10-810 /02-710

Computational Genomics

Time series analysis

nature REVIEWS GENETICS

Review

Nature Reviews Genetics **13**, 552-564 (August 2012) | doi:10.1038/nrg3244

ARTICLE SERIES: Study designs

Studying and modelling dynamic biological processes using time-series gene expression data

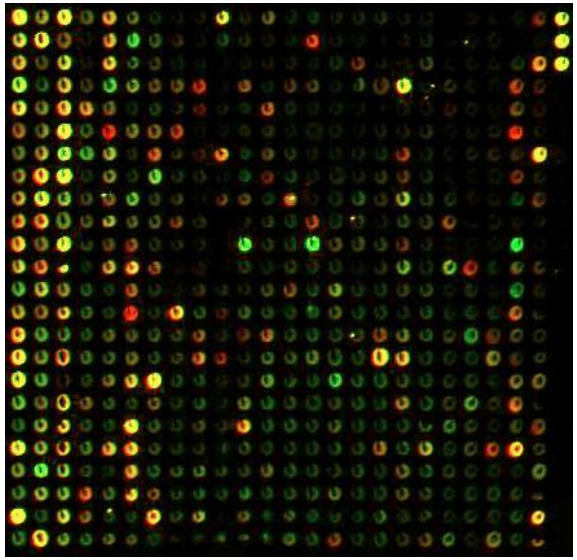
Ziv Bar-Joseph[1], Anthony Gitter[2] & Itamar Simon[3] About the authors

top

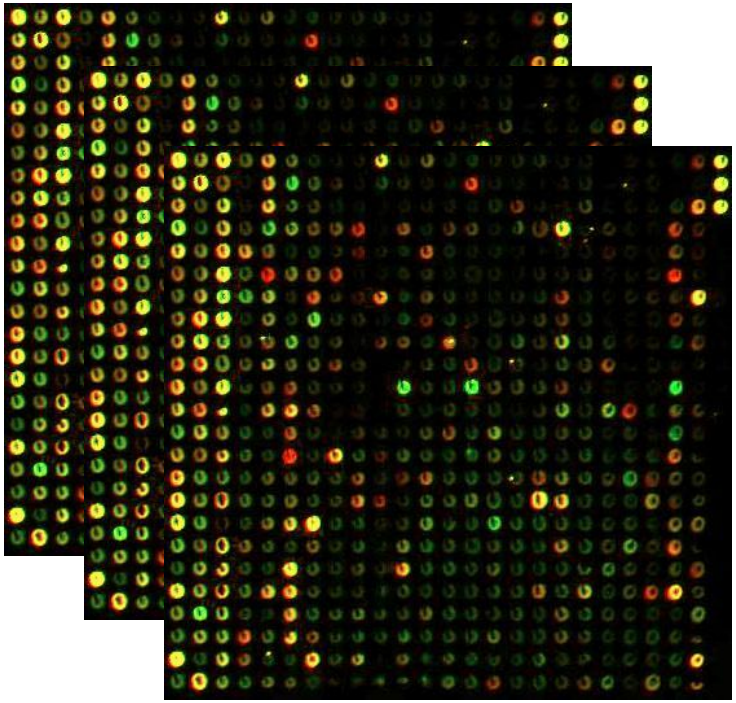
Biological processes are often dynamic, thus researchers must monitor their activity at multiple time points. The most abundant source of information regarding such dynamic activity is time-series gene expression data. These data are used to identify the complete set of activated genes in a biological process, to infer their rates of change, their order and their causal effects and to model dynamic systems in the cell.

Expression Experiments

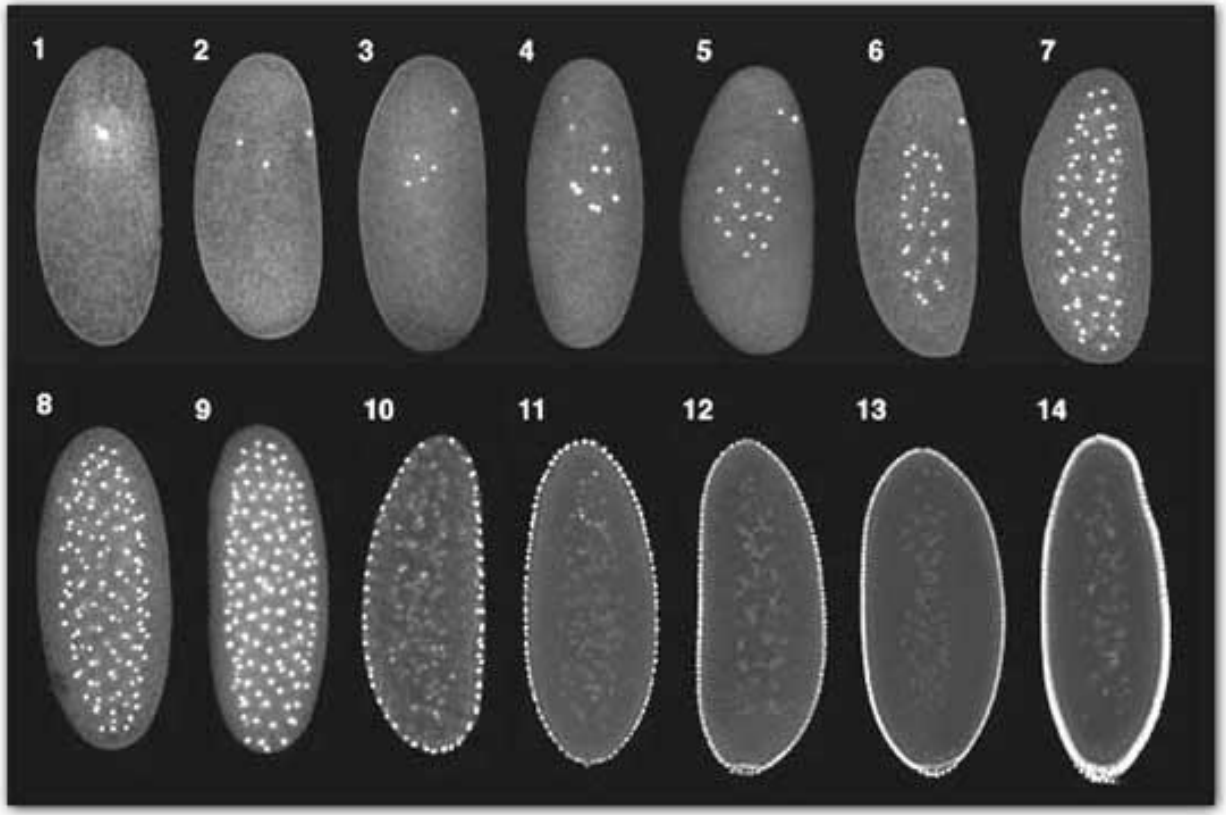
Static: Snapshot of the activity in the cell



