Pacific Symposium on Biocomputing 3:42-53 (1998)

CLUSTER ANALYSIS AND DATA VISUALIZATION OF LARGE-SCALE GENE EXPRESSION DATA

GEORGE S. MICHAELS, DANIEL B. CARR
Institute for Computational Sciences and Informatics, George Mason University, Fairfax, VA 22030 (http://www.science.gmu.edu/~michaels/Bioinformatics/; gmichael@osf1.gmu.edu; dcarr@gmu.edu)

MANOR ASKENAZI
Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501-8943

STEFANIE FUHRMAN, XILING WEN, ROLAND SOMOGYI
Molecular Physiology of CNS Development, LNP/NINDS/NIH, 36/2C02, Bethesda, MD 20892 (http://rsb.info.nih.gov/mol-physiol/homepage.html; sfuhrman@codon.nih.gov; lingling@codon.nih.gov; rolands@helix.nih.gov)

The discovery of any new gene requires an analysis of the expression context for that gene. Now that the cDNA and genomic sequencing projects are progressing at such a rapid rate, high throughput gene expression screening approaches are beginning to appear to take advantage of that data. We present a strategy for the analysis for large-scale quantitative gene expression measurement data from time course experiments. Our approach takes advantage of cluster analysis and graphical visualization methods to reveal correlated patterns of gene expression from time series data. The coherence of these patterns suggests an order that conforms to a notion of shared pathways and control processes that can be experimentally verified.

Genetic networks and large scale gene expression mapping

With the advent of the Human Genome Project and other genome sequencing efforts, we are now faced with the challenge of developing a functional genomics and new methods of data analysis. Molecular biology has traditionally focused on the study of individual genes considered in isolation as a method for determining gene function. In order to determine the principles underlying complex biological processes, such as development, however, we must also examine the expression patterns of large numbers of genes in parallel, taking into consideration temporal, as well as anatomical, patterns. Large-scale temporal gene expression patterns may provide a means for inferring causal links between genes expressed

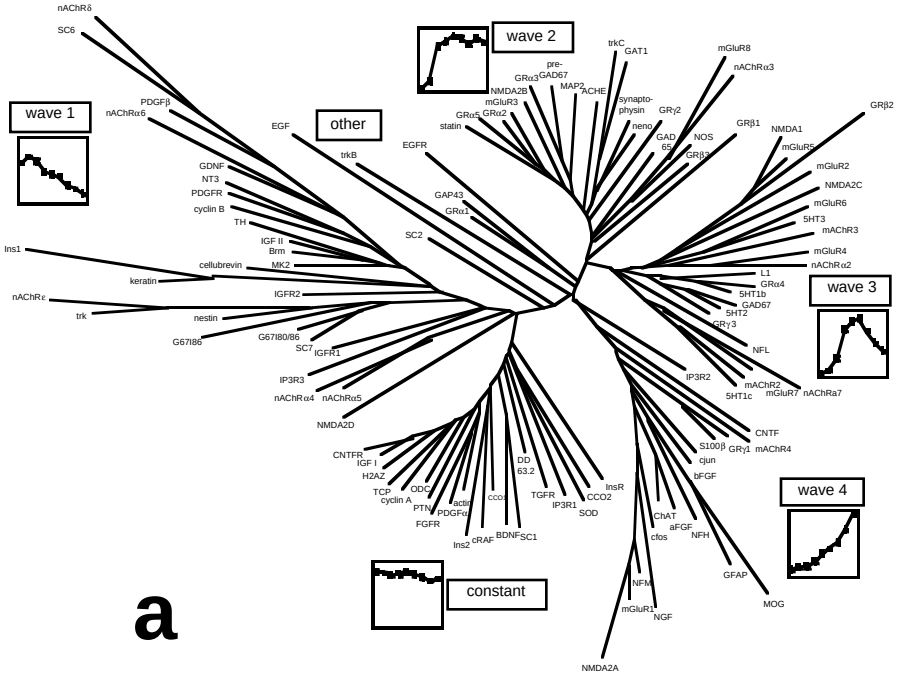
over the course of phenotypic change. Statistical and information theoretic measures may be useful for the analysis of such data.

