Implementation of a gene expression index calculation method based on the PDNN model

Henrik Bjørn Nielsen*, Laurent Gautier and Steen Knudsen

Center for Biological Sequence Analysis, BioCentrum-DTU Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

Received on March 9, 2004; revised on September 14, 2004; accepted on September 27, 2004
Advance Access publication October 27, 2004

ABSTRACT
Summary: Gene expression index calculations from Affymetrix GeneChips have been dominated by the Affymetrix MAS, dChip, and RMA methods. A new method to estimate the gene expression value utilizing the probe sequence information named position-dependent nearest-neighbor (PDNN) has been suggested by Zhang et al. (2003). Here we describe an open source implementation of the PDNN method for the statistical language R.


Contact: hbjorn@cbs.dtu.dk

INTRODUCTION

Short oligonucleotide microarrays, such as the GeneChip from Affymetrix, use multiple probes per targeted transcript. This type of microarray has shown that the probe signals are not always consistent between different probes. This inconsistency is only marginally due to noise in the measurements. The main differences in the signal are due to the differences in the probes’ properties.

In order to obtain a single expression index value representing a gene expression, several data processing methods have been developed. The most widely used are the MAS v.5, the dChip, and the RMA [Affymetrix, Inc., Santa Clara, CA, USA; Li and Wong, 2001; Irizarry et al., 2003]. Zhang et al. (2003) developed a position-dependent nearest-neighbor (PDNN) model over the probe signals that enables estimation of a gene expression index. The variation in expression index values between experiments was shown to be superior to both the MAS v.5 and the dChip methods (Zhang et al., 2003). However, our studies show that the results are comparable to the dChip method, when the latter uses PM only (Li and Wong, 2001; Gautier et al., 2004). Each plot represents 12,474 genes through 14 experiments (data is from the ‘1532 series’ in the Human Latin square dataset, available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx).

It should be noted that the PDNN model assumes that the majority of probes are designed specifically for their target, and that they will only capture specific signals from the target. Here we describe the implementation of the PDNN method, as a package named affyPDNN for the statistical language R.

IMPLEMENTATION

The PDNN model minimizes
\[
F = \frac{\sum (\ln \hat{I}_{ij} - \ln I_{ij})^2}{M}
\]

where \(\hat{I}_{ij}\) and \(I_{ij}\) are the estimated and the observed probe intensity of the \(i\)th probe in a probe set targeted for gene \(j\), respectively, and \(M\) is the number of probes. \(\hat{I}_{ij}\) is estimated using the following equation:
\[
\hat{I}_{ij} = \frac{\sum \left( \frac{(I_{ij} - B - (N^* / 1 - e^{k_j}) \lambda_{ij})}{1 - e^{k_j}} \right)}{1 - e^{k_j} \sum \left( \frac{1}{1 - e^{k_j}} \lambda_{ij} \right)} + \frac{N^*}{1 - e^{k_j}} + B
\]
Fig. 2. F landscape for a series of values for \(N^*\) and \(B\). The plot is based on the 94394hgu95a11.cel file from the '1532 series' in the Human Latin square dataset, available at http://www.affymetrix.com/support/technical/sample_data/datasets.affx.

where \(B\) is the global background, \(N^*\) is the amount of RNA molecules that contribute to non-specific binding (NSB), and

\[
\lambda_{ij} = \sqrt{I_{ij}(1 - e^{E_{ij}})}
\]

If we consider a probe as a string of bases \(\{b_1, b_2, \ldots, b_{25}\}\), and \(\omega_k\) as position-dependent weights, the target-specific free energy \(E\) and the average free energy for NSB, \(E^*\), can be calculated as a weighted sum of stacking energies (\(\varepsilon\)):

\[
E_{ij} = \sum \omega_k \varepsilon(b_k, b_{k+1})
\]

\[
E^*_{ij} = \sum \omega^*_k \varepsilon^*(b_k, b_{k+1})
\]

The minimization of \(F\) is done by permuting the \(B\) and \(N^*\) values and recalculating \(\hat{I}\). We investigated the \(F\) landscape for a series of different chip types and found it to be smooth and with only one minimum, i.e., the global minimum (Fig. 2). Therefore a steepest descent method for minimizing \(F\) was implemented.

In order to calculate \(\hat{I}\) fast for \(N^*\) and \(B\) values, \(N^*\) and \(B\) were isolated in Equation (1). Thus, \(\hat{I}\) can be calculated as:

\[
\hat{I}_{ij} = k_1 - k_2 - N^*k_3
\]

where \(k_1, k_2\) and \(k_3\) are constant matrices:

\[
k_1 = \sum \left\{ \sum (1/(1 - e^{E_{ij}}) \lambda_{ij} \right\} (1 - e^{E_{ij}})
\]

\[
k_2 = \sum \left\{ \sum (1/(1 - e^{E_{ij}}) \lambda_{ij} \right\} - 1
\]

\[
k_3 = \sum \left\{ \sum (1/(1 - e^{E^-_{ij}}) \lambda_{ij} \right\} (1 - e^{E^-_{ij}}) - \frac{1}{(1 - e^{E^-_{ij}})}
\]

Zhang et al. (2003) suggest an outlier rejection criterion for probes where \(I\) deviates more than three standard deviations from \(\hat{I}\). We tested this criterion and found that the overall variation between replicates diminished when the criterion was less strict (data not shown). However, it cannot be excluded that chip hybridizations of dubious quality will benefit from such a criterion. As a consequence we implemented the criterion to be user-specifiable.

The package was developed as an add-on for the affy package (Gautier et al., 2004) for the analysis of Affymetrix Gene chips and can be used in this context. This not only allows the scientist to use the PDNN method in concert with a series of normalization and higher-level analysis tools, but also allows easy comparison between different gene expression index calculations methods, like dChip, MAS v.5, AverageDifference, etc.

REFERENCES