BMC Bioinformatics

Research article

Correlation-maximizing surrogate gene space for visual mining of gene expression patterns in developing barley endosperm tissue

Marc Strickert*†[1], Nese Sreenivasulu†[1], Björn Usadel[2] and Udo Seiffert[1]

Address: [1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Gatersleben, Germany and [2]Max-Planck-Institute of Molecular Plant Physiology, 14424 Potsdam, Germany

Email: Marc Strickert* - stricker@ipk-gatersleben.de; Nese Sreenivasulu - srinivas@ipk-gatersleben.de; Björn Usadel - usadel@mpimp-golm.mpg.de; Udo Seiffert - seiffert@ipk-gatersleben.de

* Corresponding author †Equal contributors

Published: 22 May 2007

BMC Bioinformatics 2007, **8**:165 doi:10.1186/1471-2105-8-165

Received: 3 January 2007
Accepted: 22 May 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/165

Abstract

Background: Micro- and macroarray technologies help acquire thousands of gene expression patterns covering important biological processes during plant ontogeny. Particularly, faithful visualization methods are beneficial for revealing interesting gene expression patterns and functional relationships of coexpressed genes. Such screening helps to gain deeper insights into regulatory behavior and cellular responses, as will be discussed for expression data of developing barley endosperm tissue. For that purpose, high-throughput multidimensional scaling (HiT-MDS), a recent method for similarity-preserving data embedding, is substantially refined and used for (a) assessing the quality and reliability of centroid gene expression patterns, and for (b) derivation of functional relationships of coexpressed genes of endosperm tissue during barley grain development (0–26 days after flowering).

Results: Temporal expression profiles of 4824 genes at 14 time points are faithfully embedded into two-dimensional displays. Thereby, similar shapes of coexpressed genes get closely grouped by a correlation-based similarity measure. As a main result, by using power transformation of correlation terms, a characteristic cloud of points with bipolar sandglass shape is obtained that is inherently connected to expression patterns of pre-storage, intermediate and storage phase of endosperm development.

Conclusion: The new HiT-MDS-2 method helps to create global views of expression patterns and to validate centroids obtained from clustering programs. Furthermore, functional gene annotation for developing endosperm barley tissue is successfully mapped to the visualization, making easy localization of major centroids of enriched functional categories possible.

Background

The essence of gene expression analysis is similarity-based screening and structuring of hybridization data. Several methods exist to realize the workflow of raw array data preprocessing, background correction, filtering, clustering and/or classification to identify preferentially expressed genes and to recognize over-represented functional groups using annotation information [1,2]. The quality of each step in that processing pipeline should be validated, though. In this work, a faithful visualization technique for comparative data displays is presented for assisting in validation. Typical questions arising during expression anal-

ysis, addressed by such visualization, are: on one hand, how are hybridization experiments related to each other, and are replication experiments consistent with previously taken data? On the other hand, can correspondence be found between gene-specific expression patterns, and are centroids of gene expressions – such as obtained from k-means or neural gas clustering – located appropriately? Last not least, can typical data clusters be identified by appropriate display, either for experiments or for gene expression patterns?

Principal component analysis (PCA) – often realized as singular value decomposition (SVD) – is the standard technique to create low-dimensional displays of high-dimensional data [3]. Once eigenvectors are calculated, fast linear mappings on the principal components are possible that explain directions of maximum variance. Thereby, Euclidean data space is implicitly assumed for variance maximization. The restriction of PCA to linear mappings of Euclidean spaces can be overcome by using more general multidimensional scaling (MDS) approaches. These assign each high-dimensional data point a low-dimensional counterpart and minimize the discrepancy of the points' relationships in high- and low-dimensional space. High-dimensional input data, for example, might be compared by Minkowski metrics or by Pearson correlation similarity. The low-dimensional output space should be Euclidean – this allows a visual interpretation of close points as representing similar input data, and distant points as indicating dissimilarities. Since, for such view, *high* similarity is expressed by *small* values and vice versa, this inverse interpretation is sometimes referred to as dissimilarity in the literature.

For gene expression data, *correlation* similarity is very useful, because dense clusters of displayed points then do coincide with highly correlated expression vectors. In coexpression-related analysis, time series of gene expressions should be clustered if their temporal profiles are similar, while discorrelated dynamics should be separated. Hierarchical clustering [4], k-means [5], and self-organizing maps (SOM) [6] usually facilitate the grouping task. Some problems remain, though: in hierarchical clustering the resulting ordering is not unique and the corresponding large tree is difficult to access visually; both k-means and SOM induce data abstractions by setting a debatable number of centroids; by choosing additional free parameters for the architecture and learning process, SOM can be used for cluster visualization, but faithful SOM training requires an appropriate choice of parameters – only then, similar clusters do commonly correspond to adjacent SOM centroids. Since the vector quantization in SOM provides a mapping of input vectors to a corresponding centroid, their individuality gets lost which complicates outlier identification. Other authors have pointed out the need for a visual inspection of the gene space for comparison and validation of clustering results. The microarray latent visualization and analysis package (MILVA) is designed for mapping the gene space to a two-dimensional display using either generative topographic mapping (GTM) or the NeuroScale method [7]. Due to its built-in functional mapping, the software is very well suited for smooth interactive gene explorations. However, it requires prior assumptions to estimate density models from the available high-dimensional data for characterizing the underlying data manifold. An embedding technique for dealing with non-metric data relationships is nMDS [8]. This fast multidimensional scaling approach relies on heuristic reconstruction of rank relationships between input data and their corresponding points in the two-dimensional display. These existing data visualization tools are very useful for interacting with the data. Still, there is further need to improve data displays, especially in gene expression studies, for extracting reliable sets of coexpressed genes and for visually assessing relationships between functional categories of coexpressed genes.

A first version of high-throughput multidimensional scaling (HiT-MDS), realizing metric MDS based on a mathematical cost function formulation, has been proposed in the authors' previous work for Euclidean gene space reconstruction [9]. In a more recent study [10], a comparison of HiT-MDS to an algebraic MDS approach and to the free XGvis system [11] is given. It turns out that it is generally problematic to compare a method optimized for a specific cost minimization with a method aiming at other visualization cost criteria. Thus, a pragmatic rating is 'value by usefulness' which strongly depends on biologically informative displays and somewhat also on computing time. In the present study, two substantial extensions of HiT-MDS are described leading to HiT-MDS-2: one extension corresponds to an improvement of the MDS cost function without changing the original embedding quality, the other corresponds to the utilization of non-Euclidean measures for input data, namely, powers of Pearson correlation, for the visual exploration of regulatory patterns in temporal gene expression profiles. Here, we demonstrate the HiT-MDS-2 tool for improved assessment of quality and reliability of centroids of temporal gene expression profiles, and for pointing out visual relationships between functional categories of coexpressed genes. This allows to identify robustly the key regulatory genes in sets of transcriptionally co-regulated genes, such as from developing endosperm tissue in barley.

