Open Source Clustering Software

Michiel J.L. de Hoon mdehoon@ims.u-tokyo.ac.jp
Seiya Imoto imoto@ims.u-tokyo.ac.jp
Satoru Miyano miyano@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo,
4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: open source software, clustering, gene expression data

1 Introduction

Clustering techniques are widely used in gene expression data analysis. By grouping genes together based on the similarity in their gene expression profile, we may find functionally related genes, and potentially the function of presently unknown genes.

Several programs are currently available to analyze gene expression data. The Java code GeneCluster [1, 2, 7] implements two-dimensional Self-Organizing Maps (SOMs) [3, 4]. Unfortunately its usage and extensibility is restricted by the license, the source code is not available, and Java may be too slow for large calculations. The widely used Cluster/TreeView code [5], written in C++ for the Microsoft Windows platform, focuses on hierarchical clustering methods, while one-dimensional SOMs, principal component analysis, and k-means clustering are also implemented. We note though that the implementation of the k-means algorithm does not provide automatic repetitions of the clustering calculation starting from different random initial clusterings, and may therefore not be able to find the optimal k-means clustering result. The source code for Cluster/Treeview, minus the routines that are covered by the Numerical Recipes license [6], is available for the academic and non-profit community.

Clustering routines suitable for usage with scripting languages such as Python [8] or Perl [9] are often not available. Scripting languages are heavily used in other fields of bioinformatics, as they provide a flexible, platform-independent, and easily extendible basis for data analysis.

2 The C Clustering Library

We have implemented the main hierarchical (pairwise single-, average-, maximum-, and centroid-linkage) clustering techniques, SOMs on a 2D rectangular grid, and the k-means clustering algorithm as a library of C routines. The similarity between gene expression data can be measured by the Pearson and uncentered correlation, the Euclidean distance, the harmonically summed Euclidean distance, Spearman's rank correlation, and Kendall's τ. This makes all the clustering routines commonly used in gene expression data analysis available in a single open-source library. The C Clustering Library is written in ANSI C and is platform-independent. It was released under the GNU Lesser General Public License, and can be downloaded from our website [10].

2.1 Cluster 3.0 for Windows, Macintosh, and Linux

We have replaced the core numerical routines in Cluster/TreeView, which make use of routines in Numerical Recipes [6], by the open-source C Clustering Library, thereby enhancing the capabilities of Cluster. In particular, we implemented automatic repetitions of the k-means clustering algorithm. The complete source code to Cluster is available at our website [10]. Separating the GUI-code from

the core numerical routines enabled us to easily port the code to the Mac OS X and Linux platforms. Cluster 3.0 for Windows, Mac OS X, and Linux can be downloaded from our website [10].

We further note that recently a platform-independent Java version of TreeView became available [11]. Java TreeView can visualize hierarchical and k-means clustering results.

2.2 Using the C Clustering Library with Python

Scripting languages can be used for data analysis by issuing a series of commands to an interpreter, either manually one by one or as a script file. Examples of scripting languages commonly used in bioinformatics are Python [8], Perl [9], and Ruby [12], which are all in the public domain. Scripting languages typically contain built-in features for file input and output, data filtering, preprocessing and postprocessing, database access, and visualization. The entire data analysis can be captured in one script, thereby simplifying replication. An example of such a script for gene expression data analysis is available from our website [10]. In addition, scripting languages are very suitable for batch-mode operation, as needed for bootstrap calculations.

We used Pyfort [13] to generate a Python interface to the C clustering library. This allows us to combine the flexibility of a scripting language with the speed of C. Interfaces to other scripting languages may be generated using SWIG [14]. Pycluster is available at our website [10].

References

[1] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., and Golub, T.R., Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.

[2] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., and Lander, E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286:531-537, 1999.

[3] Kohonen, T., The Self-Organizing Map, *Proceedings of the IEEE*, 78:1464–1480, 1990.

[4] Kohonen, T., *Self-Organizing Maps*, Springer Verlag, 2001.

[5] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[6] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in C: The Art of Scientific Computing*, Cambridge University Press, 1992.

[7] `http://www-genome.wi.mit.edu/cancer/software/genecluster2/gc2.html`

[8] `http://www.python.org`

[9] `http://www.perl.org`

[10] `http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster`

[11] `http://genome-www.stanford.edu/~alok/TreeView`

[12] `http://www.ruby-lang.org`

[13] `http://pyfortran.sourceforge.net`

[14] `http://www.swig.org`