



## 2HAPI: a microarray data analysis system

J. Lynn Fink<sup>1</sup>, Scott Drewes<sup>1</sup>, Hiren Patel<sup>2</sup>, John B. Welsh<sup>3</sup>,  
Daniel R. Masys<sup>4</sup>, Jacques Corbeil<sup>4</sup>, and  
Michael R. Gribskov<sup>1, 2, \*</sup>

<sup>1</sup>San Diego Supercomputer Center, <sup>2</sup>Department of Biology, University of California, San Diego, 9500 Gilman Drive, San Diego, CA 92093-0537, <sup>3</sup>Johnson & Johnson Pharmaceutical Research & Development, L.L.C., 3210 Merryfield Row, San Diego, CA 92121, <sup>4</sup>School of Medicine, University of California, San Diego, 9500 Gilman Drive, San Diego, CA 92093-0602 and <sup>5</sup>Veterans Medical Research Foundation, 3350 La Jolla Village Drive, San Diego, CA 92161, USA

Received on February 10, 2003; revised and accepted on April 2, 2003

### ABSTRACT

**Summary:** 2HAPI (version 2 of High density Array Pattern Interpreter) is a web-based, publicly-available analytical tool designed to aid researchers in microarray data analysis. 2HAPI includes tools for searching, manipulating, visualizing, and clustering the large sets of data generated by microarray experiments. Other features include association of genes with NCBI information and linkage to external data resources. Unique to 2HAPI is the ability to retrieve upstream sequences of co-regulated genes for promoter analysis using MEME (Multiple Expectation-maximization for Motif Elicitation).

**Availability:** 2HAPI is freely available at <http://array.sdsc.edu>. Users can try 2HAPI anonymously with pre-loaded data or they can register as a 2HAPI user and upload their data.

**Contact:** [gribskov@sdsc.edu](mailto:gribskov@sdsc.edu)

### INTRODUCTION

cDNA and oligonucleotide arrays enable the simultaneous measurement of the expression levels of genes on a genomic scale. Because the data sets generated by microarrays are often large and have high dimensionality, computational approaches to data analysis are a necessity. Microarray data analysis includes several approaches: verifying known gene expression patterns; examining the promoter regions of co-regulated genes; clustering expression patterns of co-regulated genes; and sorting information about these genes. Generally, these approaches are handled by distinct pieces of software. For example, some software packages are available that perform clustering (Cluster; Eisen *et al.*, 1998, GeneCluster; Tamayo *et al.*, 1999, etc.) while others exist in order to annotate genes represented on microarrays to place them in a biological

context, such as DRAGON (Bouton and Pevsner, 2000). 2HAPI facilitates exploration and mining of microarray data by combining these aspects into an integrated analytical environment.

### SYSTEM AND IMPLEMENTATION

2HAPI consists of a relational database and an HTML/Perl-CGI interface with integrated clustering algorithms and links to external resources. Data sets are uploaded as tab-delimited text files in which the rows correspond to genes and columns correspond to experimental groups (i.e. time points or experimental conditions). Uploaded data sets are stored as tables in the database and can be accessed through the HTML interface.

The 2HAPI database also contains tables with reference information extracted from PubMed and Entrez for each probe containing a GenBank accession number. Information for the genes represented on several Affymetrix chips is already available in 2HAPI. This information includes the Entrez description of the gene or transcript identified by the accession number, PubMed unique identifiers corresponding to the literature describing the initial characterization of the gene or transcript, and MeSH terms assigned to the PubMed articles.

2HAPI is currently hosted on a Sun Enterprise™ 420R server running the Apache web server (with mod\_perl), the relational database MySQL, and Perl.

### Data processing

Pre-processing data prior to clustering can often improve the performance of the algorithm and the interpretability of the resulting clusters (Šašik *et al.*, 2002). 2HAPI can filter out genes whose overall expression level does not change by a user-specified threshold. Normalizing the data after filtering is also recommended prior to presenting expression data to a clustering algorithm.

\*To whom correspondence should be addressed.

2HAPI normalizes expression patterns such that each expression pattern, or vector, has a mean of 0 and a variance of 1 across all data points.