Results

Data of developing barley endosperm tissue

In order to demonstrate its benefits, the presented HiT-MDS-2 algorithm has been applied to an expression data set obtained from a 12 k seed array (11786 genes) of

developing barley grains [12]. The pursued hybridization experiments produced comprehensive transcriptome data covering all major events of endosperm development from 14 time points corresponding to a time span of 0 to 26 days after flowering (DAF), in two day intervals. The HiT-MDS-2 algorithm is used to address three major questions: 1. How are the experiments, representing transient development of endosperm tissue, characterized with respect to their transcriptome similarity of specifically expressed genes? 2. Which are the main regulatory genes, represented in a set of transcriptionally co-regulated genes in developing endosperm? And, finally, 3. what is their role in explaining temporal differentiation of endosperm tissue?

The 12 k gene expression data set, prepared as discussed in the methods section, is considered from its two fundamental views, one corresponding to individual hybridization experiments each involving 4824 filtered genes, the other corresponding to individual genes with expression values sampled at 14 time points. The embedding-based analysis is thus carried out for (a) experiment grouping and (b) gene profile inspection. Supplemental material is online available [13].

Experiment grouping

Visual experiment validation is obtained by embedding their pairwise correlations $(1 - \mathbf{r}(\mathbf{x}^i, \mathbf{x}^j))$, where $\mathbf{x}^i$ and $\mathbf{x}^j$ are experiments i and j, each containing expression values of 4824 genes. The scatter plot given in Fig. 1 was calculated within 0.5s on a 3 GHz P4 processor with 750 cycles of the data set. The inter-distances of the displayed points correlate at a very high level of $\mathbf{r}_L^2(\mathbf{D},\mathbf{D}) = 0.990$ with the inter-similarities of the original input data. Thus, the visualization represents almost perfectly the relationships in the 4824-dimensional correlation space of the input data. After display normalization, the zero origin demarcates a critical point for the interpretation of symmetry breaks. As a result, axis 1 can be easily associated with temporal development, axis 2 corresponds to systematic differences in both independent series (Fig. 1). The time domain can be described as follows: (i) the initial experiments at 0 DAF are not in the same line as subsequent time points – this slight orthogonal displacement corresponds to the early fertilization event with its unique gene expression, (ii) transcriptional changes during pre-storage phase are slow until day 4, (iii) between 6 to12 DAF (intermediate phase) a strong transcriptional reprogramming takes place, and (iv) the late stage of 16 to 26 DAF (storage phase) is characterized by a saturation process, indicated by a higher point density on the right, with diminishing transcriptional regulation.

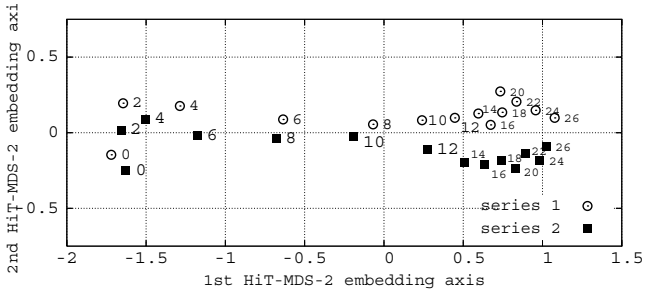


Figure 1
Embedded relationships of cDNA array experiments. HiT-MDS-2 visualization of inter-relationships between cDNA array experiments from two independent series of developing barley endosperm tissue. Experiments with 4824 selected $\log_2$-normalized genes are compared by (1-Pearson correlation)^p at power p = 1. Numbers denote days after flowering (DAF). From left to right a clear temporal order is found, corresponding to pre-storage (0–4 DAF), intermediate (6–12 DAF), and storage (14–26 DAF) phases of endosperm development. Day zero, related to the fertilization event, is systematically separated from rest of the early stages. While a relative delay of roughly two days is found between both experimental series during intermediate stages, late stages become more tightly linked (14–26 DAF). Embedding axis 2 separates the two series. Slight systematic differences of series 1 and 2 result from low and high phosphor image scanning resolutions, respectively, and thus from different dynamic ranges.

Although embedded experiments are arranged in a consistent manner, not showing major outliers, series 2 is further considered in the following: it exhibits a smoother temporal transition between 6 DAF to 12 DAF than series 1, and, in addition, a better dynamic signal range was found, because the underlying phosphor images were scanned at higher resolution than those of series 1.

Gene profile inspection

HiT-MDS-2 scatter plots for the visual validation of clusters of gene expression patterns

Dealing with thousands of temporally regulated genes is a crucial task. Tools for intuitive inspection of the gene space help to identify coexpressed gene sets associated with biological processes occurring during development. The ESTs selected for the 12 k seed array fabrication were taken from cDNA libraries specific to pre-storage and storage phase of developing seeds. This selection leads to pronounced temporal gene regulation, which results in a bipolar sandglass shape in the corresponding HiT-MDS-2 display of embedded expression data. This shape represents genes with up- and down-regulation, corresponding to pre-storage and storage phase, respectively (Fig. 2). Start and end of development, and the temporal transition between the phases have been characterized in Fig. 1

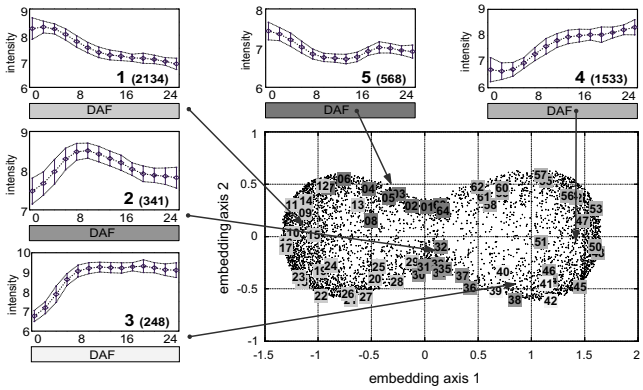


Figure 2

Gene expression correlation space obtained by HiT-MDS-2. Surrogate gene space of developing barley endosperm tissue. A total number of 4824 temporally regulated, log-normalized gene profiles with 14 time points are embedded using powers of $p = 8$ of Pearson correlation for profile comparison. The bipolar sandglass shape of points is labeled by 64 centroids obtained from correlation-based neural gas (NG) clustering which yields a good density-related covering of the data space. The clusters fall into five biologically meaningful regulatory different patterns: each of the five small panels displays the average and standard deviation of aligned genes belonging to the marked centroids – the number of genes within each category 1–5 are given in parentheses. The HiT-MDS-2 embedding shows a consistent spatial arrangement of clusters within the manually selected categories (apart from clusters 8, 13, and 51); pre-storage phase (cluster group 1), intermediate phase (cluster groups 2 and 5, anti-correlated to each other), and storage phase (cluster groups 3 and 4). Since the exponent $p = 8$ magnifies small differences, the resulting large spatial variability for down- and up-regulation represented by cluster groups 3 and 5 do still correspond to only small standard deviations. The small asymmetric bias of the gene cloud to the right indicates more subtle patterns found in up-regulation rather than in down-regulation, which underlines the benefit of the visualization for screening and validation.

in the previous section. As explained below, by using Fig. 2 a set of 340 genes with intermediate regulation can be detected, which is responsible for the observed transition event.