Time series: Multiple arrays at various temporal intervals

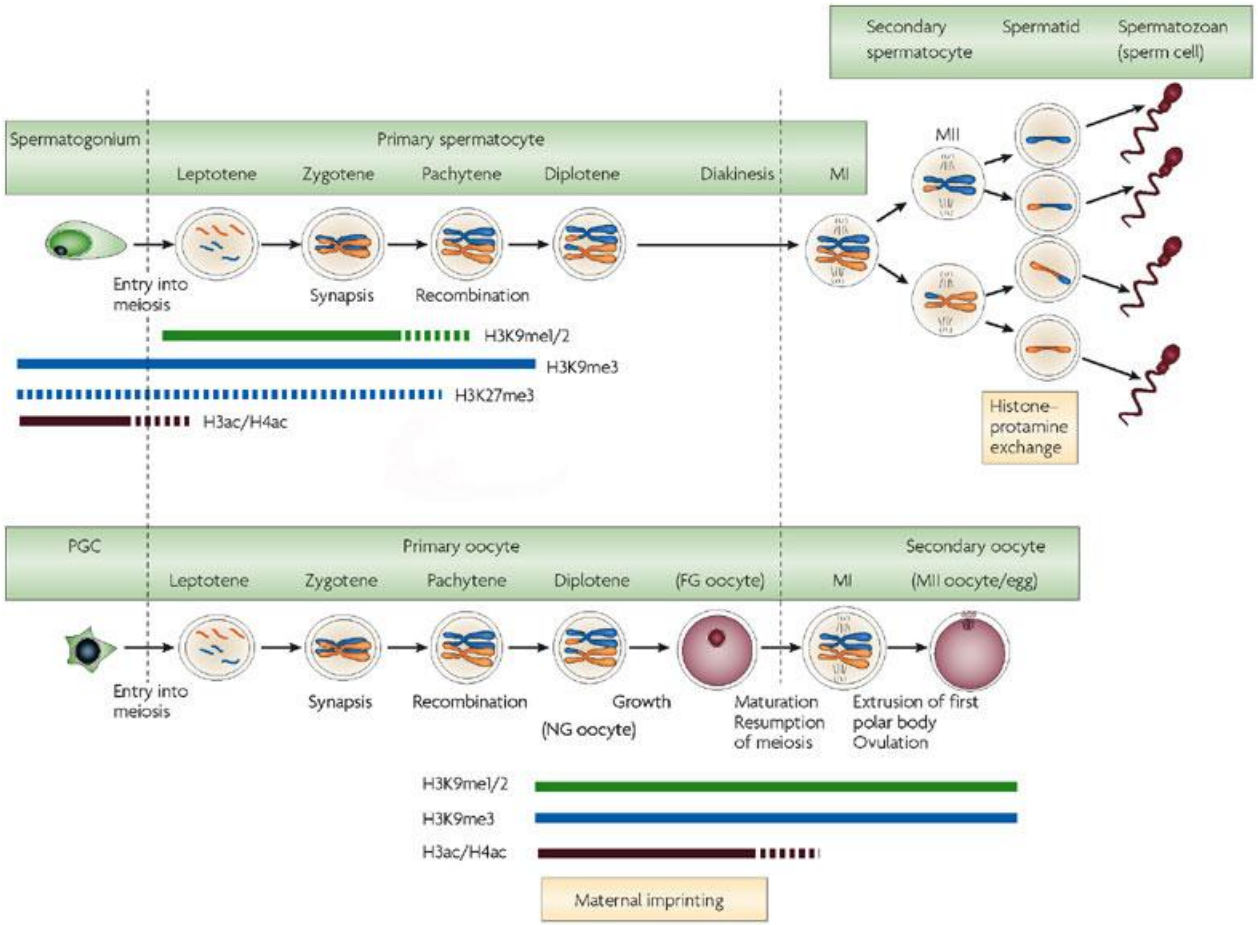


Time Series Examples: Development

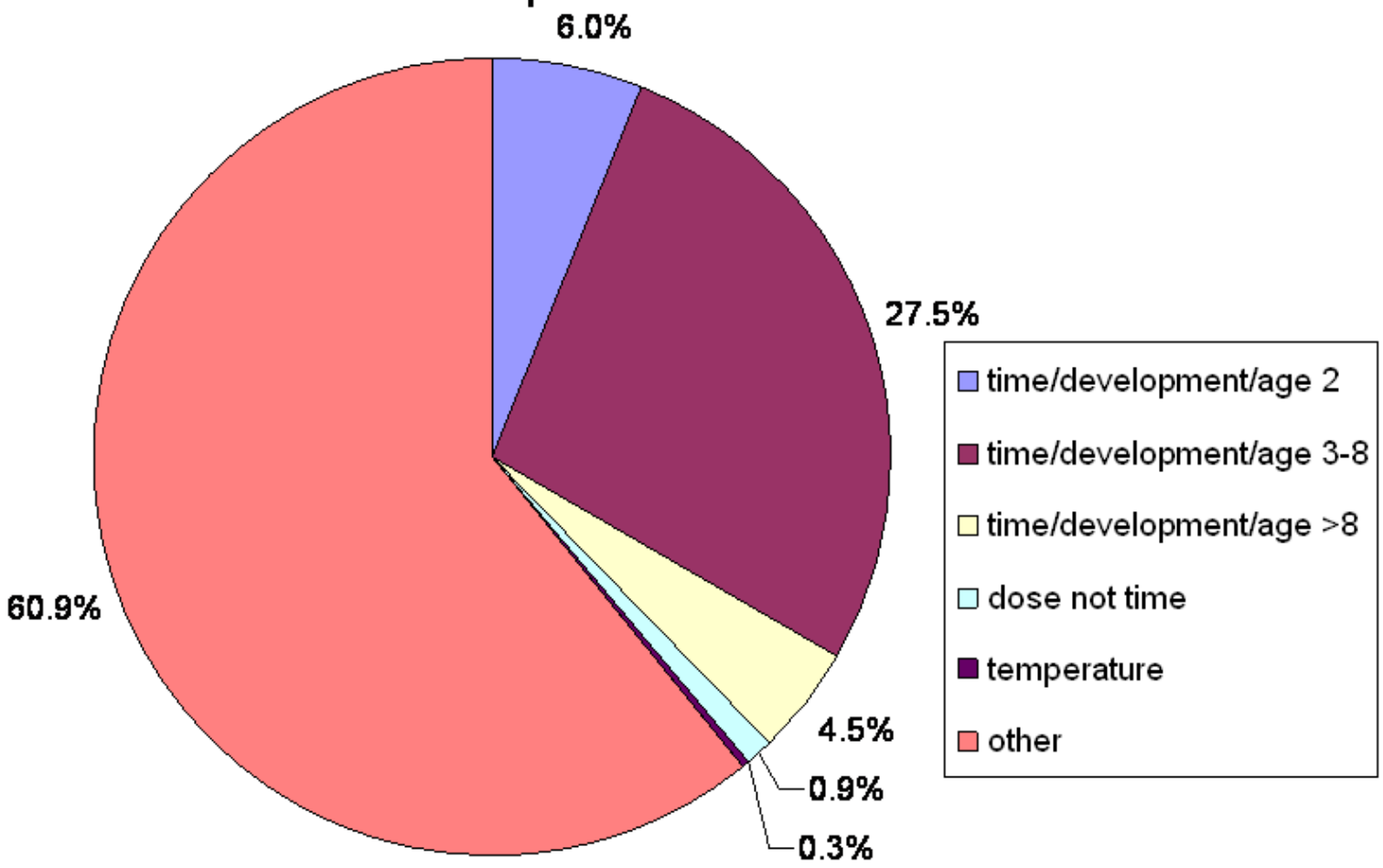


Development of fruit flies [Arbeitman, Science 02]

