AN INFORMATION RETRIEVAL PERSPECTIVE ON VISUALIZATION OF GENE EXPRESSION DATA WITH ONTOLOGICAL ANNOTATION

Jaakko Peltonen[1]*, Helena Aidos*[1]*, Nils Gehlenborg*[2]*, Alvis Brazma*[2]*, and Samuel Kaski*[1]

[1]Helsinki University of Technology, Department of Information and Computer Science,
P.O. Box 5400, FI-02015 TKK, Finland
[2]European Bioinformatics Institute, Wellcome Trust Genome Campus,
Cambridge CB10 1SD, United Kingdom
{jaakko.peltonen, helena.aidos, samuel.kaski}@tkk.fi, {nils,brazma}@ebi.ac.uk

ABSTRACT

High-dimensional data are often visualized by dimensionality reduction methods whose goals are not directly related to visualization. We use a recent formalization of *visualization as information retrieval* and apply that formalism to data with *structured annotations*: we analyze gene expression data with annotations from the Gene Ontology (GO). We show that using the GO information in visualization yields better retrieval with respect to known ontological relationships and allows discovery of data properties not explained by the ontology.

Index Terms— dimensionality reduction, gene ontology, information retrieval, structured annotation, visualization

1. INTRODUCTION

Analysis of high-dimensional data often begins with visualization. Recently a novel formalization of visualization as an *information retrieval task* has been given, where the analyst retrieves neighborhood relationships of points from the visualization [1]. Based on this task, the Neighbor Retrieval Visualizer method (NeRV; [1]) optimizes the visualization according to well-defined information retrieval measures.

Often, *annotation* (existing knowledge of the analyst) is available coupled to each data point; often each annotation is *structured* e.g. as a graph. Visualization can show how annotation is related to data features, or show which properties of data are not represented in the annotation. In particular, we study gene expression data, where genes have been classified with terms from the *Gene Ontology* (GO), representing knowledge about the function of genes, the processes they are involved in, and where their activity is localized in the cell.

Many supervised projection methods are unable to exploit structured annotation: supervised methods may, e.g., assume data points have a single label, whereas most genes are labeled with several ontology terms that are related through a graph structure. We extend NeRV to data with structured annotation for each data point, using GO annotation of gene expression profiles as a case study. From the methods point of view, we extend the well-performing NeRV to analyse graph-structured annotation. From the application point of view, we give novel methods for visualizing gene expression. Our methods have a unique *information retrieval interpretation*.

2. NEIGHBOR RETRIEVAL VIZUALIZER (NERV)

How should one define an optimization goal for a visualization? A recent rigorous answer formalized visualization as an *information retrieval task* [1]. Looking at a point on a scatterplot, the analyst *retrieves neighbors* of that point. Retrieval performance can be measured by information retrieval criteria: *precision* describes what proportion of all retrieved points were really neighbors according to original high-dimensional features or according to an expert criterion; *recall* describes what proportion of such neighbors were retrieved.

With these criteria, one can directly optimize dimensionality reduction to maximize performance of information retrieval: the *Neighbor Retrieval Visualizer* (NeRV; [1]) optimizes output coordinates of data items to allow neighbors to be retrieved from the visualization. NeRV optimizes a tradeoff $C_{NeRV} = \lambda \sum_i D_{KL}(p_i, q_i) + (1-\lambda) \sum_i D_{KL}(q_i, p_i)$ where p_i and q_i are probabilistic neighborhoods around the point i in the input and output space, respectively. The two kinds of Kullback-Leibler divergences D_{KL} generalize recall and precision [1]; the parameter λ sets the desired tradeoff between precision and recall. The cost C_{NeRV} is minimized with respect to output-space coordinates $\mathbf{y}_i$ of data points; see [1] for details. The resulting nonlinear embedding is *optimized for the information retrieval task of visualization.*

It is also possible to optimize a *linear projection* $\mathbf{y}_i = \mathbf{W}\mathbf{x}_i$ from original high-dimensional coordinates $\mathbf{x}_i$, by optimizing the cost C_{NeRV} with respect to the projection matrix $\mathbf{W}$. We call this method *linear NeRV* [2]; it is less flexible than NeRV but easy to interpret: each visualization axis is a

JP, HA, and SK belong to HIIT and AIRC. Authors were supported by PASCAL2, an EMBL fellowship, and Academy of Finland decision 123983.

linear combination of original high-dimensional data features.