The sandglass shape with its well-spread points results from power transformations of correlations, which magnifies subtle dissimilarities in highly correlated genes. In the presence of many coexpressed genes, powers applied to the input similarities shift the corresponding histogram towards zero; this leads to focus on a good reconstruction – and thus a fair differentiation – of highly correlated, i.e. with near zero dissimilarities, rather than of obviously discorrelated genes. A power of $p = 8$ applied to the input similarities, i.e. $(1 - \mathbf{r}(\mathbf{x}^i, \mathbf{x}^j))^8$, is a good choice for clearly separating between up- and down regulated expression patterns during pre-storage and storage phase. Setting $p = 8$ is a compromise for spreading highly correlated genes and for giving space also to intermediate regulations. Comparative density plots for exponents of $p = 24, 4, 1, 0.25$ are available extra [see Additional file 1]. These results indicate how powers of correlations help to emphasize the specific relationship structure in the set of genes. Similar findings are also reported by Zhou et al. for shortest path analysis in gene expression data [14].

For the number of 4824 genes, 100 data cycles are sufficient to get a high-quality display shown in Fig. 2. Overall, the HiT-MDS-2 embedding procedure applied to transcriptome data of endosperm development yields a faithful arrangement of genes with their typical temporal expressions. These are clearly divided into sets with expressions of pre-storage (cluster group 1), intermediate (cluster group 2), and storage phase (cluster group 3 and 4). The corresponding temporal expression patterns are revealed by browsing the scatter plot from the left to the right side of the bipolar sandglass shape (Fig. 2). In addition, we also noticed very interesting patterns showing dominant expression values in the pre-storage phase with drastic decrease in the intermediate stage, followed by an increase of expression levels during the storage phase (Fig. 2, cluster group 5). These results indicate that the non-linear data embedding technique of HiT-MDS-2 is a useful tool for identifying not only the major global patterns occurring during temporal development; also informative minor patterns that could be easily missed in noisy subsets of gene expression data show up as scattered point sets.

We further examined whether the non-linear 2D representation of the gene space obtained by HiT-MDS-2 is also useful for the validation of centroids from existing gene expression clustering algorithms. The neural gas (NG) clustering method according to [15] has been employed using Pearson correlation for centroid computation [16]. A number of 64 NG centroids has been embedded together with the gene expression data using HiT-MDS-2. The result displayed in Fig. 2 shows that the 64 centroids are well distributed among the embedded data, demonstrating that these clusters represent a continuum of data. Thereby, centroid numbers 1 to 8 and 63 to 64 depict similar expression patterns in neural gas clustering, which can be easily validated based on physical co-localization of centroid positions in the HiT-MDS-2 gene space plot (Fig. 2). Redundancy of the 64 centroids has been removed by summarizing them manually into the five shown major developmental patterns. These have been obtained by browsing and grouping temporally similar expressions, located at high-density peripheral regions of the bipolar embedding structure. From a global point of view, sets of coexpressed genes are identified reflecting the major cellu-

lar physiological events happening during endosperm development [see Additional file 2]. In conclusion, the output generated by HiT-MDS-2 provides faithful visualization of cluster relationships. This is a very helpful tool for the definition and validation of major centroids of gene expression profiles and for the assignment of their developmental patterns.

HiT-MDS-2 scatter plots for the visualization of relationships between functional categories of temporally coexpressed genes

In recent years, it has become general practice to subject high-throughput gene expression data to clustering methods and to browse the obtained clusters for finding representations of statistically significant functional categories of genes. Analysis by hierarchical clustering or k-means is usually complicated in the presence of high-dimensional input data and noisy outliers, the latter also affecting the interpretation of SOM clustering results. Statistical tests such as Fisher's exact test, ANOVA based global test, or gene set enrichment analysis (GSEA) produce useful hypotheses about significant transcriptional regulation [17], but they require that preconditions like certain data distributions are fulfilled and that test parameters are chosen carefully. Here, the neural gas clustering method is used with Pearson correlation similarity measure for computing cluster centroids. This method is known to yield consistent high-quality clusters, regardless of centroid initialization [15]. As with other centroid-based methods, though, the number of centroids required for deriving biological meaningful functional categorization can be hardly assessed in advance and induces additional data validation steps. By its correlation-preserving embedding facility, HiT-MDS-2 provides visual support of correlation structures and centroids by screening the spatial neighborhood of candidate genes to inspect whether they belong to clusters of certain functional categories. Here, we used manually annotated functional categories available for the 12 k barley seed array (N. Sreenivasulu and B. Usadel, unpublished data). The annotations are mapped to the embedding output of HiT-MDS-2 and get associated with corresponding expression profiles representing major developmental patterns of coexpressed genes. This mapping allows an easy transfer of biological information to the outcome of array experiments. Thereby, two levels of information are generated concerning (i) the identification of major pathways active in a particular stage of development, and (ii) the extraction of key regulators within transcriptionally co-regulated sets of genes.

(i) The mapping of individual super-pathway information to the genome-wide graphical representation of the transcriptional response during plant ontogeny yields immediate hints about the occurrence of key biological processes during particular stages of development. For instance, this method, applied to transcriptome data of endosperm development, indicates that the abundance of genes related to photosynthesis, minor carbohydrates, and also for early steps of starch biosynthesis is characteristic of the intermediate stage (Fig. 3, cluster 2a) [see Additional file 3]. Clusters 2a-4b, described by the encircled regions, have been manually selected for focussing on (intermediate and storage) up regulation. These are related to the onset of storage events according to the down-stream pathway of starch metabolism (cluster 2b), storage proteins/protease inhibitors (cluster 3a and 3b) and TAG biosynthesis genes (cluster 3b, 4a and 4b). Such systematic activation of consecutive pathways reflects major physiological events happening in developing endosperm tissue. For instance, the end of the cell division phase is marked by an intermediate stage which is characteristic of the starch accumulation initiation. During this phase, coexpressed pathway genes are noticed that show tight physiological links to the photosynthesis-associated, ATP-producing energy metabolism, and to the production of carbon skeletons for synthesis of seed storage products. This initiation is followed by an accumulation of storage proteins at the peak of storage processes and lipid accumulation. As illustrated, such a mapping of functional information allows a serviceable transfer of biological knowledge to the outcome of array experiments.