Toward that end, Wen et al. (1) have generated a limited gene expression matrix for rat cervical spinal cord. Using a reverse-transcription polymerase chain reaction (RT-PCR) protocol (2), they have assayed the expression of 112 genes (mRNA levels; normalized to maximal expression level) over nine developmental time points (E11, E13, E15, E18, E21, P0, P7, P14, and P90 or adult; E=embryonic, P=postnatal). Included in the list are genes considered important in CNS (central nervous system) development covering major gene families:

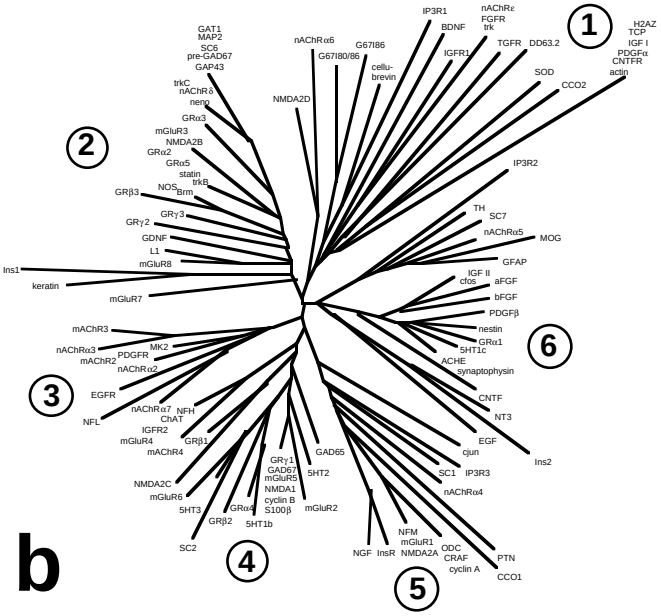
- neurotransmitter metabolizing and synthesizing enzymes, relating to
  - GABA
  - acetylcholine
  - catecholamines
  - nitric oxide
- ionotropic neurotransmitter receptors
  - $GABA_A$ receptors
  - NMDA receptors
  - nicotinic acetylcholine receptors
  - 5HT [serotonin] receptors
- metabotropic neurotransmitter receptors
  - metabotropic glutamate receptors
  - muscarinic acetylcholine receptors
  - 5HT [serotonin] receptors
- neurotrophins and their receptors
- heparin-binding growth factors and their receptors
- insulin and insulin-like growth factor (IGF) family and their receptors
- intracellular calcium channels / receptors (IP_3 receptors)
- cell cycle proteins
- transcriptional regulatory factors
- novel genes or expressed sequence tags (ESTs)
- housekeeping genes

We included genes for established “neuroglial marker” proteins as well, in order to correlate expression time series to phenotypic differentiation.

We have conceptualized these genes as participants in a genetic network. Theoretical studies of genetic networks, such as Boolean network models (3, 4), are helping us to grasp the principles of complex dynamics and to develop and test analytical tools for inference of genetic networks from gene expression data (5, 6). Here, we present two clustering methods for analyzing this collection of 9 x 112 data points. These methods represent a first step toward determining an interaction or “wiring” diagram for the genetic network of the developing mammalian CNS.


wave 1
wave 2
wave 3
wave 4
other
constant
nAChRδ
SC6
PDGFβ
nAChRα6
EGF
EGFR
trkB
GDNF
NT3
PDGFR
cyclin B
TH
IGF II
Brm
MK2
cellubrevin
keratin
IGFR2
nAChRε
trk
nestin
G67I86
G67I80/B6
SC7
IGFR1
IP3R3
nAChRα4
nAChRα5
NMDA2D
Ins1
GAP43
GRα1
SC2
trkC
GAT1
mGluR8
nAChRα3
pre-GAD67
GRα3
MAP2
AChE
NMDA2B
mGluR3
GRα2
GRα5
statin
synapto-physin
GRγ2
neno
GAD 65
NOS
GRβ3
GRβ1
NMDA1
mGluR5
GRβ2
mGluR2
NMDA2C
mGluR6
5HT3
mAChR3
mGluR4
nAChRα2
L1
GRα4
5HT1b
GAD67
5HT2
GRγ3
NFL
IP3R2
5HT1c
mAChR2
mGluR7
nAChRa7
CNTF
S100β
GRγ1
mAChR4
cjun
bFGF
ChAT
aFGF
cfos
NFH
GFAP
MOG
NFM
mGluR1
NGF
NMDA2A
CNTFR
IGF I
H2AZ
TCP
cyclin A
ODC
PTN
PDGFα
FGFR
actin
CCO1
Ins2
cRAF
BDNF
SC1
DD 63.2
TGFR
IP3R1
CCO2
SOD
insR