Epigenetics time series



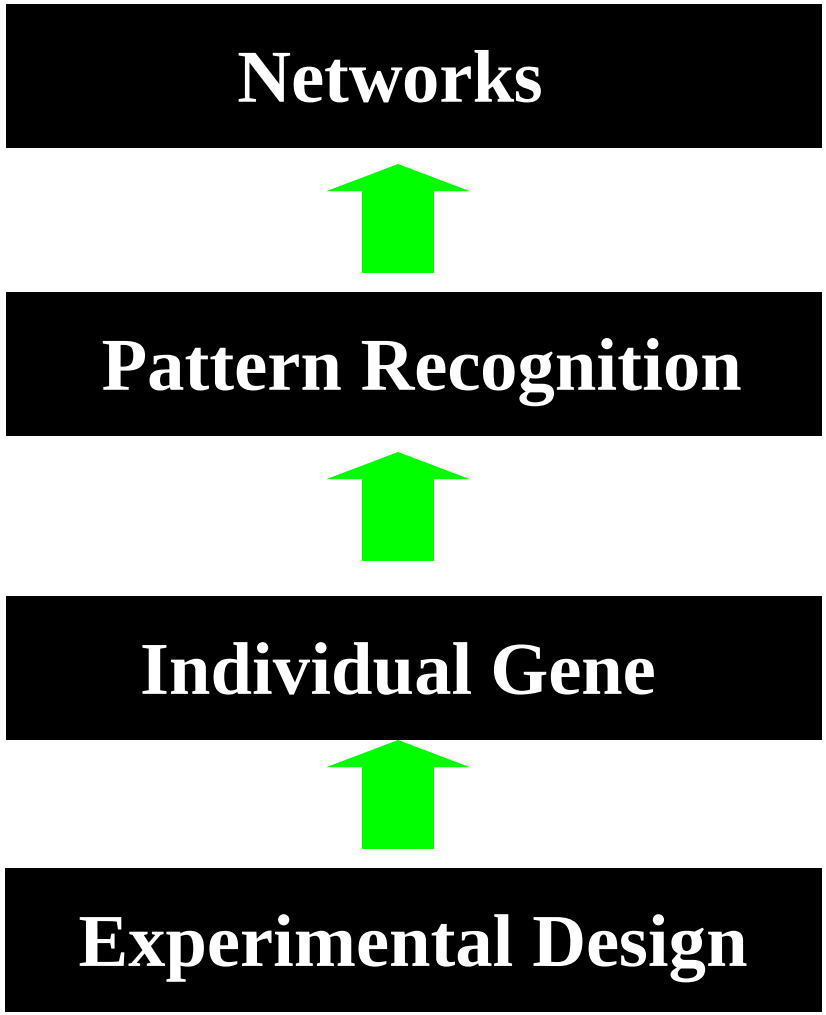
Nature Reviews | Genetics


Distribution of Microarray Data Sets in the Gene Expression Omnibus
6.0%
27.5%
4.5%
0.9%
0.3%
60.9%
time/development/age 2
time/development/age 3-8
time/development/age >8
dose not time
temperature
other

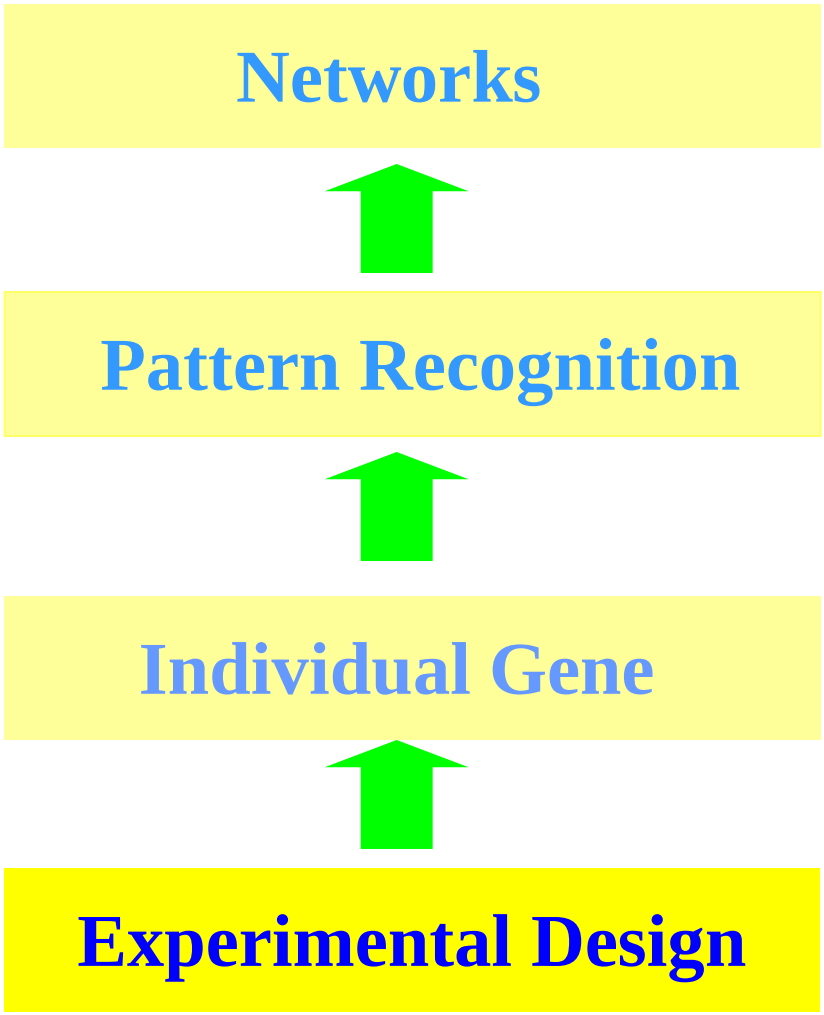
Unique features of time series expression experiments

- Autocorrelation between successive points.
- Can identify complete set of acting genes.
- Allows to infer causality.

