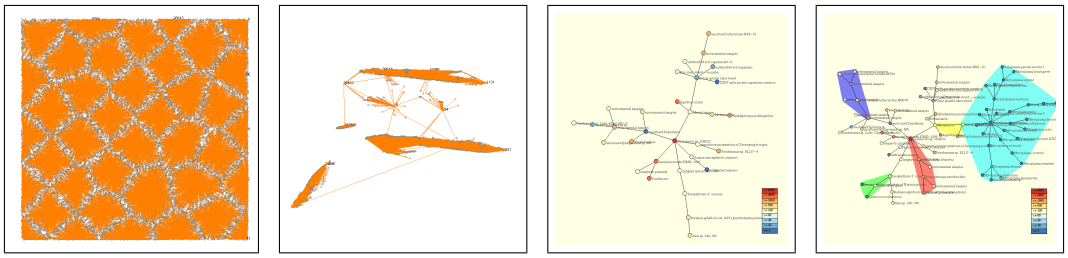
Creating a Knowledge Structure in Real-Time

Dong Hyun Jeong, Remco Chang, and William Ribarsky
Charlotte Visualization Center, UNC Charlotte *



(a) (b) (c) (d)

Figure 1: *Genomic data of various organisms (about 40,000) are plotted randomly (a) and based on using the first two principal components (b). The linkages in both (a) and (b) are created by applying a hierarchical clustering method. (c) Representative organisms chosen from the cluster are displayed and (d) can be interactively expanded based on the user's focus of interest.*

1 Introduction

Genomic data of various organisms have long been studied to understand their complex nature. People found that visualization techniques are useful to unveil and understand hidden knowledge. However, the amount of genomic data along with knowledge discovered about them have been too large and complex for such analysis to be performed in real-time. Hence, we have defined two different terms of knowledge (tacit and explicit) and proposed in our previous work that representing knowledge artifacts (explicit knowledge) is beneficial in understanding genomic data [Jeong et al. 2008].

In this paper, we present a method that creates explicit knowledge from genomic data in real-time. The explicit knowledge we define here is semantic information [Gilson et al. 2008], which should depend on the input data and the user's interest. Although it has been known that explicit knowledge can possibly be extracted by applying machine learning algorithms [Gómez-Pérez and Manzano-Macho 2004], extracting explicit knowledge from genomic data is highly computational. Hence, we design an approach based on applying both Principal Component Analysis (PCA) and hierarchical clustering method sequentially to extract representative organisms. To place the elected organisms in visual space, a force-based placement is used. Based on our approach, we found that the user is able to interactively analyze genomic data in real-time.

2 Our Approach

In biological data, the taxonomy structure (simply called taxonomy) shows biological relations among organisms. However, the taxonomy is limited in its ability to show detailed semantic relations. To help overcome this problem, we apply a statistical approach to create a structure that represents the relations among data. Our statistical approach is to build a graph topology (nodes and links), which represents the relations among the genomic data. The data we used include both annotations created by researchers and the relative information to the annotations such as paper publications (published year and number of citations).

Based on more than 40,000 genomic entities, we applied three methods: PCA, a hierarchical clustering, and a force-directed drawing algorithm. Given an $n \times m$ data matrix, PCA uses the first eigenvectors of the $m \times m$ covariance matrix as the axes of the lower k-dimensional space. The first two principal components are selected to represent each organism. An approximation approach, on-line SVD [Brand 2003], is used to compute principal components in real-time. The data matrix can be computed in $O(mnr)$ time for $r \leq \sqrt{\min(m,n)}$. Within the principal components, a performance guaranteed hierarchical clustering technique proposed by Dasgupta and Long [Dasgupta and Long 2005] is applied to build a hierarchical tree structure. It uses k-center algorithm and levels of granularity to approximately create a hierarchical clustering. It guarantees $\Omega(logk)$ lower bound on the approximation with small values of k. When the k is not defined, the complete linkage is computed in $O(n^2)$. However, representing the full graph topology in a limited screen space is not feasible. Also it is not possible to support interactive analysis of the graph topology. Therefore, representative organisms are selected based on a distance-based approach. The force-directed drawing algorithm, GRIP [Gajer and Kobourov 2000], is used to efficiently organize each node (organism) to increase the user's awareness and understanding. Unnecessary or infeasible information is activated and deactivated depending on the user's interest. This approach is useful to highlight and represent important knowledge instead of representing all information without knowledge relations. Using this approach, we can create a knowledge structure for 40,000 genomic entities in less than 3 seconds.

*e-mail: {dhjeong, rchang, ribarsky}@uncc.edu

References

BRAND, M. 2003. Fast online svd revisions for lightweight recommender systems. In *SDM 2003*, 83–91.

DASGUPTA, S., AND LONG, P. M. 2005. Performance guarantees for hierarchical clustering. *Journal of Computer and System Sciences 70*, 4, 555–569.

GAJER, P., AND KOBOUROV, S. G. 2000. Grip: Graph drawing with intelligent placement. In *Graph Drawing 2000*, Springer-Verlag, 222–228.

GILSON, O., SILVA, N., GRANT, P. W., AND CHEN, M. 2008. From web data to visualization via ontology mapping. *Comput. Graph. Forum 27*, 3, 959–966.

GÓMEZ-PÉREZ, A., AND MANZANO-MACHO, D. 2004. An overview of methods and tools for ontology learning from texts. *Knowl. Eng. Rev. 19*, 3, 187–212.

JEONG, D., CHANG, R., AND RIBARSKY, W., 2008. An alternative definition and model for knowledge visualization. IEEE Visualization 2008 Workshop on Knowledge Assisted Visualization, Oct.