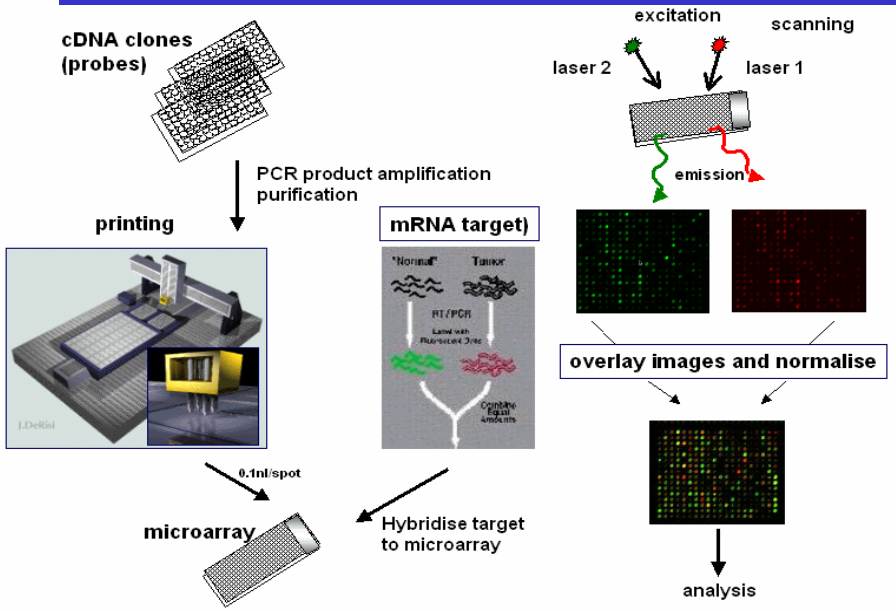
Interactive Exploration of Coherent Patterns in Time-series Gene Expression Data

Daxin Jiang Jian Pei Aidong Zhang
Computer Science and Engineering
University at Buffalo

Microarray Technology



http://www.ipam.ucla.edu/programs/fg2000/fgt_speed7.ppt

Microarray technology

- Monitor the expression levels of thousands of genes in parallel

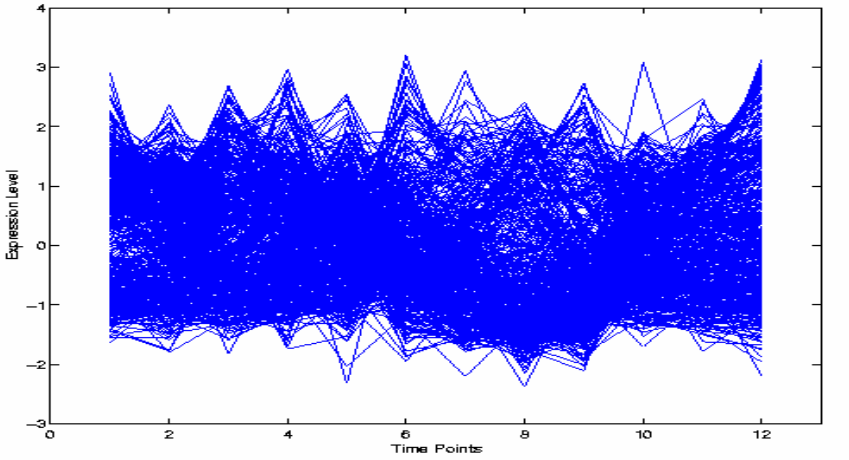
Gene Expression Data Matrix

- Each row represents a gene G_i ;
- Each column represents an experiment condition S_j ;
- Each cell X_{ij} is a real value representing the gene expression level of gene G_i under condition S_j;
  - $X_{ij} > 0$: over expressed
  - $X_{ij} < 0$: under expressed
- A time-series gene expression data matrix typically contains $O(10^3)$ genes and $O(10)$ time points.