Time Series Expression Analysis

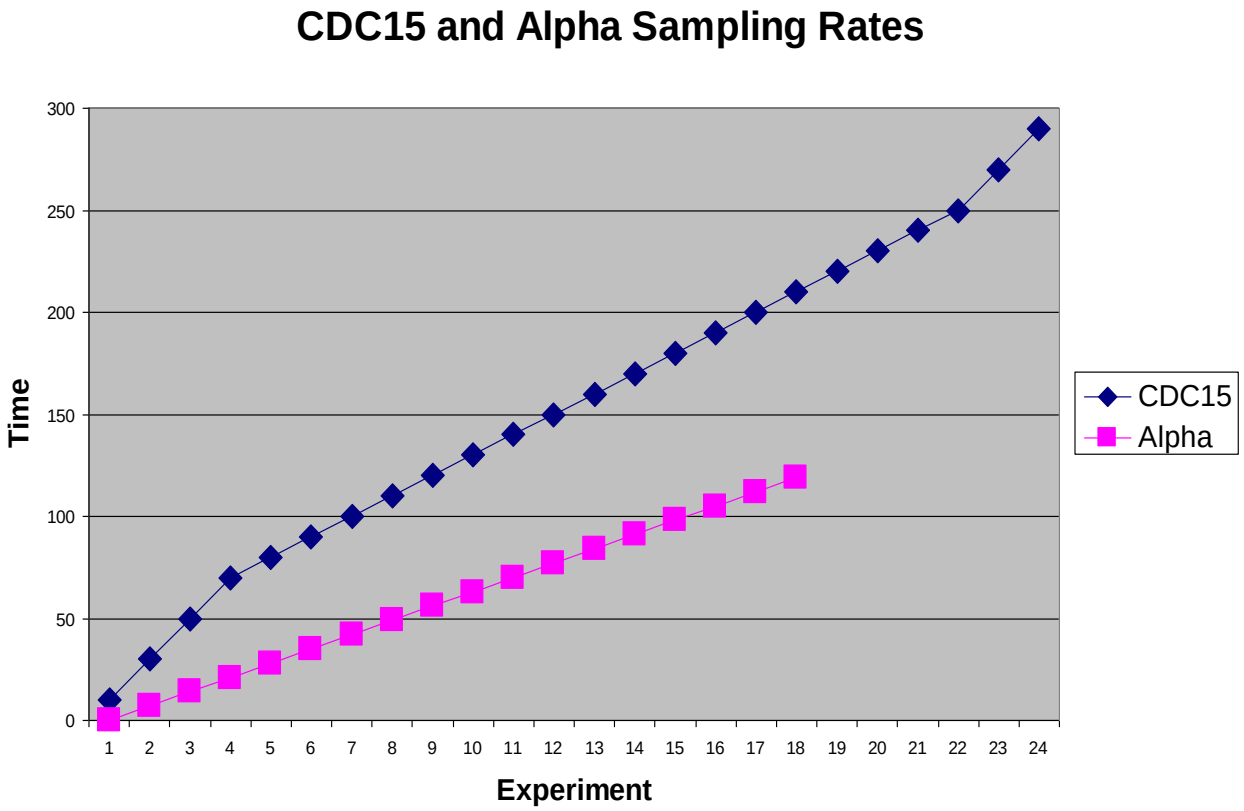


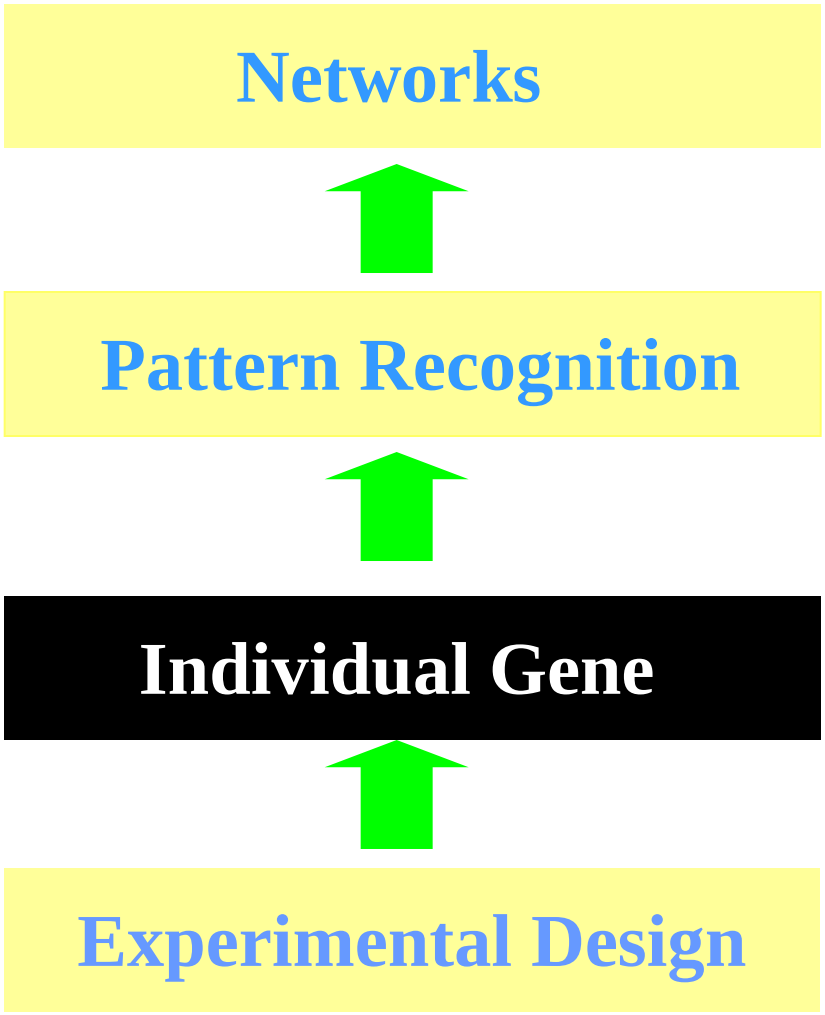
Computational	Biological
information fusion	dynamic regulatory networks
clustering, classification	function, response programs
normalization, miss. values, interpolation	alignment, diff. expressed genes
sampling rates, duration	transcription, decay rates


Networks
Pattern Recognition
Individual Gene
Experimental Design

Sampling Rates

- Non uniform
- Differ between experiments




Networks
Pattern Recognition
Individual Gene
Experimental Design

Issues to address

- Continuous representation
- Identifying differentially expressed genes
- Synchronization

Yeast Cell Cycle Datasets

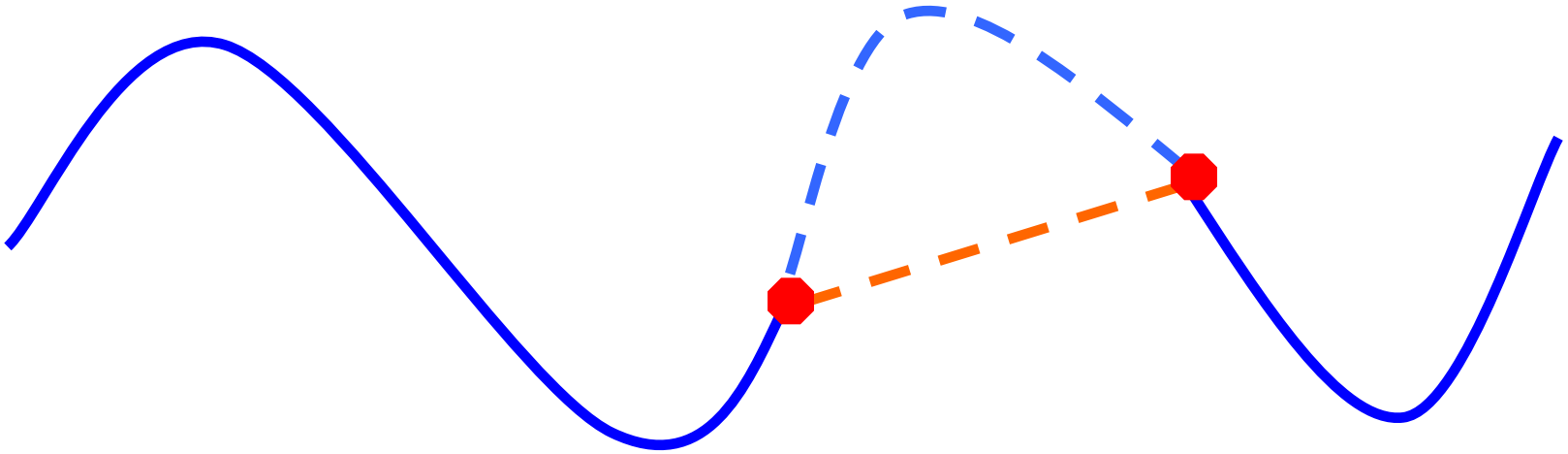
Dataset	Method of arrest	Duration	Cell cycle length	Sampling	Repeats
alpha (Spellman 98)	alpha mating factor	0-119m	64m	every 7 minutes	1
cdc15 (Spellman 98)	temp. sensitive cdc15	10-290m	112m	ev. 20m for 1 hr, ev. 10m for 3 hr, ev. 20m for final hr	1
cdc28 (Cho98)	temp. sensitive cdc28	0-160m	85m	every 10 minutes	1
fkh1/fkh2 knockout (Zhu00)	alpha mating factor	0-215m	105m	every 15m until 165m then after 45m	2
yox1/yhp1 knockout (Pramila02)	alpha mating factor	0-120m	60m	every 10 minutes	1

Representing time series expression data

- We are capturing a continuous process with a few samples.
- We need a way to convert our samples for each gene to an expression profile.
- Some simple techniques:
  - Linear interpolation
  - Spline interpolation
  - Functional assignment

Standard interpolation

If we have missing values and noise linear interpolation will fail to reproduce an accurate representation.



Cubic Splines

- Piecewise cubic polynomials satisfying continuity and smoothness constraints.

- B-splines represents the splines as a linear combination of basis functions, where the coefficients are the spline control points.

$$Y_i(t) = S(t)F$$

- When faced with noise and missing values, splines overfit the data.

Many of the genes are co-expressed. Thus, we use classes of similarly expressed genes to constrain spline assignment, and overcome noise and missing data.

Continuous representation: The power of co-expression

Many of the genes are co-expressed, we can use co-expressed genes to overcome noise in individual gene

Q: How can we identify the set of co-expressed genes?

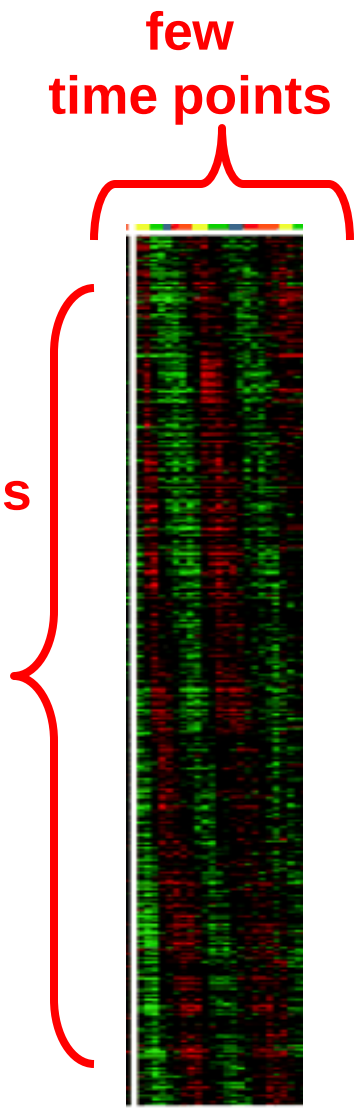
A: **Clustering**

Q: How do we use the cluster genes?