(ii) Browsing the subspaces of the HiT-MDS-2 plot helped to identify key regulatory genes situated closer to major pathway genes, such as in case of starch, storage proteins, and oleosins. The highly correlated gene sets were extracted and compiled in a supplemental table [see Additional file 3]. As exemplary approach we discuss the coexpressed regulators of starch and storage protein transcripts in the following. The prominent transcription factors expressed during the intermediate development phase of endosperm tissue include 3 members of C3H/C3HC4, 2 chromatin remodeling factors, 1 bZIP, 1 ABI3/VP1, and 5 unclassified transcription factors. These are tightly coexpressed along with genes for photosynthesis, minor carbohydrates, as well as ADP-glucose pyrophosphorylase (AGPase) and sucrose synthase transcripts related to starch metabolism genes (cluster 2a) [see Additional file 3]. Among those regulators we noticed well-characterized regulatory factors, such as ABA response element binding factors (ABF3) from the bZIP family, and the abscisic acid insensitive protein 3 from the ABI3/VP1 family, of which its homologues are supposed to participate in promotion of reserve accumulation in dicots [18]. The correlation structure in the subspace of cluster 2b is related to expression of SNF1, bZIP transcription factor ABI5, MYB transcription factor, YABBY family transcription factor, Squamosa promoter binding factor, Auxin response factor and four unclassified transcription factors along with down-stream branching enzymes of starch metabolism

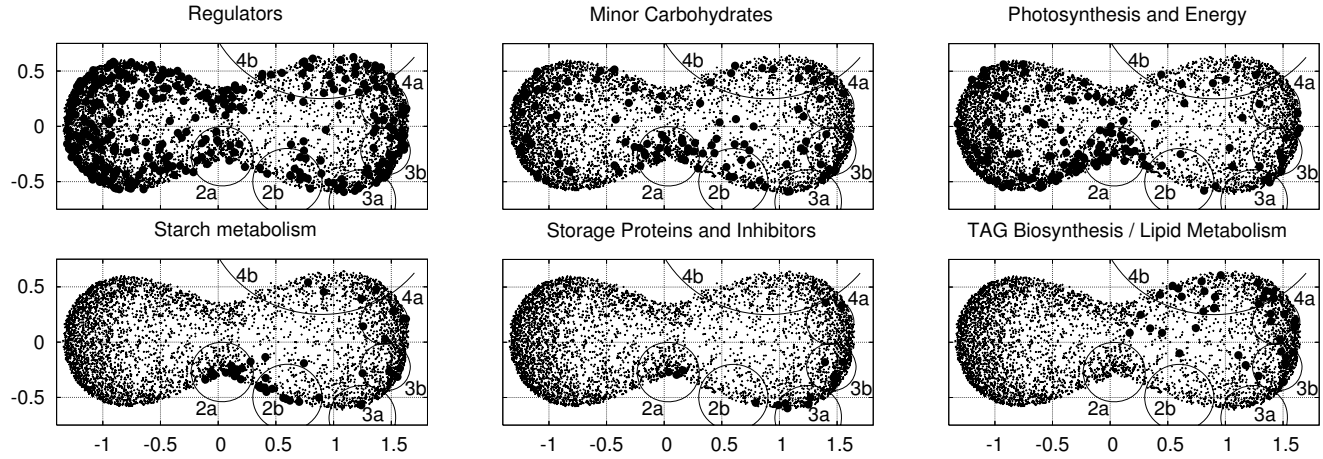


Figure 3

Visual mapping of functional gene categories. Mapping of six major functional categories to the HiT-MDS-2 scatter plot of genes. These categories can be directly related to the five prominent temporal patterns, cluster group 1–5, of gene expression in Fig. 2. Here, the focus is put on manually picked subspaces 2a & 2b of genes related to cluster group 2, 3a & 3b of group 3, and 4a & 4b of group 4. By browsing these subspaces defined by the encircled regions, key regulators can be identified that are closer to major genes of the storage pathway, storage proteins and inhibition, and TAG/lipid metabolism related genes. A list of coexpressed genes corresponding to the regions 2a/b, 3a/b, and 4a/b is provided extra [see Additional file 3]. Corresponding gene profiles are provided in a supplemental figure [see Additional file 6].

and genes controlling minor carbohydrates [see Additional file 3].

As a main highlight, we observed expression of ABA response element binding factors (ABF3, ABI5 and ABI3) coexpressed along with SNF1 and starch biosynthetic genes during the intermediate stage and in the first storage peak of endosperm development. We found ABRE elements in the SNF1 kinase promoter region [12] which indicates a positive role of ABA in triggering these regulators. As recently demonstrated, ABA positively interacts with sugar signaling pathways in controlling key starch biosynthesis genes via SNF1 kinase [19]. Based on the correlative evidences, we also propose that SNF1 expression in endosperm is mediated by ABA via ABF3/ABI5, ABI3, which in turn might be responsible in regulating key genes of starch biosynthesis such as sucrose synthase and ADP-glucose pyrophosphorylase [12].

Another set of transcription factors, preferentially coexpressed along with transcripts of the hordein storage protein and the protease inhibitor during the main storage phase of endosperm development, includes 8 chromatin remodeling factors, 3 NAC, 2 DOF, and 9 unknown transcription factors. It was shown recently that two DOF transcription factors (SAD and BPBF) serve as activators of B1 storage protein genes during the maturation phase [20]. In the present study we also noticed (a) coexpression of two DOF family members, SAD and BPBF transcription factors along with hordein storage protein transcripts, and (b) in our recent study [12] we found enrichment of prolamin box cis-elements in upstream sequences of rice prolamin class storage protein genes (D, B1 and B3 hordeins). These evidences again point out that our detailed bioinformatics analysis of co-regulation of transcription does not only enhance our comprehensive knowledge of the developmental phenomena at gene regulation level, but it also helps to get initial glimpse of the systemic description of gene regulatory networks and their dynamics.

Discussion

The validation of temporal gene expression centroids obtained by commonly used unsupervised clustering methods is a nontrivial task [4-6]. Since clustering results depend on the choice of method, the similarity measure, and the number of centroids, the assignment of expression profiles to clusters of interest does profit from faithful visual assistance. The proposed HiT-MDS-2 data embedding tool is designed to meet this purpose. Its versatile visualization abilities can be used to validate the results of centroid-based clustering methods, as has been demonstrated in the present study for the iterative neural gas clustering approach.