	A	B	C	D	E	F	G	H
1	-0.26958	-1.11968	-1.61092	-0.01726	-0.91858	-0.39625	-0.59544	0.223919
2	-1.22081	-0.61064	-1.4015	0.408471	-0.32483	-0.40845	0.667481	-0.99353
3	0.65037	-0.93969	-1.87056	-1.75382	-0.985	-0.28842	-0.37445	1.216714
4	-0.31764	-1.90086	-1.68162	-1.07273	-0.54844	0.243838	0.064647	-0.22084
5	0.039959	-1.07316	-2.0859	0.312404	-0.82065	-0.6364	-0.74902	-0.32929
6	-1.42209	0.214419	-1.30109	-0.20728	-0.42058	-0.95167	-0.85251	-1.84193
7	-0.5303	1.209095	-1.23959	0.355066	-0.5467	-0.60259	-0.31395	-0.57044
8	-0.75512	-0.71537	-1.0849	-0.00374	-0.68283	-0.37029	-0.42753	0.593289
9	-0.47879	-1.02277	-1.01239	-0.33202	0.33603	0.882159	0.979181	-0.2768
10	-0.72888	-0.64282	-1.04118	0.13236	-0.18425	-0.17339	0.524335	-0.97261
11	-1.54334	-0.90966	-1.23333	0.801135	-0.35393	-0.0496	0.422014	-1.39041
12	0.734163	0.336594	-1.74851	0.177696	-0.36607	-1.01617	-0.91572	0.285304
13	-0.08299	-1.03087	-1.56255	-0.3983	0.028534	-0.00045	-0.63836	-0.70778
14	-0.75259	-0.88778	-1.20852	0.703398	-0.79132	-1.07037	0.257377	-0.10077
15	0.811	0.130058	-1.16391	-0.16098	-1.10368	-1.23961	-0.92942	0.503864
16	-1.50545	-0.22578	0.446751	-1.05506	-1.52191	-1.15962	-1.14207	-1.33865
17	-1.88456	-0.19605	0.822872	1.34748	-1.09371	-0.4543	-0.54614	-1.05905
18	-0.71679	-0.97829	-0.95782	0.964454	0.420057	0.314381	0.7907	-0.43544
19	-0.15942	-0.59816	-1.86775	1.819711	0.429999	0.167825	0.447172	-0.46287
20	-0.65254	0.467307	-0.11772	1.380863	0.0606	0.214798	0.317073	-0.63938
21	0.125219	-0.87144	-1.40036	0.776659	-1.84147	-0.20126	-0.79696	0.58284
22	-1.03492	0.206928	-1.18701	0.951558	-0.84569	-0.94715	-0.57483	-0.83308
23	-0.88779	-0.8295	-1.18472	0.869415	-0.42327	-0.39738	-0.26231	-1.38634
24	-0.61967	-0.727	-1.27885	0.265871	-0.05476	-0.2753	-0.30522	-0.82088
25	-1.26181	0.630099	-0.16568	0.127351	0.021887	0.022302	0.12727	-0.75273
26	-0.9317	-0.59289	-0.70058	-0.64865	-0.244	-0.03328	0.003402	-0.96184
27	-1.00066	-0.79028	-1.00607	-0.04573	-0.24725	-0.12933	-0.0095	-0.70748
28	-0.09831	-1.48009	-1.5084	-0.07404	-0.02446	0.068175	0.217438	0.411357
29	0.328346	-0.62071	-1.81974	-0.20051	-0.72874	-0.74451	-0.28102	0.400385
30	-0.89746	0.069834	-0.47114	0.831606	0.208512	-0.16735	0.506605	-0.18187
31	-0.61731	-0.35206	-0.65678	0.919952	0.129398	-0.07423	0.645191	-0.11715
32	-1.11754	1.056664	0.351571	-0.1779	-0.66105	-0.89209	-0.28012	-0.80815
33	-0.46696	-1.18017	-1.16218	0.215408	0.104611	0.42643	0.768182	-0.60385
34	-1.68415	-0.46408	-0.38539	0.286911	-0.03672	0.021101	0.691408	-1.03612
35	-0.44974	-1.17955	-1.25839	-0.23573	0.168901	0.036602	0.788574	-0.2853
36	0.63699	0.993603	-0.44663	-0.20089	-1.1892	-1.19508	-1.35657	0.3804
37	0.180332	0.193894	-0.14022	-0.39459	-0.86483	-0.71266	-0.54127	0.056965
38	0.031928	-0.21662	-1.08205	0.13115	-0.72443	-0.82681	-0.84647	0.709053
39	1.346822	0.313961	-1.21197	-0.62689	-1.07646	-1.08814	-1.34569	0.736034
40	0.360505	-1.2221	-1.43524	0.394193	-0.08287	-0.24353	0.203362	0.534584
41	-1.62892	0.158886	-0.28768	1.263494	0.627216	0.777965	1.095598	-1.36229
42	-1.28297	-1.01468	-0.49856	2.444438	-0.25267	-0.63069	0.04216	-0.10562
43	-1.0645	-0.10843	0.376248	0.125929	-0.44348	-0.03026	0.263266	-0.4077