A: Instead of average representation extract **shape information** (co-variance matrix)

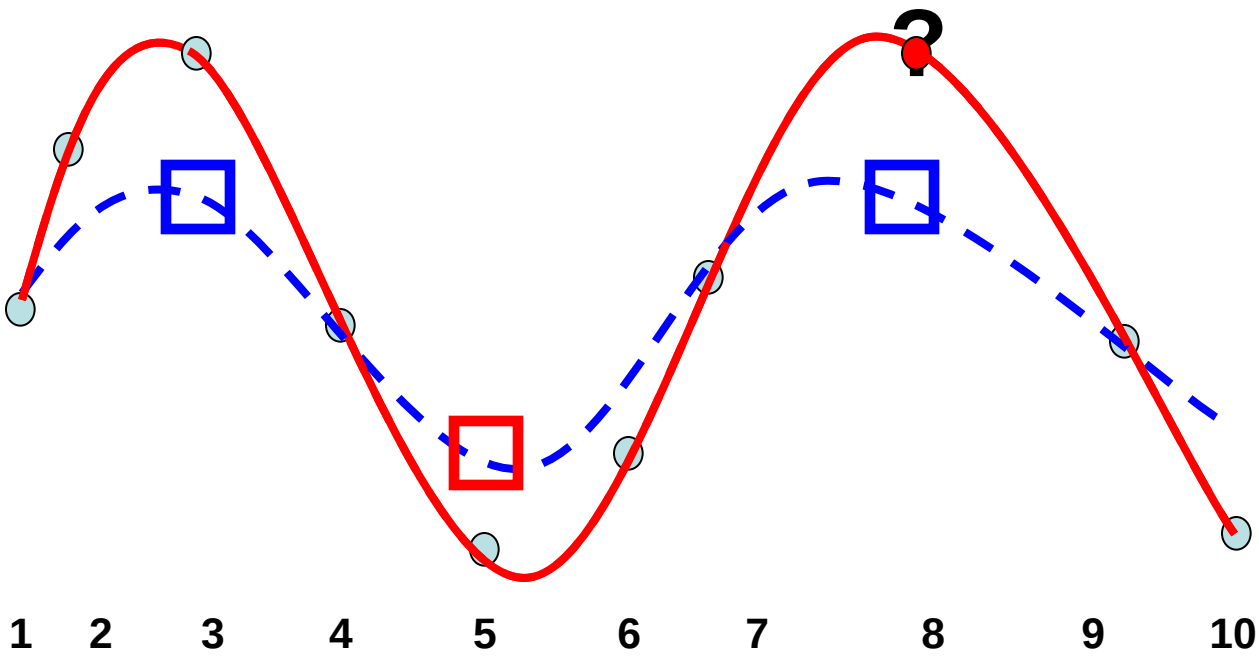
Q: Covariance matrix is very big, what about overfitting?

A: Use **dimensionality reduction** methods (splines)



Bar-Joseph *et al J. comp. bio.* 2003

A mixed effects model



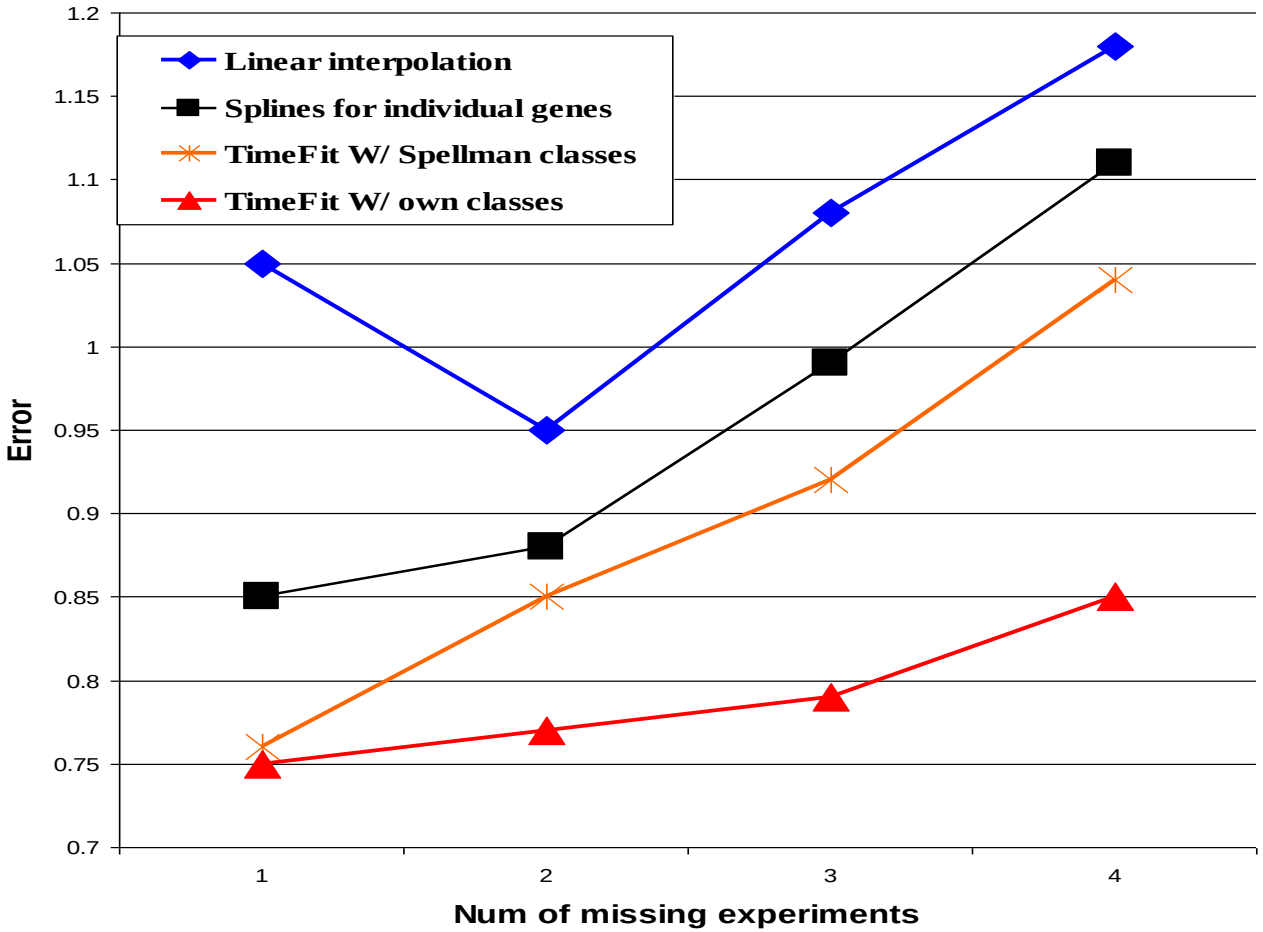
Class covariance matrix

	1	3	5	8	10
1	.3	.5	-.5	.5	-.5
3	.5	.3	-1	1	-1
5	-.5	-1	.3	-1	1
8	.5	1	-1	.3	-1
10	-.5	-1	1	-1	.3

Class average expression profile

Comparing Interpolation Methods

Holding out time points and using each method to predict missing data



Issues to address

- Continuous representation
- Identifying differentially expressed genes
- Synchronization

Issues to address

- Continuous representation
- Alignment
- Identifying differentially expressed genes
- Synchronization

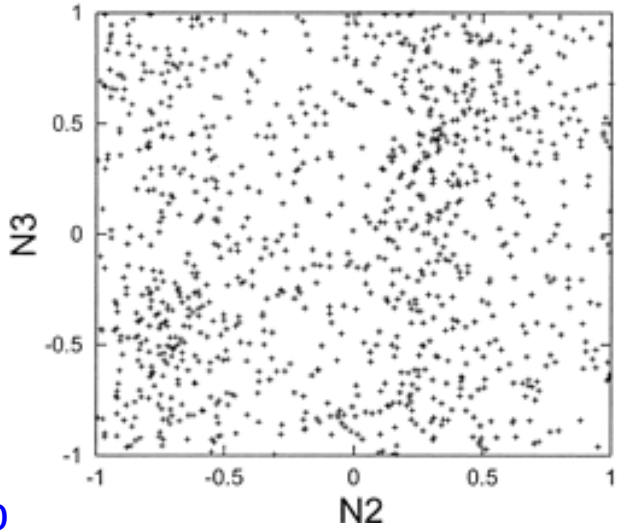
Cell cycle expression: time line

- 1997, 1998 – budding yeast
- 2000 - bacteria
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse

Cho *et al*, *Nature Genetics* 2000

Cell cycle expression: time line

- 1997, 1998 – budding yeast
- 2000 - bacteria
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse
- 2002 – human data is noise!