Linear NeRV also allows a new kind of analysis: when input neighborhoods p_i are derived from an expert distance measure, the linear projection of data features is optimized for retrieval of expert neighborhoods; the projection *reveals the relationship* between the features and the expert knowledge. We call this case 'supervised linear NeRV' (SL-NeRV). When expert knowledge is not used, i.e, input neighborhoods are directly computed from high-dimensional features, we call this case 'unsupervised linear NeRV' (UL-NeRV).

We now extend NeRV to data with structured annotation.

2.1. NeRV for Data with Structured Annotation

Consider data where each data item (e.g. gene) has a feature vector (e.g. expression profile) and a structured annotation (e.g. the graph-formed set of ontology labels of the gene): each item is annotated with a (different-sized) set of class labels. The labels come from a hierarchy (a directed acyclic graph): there is a root class (root node) with several children, and each class (node) under the root may have several parent and child classes. The Gene Ontology is such a graph.

We can create unsupervised visualizations of the data by nonlinear or linear NeRV: compute input neighborhoods based on (here Euclidean) distances between expression profiles, ignoring annotation, and apply NeRV. Such visualizations are good for retrieving neighbors in the feature space.

We will use annotations in the analysis to complement unsupervised visualizations, in two ways: by visualizing *similarities (neighborhoods) of the annotations themselves* without considering the feature vectors, and by visualizing the *relationship* of features and annotations. Let S_i be the set of of class nodes occupied by data item i. We measure annotation distance between data items as the *Jaccard distance*: $J(S_i, S_j) = (|S_i \cup S_j| - |S_i \cap S_j|)/|S_i \cup S_j|$ where $|S|$ is the size of S. This distance compares the number of nodes where i and j differ to the total number of nodes occupied by i and j. Weighting could be applied when computing $|S|$ to emphasize parts of the hierarchy; we used no weighting.

To visualize regularities in annotation, we give the Jaccard distances as input distances to nonlinear NeRV; it visualizes which data items are neighbors in terms of annotation. **To visualize the relationship between feature vectors and the expert knowledge (the annotation)**, we use supervised linear NeRV (a previous supervised NeRV [3] is not feasible with vast numbers of possible annotations.) We give the Jaccard distances (expert knowledge) as input distances, and optimize a linear projection of feature vectors. SL-NeRV optimizes a projection so that neighbors in the visualization correspond to annotation neighbors: e.g. a projection of expression profiles where neighbors have similar GO annotation.

Influence of original data features. To further analyze the SL-NeRV visualization, we estimate how much each original feature d contributes to the projection, as $Influence(d) = C \cdot \text{Var}\{x_d\} \sum_k w_{k,d}^2$ where $w_{k,d}$ is the weight of feature d along the kth projection direction, $\text{Var}\{x_d\}$ is the variance of data along feature d, and C normalizes the influences so they sum to one over features. For SL-NeRV, influence values tell which features (gene expression measurement conditions) have strongest relationships to the annotation (GO information). Lastly, to analyze whether there is structure in the data not explained by the annotation, we will plot the influence values of SL-NeRV together with the influence values of a comparable UL-NeRV: features with low influence for SL-NeRV but high influence for UL-NeRV contain structure not explained by the annotation.

