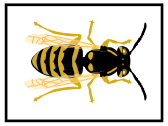
Module 4 (Functional Genomics/FG) recall Lecture 2 *data analysis workflow*



Biological system / state

Transcriptome

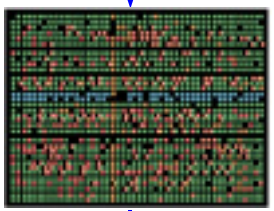


Image Analysis

Data matrix
D = N × M

Gene	P1-1	P3-1	P5-1	P7-1	P10-1
Csrp2	-2.4	74.6	25.5	-30.7	14.6
Mxd3	126.6	180.5	417.4	339.2	227.2
Mxi1	2697.2	1535	2195.6	3681.3	3407.1
Zfp422	458.5	353.3	581.5	520	348
Nmyc1	4130.3	2984.2	3145.5	3895	2134.3
E2f1	1244	1761.5	1503.6	1434.9	487.7
Atoh1	94.9	181.9	268.6	184.5	198
Hmgb2	9737.9	12542.9	14502.8	12797.7	8950.6
Pax2	379.3	584.9	554	438.8	473.9
Tcfap2a	109.8	152.9	349.9	223.2	169.1
Tcfap2b	4544.6	5299.6	2418.1	3429.5	1579.4

Math formulation
Data representation

Map data into metric/measure space, model appropriate to biological question

Normalization
Replicates

Correct for noise, variation arising not from bio-relevant transcriptome program

Un/supervised math techniques. E.g., clustering, networks, graphs, myriad computational techniques guided by overiding scientific question !

Uncover regularities / dominant variance structures in data

Chance modeled by null hypothesis
Statistics
Permutation analyses

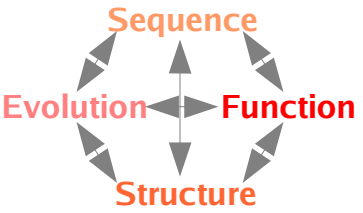
Likelihood of regularities arising from chance alone

Prediction. Inferential statistic.
Minimizing an energy functional
Correlation vs causality
Figure of Merit

Do regularities reflect biological system – state?

Analysis / Modeling

Big Picture



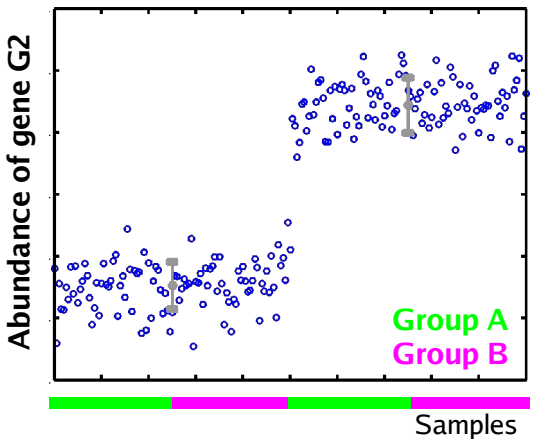
FG: recall Lectures 1,2 two archetypal FG questions

- What *function* does a given molecule X have in a specific biological system / state?
  - *Function* at its most basic refers to – microscopic chemical (inter)actions of X with other (un / known) molecules. And processes downstream of X which might eventually snowball / scale (in time, space) into a phenotypic / macroscopic event.
  - Mapping from sequence to function.
- Which molecules (interactions) characterize / modulate a given biological system / state?
  - *Regularities* (if they exist at all) in transcriptome data may be molecular reflections of the system state.
  - Mapping from regularities to system state.

Sequence / Gene → **Interactions between genes** → **Regularities** → **Function** → **System state**

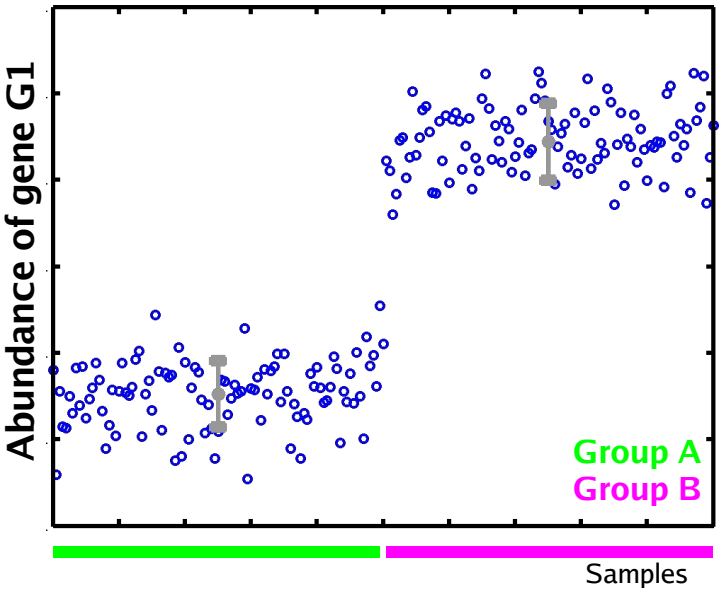
FG: recall Lecture 2 *regularities in data*