Gene expression data matrix

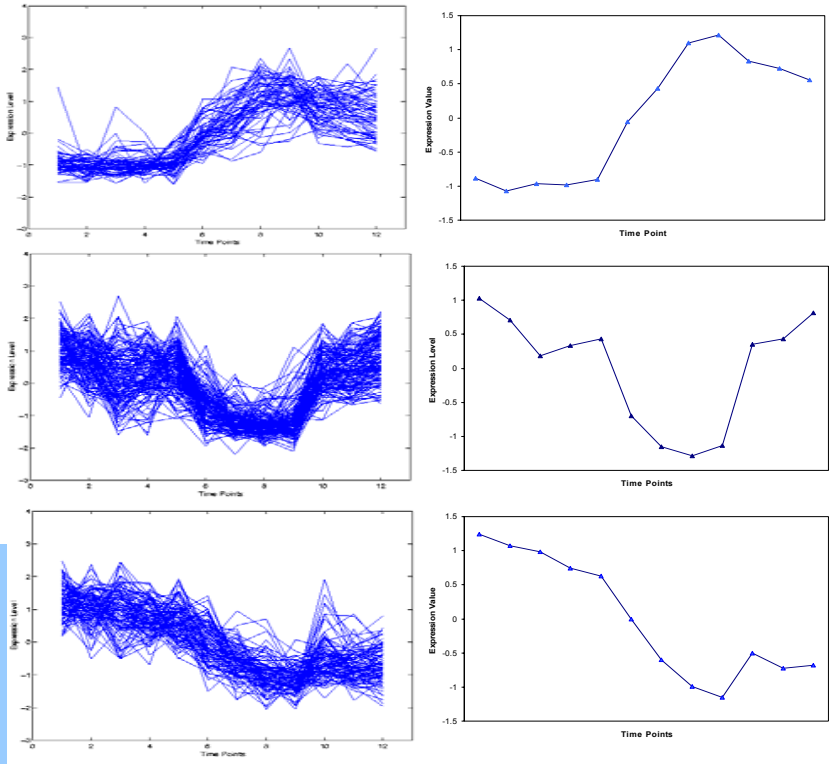
Coherent Patterns and Co-expressed Genes