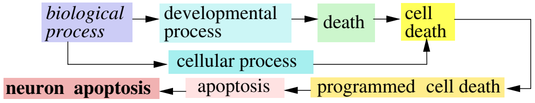

Fig. 1. One of the 19 GO true paths for human gene *AIFM1*. Root note in italics; annotated term in bold. A node can have multiple children and parents. Colors are for clarity only.

Technical note. To avoid overfitting SL-NeRV, we regularize Jaccard distances by Euclidean distances between feature vectors, as $d(\mathbf{x}_i, \mathbf{x}_j) = \beta J(S_i, S_j) + (1-\beta)||\mathbf{x}_i - \mathbf{x}_j||$, where β is chosen by cross-validation; we apply the *area under the precision-recall curve* (an information retrieval statistic) to visualization, as the criterion to find the best β.

3. EXPLORATION OF GENE EXPRESSION DATA

Exploring genes visually, with expression profiles as features, can yield hypotheses for targeted analyses or lab experiments. E.g. heatmaps [4] work poorly when the number of genes and conditions is large, and yield limited insight into global data structure; we use scatterplots which can yield hypotheses e.g. that a set of genes is co-regulated. Trustworthiness and continuity [6] of dimensionality reduction methods have been evaluated for visualizing gene expression data [5]; several methods had problems, but NeRV performed well [1].

The *Gene Ontology* (GO) includes three ontologies of terms describing genes and their products: *molecular functions* (MF), *biological processes* (BP) and *cellular components* (CC). Each ontology is a directed acyclic graph: a class node may have many parents and children. Classes follow a *true path rule*: a gene belonging to some class node must belong to *all* parent classes of the class, all grandparents, and so on up to the root of the ontology; Figure 1 shows an example. A gene with multiple annotation terms belongs to the union of their true paths. Thus, for each gene i we do not simply use its GO terms as the label set S_i, rather **we use the union of its true paths as S_i when computing Jaccard distances**.

We first perform a noise tolerance study, on part of the Novartis SymAtlas [7]: expression of 627 genes in 33 human tissues, and GO annotations from the MF ontology. In exper-

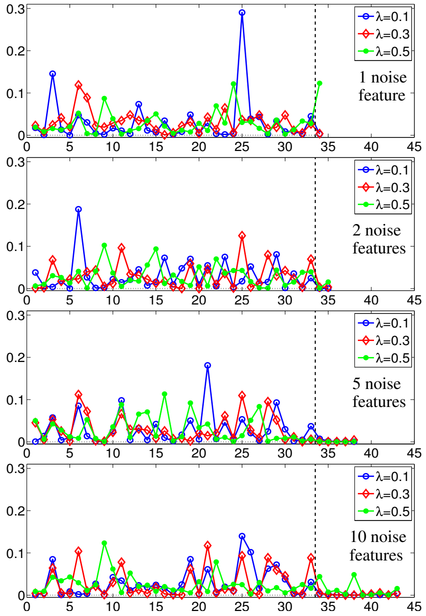

Fig. 2. SL-NeRV ignores noise features (right of the dashed line). Relative influences are shown: visualizations ignore features with low influence. λ sets the tradeoff in C_{NeRV}.

iments (not shown), we found that relationships between expression and annotations are here hard to detect; we show that even then, SL-NeRV ignores features unrelated to GO annotation. We add 1, 2, 5, or 10 'noise features' whose marginal distribution matches a real tissue, but permuting values so expression of noise features is independent of annotation.

Visualizations that reveal expression-annotation relationships should ignore noise features. Figure 2 shows influence plots for SL-NeRV; it successfully sets low weights to noise features except some fluctuation with $\lambda = 0.5$ which needs future study. In the rest of the paper we use $\lambda = 0.1$ which gave stable good performance. (We also tried a scenario where noise features come from a mixture of 10 multivariate Gaussians, so they have a cluster structure unrelated to annotation. Results were similar in all but the most difficult structured noise tests.) In summary, SL-NeRV can rule out expression-ontology relationships not supported by the data.

