Summary of discussion for

Deterministic projection by growing cell structure networks for visualization of high-dimensionality datasets

by Jason W.H. Wong & Hugh M. Cartwright.
Journal of Biomedical Informatics 38 (2005) 322–330

INLS 279: Bioinformatics Research Review
Presented by Noel Fiser
September 20, 2005

The purpose of this article was to highlight a new system of dimensionality reduction to allow the two-dimensional visualization of high dimensionality biological data sets. It is the authors' intention to simplify such complex data sets as spectrographic data so that new clusters can be quickly identified for closer study. In order to accomplish this they decided that for various reasons of unstable results they could not use random or deterministic projection, but developed an enhanced algorithm for dimensionality reduction. The authors then presented some examples of this algorithm compared to the less effective methods, first using a standard cube dataset, then a dataset based on the composition of wine, and finally a clinical proteomics dataset. As the datasets became more advanced, the basic algorithms failed to indicate some areas of clustering as well as the proposed algorithm. Also, these deterministic projection do not require preparation of the data in order to find "useful" points from which to start the analysis, unlike random projection which can be skewed considerably by poor node choices early in the analysis.

We found the algorithm compelling and could see its benefit for dimensionality reduction. There was a question, however, of how beneficial this algorithm might be when applied to more and more dimensions and how best to analyze those results (since in the clinical proteomics dataset it was sometimes hard to see how the clustering of the deterministic results was markedly better than the random projection results). Further, there is a reason mass spectrum results are so ubiquitous, as this standard two-dimensional method of analysis already shows the general areas of high clustering without the need for further analysis: the peaks in the mass spectrum are the "hot spots." Therefore, we questioned whether this method would be the best for the day-to-day kinds of analysis, since the purported value of this method is in finding otherwise obscure clusters for further study. This is particularly crucial as datasets get larger, because this algorithm slows down considerably with more complex datasets, requiring more processing power on more powerful machines. Still, the article seemed to prove that when the dataset needs to be analyzed critically, it might be important to apply a deterministic "flattening" to it in order to determine the important clusters in the data.