Selection and ranking of genes relevant for cancer diagnosis based on the classification ability of their expression pattern

Maglietta R (1), D'Addabbo A (1), Piepoli A (2), Perri F (2),
Liuni S (3), Pesole G (3,4), Ancona N (1)

(1) Istituto di Studi sui Sistemi Intelligenti per l'Automazione, CNR,
Via Amendola 122/D-I, 70126 Bari, Italy
(2) Unità Operativa di Gastroenterologia, IRCCS, "Casa Sollievo della Sofferenza"-Ospedale,
Viale Cappuccini, 71013 San Giovanni Rotondo (FG), Italy
(3) Istituto di Tecnologie Biomediche-Sezione di Bari,CNR,
Via Amendola 122/D, 70126 Bari Italy
(4) Dipartimento di Biochimica e Biologia Molecolare - Università di Bari,
Via E. Orabona 4, 70126 Bari, Italy

Motivation

One of the main problems in cancer diagnosis by using DNA microarray data is the selection of genes whose expression is most significantly altered by the pathology by analyzing their expression profiles in normal and tumour tissues. The question we pose is the following: how do we measure the relevance of a single gene in a given pathology?

Methods

A gene is relevant for a particular disease if we are able to correctly predict the occurrence of the pathology in new patients on the basis of its expression level only. In other words, a gene is informative for the disease if its expression levels are useful for training a classifier able to generalize, that is, able to correctly predict the status of new patients. In this paper we present a selection bias free, statistically well founded method for finding relevant genes on the basis of their classification ability.

Results

We applied the method on a colon cancer data set and produced a list of relevant genes, ranked on the basis of their prediction accuracy. We found that, among more than 6500 available genes, 54 genes over-expressed in normal tissue and 77 genes over-expressed in tumour tissue showed prediction accuracy greater than 70% with p-value $p \leq 0.05$. The relevance of the selected genes was assessed a) statistically, evaluating the p-value of the estimate prediction accuracy of each gene; b) biologically, confirming the involvement of many genes in generic carcinogenic processes and particularly for the colon; c) comparatively, verifying the presence of these genes in other studies on the same data-set.

Contact email: ancona@ba.issia.cnr.it