a


1
2
3
4
5
6
IP3R1
nAChRε
FGFR
trk
BDNF
IGFR1
TGFR
DD63.2
SOD
CCO2
H2AZ
TCP
IGF I
PDGFα
CNTFR
actin
nAChRα6
G67I86
G67I80/86
cellu-brevin
NMDA2D
GAT1
MAP2
SC6
pre-GAD67
GAP43
trkC
nAChRδ
neno
GRα3
mGluR3
NMDA2B
GRα2
GRα5
statin
trkB
NOS
Brm
GRβ3
GRγ3
GRγ2
GDNF
L1
mGluR8
Ins1
keratin
mGluR7
IP3R2
TH
SC7
nAChRα5
MOG
GFAP
IGF II
cfos
aFGF
bFGF
PDGFβ
nestin
GRα1
5HT1c
ACHE
synaptophysin
CNTF
NT3
EGF
cjun
Ins2
SC1
IP3R3
nAChRα4
PTN
CCO1
mAChR3
nAChRα3
MK2
PDGFR
mAChR2
nAChRα2
EGFR
NFL
nAChRα7
NFH
ChAT
IGFR2
mGluR4
GRβ1
mAChR4
GAD65
GRγ1
GAD67
mGluR5
5HT2
NMDA2C
mGluR6
NMDA1
cyclin B
S100β
5HT3
GRα4
mGluR2
GRβ2
5HT1b
SC2
NGF
InsR
NFM
mGluR1
NMDA2A
ODC
CRAF
cyclin A

b

Principles of clustering based on euclidean and mutual information distance matrices

Gene expression data from RT/PCR experiments (1, 2) serve as the basis for the analysis conducted below. We clustered genes according to their patterns of expression over the nine developmental time points using the FITCH software (7) and the euclidean distance measure. The FITCH software, while developed for reconstruction of phylogenetic trees, is essentially a general clustering program. We have compared the performance of FITCH to other standard clustering algorithms using this data set. We found that its performance was best with respect to conserving known similarities in gene expression patterns, although its scalability is limited but sufficient in this particular case. The euclidean distance measure quantitates the differences between expression histories of pairs of genes. The euclidean distance between a pair of points is simply the square root of the sum of the squared distances in each dimension. We determined the euclidean distance of the combined, 17 dimensional vector of 9 expression values (ranging between 0 and 1) and 8 slopes (ranging between -1 and +1; slopes were calculated based on a reduced time interval of 1, not taking into account the variable measurement intervals). We chose to include slopes to take into account offset but parallel trajectories. The pair-wise distances were entered into a 112 gene x 112 gene distance matrix, which served as the input for the FITCH clustering program. We used the default parameters for FITCH, except for setting the P parameter to zero, required for implementing the least squares method appropriate for data with expected linearly proportional error.

As a second, further-reaching distance measure, we used the concept of mutual information (8, 9, 10), also referred to as rate of transmission, which quantifies the reduction in the uncertainty of one random variable given knowledge about another random variable. In our case, we used normalized mutual information as a measure of the extent to which knowledge about the expression levels of one gene reduces the uncertainty about the expression levels of another given gene. In order to calculate the information entropy of each gene expression sequence, the

Figure 1 (see previous page)**.** Clustering trees. All genes shown were clustered as described in the text over the nine time points, embryonic days 11, 13, 15, 18, 21, and postnatal days 0, 7, 14, and 90. For both trees, common branch points indicate close correlations in expression patterns among the genes. (**A**) Euclidean distance tree. Genes cluster into five basic expression patterns (waves 1 through 4 and constant). For each cluster, the average expression pattern for all genes in the cluster is shown as an inset. (**B**) Mutual information tree (based on normalized mutual information distance measure). Since calculations were based on pair-wise comparisons between expression levels at *multiple* time points, both trees are somewhat distorted, being two-dimensional projections of multi-dimensional maps; therefore, line lengths in this figure do not accurately reflect degrees of correlation between distant expression patterns.