Parallel Coordinates for a gene expression data

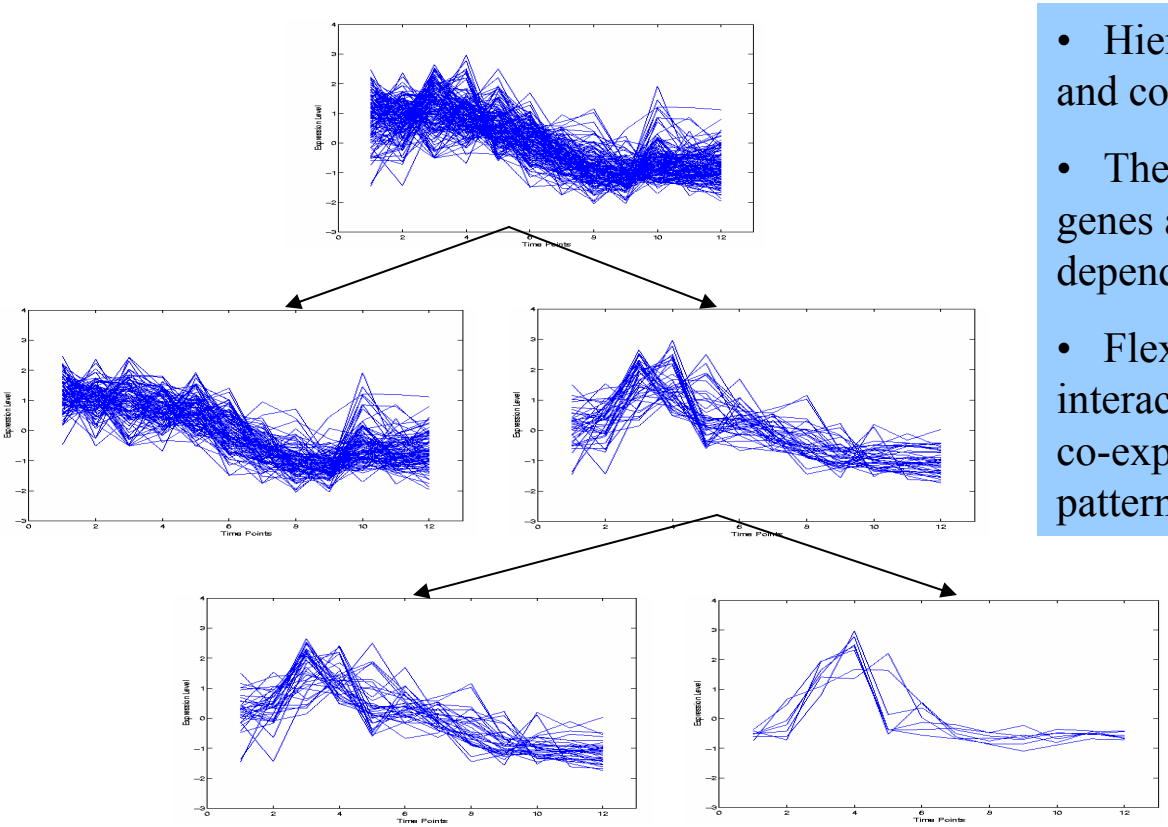
Why coherent patterns and co-expressed genes interesting?

- Co-expression may indicates co-function;
- Co-expression may also indicates co-regulation
- Coherent patterns may correspond to important cellular process



