Multi-mapping Bayesian analysis of whole-transcript Affymetrix expression arrays

Ernest Turro, Alex Lewin and Sylvia Richardson

Department of Epidemiology and Public Health, Imperial College London

Contact: ernest.turro@ic.ac.uk

Summary

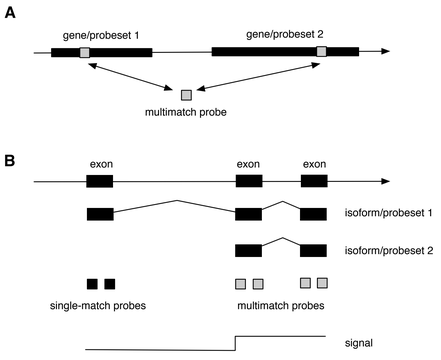
We present a fully hierarchical Bayesian method to analyse whole-transcript Affymetrix arrays that (a) makes use of the new background probe structure and (b) splits the signal captured by probes that match multiple transcripts in a coherent way. The implementation uses shared-memory parallelism to achieve considerable speedups. The method performs well on simulated and real data.

Whole-transcript expression arrays

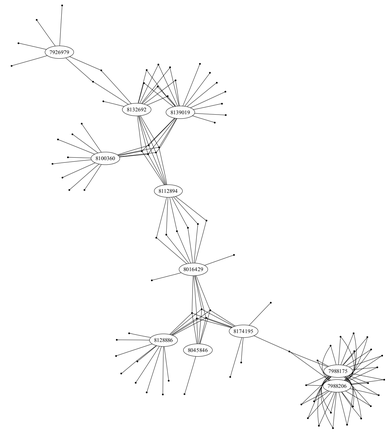
Affymetrix has recently developed whole-transcript expression arrays that interrogate exons along the whole length of each gene, as opposed to previous arrays, which interrogated only the few hundred bases near the 3' end of each gene.

Unlike 3' arrays, the whole-transcript arrays use groups of several hundred "background probes" sharing identical GC content to help quantify the degree of non-specific hybridisation at the Perfect Match (PM) probes.

Some probes bind to more than one transcript. This is because they have a sequence that matches multiple target transcripts/probesets and therefore capture signal from more than one transcript (cf. A and B below).



To estimate the abundance of each transcript accurately, the probe signal should be split according to the mapping structure between probes and probesets. The figure to the right is an example of such a structure, where ellipses represent probesets and dots represent probes. Probesets are linked to their component probes by edges. The number of edges attached to a probe reflect the number of transcripts it measures.



Multi-mapping Bayesian Gene eXpression

We developed a new model that (a) accounts for the use of global GC-content specific background probes and (b) models the complex mapping between transcripts and probes explicitly. Thus signal from probes which match several transcripts is split between them in a logical way.

Perfect match probe intensities, PMs, are modelled as arising from specific hybridisation, S (the signal), and from non-specific hybridisation, H, both of which are probe (j), condition (c) and replicate (r) specific:

$$PM_{jcr} = S_{jcr} + H_{jcr}. \tag{1}$$

The signal S_{jcr} at each probe is modelled on the log scale to account for multiplicative error. To summarise expression at the transcript level, the mapping of probes to transcripts must be taken into account:

- If a probe captures signal from a single transcript, then we use the simple model

$$\log(S_{jcr}) \sim N(\mu_{gc}, \sigma^2_{gc}), \tag{2}$$

where μ_{gc} is the log expression measure and σ^2_{gc} is an error term.

- If a probe captures signal from multiple transcripts, we assume that the contribution of each transcript to the signal and its variance is additive on the real scale:

$$\log(S_{jcr}) \sim N\left[\log \sum_{g \in G(j)} e^{\mu_{gc}}, \log \sum_{g \in G(j)} e^{\sigma^2_{gc}}\right], \tag{3}$$

where $G(j)$ returns the indices of transcripts matched by probe j.

The non-specific binding parameter, H, log-transformed, follows a GC content (k) and array (r) specific normal distribution:

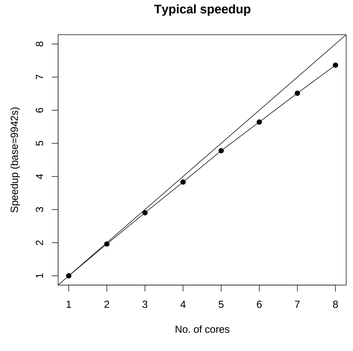
$$\log H_{jcr} \sim N(\widehat{\gamma_{k(j)cr}}, \widehat{\delta^2_{k(j)cr}}), \tag{4}$$

where $k(j)$ returns the GC-content category of probe j. The mean and variance, $\widehat{\gamma_{k(j)cr}}$ and $\widehat{\delta^2_{k(j)cr}}$ respectively, are estimated empirically from the logarithms of the GC-content specific background probes.

The model is estimated in a fully Bayesian framework, thus providing a principled measure of uncertainty in the posterior distribution of the log expression measure, μ_{gc}.

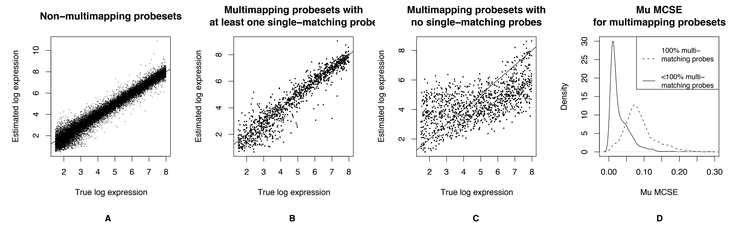
Parallel implementation

Samples from the posterior distributions of the parameters are obtained using a Markov chain Monte Carlo algorithm, implemented in C++. The programme makes use of shared-memory parallelism to obtain significant speed-ups on multi-core computers. On an 8-core computer, runtimes were reduced from 2h45 to only 22 minutes in a single-array analysis, while an analysis of a large data set consisting of 33 arrays spread over 9 conditions was achieved in under 8 hours.



Simulation study

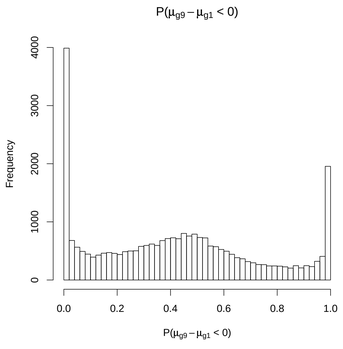
We generated data from the model and evaluated our parameter estimates. The results are shown in plots A to D below.



- A: the log expression measure, μ_{gc} is estimated well for non-multi-mapping transcripts.
- B: the signal from multi-mapping probesets with at least one single-matching probe is also recovered well.
- C: there is shrinkage towards the mean at low and high levels of μ_{gc} for probesets with no single-matching probes.
- D: the estimates of μ_{gc} for multimapping probesets with no single-matching probes (i.e. with high shrinkage) have a higher Monte carlo Standard Error than those with at least one single-matching probe.

Performance on real data

The Affymetrix Gene 1.0 ST Array Data Set contains the results of Human Gene array experiments on 9 mixtures of brain and heart tissues ranging from a 0/100% to a 100/0% brain/heart mixture. The histogram on the right shows the probability of a transcript being under-expressed in the pure brain relative to the pure heart sample. The two peaks represent brain- and heart-expressed genes respectively.



As shown below, we found that the intensities of brain-expressed transcripts follow a consistent upward trend as the brain sample proportion increases and a consistent downward trend as the heart sample proportion decreases, as expected. Transcripts occurring in equal abundance in the two pure conditions have flat intensities across mixture levels.