continuous data was transformed by binning the expression levels into three discrete, equidistant levels for statistical purposes. The 9 measurement time points can allow for no more than 9 co-expression levels; clearly, more data will be required to generate robust binning. We present this approach mainly to demonstrate the essence of our methodology. The information entropy, H, of each gene expression series, I, was calculated from the probabilities, P(i), of the occurrence of one of the three expression levels over the nine measurement points:

$$H(I) = -\sum[P(i)*\log P(i)].$$

The joint entropy for each gene expression pair, H(I,J), was calculated analogously:

$$H(I,J) = -\sum[P(i,j)*\log P(i,j)].$$

The mutual information of each pair, M(I,J), was determined according to the following definition:

$$M(I,J) = H(I)+H(J)-H(I,J).$$

Since the degree of mutual information is dependent on how much entropy is carried by each gene expression sequence, a gene expression sequence pair exhibiting low information entropies will also have low M, even if they are completely correlated. Therefore, we normalized M to the maximal entropy of each of the contributing sequences (numerical range: 0-1), giving a high value for highly correlated sequences, independent of the individual entropies:

$$M_{norm} = M(i,j) / \max [H(i), H(j)].$$

We used 1- M_{norm} as a distance measure for the distance matrix used for clustering, since maximal coherence must correspond to minimal distance. Normalized mutual information may also be expressed in terms of conditional entropies:

$$M_{norm} = \max [H(I|J)/H(I), H(J|I)/H(J)].$$

Unlike euclidean distance, this method also recognizes negatively and non-linearly correlated data sets as proximal.

Analysis of developmental gene expression trajectories

One may conjecture that genes which a) share common control inputs, b) operate together (e.g. proteins that are part of a metabolic or signaling pathway or signaling network), or c) are members of the same gene sequence family, might be regulated in a largely parallel fashion. A goal of this work is to determine whether genes within these categories exhibit overlapping mRNA expression trajectories or control patterns. Examination of trajectories in Boolean network models shows a good correspondence between clusters of expression time series and overlapping regulatory inputs (5).