Examples of co-expressed genes and coherent patterns in gene expressiond data

Hierarchies of Co-expressed Genes and Coherent Patterns

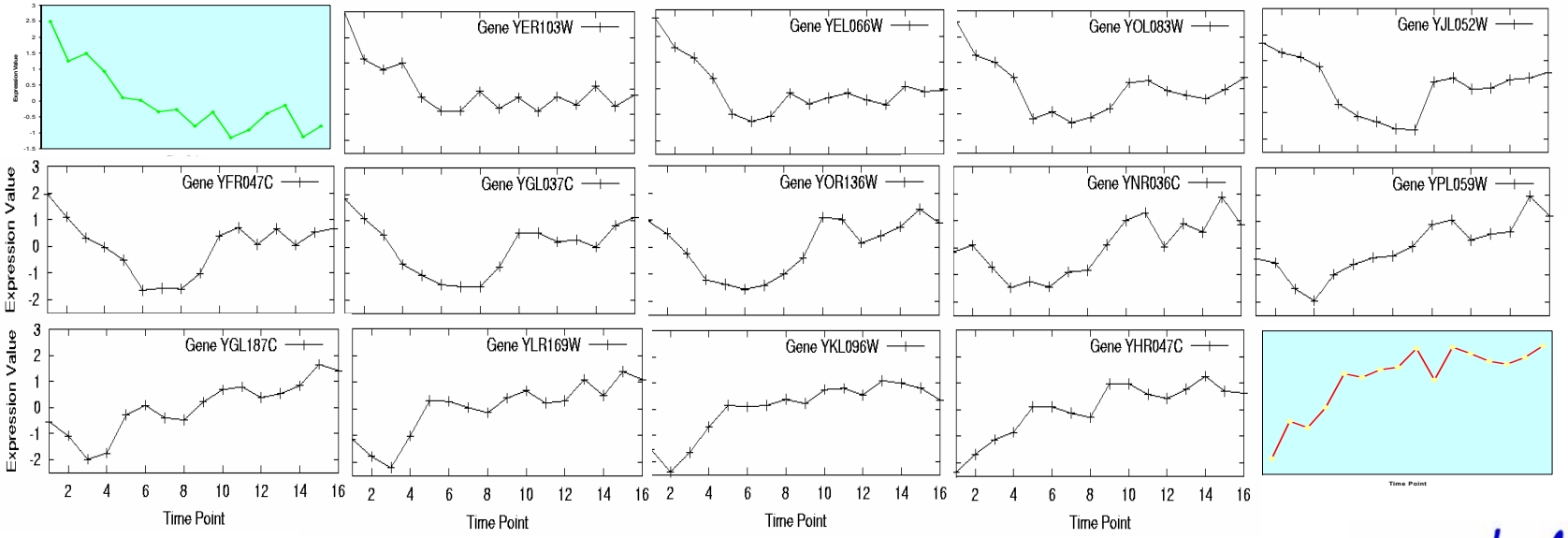


- Hierarchies of co-expressed genes and coherent patterns are typical
- The interpretation of co-expressed genes and coherent patterns mainly depends on the domain knowledge
- Flexible tools are needed to interactively unfold the hierarchies of co-expressed genes and derive coherent patterns

High Connectivity of the Data

- Groups of co-expressed genes may be highly connected by a large amount of "intermediate" genes
- Two genes with completely different patterns can typically be connected by a "bridge"
- It is often hard to find the clear borders among the clusters

Two genes with complete different patterns connected by a "bridge"



Distance Measure

- **We measure the similarity and distance between two genes (objects) as follows**

$$similarity(O_i, O_j) = d_P(O_i', O_j')$$

$d_P(O_i', O_j')$ Is the Pearson's Correlation Coefficient between O_i' and O_j'

$$diatance(O_i, O_j) = d_E(O_i', O_j')$$

$d_E(O_i', O_j')$ Is the Euclidean distance between O_i' and O_j'

O' is the transformation of object O by transforming each attribute d as $O_d' = \frac{O_d - \eta_O}{\sigma_O}$, η And σ are the mean and the standard deviation of all the attributes of O, respectively.

- **The similarity and distance measure defined above are consistent, i.e., given objects O_1, O_2 , O_3 ,and O_4, similarity(O_1,O_2) > similarity(O_3,O_4) if and only if distance(O_1,O_2) < distance(O_3,O_4)**

Definition of Density

- **We choose the density definition by Denclue[1]**
  - **The Gussian influence function**

$$f(O_i, O_j) = e^{-\frac{d(O_i, O_j)^2}{2\sigma^2}}$$

$d(O_i,O_j)$ is the distance between Oi and Oj, and σ is a parameter

  - **Given a data set D**

$$density(O) = \sum_{O_j \in D, O_j \neq O} f(O, O_j)$$

- [1] Hinneburg, A. et al. An efficient approach to clustering in large multimedia database with noise. *Proc. 4th Int. Con. on Knowledge discovery and data mining*, 1998.

Attraction Tree

- Genes with high density "attract" other genes with low density
- The "attractor" of object *O* is the object with the largest attraction to *O*

$$Attractor(O) = \arg \max_{O_j \in A(O)} f(O_j, O)$$

$$where \;\; A(O) = \{O_j \mid density(O_j) > density(O)\}$$

- We can derive an attraction tree based on the attraction between the objects

$$Parent(O) = \begin{cases} nil & if \;\; Attractor(O) = O \\ Attractor(O) & otherwise \end{cases}$$

- The weight for each edge $e(O_i,O_j)$ on the attraction tree is defined as the similarity between O_i and O_j.

