Filtering and Identifying non-reliable probes in Affymetrix GeneChip® platforms

Some probes of oligonucleotide microarrays show different responses to a signal even if they are supposed to detect the same signal, e.g. they do not detect a signal at all. Such “bad” probes will lead to noisy measurements and should be identified to design curated microarray platforms and to improve post processing. Probe sets are called “informative” if the majority of the probes are consistent in terms of intensity (they increase simultaneously and decrease simultaneously). A probe outlier in an informative probe set (IPSs) is identified as a “bad” probe, because it fails to detect a signal confirmed by the other probes. IPSs are detected by a probabilistic factor analysis model called “FARMS” which allows to identify “bad” probes by computing a signal-to-noise ratio for every probe. Those “bad” probes are analyzed with respect to their annotation, their position in the target sequence, and their probe sequence. The latter is done by support vector machines (SVMs) using spectrum kernels with and without position dependence.

Results: On datasets from six different platforms, spectrum kernels are able to distinguish “bad” from well-performing probes with accuracies in the range of 60–75%, where the position plays a minor role. The SVM generalizes well across platforms, organisms, and experiments, i.e. model selection on one platform and testing on another. As a first result, it was found that over-representation of adenine in a probe sequence indicates a “bad” probe.