Benchmarking the ordering of microarray data

Alexandre Mendes, Regina Berretta and Pablo Moscato
Newcastle Bioinformatics Initiative (NBI)
School of Electrical Engineering and Computer Science, The University of Newcastle, NSW, Australia

Abstract:
Bioinformatics is the “first defense” against the deluge of data that is expected to overwhelm both researchers and decision makers in the 21st century as Biology and Medicine will increasingly depend on information extraction from databases and become large-scale, data-driven sciences. One of the most common problems in microarray studies involves the visualisation of the instances, so that meaningful information can be extracted from it. In this work, we tackle the problem of finding a permutation of the genes names list such that genes with similar expression patterns should be relatively close in the permutation. This problem is commonly known as gene ordering. To help illustrate some aspects of the performance of the algorithms and provide a complementary framework for the analysis, we use images corrupted with known levels of noise. Moreover, as the test images have known high-quality solutions, they facilitate in some cases the assessment of the methods and help the software development, validation and reproducibility of results. Four algorithms were tested: a traditional hierarchical clustering algorithm that is available online at the European Bioinformatics Institute; the CLICK algorithm developed at Tel Aviv University and that uses statistical information to assign genes into clusters; Eisen’s hierarchical clustering; and a memetic algorithm developed by the Newcastle Bioinformatics Initiative.

In this study we analyse a problem that can generically be labeled as gene ordering. More concretely, we are addressing the basic problem of visualising in two-dimensions a large matrix of gene expression information. In some way, we are responding to the need of several biologists who are unsatisfied with the outputs of current commercial and public domain packages (e.g., GeneSpring, Eisen’s Cluster, CLICK). These packages provide interesting and sometimes statistically sound clustering results (for instance, hierarchical clustering and others) but they do not explicitly optimise their output in the way we do in the memetic algorithm. In addition, one of the objectives of this work is to benchmark our results and to introduce instances of the basic problem that would allow to study the robustness to the presence of errors in the measurements.

In Figure 1, we show a series of tests using the so-called Lena instance, with 5,120 rows and 512 columns. The performance of the three algorithms tested varies considerably. Beginning with the Hierarchical clustering from the European Bioinformatics Institute, the method managed to retrieve the most conspicuous features of the image, but with many discontinuities. If microarray data were used, that would translate into similar genes being placed apart, in several groups instead of a large single one. However, the method was able to retrieve the main features even with 40% noise, which indicates good robustness. The memetic algorithm not only retrieved the main features, but also had very few discontinuities. Only the worst case, with 40% noise, the method lost part of its whole picture viewing capability. The CLICK algorithm, has always separated the genes into well-defined clusters, with very distinctive profiles. However, in the presence of noise, the number of clusters diminish, and they become larger as the statistical measures fail to discriminate between the profiles. Finally, the problem of similar genes being assigned to different groups is present again.

The next set of images, shown in Figure 2, is based on an Opera House photo, with 979 x 469 pixels. In this case, we extend the tests to the ordering of rows and columns (i.e. genes and experiments). Similarities between adjacent rows (and columns) in the original image are not as evident as in Lena’s, and the hierarchical clustering once again finds many blocks with similar profiles, but also too many discontinuities. Worth of mention is the resilience of the method to noise. The method maintains the same level of performance independently of the noise level. The memetic algorithm managed to obtain a much clearer picture, with less discontinuities. It appears to better satisfy the main goal of having similar ‘genes’ and ‘experiments’ placed closer in the sequence.

The results with the images indicate that the memetic algorithm would be more successful in finding large groups of co-regulated genes, or samples with similar expression profiles, than hierarchical clustering alone. But how would it perform with microarray images? To answer this question, we present Figure 3, where we compare the results from Eisen’s hierarchical clustering and our memetic algorithm in terms of clusters of genes found. In Figure 3, we present correlation matrices corresponding to both solutions. To identify groups of co-regulated genes, we defined a correlation threshold of 0.7. If the correlation for a given pair of genes is higher than this threshold, it is shown as a green dot in the matrix. If the correlation is lower, it appears as a red dot. The memetic algorithm was able to find better defined blocks of correlated genes, represented by clusters of green dots. In Eisen’s hierarchical clustering solution, the green dots are sparser. For more information regarding the memetic algorithm, please refer to the method’s most recent paper published in Biosystems (Moscato et al., 2006).

References:

Acknowledgements:
This research was funded by an NHM Early Career Researchers in Cancer grant, sponsored by Mrs. Jenine Thomas, HMRI Life Governor.

Contacts:
Alexandre Mendes: Alexandre.Mendes@newcastle.edu.au
Regina Berretta: Regina.Berretta@newcastle.edu.au
Pablo Moscato: Pablo.Moscato@newcastle.edu.au