Coherent Pattern Index Graph

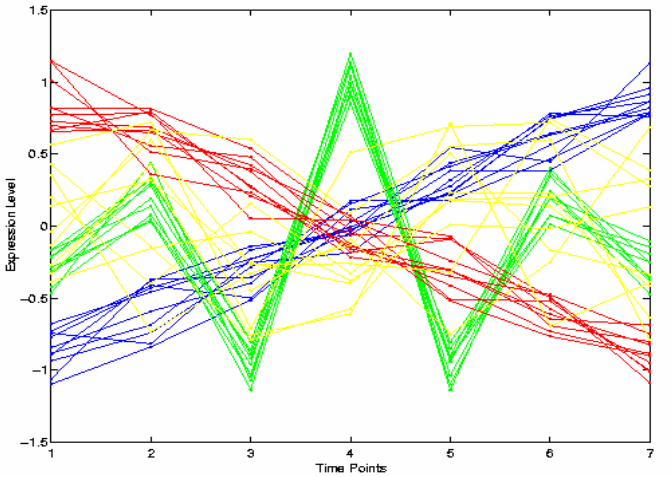
- We search the attraction tree based on the weight of edges and order the genes in the "index list"
- For each gene g_i in the index list $g_1 \ldots g_n$, the "coherent pattern index" is defined as

$$CPI(g_i) = \sum_{j=1}^{p} Sim(g_{i+j}) - \sum_{j=0}^{p-1} Sim(g_{i-j})$$ where p is a parameter,

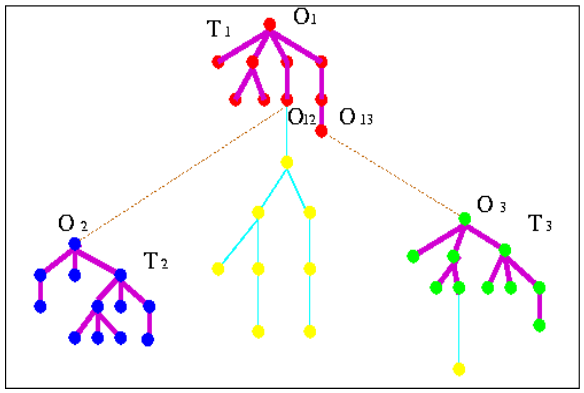
Sim(gi) is the similarity between gi and its parent gj on the attraction tree. Sim(gi) is set to 0 if i≤1 or I>n.

- The graph plotting the coherent pattern index value w.r.t. the index list is called the "coherent pattern index graph"
- A pulse in the coherent pattern index graph indicates a coherent expression pattern