- *Regularities* refer to dominant variance structures or coherent geometric structures intrinsic to data with respect to a particular measure / metric space.
- An observed pattern may be regarded a regularity if the pattern can be correlated to *a priori* scientific knowledge. Caveat: bias (*supervised* analysis to additionally test for bias)
  - Eg. in a 2-group comparison study, k of N genes were found to be differentially expressed between groups. "Step" function pattern for each gene is a regularity.
- Statistical likelihood of obtaining regularities given the data distribution
  - *A priori* knowledge/assumptions of underlying distributions to form relevant null hypotheses. Internal correlations and structural assumptions to reduce theoretical degree of freedom – modifying null hypothesis
  - Multiple testing. Bonferroni-type corrections: (type 1 error / false +) $\alpha \rightarrow \alpha/(\#$ of times test applied)
- Correspondence of regularities to biologically relevant programs
  - Eg. in 2-group study, step pattern reflects biologic difference?



Yes?

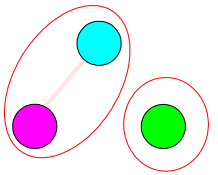
No?



Module 4 FG Lecture 3 outline: Modeling / conceptualizing implicit regularities / functional relationships in transcriptome datasets

- Today's viewpoint is genes in sample space. **Objective**: To induce / impose structure or granularity on the dataset D = N genes × M samples in a principled way in keeping with a biological question

- Functional / ontogic similarities and clustering configurations.



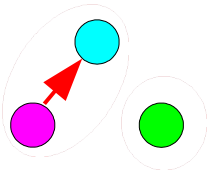
Some structure

Cluster

"Correlation"

  - What is a cluster? Clustering dogma
  - Survey canonical clustering principles. (Non-)Hierarchical, (Non-)Parametric, Global / Local cost criteria, Agglomerative / Divisive.

- Infering networks of biomolecular interactions
  - Biomolecular interactions as (time in/dependent) dynamical systems. Discrete (eg. Binary / Boolean) / Continuous. Deterministic / Stochastic. Asymptotic behaviour.
  - 3 properties – Feedback, Redundancy, Modularity
  - Forward / Backward modeling



More structure

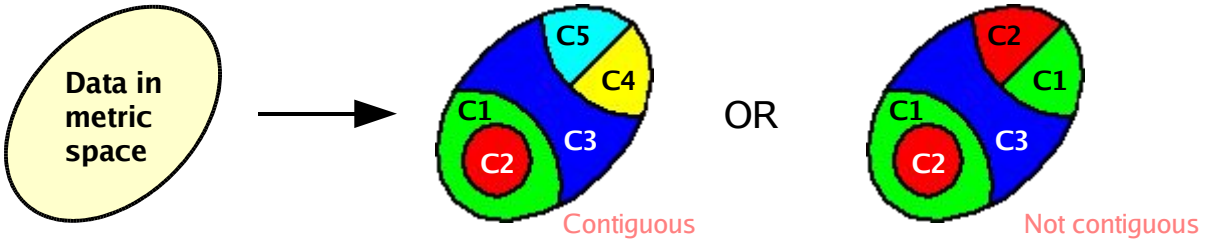
Network

"Causality"

- Figure/s of merit in clustering and modeling interactions
  - Correlation vs. causality. Receiver operating characteristic (ROC) analysis.
  - Lessons from metabolic network models, flux balance analysis

FG: Functional / ontologic similarities and clustering configurations

- **View**: Genes in a (sample-wise) metric space. **Goal**: Partition the genes in sample space into disjoint subsets (*clusters*, usually optimizing some non-negative global / local cost criterion). This procedure is called *clustering*.



  - **Why do this**? To obtain mesoscopic / macroscopic scale *summary structure* for data (reductionist). Clusters *might* reflect distinct bio relevant system programs / dynamics.
  - Clusters are one type of regularity.

- Basic assumptions
  - At least one data partition exists
  - Cost criteria encodes *a priori* assumptions about cluster structure (technical)
  - Sometimes, cost criteria encodes *a priori* knowledge of program characteristics underlying the biological system.