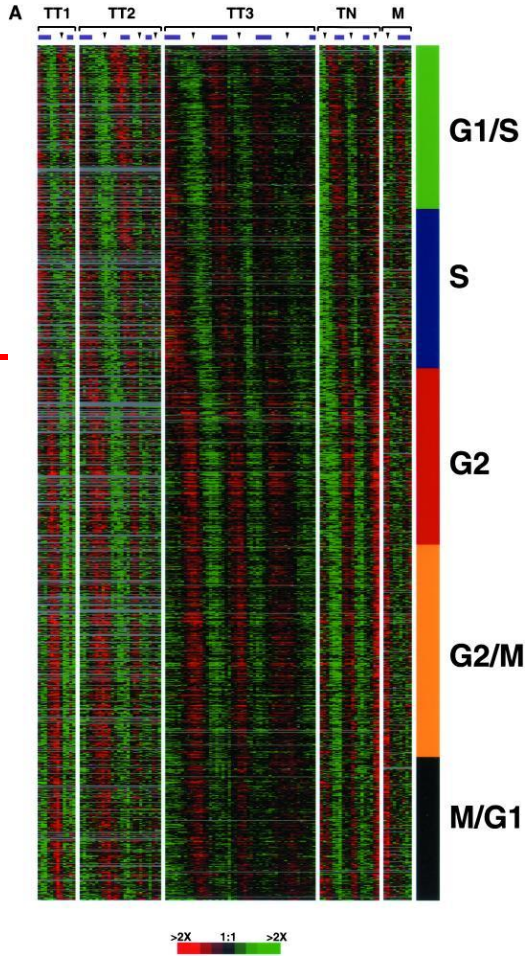
Reproducibility of peak between two repeats

Shedden & Cooper, PNAS, 2002

Cell cycle expression: time line

- 1997, 1998 – budding yeast
- 2000 - bacteria
- 2000 – plants
- 1999, 2000 - human
- 2001 – mouse
- 2002 – human data is noise!
- 2002 – Cancer cell cycle expression

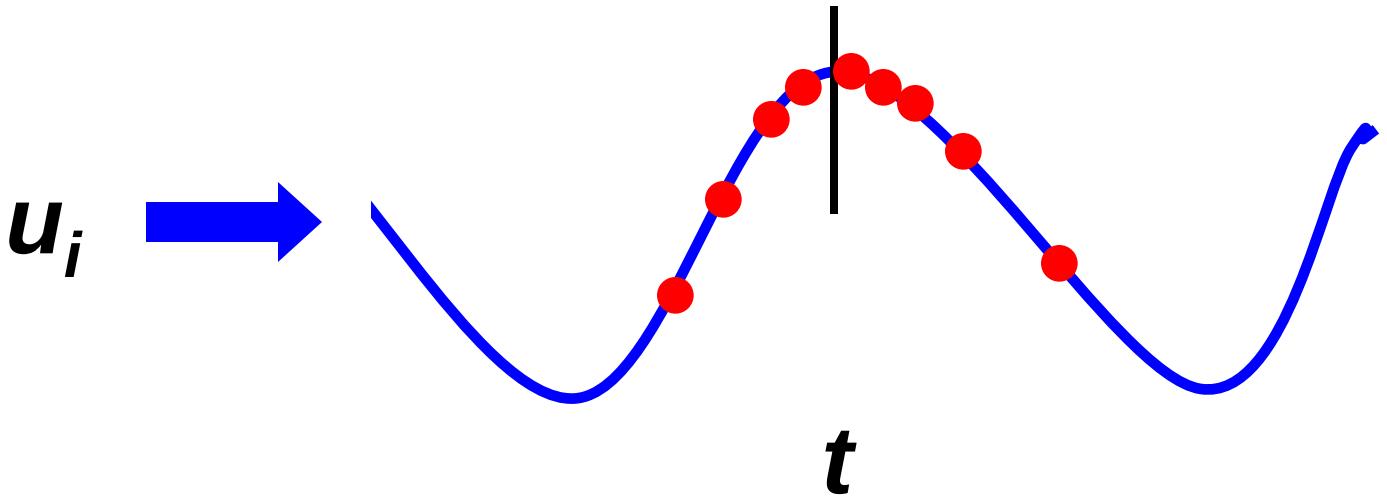
Can we compare cancer and normal expression programs?



Whitfield et al, *MBC*, 2002

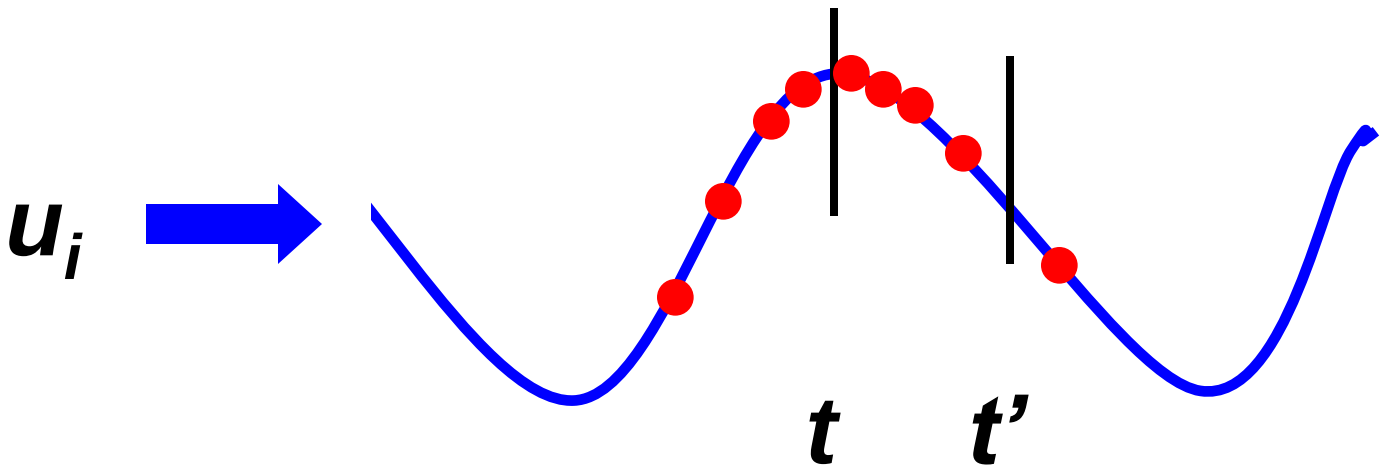
Main problem: Population effects

- Microarray experiments profile population of cells.
- Cells are artificially synchronized, not all cells are arrested.
- Even for those that are, synchronization is lost over time.

