By interpreting point measurements as a stochastic process we are able to derive its exact distribution and give a characterisation of not re-observing an extremum as the result of noise and/or smoothing. More specifically, the contributions of this paper are the following:

- Under the assumption of Gaussian distributed additive noise, the repeated measurement of two subsequent points in the scale-space representation of a one-dimensional discrete signal is shown to have a bivariate Gaussian distribution (Section 4).
- Using the criteria for detecting local extrema in the scale-space representation, the probability that an extremum is not re-observed because of noise and/or smoothing is phrased in terms of integrating the tails of a bivariate Gaussian distribution that cross the horizontal and vertical axis, respectively (Section 4).
- We apply the method of Owen (1956) for the integration of a bivariate Gaussian distribution (Section 5) and provide an algorithm for the procedure (Section 7).

In summary, the paper uses a methodological approach for detecting local extrema in a signal with Gaussian distributed noise that is accompanied with a precise quantification of the measurement quality of the local extrema that can be computed with the provided algorithm.

The remaining sections discuss in more detail the motivations behind our work (Section 2), related work (Section 6), and conclusions and further work (Section 8).

2. Motivation

One of the major research directions in bioinformatics is the identification of functional relations between gene expression profiles based on extracted features. For example, Fig. 1 shows that when the pol32-profile has a local extremum at time n, the rad51-profile has a similar extremum at time $n+1$, strongly suggesting a functional relation between the pol32 and rad51 gene. However, in

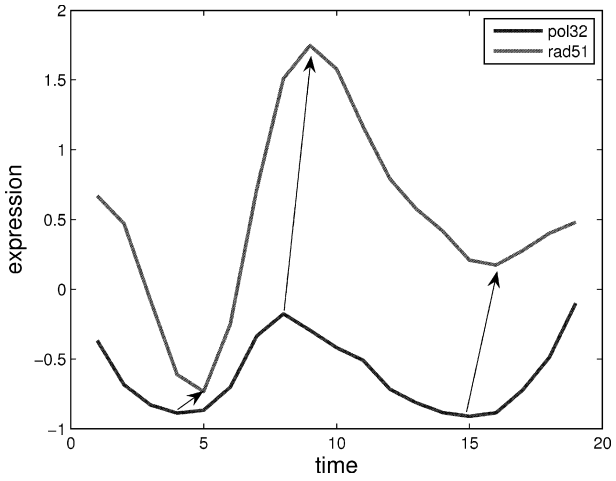

Fig. 1. Functional relations based on local extrema between gene expression profiles in yeast.

contrast with the smooth profiles shown in Fig. 1, gene expression profiles inherently contain a lot of noise making it difficult to clearly distinguish features in such profiles. The aim of our work is to have a clear underlying methodology for obtaining features (i.e., local extrema) from gene expression profiles for predicting functional relations between genes, which is accompanied with a precise characterisation of not being able to observe the feature due to noise present.

Here, we focus on locating local extrema in gene expression profiles. By discretising gene expression levels on the basis of local extrema we focus on the most likely time points at which a gene (and eventually its associated protein) is active (local maximum) and inactive (local minimum). As local extrema are invariant under scaling and vertical shifting (dosage effects Goldenthal et al., 2004), this representation effectively captures the global dynamics of the time-dependent data, i.e., local extrema allow the prediction of functional relations between a transcription factor with small absolute changes in expression ratio, and a target gene, because the amplitude is disregarded (Egmont-Petersen et al., 2004).

We start the analysis of local extrema from the theory of scale-space (Section 3), which allows one to analyse signals at different scales. At each scale a different amount of smoothing is applied, resulting in a simplification of the signal as spurious structures (i.e., local extrema) as the result of noise is removed. This is shown in Fig. 2 for the gene expression profile of the swi4 gene along the scale-space dimension.

After formulating criteria for detecting local extrema in the scale-space representation of a gene expression profile, the next step is to precisely characterise the probability of not detecting a local extremum due to noise. This allows one to focus on those profiles that have the most likely correct measurements when determining local extrema for the prediction of functional relations based on those local extrema.