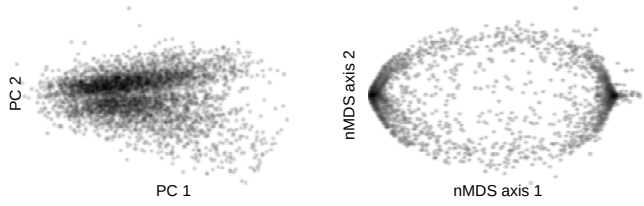
Moreover, HiT-MDS-2 scatter plots can be used for browsing interrelated temporal gene expression patterns (tightly coexpressed genes), and also the relationships between functional categories of coexpressed genes can be easily screened. Such a co-visualization of genes, exhibiting

characteristic regulatory patterns, and their functional assignments is the major benefit of the nonlinear surrogate data representation realized by HiT-MDS-2.

An additional study has been carried out in order to demonstrate the generality of HiT-MDS-2 also for other data sets. We switched from the 12 k seed array containing EST clones selected from developing seed cDNA libraries (see results section) to 22 k Barley 1 Affymetrix chip in which oligos are compiled from at least 84 cDNA libraries encompassing various stages of plant ontogeny. This Affymetrix data set covers stages of developing endosperm tissue at 4, 8, 16 and 25 DAF in two replicate series. We applied two gene filtering criteria to the data set with (a) gene profiles with Pearson correlation greater 0.8 between the two available replicates and (b) at least 2-fold change between minimum and maximum expression values at 4, 8, 16 and 25 DAF. The filtered gene set contains 3031 differentially expressed high-quality genes. As shown in an additional figure, HiT-MDS-2 embedding of these genes produced a sandglass shape similar to Fig. 2 for 12 k seed data set [see Additional file 4]. Furthermore, clear global patterns of up-, down- and intermediate regulation are identified by browsing the obtained gene space [see Additional file 5]. This result confirms that the application of HiT-MDS-2 is not restricted to one specific data set but that it can be transfered to Affymetrix data as well. Thus, regulatory pattern structures revealed by HiT-MDS-2 are no artefacts of data selection, but they do reflect inherent properties of barley endosperm development.

Comparison of Hit-MDS-2 with related visualization tools

Despite of the growing number of unsupervised clustering tools for gene expression data, currently only few visualization techniques offer intuitive validation of the clustering results. HiT-MDS-2 provides great flexibility in the choice of similarity measure, and also the dimensionality of the visualization can be chosen freely. One major advantage over SOM visualization is that the genes keep their individuality in the scatter display, which can be visually clustered on demand. Likewise, expression data and centroids from specific clustering methods can be embedded simultaneously for validation purposes. A standard data projection method like PCA puts too many constraints on the data similarity measure and on the modeling quality of surrogate data. By nature, PCA is restricted to the domain of Euclidean input spaces where variance is a properly defined concept [3]. Projection results of PCA are given in the left panel of Fig. 4. The density image displays the projection of the 4824 genes to the second principal component (PC2) against the projection to the first principal component (PC1). Two separated regions are revealed, the upper region corresponding to down-regulated gene profiles, the lower high-density region to up-regulated gene profiles. In contrast to correlation-based



PCA projection. nMDS embedding.

Figure 4
Visualization of the gene space by other methods. Density plots by PCA and nMDS, dark shading denoting high gene densities. For comparison, a much better and uniform visual spreading of the genes is provided by HiT-MDS-2, as displayed in Fig. 2.

methods, the separation is not very strong, but the different structure of high-density regions indicates different regulatory characteristics specific to up-and down-regulation. The PCA result is complemented to the the much more advanced non-metric MDS (nMDS) method of Taguchi and Oono [8] shown in the right panel of Fig. 4. In comparison to PCA, many more details of the expression profile correlation structure is captured by the nMDS method. Like HiT-MDS-2 a bipolar structure appears, representing patterns of down-regulation at the left pole and up-regulation at the right pole. This density plot of nMDS is indeed very similar to the one obtained by HiT-MDS-2 for exponent $p = 1$ given in a supplemental figure [see Additional file 1]. However, since nMDS turns the implemented Pearson correlation input similarities $(1 - \mathbf{r}(\mathbf{x}^i, \mathbf{x}^j))$ by a sorting operation into their ranks, there is no difference to the monotonic eighth power wrapper $(1 - \mathbf{r}(\mathbf{x}^i, \mathbf{x}^j))^8$. Compared to PCA and nMDS, the display of HiT-MDS-2 in Fig. 2, based on powers of correlation ($p = 8$), exhibits the characteristic bipolar sandglass shape representing not only magnified areas of up- and down-regulation, but also distinct intermediate regulation. A supplemental figure shows how the choice of exponent p can be used to emphasize specific correlation structure [see Additional file 1]. In principle, the XGvis system [11] is able to yield similar embedding results, but it requires that the similarity matrix is computed in advance as input to XGvis.

Regarding the computation efficiency, the HiT-MDS-2 algorithm is outstanding in the domain of metric MDS: it takes only 14 minutes and 21 seconds (861s) for 100 data cycles on a 3 GHz P4 processor for embedding the 4824 genes, while XGvis, for example, requires more than 4 hours for a comparable result. The nMDS approach, pursuing non-metric optimization, generated the displayed embedding within only 18 cycles using a relatively short time of 17 minutes and 21 seconds (1041s). Although, the PCA computation took less than 2 seconds on the ref-

erence PC with a 3 GHz Pentium 4 processor, the visualization cannot be used for screening temporal ordering gene expressions and is, hence, worthless for our purposes. Complementary to the visualization of the gene space, HiT-MDS-2 can also be used to display and evaluate hundreds of hybridized cDNA arrays without significant time requirement.

Conclusion

HiT-MDS-2 allows creating faithful surrogate spaces, such as 2D scatter plots with Euclidean metric, from input spaces with custom data similarity measures. Fast convergence of the reconstructed space is obtained by stochastic optimization of an efficient correlation-based comparison of source and target space. For source data comparison, it has turned out that very useful graphical outputs are obtained when the short 14-dimensional expression time series of our 4824 genes are compared by 8th power of (1 - $\mathbf{r}(\mathbf{x}^i, \mathbf{x}^j)$).