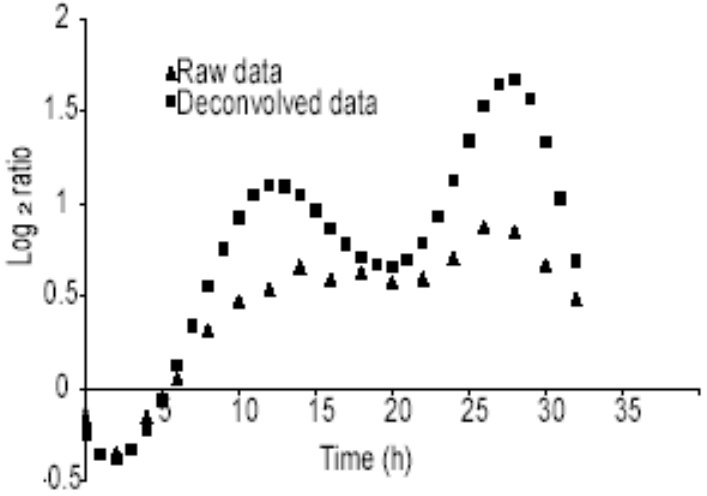


Data integration to overcome synchronization loss

- We learn a synchronization loss model from independent measurements
- Using this model we estimate the proportion of cells at time ***t'*** when the real time is **t**
- We re-distribute the values measured for each gene according to the number of cells at this time

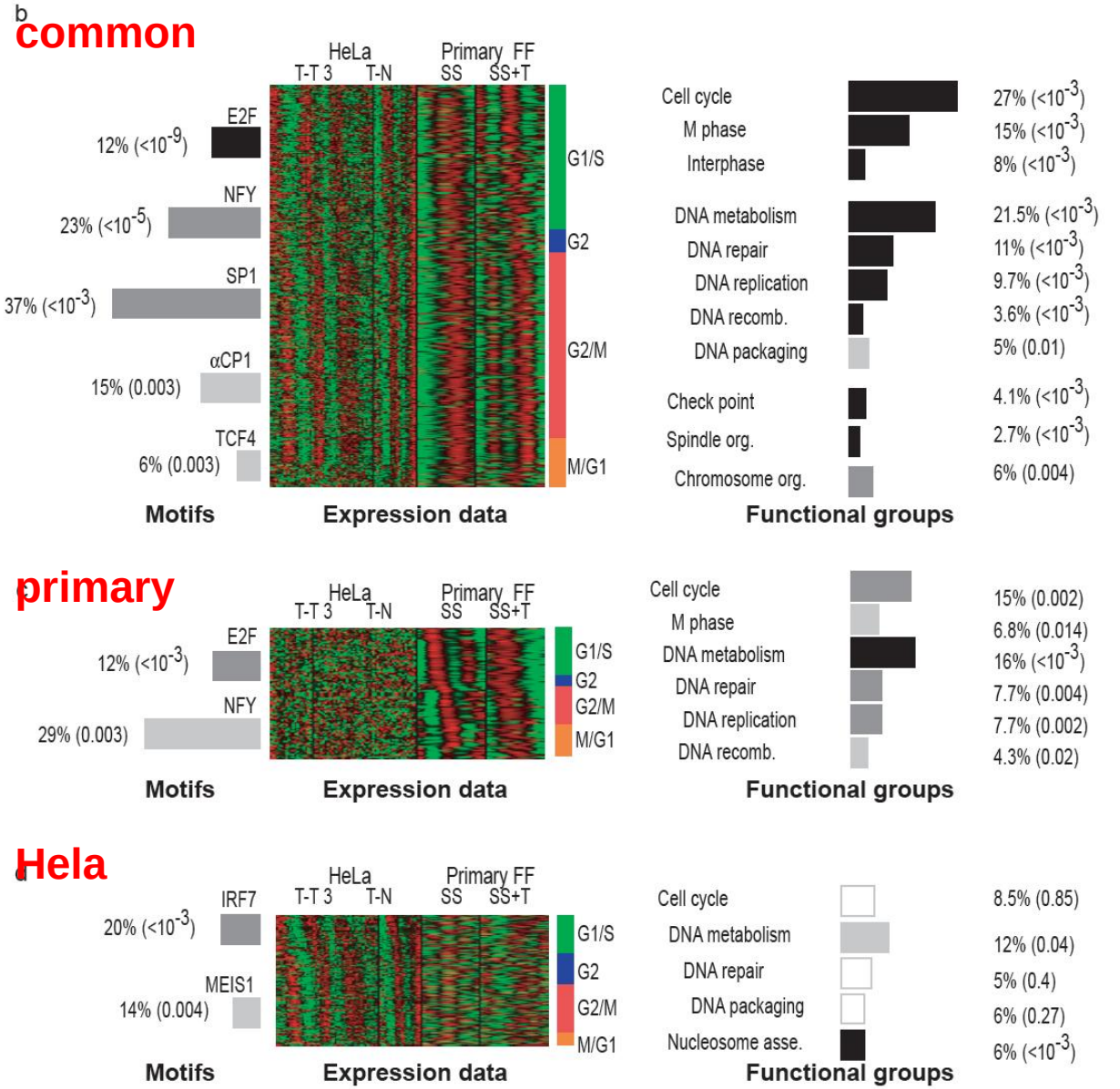


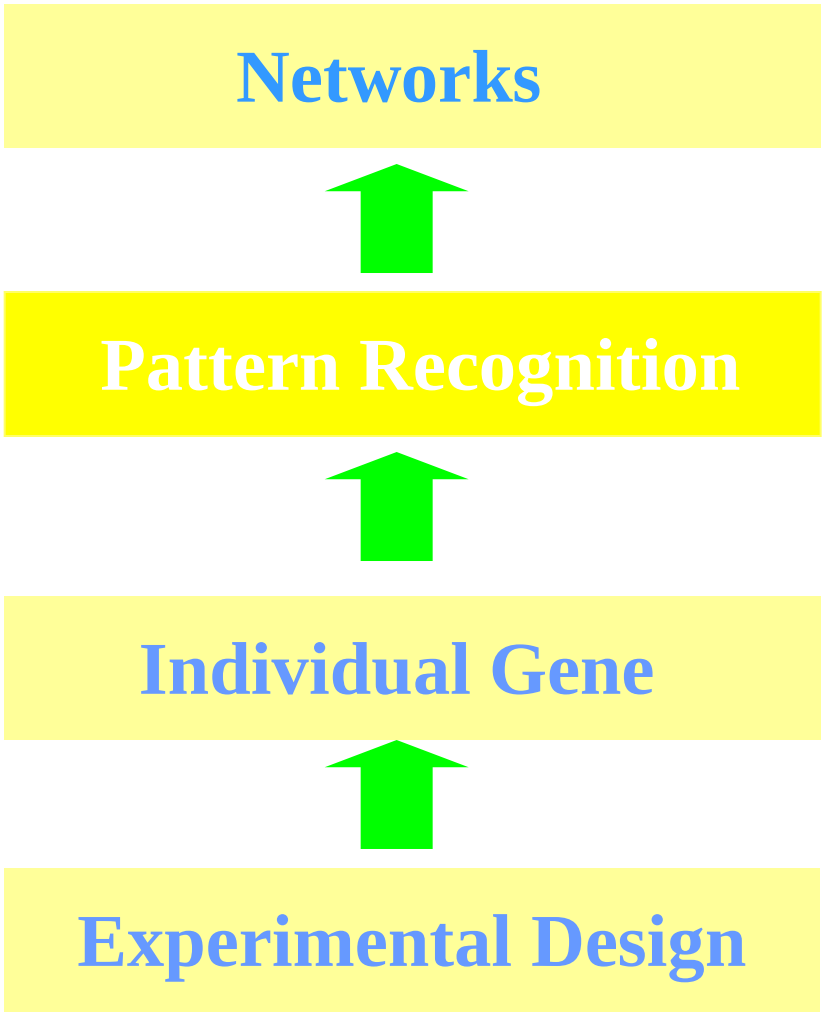
Re-Synchronization: Birc5 measured vs. corrected



Results for human expression data

Validation by PCR



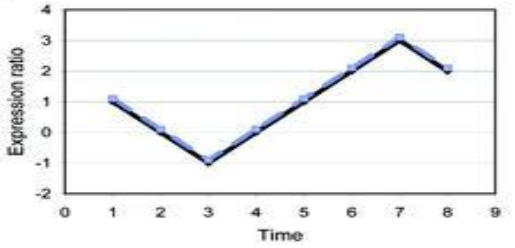

Networks
Pattern Recognition
Individual Gene
Experimental Design

Clustering

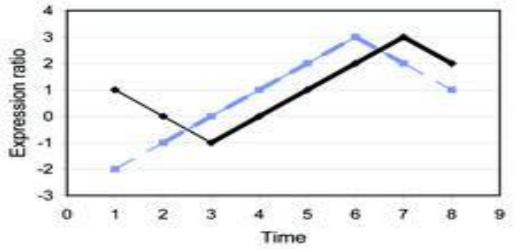
- Handling non uniform sampling rates.
- Identifying relationships between genes based on expression profiles.
- Determining relationships between clusters.

