Bioinformatics at Leibniz Institute of Plant Genetics and Crop Plant Research

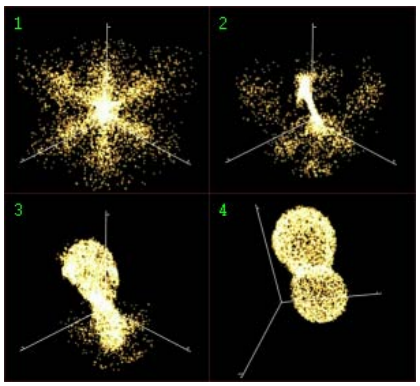

Figure 1 Four different stages of the distance matrix reconstruction procedure. Starting from random initialization (1), an axis is formed in (2), leading via fine tuning (3) to the converged set of 3D points (4) that approximates the relationships between the original temporally regulated gene expression time series, each covering 14 time points.

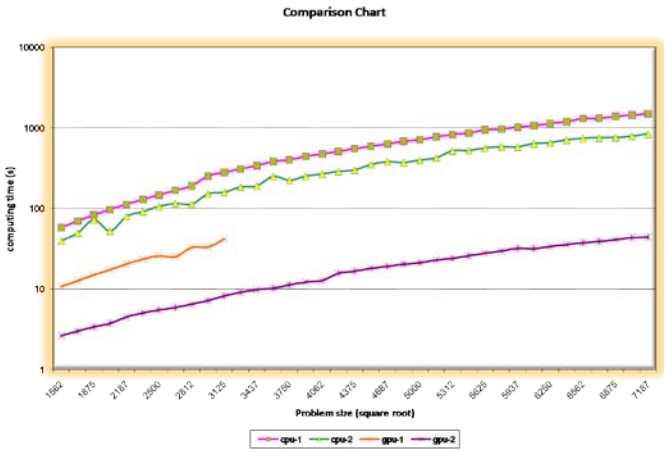

Figure 2 Performance comparison chart for pure MATLAB and Jacket enhanced execution of the HiT-MDS procedure.

cpu-1: AMD Opteron(tm) Processor 2356 with single core operation at 2.3 GHz**;**
cpu-2: AMD Opteron(tm) Processor 8222 with 16-core at 3 GHz each (full multi-threading enabled in MATLAB)**;**
gpu-1: GeForce 9600 GT with 512MB memory and 64 cores at 1.60 GHz (line ends because of memory limitations);
gpu-2: Tesla C1060 with 4GB memory and 240 cores at 1.30 GHz.

The problem size corresponds to the edge length of the involved square source distance matrix to be processed by HiT-MDS. Jacket shows 20 to 35X speed up consistently on various problem sizes.

Multidimensional scaling (MDS) is a general computing technique to turn a distance matrix into a set of reconstructed points with pair-wise relationships approximating the original distances by points located in a usually low-dimensional space. Three typical applications of MDS are:

1. Dimension reduction. Distances from high-dimensional data are taken and reconstructed in a 1/2/3-dimensional space for visualization.
2. Metric conversion. For example, a dissimilarity measure like 1-correlation can be turned into a representation in the Euclidean space.
3. Relationship inference. If only dissimilarities between a (limited) number of entities are given, MDS can be used to infer an estimation of the complete data set.

The algorithm provided here (HiT-MDS) is optimized for dealing with data sets of up to around 10000 points from high-throughput experiments. In contrast to most existing MDS schemes, it does not reconstruct exact distances, but it creates a set of points with a distance matrix that is in maximum correlation with the original (source) matrix. For reaching this aim, points are moved around according to a simplified optimum gradient in the target space such that the desired correlation gets maximized. The process of optimizing these several thousand parameters is based on a gradient of nested derivatives of correlation wrt. distances wrt. point coordinates [1]. This task is perfectly suited for utilizing GPU hardware, because the involved large distance matrices can be efficiently processed in parallel.

HiT-MDS can be considered as nonlinear alternative to principal component projections from PCA. Point neighborhoods of reconstructed points are better preserved in HiT-MDS than in PCA. The result of MDS algorithms is not unique, because rotations of data clouds are not reflected in a distance matrix and thus cannot be resolved. Compared to PCA, HiT-MDS can be also used for non-Euclidean input spaces and for sparse distance matrices.

Visualization of incomplete relationship information and high-dimensional data tables is an important task in bioinformatics. For example, metabolic pathways of sparsely specified data can be processed, and array-based gene expression data can be visually inspected by faithful low-dimensional displays. Figure 1 shows four stages of the point optimization of 4824 temporally regulated genes in the development of barley endosperm tissue over 14 time points. The final sandglass shape of the 3D scatter plot results from the main fraction of either up- or down-regulated (thus anti-correlated) genes that get arranged on opposite sides in the Euclidean space. The plot can be used as part of an interactive browser of high-dimensional data for the identification of outliers and main regulatory patterns. A comparison of runtimes for different sizes of the source distance matrix (problem size) is shown in Figure 2.

[1] Marc Strickert, Nese Sreenivasulu, Björn Usadel, and Udo Seiffert: Correlation-maximizing surrogate gene space for visual mining of gene expression patterns in developing barley endosperm tissue, BMC Bioinformatics 2007, 8:165doi:10.1186/1471-2105-8-165
[http://www.biomedcentral.com/1471-2105/8/165]