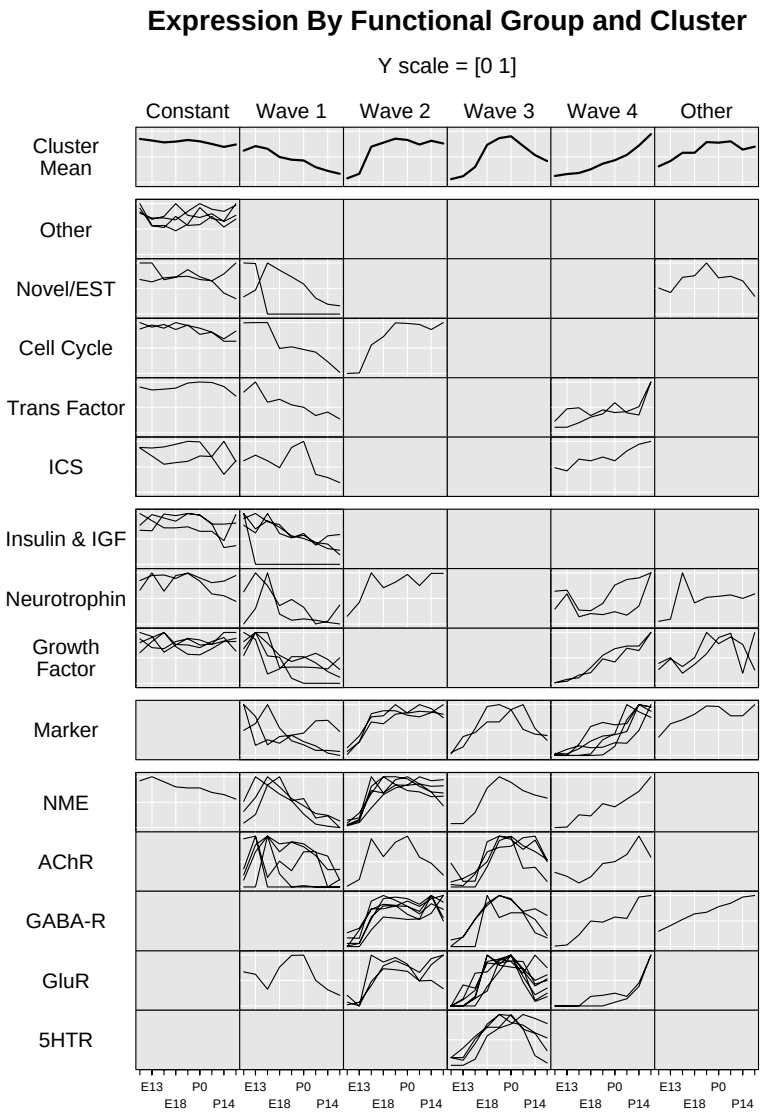


Figure 2. Groupings of developmental gene expression time series. Cluster mean patterns are shown in top panel. Individual data series are shown in remaining panels.

According to the euclidean distance measure, which we used to capture positive correlations between temporal gene expression patterns, the genes cluster into four major waves of expression and a largely invariant group (Fig. 1). The members of each wave may share inputs from other genes in the genetic network. Wave 1, characterized by high levels of expression occurring only during early developmental stages in 27 genes (E11-E15), includes members of diverse gene families (Fig. 1). Two novel genes, SC6 and SC7, closely cluster to individual members of wave 1, providing a context for their possible roles in terms of functional kinship to known genes. Genes in waves 2 (20 genes) and 3 (21 genes) steeply increase from E13 to E15, and E15 to E18, respectively. Interestingly, the members of wave 2 remain at their plateau expression level at the completion of development, while for wave 3 we generally observe a pattern of transient overexpression. Both waves 2 and 3 are notably confined to neurotransmitter signaling. Finally, wave 4 (17 genes) clusters genes that primarily increase during postnatal development, belonging to several functional families. Twenty-one genes showing largely constant expression (*constant* group) originate from diverse families, however, strictly excluding the neurotransmitter signaling genes and neuro-glial markers.

Comparison of the developmental expression time series for the GABA, nicotinic and muscarinic acetylcholine, and glutamate receptor genes with the phylogenetic trees for these gene families reveals a poor correlation between phylogeny and ontogenetic expression patterns (the same also applies to the data on peptide signaling gene families studied here). This suggests that sequence homologies within gene families are not tightly coupled to the timing of gene regulation over the course of development.

Interestingly, tyrosine hydroxylase (TH), insulin 1 (Ins1), and insulin-like growth factor II (IGFII), which are located on the same human cytogenetic band, 11p15.5 (11), are closely clustered in euclidean distance wave 1. Further, TH and IGFII are in the same subcluster (Fig. 1, a). TH, Ins2, and IGFII are also in close proximity to one another on mouse chromosome 7 (12). This suggests that the expression of some genes may be regulated in parallel due to their proximity on the chromosome.

We used mutual information to capture any correlations (positive, negative and non-linear) between expression time series. Since some genes may share inputs, but respond differently to those inputs, only mutual information is able to identify their coordinated changes in an unbiased fashion. The tree of normalized mutual information clusters (Fig. 1, b) therefore captures potential functional relationships between genes that partially overlap with, but also go beyond those suggested by euclidean clustering.