Time Shifted and Inverted Profiles

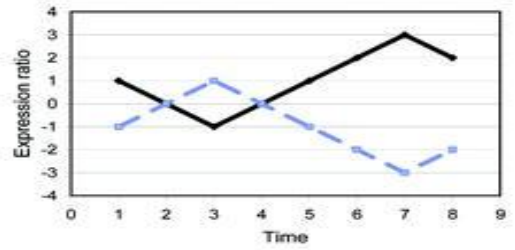
A



B



C



D

		1	0	-1	0	1	2	3	2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	0
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$$E_{i,j} = \max(E_{i-1,j-1} + x_i \; y_j, 0)$$

E

		-2	-1	0	1	2	3	2	1
	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	2	3	2	1
0	0	0	0	0	0	1	2	3	2
-1	0	2	1	0	0	0	0	0	2
0	0	0	2	1	0	0	0	0	0
1	0	0	0	2	2	2	3	2	1
2	0	0	0	0	4	6	8	7	4
3	0	0	0	0	3	10	15	14	10
2	0	0	0	0	2	7	16	19	16

$$E_{i,j} = \max(E_{i-1,j-1} + x_i \; y_j, 0)$$

F

		-1	0	1	0	-1	-2	-3	-2
	0	0	0	0	0	0	0	0	0
1	0	1	0	0	0	1	2	3	2
0	0	0	1	0	0	0	1	2	3
-1	0	0	0	2	0	0	0	0	2
0	0	0	0	0	2	0	0	0	0
1	0	1	0	0	0	3	2	3	2
2	0	2	1	0	0	2	7	8	7
3	0	3	2	0	0	3	8	16	14
2	0	2	3	0	0	2	7	14	20

$$D_{i,j} = \max(D_{i-1,j-1} - x_i \; y_j, 0)$$

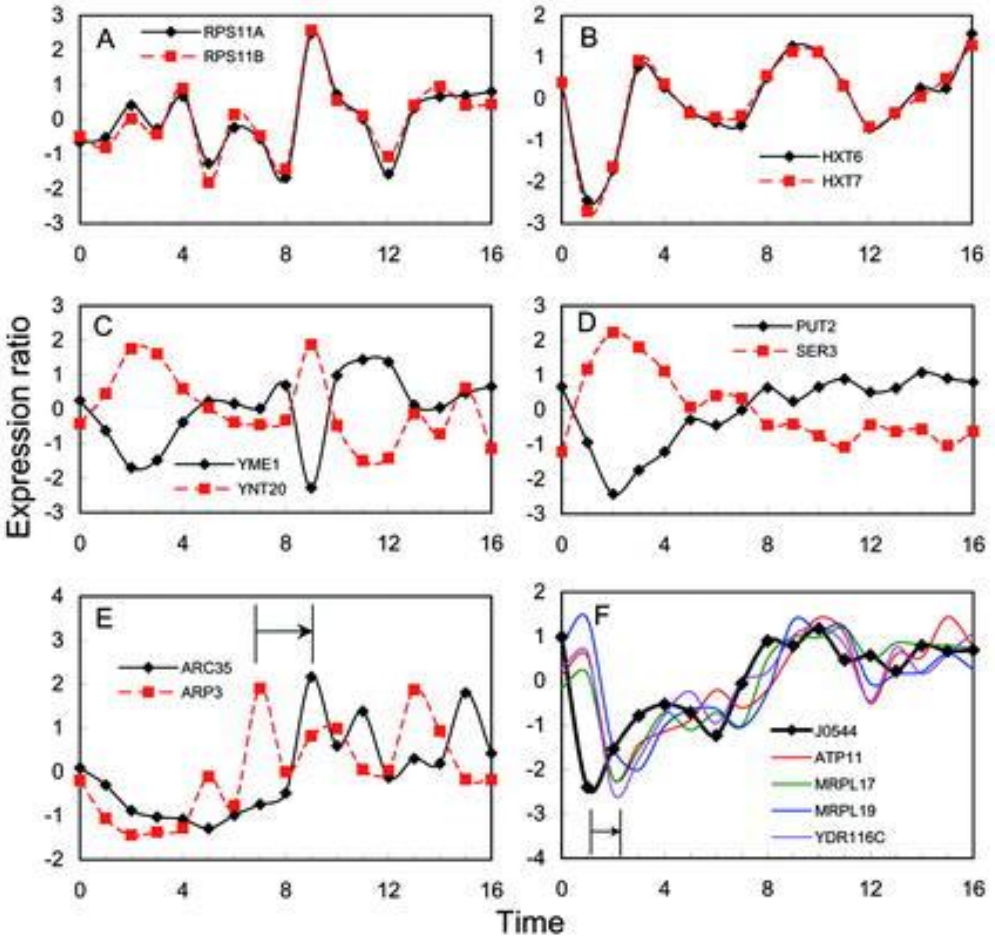
Qian *et al* Journal of Molecular Biology 2001

Results

Simultaneous expression profile relationships:

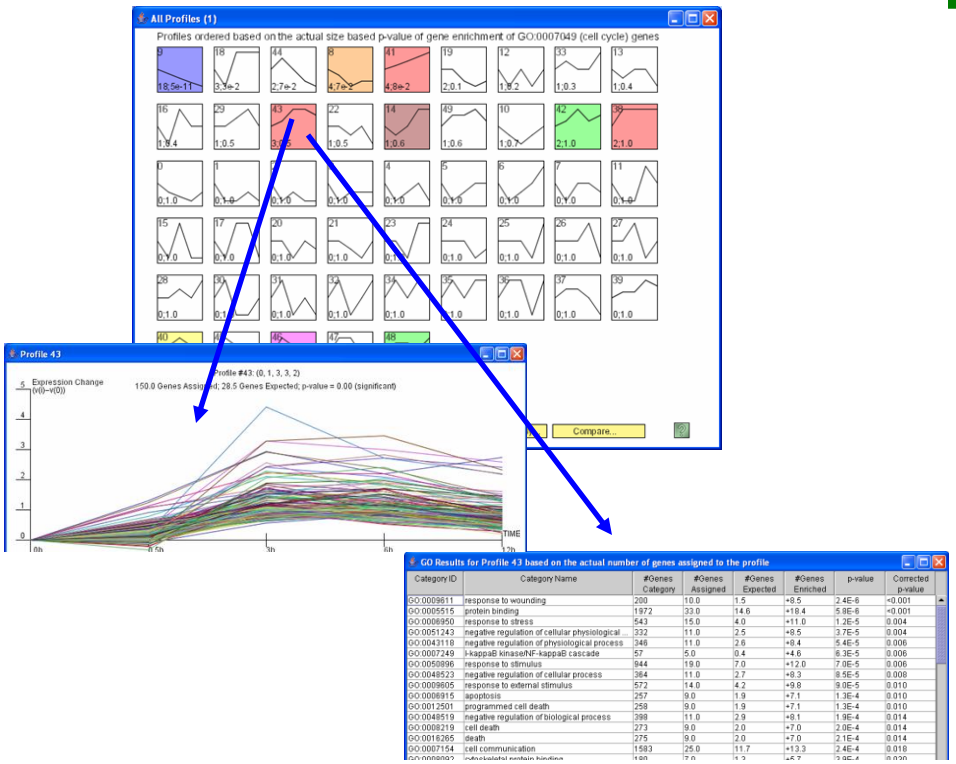
Inverted expression profile relationships:

Time delayed expression profile relationships



Time series clustering methods

STEM: Clustering time series data.



Optimal leaf ordering