- What is a *"cluster"*? From Jain & Dubes, *Algorithms for Clustering Data* 1988,

"There is no single best criterion for obtaining a partition because no precise and workable definition of "clusters" exists. Clusters can be of any arbitrary shapes and sizes in a multidimensional pattern space. Each clustering criterion imposes a certain structure on the data, and if the data happens to conform to the requirements of a particular criterion, the true clusters are recovered"

FG: Functional / ontologic similarities and clustering configurations

- Clustering dogma: **Co-cluster** (to be in a common cluster) ⇔ **Co-regulated** ⇔ **Co-functional**
  - Exceptions: time lag exists between effector and affected, many correlated processes have no common higher level ontology. Aspect of more fundamental problem of *correlation* vs. *causality* (later)
- *Different* metric space + clustering algorithm combinations → *different* partitions of a dataset.
  - Each combination emphasizes a *different* regularity in the dataset. Cost criterion often implicit in metric space / clustering algorithm.
  - Partitional clustering = assumes a pre-set # of "original" clusters.
  - Parametric = assumes some knowledge of cluster structure encoded in a global criterion eg. cluster is radially symmetric.
  - Non-parametric = no prior knowledge of cluster structure, use a local criterion to build clusters via local data structure, eg. Regions of high gene density in sample space)
- Gross taxonomy of clustering algorithms
  - Sequential
  - Spectral (more a transformation than a clustering algorithm)
  - Hierarchical
  - Optimizing a cost criterion (includes graph-based)
- Sequential clustering. Pick initial singleton (1st seed) at random. Find objects most similar (< a similarity threshold) to initial object, then consecutive objects most similar to those iteratively.Otherwise, pick singletons not previously picked (nor sequentially dependent to previous seeds) as new seed. Sequentially dependent objects form a cluster. Sensitive to choice of 1st seed.

FG: Functional / ontologic similarities and clustering configurations

- Spectral Clustering. S= N $\times$N pairwise similarity matrix (eg. S = covariance matrix). D = diagonal with diagonal entries of S. L = $D^{-½}$ $SD^{½}$. X = column of *k* eigenvectors of L normalized to length 1. Each row of X is in $\mathbb{R}^k$. Form *k* clusters via any other canonical clustering algorithm (cost optimization, hierarchical).

Figures removed due to copyright considerations.

Figures from Ng, Andrew Y., Michael Jordan, and Yair Weiss. "On Spectral Clustering: Analysis and an algorithm." *NIPS* 14 (2002). Courtesy of Andrew Y. Ng, Michael Jordan, and Yair Weiss.

FG: Functional / ontologic similarities and clustering configurations

- Hierarchical clustering. A hierarchy of clustered objects. No prior assumption about # of clusters
  - Agglomerative (bottom up). Start with singleton clusters. @ each iteration, fuse cluster pairs that are most similar to form one (single, complete or average linkage*)
  - Divisive (top down). Start with 1 cluster of all objects. At each iteration, divide each parent cluster into 2 most dissimilar subsets within the parent. Computationally more expensive than agglomerative.
  - *Linkage = deciding on a vector of features to represent a branch / cluster.
- Hierarchical clustering limitations. N^2 computation time. Unstable, >1 different trees from 1 run.
- Illustration of hierarchical clustering (recall iconic example 1 Lecture 1 Alizadeh et al , Nature 2000)
  - Correlation coefficient as measure of similarity for clustering samples (samples in gene space). Agglomerative hierarchical clustering. Linkage not specified, I guess average.

Figure removed due to copyright reasons. Please see:
Alizadeh, et al. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling."
Nature 403, 6769 (Feb 3, 2000): 503-11.

FG: Functional / ontologic similarities and clustering configurations

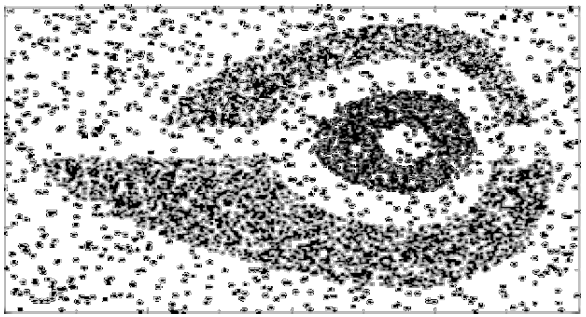
- Clustering that optimize a global / local cost criterion. Partitional = Cluster N objects into K (<N).

  Algebraically, # clusters ≤ (matrix) rank of D!!