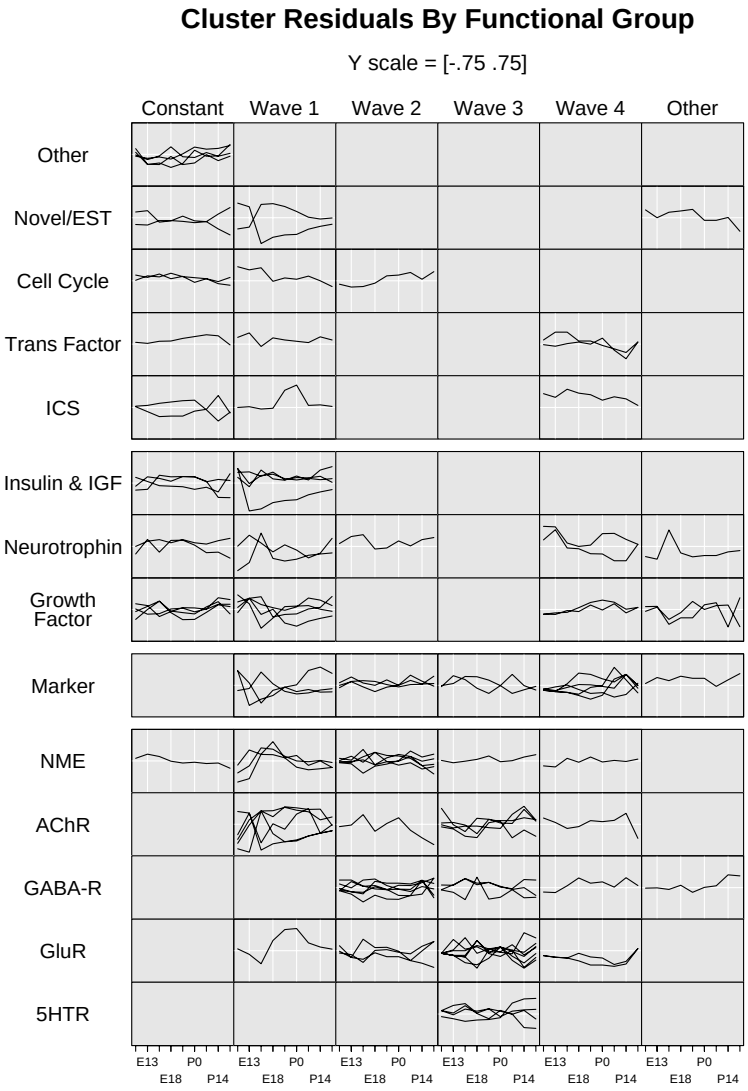


Figure 3. Residuals of individual expression patterns with respect to cluster mean. Cluster mean is subtracted from each gene expression time series. The remainders are plotted to provide a measure of how well the cluster method worked.

The quality of the clustering according to the euclidean distance measure is examined in Fig 2. In this figure, the average gene expression pattern for each cluster is shown in the top panel. The expression time series for each gene family are shown below each cluster. The graph shows all of the measured gene expression patterns, and reflects the range of patterns captured in each cluster. The tightest pattern overlap is found in waves 2 and 3, while wave 1, wave 4 and the constant groups are more diverse. However, compact subclusters (small branches) within the limbs of the euclidean tree (Fig. 1, a) expectedly correspond to highly overlapping trajectories (Fig. 2). It is interesting to note that some of the tightest groupings can be found in particular combinations of gene family and expression cluster, e.g. Marker, NME and GABA-R with wave 2, GluR and 5HTR with wave 3.

To allow a better assessment of clustering quality, we plotted the cluster residuals for all groups (Fig. 3). Residuals are calculated by subtracting each expression pattern from the mean pattern of each cluster. This normalizes the distance from each point, allowing the eye to judge variations without the bias of the absolute expression levels. This representation highlights the cluster tightness in waves 2 and 3.

The functional groups of genes listed in Figure 2 are mapped onto their corresponding expression clusters as determined by euclidean distance and normalized mutual information in Fig. 4. The neurotransmitter signaling genes map primarily to euclidean distance cluster waves 2 and 3, and to mutual information clusters 2 and 4. Similarly, peptide signaling genes map mainly to euclidean distance clusters C (constant) and wave 1, and to mutual information clusters 1 and 6.

The coherence of the ionotropic and metabotropic neurotransmitter receptor gene expression patterns is not a trivial observation. Since metabotropic neurotransmitter receptors are more closely related to peptide receptors than to ionotropic receptors in terms of sequence kinship, one could easily have expected highly parallel expression patterns for metabotropic neurotransmitter receptors and peptide receptors. We conclude from this analysis that receptor ligand class plays a more important role in determining expression patterns than gene sequence homology.

These two distinct clustering methods have a high degree of correspondence, as shown in the central horizontal band of Figure 4, where the euclidean distance clusters are mapped to the mutual information clusters. Many elements of wave 1 and wave 4 map to mutual information cluster 6, as would be expected for anti-parallel gene expression series. That is, a set of the wave 1 genes have a temporal relationship to a set of the wave 4 gene that suggest a common control process. This suggests that some collection of these genes may be participating in a common functional pathway. Fluctuations within the largely "constant" euclidean cluster suffice for mutual information to capture correlations between gene expression time series in this group, mapping them to mutual information clusters 1 and 5.