We now perform a real case study, on expression data of yeast genes measured under 300 conditions, each comparing a mutant yeast strain to a wild-type (normal) strain [8]. We left out genes with non-significant expression, genes without ontology information, and genes and conditions with too many missing values; we then had 501 data items (yeast open reading frames), whose expression is measured across 31 *deletion mutant* conditions (where a genetic sequence has been removed from the yeast). We lastly left out a subset of genes as test data. We studied all three GO ontologies: MF, BP, and CC. For brevity, we show results with CC annotations only.

Step 1: we apply nonlinear NeRV to Jaccard distances between CC annotations of genes, and hence visualize *regularities in the CC ontology annotation*; Figure 3 (top left) shows some visible regularity. **Step 2:** we apply nonlinear NeRV to Euclidean distances of gene expression profiles; Figure 3 (top middle) shows clear structure. **Step 3:** we want to see if expression-annotation relationships are visible in the NeRV visualization. This is nontrivial: we cannot simply color each annotation with some arbitrary color; with too many different annotations the result would be a jumble of colors. Instead, we *take the NeRV visualization of annotations as a colorspace map for the genes*, and apply the colors to Figure 3 (top middle). We see some potential expression-annotation relationships, but not very clear ones. **Step 4:** we use SL-NeRV to optimize a projection to reveal the expression-annotation relationships. Figure 3 (top right) shows the result: some clear groups of color are now visible. **Step 5:** we try to find *structure that is not explained by the CC annotation*. First, we verify that SL-NeRV really reveals expression-annotation relationships better than an unsupervised comparison (UL-NeRV): we measure their ability to retrieve ontology neighbors, by the area under the precision-recall curve (AUC). SL-NeRV attains AUC 0.0687 and UL-NeRV 0.0628, hence SL-NeRV performs better; note that AUC values are low for this difficult retrieval task. Next, we identify which data features contain structure not explained by the CC annotation: we compute influence plots for SL-NeRV and UL-NeRV, and find the dimensions having large influence on UL-NeRV and low influence on SL-NeRV. Figure 3 (bottom left) shows the influence plots; boxes mark dimensions with unexplained structure. **Step 6:** we lastly use nonlinear NeRV to visualize the dimensions that had unexplained structure. Figure 3 (bottom right) shows the results: we see that interesting structure remains. The result shows clear structure (some of which may be present in the previous unsupervised visualization, top middle subfigure). This structure is not explained by the CC ontology and is worth investigating in further work.

4. CONCLUSIONS

We presented methods for nonlinear and linear visualization of data with structured annotation, and applied them to gene expression data with gene ontology (GO) annotations. The Neighbor Retrieval Visualizer (NeRV) has proven a powerful nonlinear visualizer here and in previous quantitative comparisons [1, 3]. We recommend it for high-dimensional data like gene expression, for visualizing similarities between features like expression profiles or between (structured) annotations like GO annotations. To find how expression profiles are related to GO annotations, we used a linear version of NeRV; it allowed easy interpretation of influence of the features, and discovery of features having structure not explained by the annotations. Other nonlinear (e.g. Isomap) and linear methods could be used; we use NeRV which performed well in comparisons [1, 3] and has an information retrieval interpretation.