### Searching and grouping genes

Several tools are available with which to search expression data. Data can be searched by probe identifier, Entrez description (if available), MeSH term (if available), equivalent expression level (i.e. average difference or normalized fluorescence intensity), ratio between equivalent expression levels of two different data points, and absolute call (if data were generated by Affymetrix chips). The first three tools allow the user to display specific genes of known interest or groups of genes that share a gene name, function, cellular/chromosomal location, disease, etc. In particular, we have found that the MeSH term association is a useful data mining approach (Masys *et al.*, 2001). The last three tools can be used to select genes based on expression levels (e.g. genes that are highly expressed at a specific data point) or genes that display differential expression between data points.

### Clustering genes

Several clustering algorithms are integrated into 2HAPI for the purpose of grouping genes based on similarities in their expression patterns. Algorithms previously used in microarray data analysis, *K*-means (Tavazoie *et al.*, 1999) and self-organizing maps (SOMs; Tamayo *et al.*, 1999; Törönen *et al.*, 1999), are included as well as two algorithms that are new to microarray analysis, *K*-harmonic means (*KHM*; Zhang *et al.*, 1999), and growing neural gas (*GNG*; Fritzke, 1994, 1995). *KHM* and *GNG* are currently available only in 2HAPI. Once genes have been clustered by an algorithm of the researcher's choice, the resulting clusters can be viewed as plots of the centroid and range of variance and information about the individual members is easily retrieved.

### Promoter analysis using MEME

The upstream regions of clustered or otherwise-grouped genes can be presented to MEME (<http://meme.sdsc.edu>; Bailey and Elkan, 1994; Patel, 2001), a sequence motif discovery program. The user may specify a range of bases in the upstream regions of the genes represented by chip probes to search. These regions will be retrieved from an internal database of sequences from the UCSC Human Genome Working Draft (International Human Genome Sequencing Consortium, 2001; Kent and Haussler, 2001) and automatically formatted and submitted to MEME for analysis. Currently, promoter analysis is only available for human sequences.

## CONCLUSION

2HAPI is a publicly available microarray data analysis system that integrates several aspects of data analysis and visualization. This system is a useful alternative to using several stand-alone software packages or costly commercial products. Features unique to 2HAPI include the *KHM* and *GNG* clustering algorithms, automatic association with GenBank information and integration with MEME.

## ACKNOWLEDGEMENTS

This work was supported by the National Institute of Allergy and Infectious Diseases AI46237 and AI47703, the Center for AIDS Research Genomics Core Laboratory (AI36214), the Universitywide AIDS Research Program IS99-SD213 and the San Diego Veterans Medical Research Foundation (J.C.). Computational facilities and support for M.G., J.L.F., S.D. and H.P. provided by the National Biomedical Computation Resource, an NIH Research Resource (P41 RR08605-08).

## REFERENCES

- Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, pp. 28–36.
- Bouton,C.M. and Pevsner,J. (2000) DRAGON: database referencing of array genes online. *Bioinformatics*, **16**, 1038–1039.
- Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.
- Fritzke,B. (1994) Fast learning with incremental RBF networks. *Neural Process. Lett.*, **1**, 2–5.
- Fritzke,B. (1995) A growing neural gas network learns topologies. In *Advances in Neural Information Processing Systems*, Vol. 7, MIT Press, Cambridge.
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Kent,W.J. and Haussler,D. (2001) Assembly of the working draft of the human genome with GigAssembler. *Genome Res.*, **11**, 1541–1548.
- Masys,D.R., Welsh,J.B., Fink,J.L., Gribskov,M., Klacansky,I. and Corbeil,J. (2001) Use of keyword hierarchies to interpret gene expression patterns. *Bioinformatics*, **17**, 319–326.
- Patel,H.J. (2001) Identification of protein binding sites in genes sharing similar expression profiles using MEME and MAST, Thesis (M.S.), University of California, San Diego.
- Šašik,R., Iranfar,N., Hwa,T. and Loomis,W.F. (2002) Extracting transcriptional events from temporal gene expression patterns during *Dictyostelium* development. *Bioinformatics*, **18**, 61–66.
- Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareenwan,S., Dmitrovsky,E., Lander,E. and Golub,T. (1999) Interpreting

- patterns of gene expression with self-organizing maps: methods and applications to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA*, **96**, 2907–2912.
- Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J. and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nature Genet.*, **22**, 281–285.
- Törönen,P., Kolehmainen,M., Wong,G. and Castrén,E. (1999) Analysis of gene expression data using self-organizing maps. *FEBS Lett.*, **451**, 142–146.
- Zhang,B., Hsu,M. and Dayal,U. (1999) *K*-harmonic means—a data clustering algorithm. *Technical Report HPL-1999-124*. Hewlett-Packard Laboratories, Palo Alto.