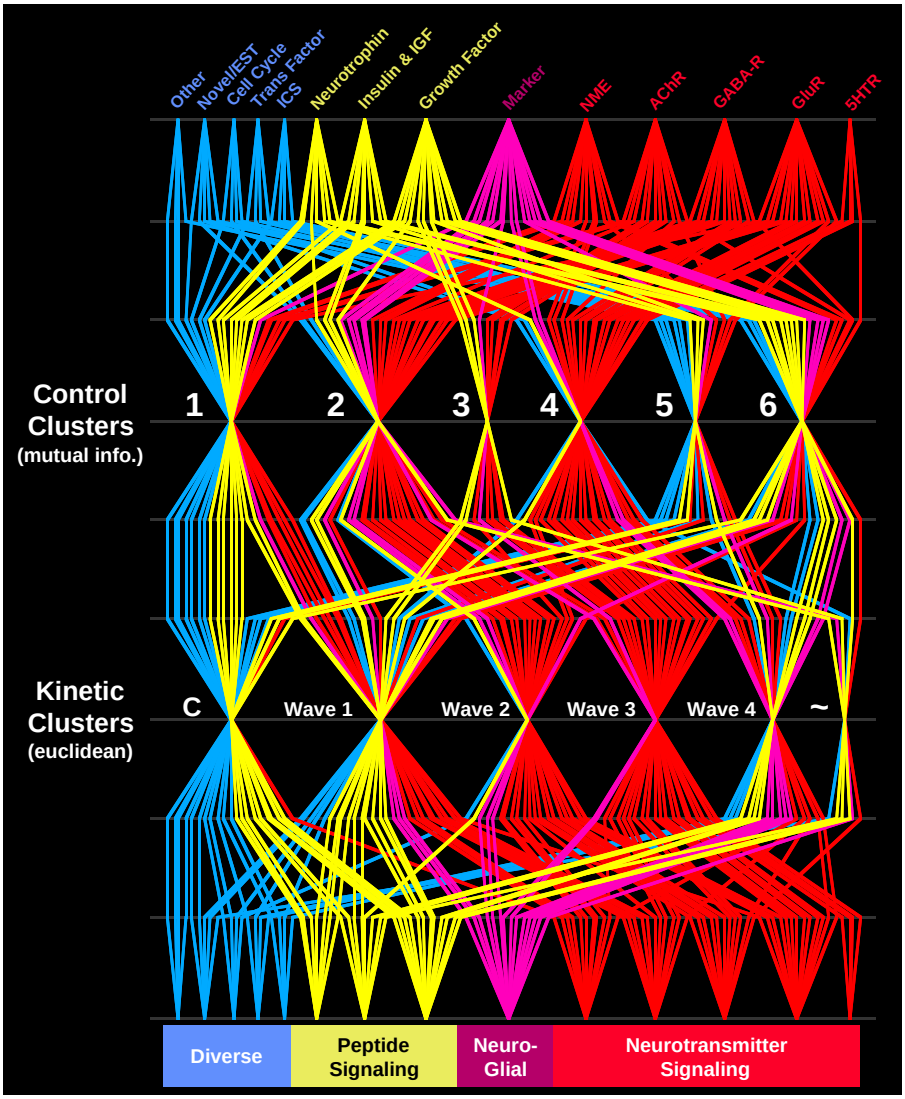


Figure 4. Cluster connectivity. Each line represents an individual gene. Lines converging on a point indicate a cluster of genes with a common property. Genes grouped according to function (top and bottom labels) are mapped to expression clusters of the euclidean distance and mutual information trees. Different colors or shades of gray correspond to the specific functional groups of genes listed at the top and bottom of the diagram, and each line corresponds to a specific gene. Clusters are numbered as in Fig. (1). For a color representation of this plot, please see: http://rsb.info.nih.gov/mol-physiol/PSB98/Clustering.html.