Resulting scatter plots of the well-distributed embedded points have been utilized in four ways: (1) for finding inter-sample correlations among experimental series; (2) for the detection of global regulatory gene expression patterns and for centroid validation; (3) for browsing the major temporal gene expression data and revealing the underlying functional pathway information; and (4) for visual mapping of regulatory genes co-localized with major functional gene categories. These features allow convenient visual screening of thousands of genes in parallel from time-course experiments. Although we have demonstrated only temporal data for screening co-responses in this study, HiT-MDS-2 can be also applied to highlight systematic differences among mutants or transgenics at multiple stages. The obtained visualizations help to get insights to massive data sets for approaching the goal of deriving new biological knowledge.

Methods

Multi-dimensional scaling (MDS) implies the optimization of free parameters $\hat{\mathbf{x}}^i \in \hat{\mathbf{X}}_{n\times d}$, i.e. locations of points $\hat{\mathbf{x}}^i = (\hat{x}_1^i, ..., \hat{x}_d^i)$ in a d-dimensional target space corresponding to $i = 1...n$ input vectors $\mathbf{x}^i \in \mathbf{X}_{n \times q}$ of dimension q. In case of the classical stress criterion, mutual distances $\hat{\mathrm{d}}_{ij} = \mathrm{d}(\hat{\mathbf{x}}^i, \hat{\mathbf{x}}^j)$ of all data pairs indexed by (i, j) should best fit the original distances $\mathrm{d}_{ij} = \mathrm{d}(\mathbf{x}^i, \mathbf{x}^j)$ in terms of the least squares $\mathrm{s} = \sum_{i<j}^{n} (\mathrm{d}_{ij} - \mathrm{d}_{ij})^2 = \min$.

Improvements of MDS (HiT-MDS-2)

A visual control, equivalent to least squares fit, is the Shepard diagram where on the d_{ij} vs. $\hat{\mathrm{d}}_{ij}$ plot all points should be located on the diagonal line of unit slope, i.e. $(\mathrm{d}_{ij} - \mathrm{d}_{ij})^2 = \min$. Although the Shepard plot is usually provided by MDS packages, it implies a misleadingly strict quality criterion: in most cases it is sufficient to maintain only the intra-distance relationships, while the scaling factor between source and target distances, i.e. the scale sizes of the corresponding point clouds, need not be unity. Thus, the strict least squares criterion can be relaxed to shift- and scale-invariant comparison by maximizing the Pearson correlation between the lower triangular source and target distance matrix:

$$\mathbf{r}_{\mathrm{L}}(\mathbf{D},\hat{\mathbf{D}}) = \frac{\sum_{i<j}^{n} (\mathrm{d}_{ij} - \mu_{\mathbf{D}}) \cdot (\hat{\mathrm{d}}_{ij} - \mu_{\hat{\mathbf{D}}})}{\sqrt{\sum_{i<j}^{n} (\mathrm{d}_{ij} - \mu_{\mathbf{D}})^2} \cdot \sqrt{\sum_{i<j}^{n} (\hat{\mathrm{d}}_{ij} - \mu_{\hat{\mathbf{D}}})^2}} =: \frac{B}{\sqrt{C \cdot D}} \in [-1;1] \tag{1}$$

$$\text{with } \mu_{\dot{\mathbf{D}}} = \frac{2}{n \cdot (n-1)} \cdot \sum_{i<j}^{n} \dot{\mathrm{d}}_{ij},\ \dot{\mathbf{D}} = \{\mathbf{D}, \hat{\mathbf{D}}\} \rightarrow \dot{\mathrm{d}}_{ij} = \{\mathrm{d}_{ij}, \hat{\mathrm{d}}_{ij}\}. \tag{2}$$

Matrix $\mathbf{D} = (\mathrm{d}_{ij})_{i,\ j\ =\ 1...n}$ contains pattern distances, and matrix $\hat{\mathbf{D}} = (\hat{\mathrm{d}}_{ij})_{i,j=1...n}$ those of the reconstructions. In principle, input and output spaces are generic. However, the target configurations $\hat{\mathrm{d}}_{ij}$ should be modeled by a Euclidean space for realizing intuitive low-dimensional spatial arrangements, such as 2D plots; furthermore, in the following, input distances d_{ij} are expressed as dissimilarities by taking powers of gene profile correlations **r**. The two measures for reconstructions and input data are

$$\hat{\mathrm{d}}_{ij} = \sqrt{\sum_{l=1}^{d} (\hat{x}_l^i - \hat{x}_l^j)^2}, \quad \mathrm{d}_{ij} = \left(1 - \mathbf{r}(\mathbf{x}^i, \mathbf{x}^j)\right)^p.$$

Thereby integer exponents $p \geq 1$ control the discrimination of input data: in this study, a large value of $p = 8$ is used for separating clusters of highly correlated gene expression profiles, while $p = 1$ emphasizes the separation of anti-correlated patterns. The choice of the real value $p > 0$ is application-specific and up to the user's desire to accentuate the reconstruction of close or distant data.

In Eqn. 1 the abbreviated shorthand fraction is a literal one-to-one correspondence to the explicit term with sums. $B = B(\hat{\mathrm{d}})$ is related to the mixed summation of both original and reconstructed distances, $D = D(\hat{\mathrm{d}})$ refers to

the dissimilarities dependent on the choices of the reconstructions $\hat{\mathbf{X}}$, and C denotes the connection to the initially calculated and thus constant input pattern distances.

Instead of maximizing $\mathbf{r}_L$ directly, minimization is performed on a very efficient stress function that inverts and stretches the domain [-1;1] of Pearson correlation for getting good convergence. In previous work, inverse power transformations of the correlation $\mathbf{r}_L$ have been considered that worked reasonably well [9]. However, an exponent parameter required there had to be chosen carefully in combination with the step size of the stochastic gradient descent. Here, new formulas are derived for Fisher's Z' wrapper of the correlation $\mathbf{r}_L$ given in Eqn. 1. This alternative transformation yields superior convergence while being more robust with respect to the choice of parameters.

The new stress function is based on Fisher's *negative* Z'-transformation:

$$s = -\frac{1}{2} \cdot \log\left(\frac{a + \mathbf{r}_L(\mathbf{D},\mathbf{D})}{a - \mathbf{r}_L(\mathbf{D},\mathbf{D})} \right), \quad a = 1 + \varepsilon .$$

Fisher's original formulation implies $a = 1$; here, however, potential singularities are prevented by $a > 1$.