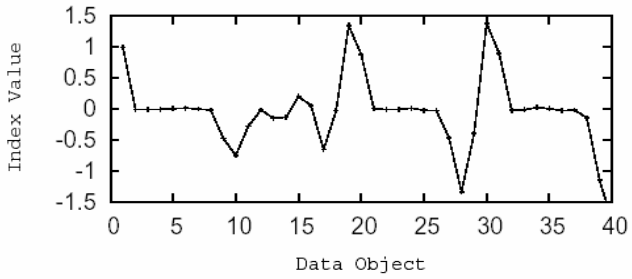
An Example



A sample data set



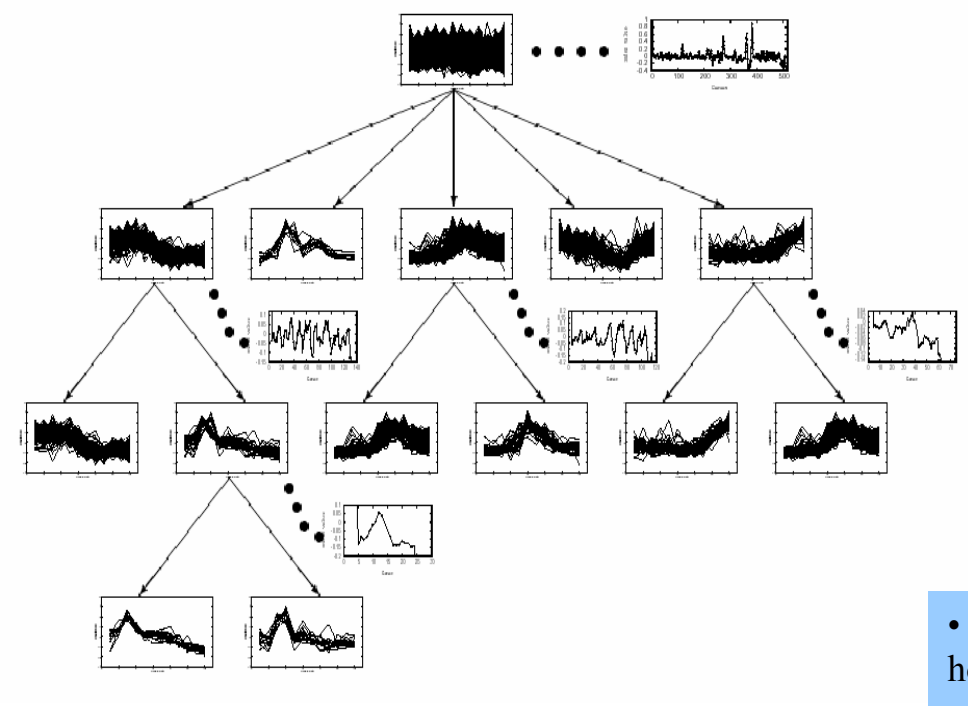
The attraction tree

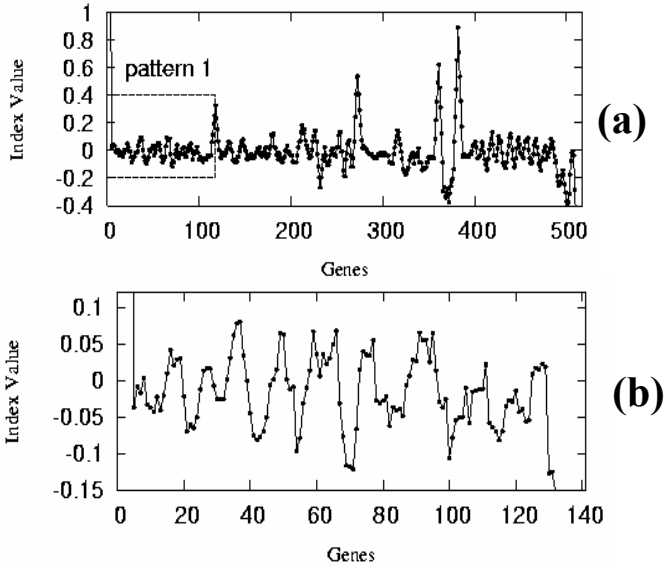


The coherent pattern index graph

- The weight of edges on the attraction tree characterizes the coherence relationship between genes (represented by purple, cyan and brown lines)
- The three pulses in the coherent graph index graph indicate the three patterns in the data set
- Genes between two neighboring pulses are co-expressed genes and share coherent patterns

Interactive Exploration -- GeneXplorer





Interactive exploration on Iyer's data[2]

- The coherent pattern index graph gives indications on how to split the genes into co-expressed groups
- Suppose the user accept the 5 pulses suggested in figure (a), and click on the 2nd pulse
- The system will "zoom in" the coherent pattern index graph for genes between the 1st pulse and the 2nd pulse (figure (b))
- The user can select clicking on the pulses in figure (b) and further split the genes until no split is necessary

[2] Iyer, V.R. et al. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

Comparison With Other Approaches

We compare the patterns discovered from the Iyer's data[2] by different approaches with the ground truth by Eisen *et al.* [3]

- GeneXplorer identifies more patterns in the ground truth and does not report any false patterns
- Pattern 5 in the ground truth is only reported by GeneXplorer
- The only pattern in the ground truth (pattern 9) missed by GeneXplorer is missed by any other method

Pattern	GeneXplorer(9)	Adapt(7)	CLICK(7)	CAST(9)
1	0.993	0.956	0.884	0.955
2	0.957	0.911	0.991	0.887
3	0.984	0.993	0.994	0.997
4	0.980	0.984	0.883	0.968
5	0.958	0.855	0.868	0.855
6	0.952	0.989	0.970	0.984
7	0.967	0.976	0.990	0.719
8	0.991	0.997	0.914	0.999
9	0.702	0.824	0.844	0.800
10	0.974	0.981	0.976	0.996

Each cell represents the similarity between the pattern reported by different approaches and the corresponding pattern in the ground truth (if any)

Conclusions:

- ☑ **The coherent pattern index graph is effective to give users highly confident indication of the existence of coherent patterns**
- ☑ **The GeneXplorer provides interactive exploration to integrate user's domain knowledge**

■ [3] Eisen M.B. et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, Vol. 95:14863–14868, 1998.