Concluding Remarks

New methods for the analysis of large-scale gene expression data will be necessary for the study of functional genomics. This will involve conceptualizing genomes as sources of information which govern development and other patterns of phenotypic change. From this perspective we hope to gain a greater understanding of the principles underlying complex biological processes.

The euclidean distance cluster analysis captures the kinetic styles of patterns of gene expression. It is remarkable that we only observe six general patterns. This is reminiscent of Turing's observation that there were only six mathematical formulations needed to describe the chemical basis of morphogenesis (13). We are equally reminded Lindenmayer's observation of six basic grammars or algorithms to describe branching processes in Anabena (14). While we do not want to draw attention to this numerological correspondence, it is important that there are a small number of general control processes that appear to be in operation in all of these cases. From other experiments it appears that six clusters are not a maximum set of kinetic of control patterns (data not shown).

Theoretically, the determination of a gene expression matrix could be expanded to include all known genes for any organism. But, that will require some advances in technology to accomplish. However, an incomplete gene expression matrix is sufficient to begin forming hypotheses concerning the regulation of developmental events as reflected in specific gene clusters. Observations of coherent expression behavior suggest common control processes in operation for groups of genes. There is yet insufficient data to adequately characterize the shared control mechanisms implied by the euclidean and mutual information analysis. However, the present analysis does help direct one's attention to a specific set of genes for further investigation. The constraints that cluster analysis places on potential interactions among genes could be incorporated into algorithms designed for exhaustive reverse engineering of genetic networks (5, 6). We intend to integrate and refine these methods also by analyzing simulated control networks that generate this same style of time series data.

We anticipate that further progress in data acquisition and level-by-level inferential analysis will contribute to the goal of reliable predictive modeling of complex biomolecular networks. Attainment of such predictive capacity will obviously play an important role in our understanding of diseases, therapeutic targeting and drug design, and potential re-engineering of small organisms.

References

1. Wen X., Fuhrman S., Michaels G.S., Carr D.B., Smith S., Barker J.L., Somogyi R. (1997) Large-Scale Temporal Gene Expression Mapping of CNS Development. Manuscript submitted for publication.
2. Somogyi R., Wen X., Ma W., Barker J.L. (1995) Developmental kinetics of GAD family mRNAs parallel neurogenesis in the rat spinal cord. J. Neurosci. 15:2575-2591
3. Kauffman S.A. (1993) The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, Inc., New York.
4. Somogyi R., Sniegoski C.A. (1996) Modeling the Complexity of Genetic Networks: Understanding Multigenic and Pleiotropic Regulation. Complexity 1(6):45-63.
5. Somogyi R., Fuhrman S., Askenazi M., and Wuensche A. (1996) The gene expression matrix: Towards the extraction of genetic network architectures. Proc. Second World Congress of Nonlinear Analysts (WCNA96), Elsevier Science.
6. Liang S., Fuhrman S., and Somogyi R. (1998) REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. Proc. Pacific Symposium on Biocomputing 1998. In press.
7. Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package), version 3.5c, distributed by the author, Department of Genetics, University of Washington, Seattle.
8. Shannon C.E. and Weaver W. (1963) The Mathematical Theory of Communication, University of Illinois Press.
9. Thomas J.B. (1969) An Introduction to Statistical Communication Theory. Wiley, New York.
10. Korber B., Farber R., Wolpert D., and Lapedes A. (1993) Covariation of mutations in the V3 loop of HIV-1: An information theoretic analysis. Proc. Natl. Acad. Sci. USA 90:7176-7180.
11. . Lucassen A.M., Julier C., Beressi J.P., Boitard C., Froguel P., Lathrop M., and Bell J.I. (1993) Susceptibility to insulin dependent diabetes mellitus maps to a 4.1 kb segment of DNA spanning the insulin gene and associated VNTR Nature Genet. 4**:** 305-310.
12. Mouse Genome Database (MGD), Mouse Genome Informatics, The Jackson Laboratory, Bar Harbor, Maine. WWW site, URL: http://www.informatics.jax.org/
13. Turing, A., (1952) The Chemical Basis of Morphogenesis. Philosophical Trans. Roy. Soc. B, 237(32):5-72
14. Lindenmayer, A. (1968) Mathematical Models for cellular interaction in development, Parts I and II. J. Theor. Biol., 18:280-315.