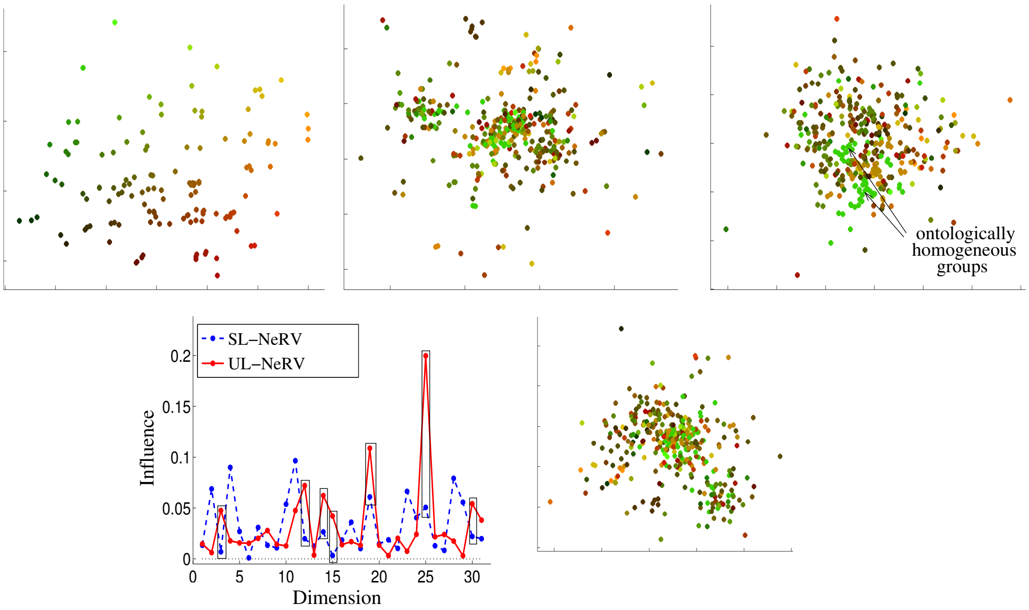

Fig. 3. Visualizing yeast genes with CC annotation. **Top left:** nonlinear NeRV shows some regularity in ontology neighborhoods. **Top middle:** nonlinear NeRV embedding of expression profiles shows clusters. Coloring from the top left plot; homogeneity of coloring would indicate expression-annotation relationships. **Top right:** SL-NeRV projection of expression profiles, optimized to reveal expression-annotation relationships; some ontologically homogeneous groups are shown. **Bottom left:** finding features with structure unexplained by the CC ontology, i.e., features with low influence for SL-NeRV and high for UL-NeRV; such features are marked with boxes. **Bottom right:** UL-NeRV visualization of expression profiles along dimensions with unexplained structure (boxes in bottom left subfigure); visible structure is not well explained by the CC annotation.

NeRV revealed structure in the data: gene groups that are similar in terms of expression profiles or GO annotations. By encoding ontological similarities as colors we found that some clusters of genes with similar expression are also similar in terms of GO annotations. We also found structure not explained by CC annotation. Our methods are from ongoing work for visualization with structured annotation, where many standard supervised methods are not suitable. Our methods in this first paper using annotations already obtain interesting results that will be compared further later.

5. REFERENCES

[1] J. Venna and S. Kaski, "Nonlinear dimensionality reduction as information retrieval," in *Proc. AISTATS*07*, 2007.

[2] J. Peltonen, "Visualization by linear projections as information retrieval," in *Advances in Self-Organizing Maps*, Berlin Heidelberg, 2009, pp. 237–245, Springer.

[3] J. Peltonen, H. Aidos, and S. Kaski, "Supervised nonlinear dimensionality reduction by neighbor retrieval," in *Proc. ICASSP 2009*, 2009, pp. 1809–1812, IEEE.

[4] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *PNAS*, vol. 95, pp. 14863–14868, 1998.

[5] J. Venna and S. Kaski, "Comparison of visualization methods for an atlas of gene expression data sets," *Information Visualization*, vol. 6, pp. 139–54, 2007.

[6] S. Kaski, J. Nikkilä, M. Oja, J. Venna, P. Törönen, and E. Castrén, "Trustworthiness and metrics in visualizing similarity of gene expression," *BMC Bioinformatics*, vol. 4, pp. 48, 2003.

[7] A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch, "A gene atlas of the mouse and human protein-encoding transcriptomes," *PNAS*, vol. 101, no. 16, pp. 6062–6067, 2004.

[8] T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffrey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend, "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, pp. 109–126, 2000.