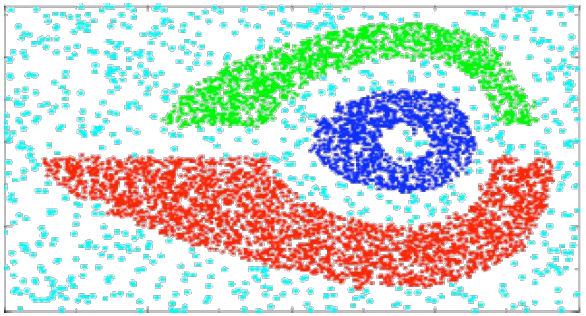
- Example clustering algorithms that optimize a cost criterion
  - k-means. Widely used benchmark to compare clustering algorithms. Start with k centroids scattered at "random" in the sample space. Each gene is assigned to its closest centroid. Recalculate each centroid based on genes assigned to it. Iterate until "distance" between consecutive generations of centroids fall below a threshold.
  - Self-organizing maps (SOM). Constrained version of k-means. Centroids linked together on a grid topology. Each iteration, pick a gene at random. This gene attracts nearest centroid and some this centroid's neighbours. Eventually random gene will only attract nearrest centroid. Same stop criterion as k-means.
  - Expectation-Maximization (EM). Like k-means. Fit a mixture of Gaussians in sample space. Each gene is assigned to each Gaussian with a different probability.
  - Graph-based. Too vast to cover here eg. minimal spanning trees.
  - Ferromagnetic (Ising Potts). From statistical mechanics. Optimize a local cost / energy criterion (local interaction potential, Hamiltonian). Start with K different spins. Spin flip (not conserved) / exchange (conserved). Like spins tend to congregate together. Macroscopic scale single phase regions and clear interface emerge at low temperature after iterations (Metropolis algorithm).
- Some applications of transcriptome clustering
  - Molecular signature for a biological system / state, eg. specific tissue at particular time
  - Functional inference – co-cluster / expressed ⇔ co-functional
  - Regulatory inference – co-cluster / expressed ⇔ co-regulated

FG: Functional / ontologic similarities and clustering configurations

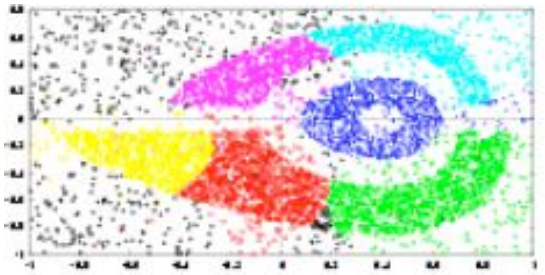
- Recall example from Leonid's intro lecture,



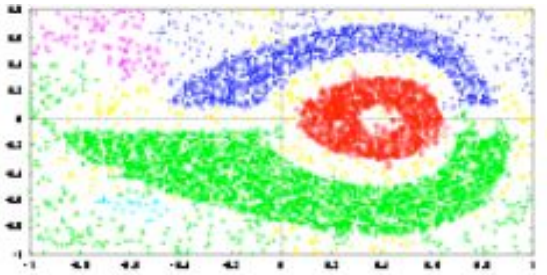
Given toy pattern



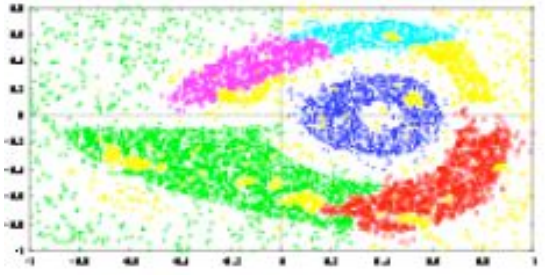
Superparamagnetic clustering (Blatt, Wiseman & Domany, 1996)



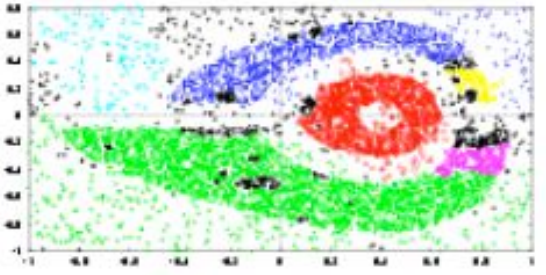
Valley seeking (Fukunaga)



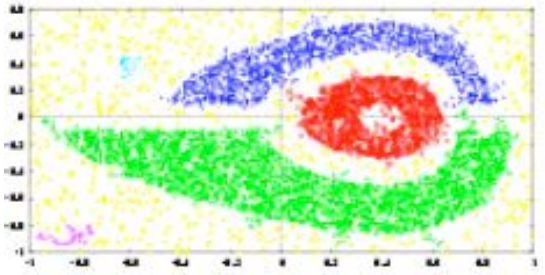
Minimal spanning tree (Zhan)