The stress function s is minimized by optimally arranging the reconstruction points $\hat{\mathbf{X}}$ in the Euclidean target space. This is achieved by a gradient descent on the stress function s, which requires finding zeros of the derivatives of s with respect to the free parameters $\hat{x}_k^i$:

$$s = -Z' \circ \mathbf{r}_L \circ \hat{\mathbf{D}} \circ \hat{\mathbf{X}} \to \min \tag{3}$$

$$\Rightarrow \frac{\partial s}{\partial \hat{x}_k^i} = -\sum_{j=1\ldots n}^{j \neq i} \frac{\partial Z'}{\partial \mathbf{r}_L} \cdot \frac{\partial \mathbf{r}_L}{\partial \hat{d}_{ij}} \cdot \frac{\partial \hat{d}_{ij}}{\partial \hat{x}_k^i} \to 0, \; i = 1\ldots n \tag{4}$$

Solutions are found by iterative updates of randomly drawn points i by $\Delta \hat{x}_k^i = -\gamma \cdot \frac{\partial s}{\partial \hat{x}_k^i}$ of step size γ into the direction of the steepest gradient of s. Although convergence to a global optimum cannot be claimed by such an approach, final point configurations have been found to be very stable and of high quality in different runs. The required derivatives of Eqn. 4 are

$$\frac{\partial Z'}{\partial \mathbf{r}_L} = \frac{a}{a^2 - \mathbf{r}_L^2}$$

$$\frac{\partial \mathbf{r}_L}{\partial \hat{d}_{ij}} = \frac{(d_{ij} - \mu_{\mathbf{D}}) \cdot D - (\hat{d}_{ij} - \mu_{\hat{\mathbf{D}}}) \cdot B}{D \cdot \sqrt{C \cdot D}}$$

$$\frac{\partial \hat{d}_{ij}}{\partial \hat{x}_k^i} = (\hat{x}_k^i - \hat{x}_k^j) / \hat{d}_{ij}.$$

The two parameters of the new HiT-MDS-2 are non-critical and they can be fixed to $\gamma = 0.1$ and $a = 1.001$ in most cases. This robustness is a substantial advantage over the first HiT-MDS formulation described in [9], where a tight coupling of an additional parameter with the learning rate γ required three to five re-runs of the algorithm with appropriately chosen parameters. As a consequence, the old version took on average four times longer to converge to the same final results like the new HiT-MDS-2.

The embedding procedure is outlined in Algorithm 1. Initially, a random projection of the high-dimensional source data is calculated and the resulting similarity matrix is correlated with the mean-subtracted original one for obtaining B, C and D. Mean subtraction from $\mathbf{D}$ does not affect the Pearson correlation $\mathbf{r}_L(\mathbf{D}, \hat{\mathbf{D}})$, but simplifies further calculations. More substantial speed-up is obtained by exploiting the symmetry of $\mathbf{D}$ and $\hat{\mathbf{D}}$ and by implementing differential updates of B and D corresponding to changes in single rows and columns of $\hat{\mathbf{D}}$ during the iterative embedding process: further details on an efficient realization of line 12 in Algorithm 1 are given in [9]. Finally, stochastic gradient descent on s maximizes the correlation iteratively by moving the target points $\hat{\mathbf{x}}^i$ into proper places until a saturated quality level is reached.

Features of HiT-MDS-2

Embedded point distances maximize correlation with source data similarities for a faithful display of relationships. Classical MDS stress applied to Euclidean data yields final configurations equivalent to the PCA projection [21]. The HiT-MDS-2 criterion, though, provides more degrees of freedom and allows thus fast convergence and improved displays. Any symmetric similarity matrix of relationships between input data can be processed, but different powers of correlation measures turn out to be preferable in the context of gene expression mining. In principle, for Euclidean target displays, MDS axes of embedded data do not carry any special meaning, because the embedding procedure is invariant to offsets, scaling, sign-flipping, and rotation; thus, there is no preferred intrinsic direction. What counts is the arrangement of

inter-point distances only. Final displays can and should be normalized

1: Read input data $\mathbf{X}$.

2: Initialize $\hat{\mathbf{X}}$ by random projection $\hat{\mathbf{X}}_{n\times d} = \mathbf{X}_{n\times q} \cdot \boldsymbol{R}_{q\times d}$.

3: Calculate input matrix $\mathbf{D}$ and subtract mean $\Rightarrow$ constant C.

4: Calculate target distances $\hat{\mathbf{D}} \Rightarrow$ initial B, D.

5: **repeat**

6: Draw a pattern index $1 \le i \le n$ from randomly shuffled list.

7: **for all** $j \neq i$ do

8: $\Delta\hat{\mathbf{x}}^i \leftarrow \Delta\hat{\mathbf{x}}^i - \frac{\partial \mathrm{s}}{\partial \hat{x}^i_k}$ { accumulate derivatives (Eqn. 4) }

9: **end for**

10: $\hat{\mathbf{x}}^i \leftarrow \hat{\mathbf{x}}^i + \gamma \cdot \Delta\hat{\mathbf{x}}^i$ { adapt location of target point }

11: Recalculate distances $\hat{\mathrm{d}}(\hat{\mathbf{x}}^i, \hat{\mathbf{x}}^j)$ influenced by new point $\hat{\mathbf{x}}^i$;

12: thereby, update differential changes in B and D.

13: **until** convergence criterion is met.

14: Postprocess: center $\hat{\mathbf{X}}$, normalize by largest dimension variance.

15: Optional coordinate rotation: project $\hat{\mathbf{X}}$ to eigenvectors.

Algorithm 1: HiT-MDS-2

for comparison purposes. Thereby, four steps of (1) mean centering, (2) variance-based rescaling, (3) PCA coordinate rotation, and (4) skewness-based sign flipping properly resolve embedding invariances while maintaining the reconstructed distance relationships. Furthermore, density plots of the embedded data can be computed by using symmetric Gaussian kernels in order to inspect the similarity densities in the data space. GNU Octave/MATLAB and R implementations as well as fast C source code of the HiT-MDS-2 algorithm are available under GPLv2 license [22].

Data preparation

A total of 330 008 gene expression values collected from 28 hybridization experiments with 12 k macroarrays, covering 14 temporal developmental points from two independent series, were considered for data processing. As a first quality criterion, gene expression values surpassing twice the background level are considered. Background subtraction is carried out for the remaining genes, followed by quantile normalization. This processing is done separately for each series to allow the comparison of signal intensities across time series. As usual, $\log_2$-transformed final expression values are considered. Cubic spline smoothing with moderate smoothing parameter has been applied to each temporal gene expression profile. A filter based on Pearson similarity has been applied to select gene profile time series that correlate at a conservative level of $r > 0.5$ between the two independent series. With this criterion, a qualified subset of 4824 out of 11786 genes has been created for analysis.

Authors' contributions

MS implemented the software and applied it to gene expression data. MS and NS designed the study and prepared the manuscript. BU and NS contributed the barley ontologies with functional categories. All authors have read and approved the final manuscript.