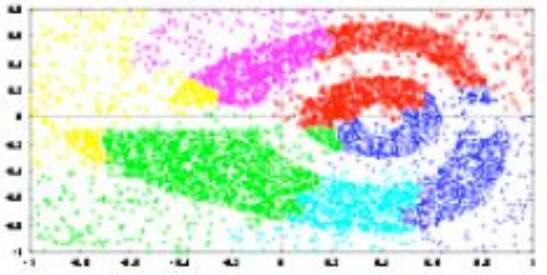
K shared neighbors (Jarvis)



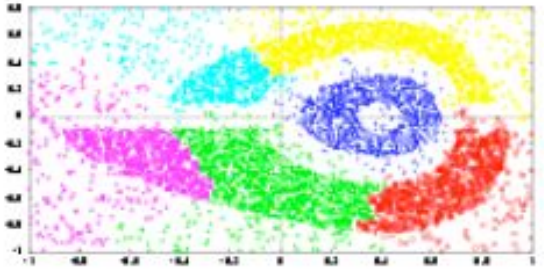
Mutual neighborhood (Gowda)



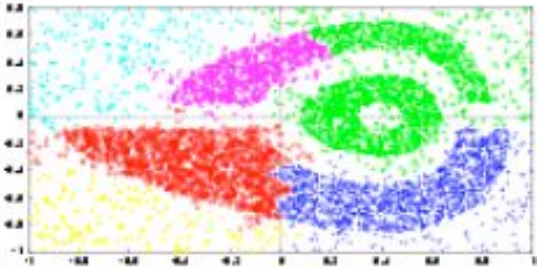
Single linkage method (Milligan,Dekel)



Complete linkage method (Milligan)



minimum variance (Ward)



arithmetic averages (Sokal)

Figures derived from: Blatt, M., S. Wiseman, and E. Domany. "Superparamagnetic Clustering of Data." *Phys Rev Lett* 76, no. 18 (1996): 3251.

FG: Infering networks of biomolecular interactions

- 3 prototypical questions
  - Given measurements of N genes under M conditions of a biological system, (how accurately) can one infer inter-gene regulations / interactions?
  - How will remainder gene profiles (thus system state) change when one gene is perturbed?
  - What are the macroscopic states of this system? ... the asymptotic behavior of this system (when M is time)?
- Recall, iconic example 2 Arkin et al. Science, 1997. Time (lag) is vital component for decoupling, infering molecular interactions.
- Reality check
  - First, characterize each gene by its M-sample profile
  - Then use this characteristic representation to infer relationship between genes (like clustering)
  - Algebraically, # algebraically independent gene profiles ≤ # samples ... in fact this is the kindergarten version of spectral clustering.

FG: Infering networks of biomolecular interactions

- Biomolecular interactions as (time in/dependent) dynamical systems.
  - Discrete (eg. binary / boolean) – system of difference equations or logic table (boolean)
  - Continuous – system of ordinary differential equations
  - Deterministic / Stochastic

Figure removed due to copyright reasons. Please see figure 1 in: D'haeseleer, P., S. Liang, and R. Somogyi. "Genetic network inference: from co-expression clustering to reverse engineering." *Bioinformatics* 16, no. 8 (Aug, 2000): 707-26.

  - Asymptotic behaviour, d/dt X = 0, X is system state characterized by transcriptome

    Deterministic binary dynamical system necessarily cyclic

  - D'haeseleer, Liang & Somogyi, *From co-expression clustering to reverse engineering*, Bioinformatics 2000

- Properties (believed to be) inherent to biological networks
  - Feedback, Redundancy, Modularity
- Forward / Backward modeling
  - Time is important input factor as time (lag) will be used to infer direction (arrow) of causality
  - Forward: Given a dynamical system representation, evolve in time and observe / compute macroscopic states, asymptotic behaviour
  - Backward (Reverse) engineering: Recall Arkin example.

FG: Figures of merit in modeling interactions

- >1 models for 1 physical system. How to pick?
  - Correlation vs. causality. As noted before, time (lags) will be used to infer direction of causality.
  - Receiver operating characteristic (ROC) analysis
  - Lessons from metabolic network models (flux balance analysis). Metabolic processes are quite well understood (ie. stochiometric parameters between metabolic agents are characterized)) – a natural test bed for modeling (Edwards & Palsson, Biotechnology & Bioengineering, 1998)