Additional material

Additional file 1

Additional HiT-MDS-2 embeddings of the expression data set containing 4824 genes, using different exponents p. Different exponents used in the data similarity measure $(1 - r(\mathbf{x}^i, \mathbf{x}^j))^p$ highlight specific correlation structures in the corresponding HiT-MDS-2 embeddings. Results for exponents p = 24, 4,1,0.25 are shown in panels a-d, respectively.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-165-S1.pdf]

Additional file 2

List of 4824 genes with 64 centroids. The table contains gene expressions of 4824 high-quality genes covering 14 developmental stages, ODAF-26DAF, in steps of two days. Expression levels are $\log_2$-transformed quantile-normalized values. Genes are assigned to 64 centroids from neural gas clustering.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-165-S2.xls]

Additional file 3

Table of functional categories. The table contains genes with manually assigned major functional categories corresponding to Fig. 3 of the manuscript.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-165-S3.xls]

Additional file 4

HiT-MDS-2 embedding of gene expressions of 3031 filtered genes from developing barley endosperm at time points 4, 8, 16, 25 days after flowering. Expression levels are taken from Barley 1 Affymetrix chip. Like in Figure 2 of the manuscript, a sandglass shape is obtained for a correlation exponent of p = 8. Since only four time points are considered, the four-dimensional expression vectors are very faithfully represented in the scatter plot. The corresponding regulation patterns of up-, down- and intermediate regulation are displayed in an extra figure [see Additional file 5].

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-165-S4.pdf]

Additional file 5

Gene expression profiles obtained from Barley 1 Affymetrix gene chip connected to highlighted regions of the HiT-MDS-2 gene space plot [see Additional file 4]. High spatial specificity is observed in the exemplary clusters 1–8 covering interesting locations in the gene space. Patterns of up- and down-regulation fall into opposite poles. Cluster number 5 shows intermediate up-regulation with quite diverse characteristic, where Pearson correlation does not yield good discrimination between peaks in the second or third temporal stage. Cluster number 3 contains genes that become active just at the last stage (day 25 after flowering).

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-165-S5.pdf]

Additional file 6

Gene expression profiles corresponding to the six sub-clusters 2a, 2b, 3a, 3b, 4a and 4b referred in Figure 3 of the manuscript. The expression profiles reflect z-score normalized $\log_2$ values. In addition to individual gene expression curves displayed in blue, their mean and standard deviation are depicted by red lines. Highest variability is observed for intermediate regulation events in cluster 2a; yet, the overall quality of coexpression in the six clusters is represented well.

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2105-8-165-S6.pdf]

Acknowledgements

The authors thank Yoshihiro Taguchi for his valuable and kind expertise on nMDS and its application. We acknowledge Heiko Miehe for creating the web site. We also want to thank the reviewers for their helpful comments. The work is supported by BMBF grants FKZ 0313115 (GABI-SEED-II) and FKZ 0313112/0313110 (GABI-MapMan).

References

1. Herrero J, Al-Shahrour F, Diaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo J, Dopazo J: **GEPAS: a web-based resource for microarray gene expression data analysis.** *Nucleic Acids Research* 2003, **31(13):**3461-3467.
2. Pelizzola M, Pavelka N, Foti M, Ricciardi-Castagnoli P: **AMDA: an R package for the automated microarray data analysis.** *BMC Bioinformatics* 2006, **7:**. doi: 10.1186/1471-2105-7-335
3. Yeung KY, Ruzzo WL: **Principal component analysis for clustering gene expression data.** *Bioinformatics* 2001, **17(9):**763-774.
4. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *PNAS* 1998, **95(25):**14863-14868.
5. Hartigan JA, Wong MA: **A K-Means Clustering Algorithm.** *Applied Statistics* 1979, **28:**100-108.
6. Kohonen T: *Self-Organizing Maps* 3rd edition. Berlin: Springer-Verlag; 2001.
7. D'Alimonte D, Lowe D, Nabney I, Mersinias V, Smith CP: **MILVA: An interactive tool for the exploration of multidimensional microarray data.** *Bioinformatics* 2005, **21(22):**4192-4193.
8. Taguchi Y, Oono Y: **Relational patterns of gene expression via non-metric multidimensional scaling analysis.** *Bioinformatics* 2005, **21(6):**730-740.
9. Strickert M, Teichmann S, Sreenivasulu N, Seiffert U: **High-Throughput Multi-Dimensional Scaling (HiT-MDS) for cDNA-Array Expression Data.** In *Artificial Neural Networks: Biological Inspirations. Part I, LNCS 3696* Edited by: Duch et al W. Springer; 2005:625-634.
10. Strickert M, Sreenivasulu N, Seiffert U: **Sanger-driven MDSLocalize – A comparative study for Genomic Data.** In *European Symposium on Artificial Neural Networks (ESANN)* Edited by: Verleysen M. D-facto Publications; 2006:265-270.
11. Buja A, Swayne D, Littman M, Dean N, Hofmann H: **Interactive Data Visualization with Multidimensional Scaling.** *Report, University of Pennsylvania* 2004.
12. Sreenivasulu N, Radchuk V, Strickert M, Miersch O, Weschke W, Wobus U: **Gene expression patterns reveal tissue-specific signaling networks controlling programmed cell death and ABA-regulated maturation in developing barley seeds.** *The Plant Journal* 2006, **47(2):**310-327.
13. **12k EST-Array** [http://pgrc.ipk-gatersleben.de/seeds/12000_EST.php]
14. Zhou X, Kao MCJ, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data.** *PNAS* 2002, **99(20):**12783-12788.
15. Martinetz T, Schulten K: **A "Neural-Gas" Network Learns Topologies.** *Artificial Neural Networks* 1991, **I:**397-402.
16. **Neural Gas Clustering with Correlation** [http://pgrc-16.ipk-gatersleben.de/~stricker/ng/]
17. Manoli T, Gretz N, Grone H, Kenzelmann M, Eils R, Brors B: **Group testing for pathway analysis improves comparability of different microarray datasets.** *Bioinformatics* 2006, **22(20):**2500-2506.
18. Finkelstein R, Gampala S, Rock C: **Abscisic acid signaling in seeds and seedlings.** *Plant Cell* 2002, **14:**S15-S45.
19. Halford N, Paul M: **Carbon metabolite sensing and signaling.** *Biotechnology Journal* 2003, **1(6):**381-398.
20. Diaz I, Martinez M, Isabel-Lamoneda I, Rubio-Somoza I, Carbonero P: **The DOF protein, SAD, interacts with GAMYB in plant nuclei and activates transcription of endosperm-specific genes during barley seed development.** *The Plant Journal* 2005, **42(5):**652-662.
21. Gower J: **Some distance properties of latent root and vector methods used in multivariate analysis.** *Biometrika* 1966, **53:**325-338.
22. **High-Throughput Multidimensional Scaling (V2)** [http://hitmds.webhop.net/]