Computational Molecular Biology and Bioinformatics

Expression Analysis

Malay Bhattacharyya

Associate Professor

Machine Intelligence Unit
Indian Statistical Institute, Kolkata

September, 2024

Microarrays

The microarray technologies (e.g., spotted arrays on glass, in-situ synthesized arrays, and self assembled arrays, etc.) assist in profiling the expressions of thousands of biomoleculaes simultaneously.

Next Generation Sequencing

The Next Generation Sequencing (NGS) technologies (e.g., Roche 454, ABI SOLiD, Illumina, etc.) provide opportunities to detect expressions of biomolecules in a wide range of settings.

Basics

Co-expression, differential expression, differential co-expression and co-expression dynamics are some popular models to reflect local patterns of gene expressions.

		# Phenotypes	
		Single	**Multiple**
Pattern for # Genes	One	-	Differential expression
	Two	Co-expression	Differential co-expression
	Three	-	Co-expression dynamics

Co-expression

The following measures are used for computing co-expression between a pair of expression vectors $(\mathbf{E}_i, \mathbf{E}_j)$ of genes.

Name	Measure	Type
Cosine	$(\mathbf{E}_i \bullet \mathbf{E}_j)/(\|\|\mathbf{E}_i\|\|\|\|\mathbf{E}_j\|\|)$	Similarity
Pearson correlation coefficient	$Cov(\mathbf{E}_i, \mathbf{E}_j)/(\sigma_{\mathbf{E}_i}\sigma_{\mathbf{E}_j})$	Similarity
Spearman's rank correlation	$\rho(Ranked(\mathbf{E}_i), Ranked(\mathbf{E}_j))$	Similarity
Root mean square	$\frac{1}{n}\sqrt{\|\|\mathbf{E}_i - \mathbf{E}_j\|\|^2}$	Distance
Minkowski	$\sqrt[p]{\|\|\mathbf{E}_i - \mathbf{E}_j\|\|^p}$	Distance
Squared Euclidean	$\|\|\mathbf{E}_i - \mathbf{E}_j\|\|^2$	Distance
City block/Manhattan	$\|\mathbf{E}_i - \mathbf{E}_j\|$	Distance
Chebyshev	$\max_t(\|\mathbf{E}_i(t) - \mathbf{E}_j(t)\|)$	Distance
Kullback-Leibler	$\sum_{t=1}^{n} e_j(t) \ln \frac{e_j(t)}{e_i(t)}$	Distance

Note: Pearson correlation coefficient is a better choice to quantify co-expression because - it is normalized and it can reflect both positive/negative dependence.

Differential expression

Fold change is a measure of quantifying differential pattern.

Student's paired t-test can be used as a measure of differential expression where the test statistic is considered as follows

$$T_i = \frac{\mu_{i1} - \mu_{i0}}{\sqrt{\frac{\sigma_{i1}^2}{n_1} + \frac{\sigma_{i0}^2}{n_0}}}, \tag{1}$$

where the expression vector corresponding to the gene i can be divided into the parts $i0$ and $i1$.

Significance analysis of microarrays (SAM) is another popular measure of differential expression. It calculates a test statistic for relative difference in gene expression based on permutation analysis of expression data and calculates a false discovery rate.

Differential co-expression

It can be mainly of two types – gap/subsitution and on/off. Let the expression vectors $E1 = [E1_0\ E1_1]$ and $E2 = [E2_0\ E2_1]$ be divided into two portions, then we can quantify these two types as follows.

- **Gap/substitution case:**

$$S(E1, E2) = |\rho(E1_0, E2_0) + \rho(E1_1, E2_1) - \alpha\rho(E1, E2)|,$$

where $\rho(E1_0, E2_0)$ and $\rho(E1_1, E2_1)$ are the class-conditional dependence measures computed with Pearson correlation coefficient and $\rho(E1, E2)$ is the overall correlation.

- **On/off case:**

$$S(E1, E2) = |\rho(E1_1, E2_1) - \rho(E1_0, E2_0)|,$$

where $\rho(E1_0, E2_0)$ and $\rho(E1_1, E2_1)$ are the class-conditional dependence measures computed with Spearman's rank correlation.

Co-expression dynamics

Suppose, three expression vectors $E1$, $E2$ and $E3$ of length n are given. Then the co-expression dynamics of $E1$ with respect to the pairs $E2$ and $E3$ can be measured as follows

$$S(E1|E2, E3) = \rho(E1, \rho(E2, E3) \cdot \mathbf{1}_n),$$

where $\mathbf{1}_n$ denotes a unit vector of length n.

Basics of clustering

Clustering is a process of grouping similar objects and segregating the dissimilar ones. We can cluster genes from its expression profiles or other features.

Note: Clustering is often termed as unsupervised classification.

k-means clustering

Input: A set of d-dimensional vectors $\mathcal{V} = v_1, v_2, \ldots, v_n$ and the number of clusters k.

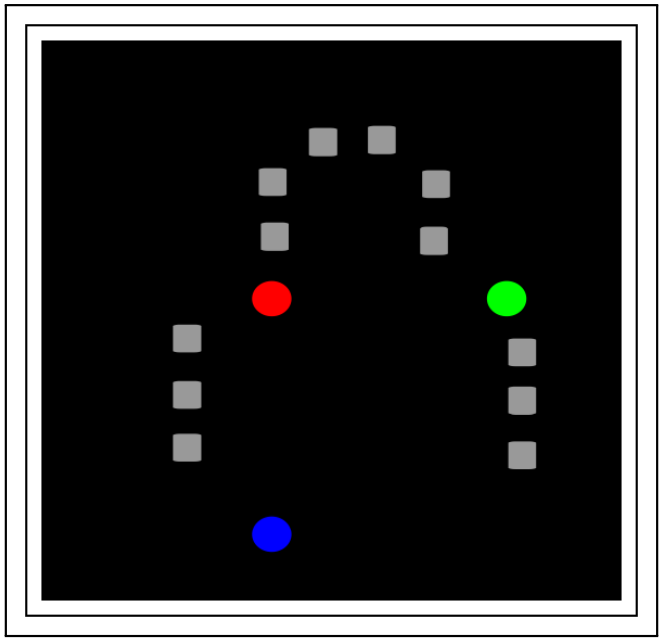
Output: The cluster centers $\mu_1, \mu_2, \ldots, \mu_k$.

1: Randomly initialize the clusters centers $\mu_1, \mu_2, \ldots, \mu_k$
2: **repeat**
3: Classify the n number of d-dimensional vectors according to nearest μ_i
4: Recompute μ_i
5: **until** No change in μ_i
6: Return $\mu_1, \mu_2, \ldots, \mu_k$

Note: The time complexity of k-means algorithm is $O(ndki)$ where n is the number of d-dimensional vectors, k is the number of clusters and i is the number of iterations needed until convergence.

k-means clustering

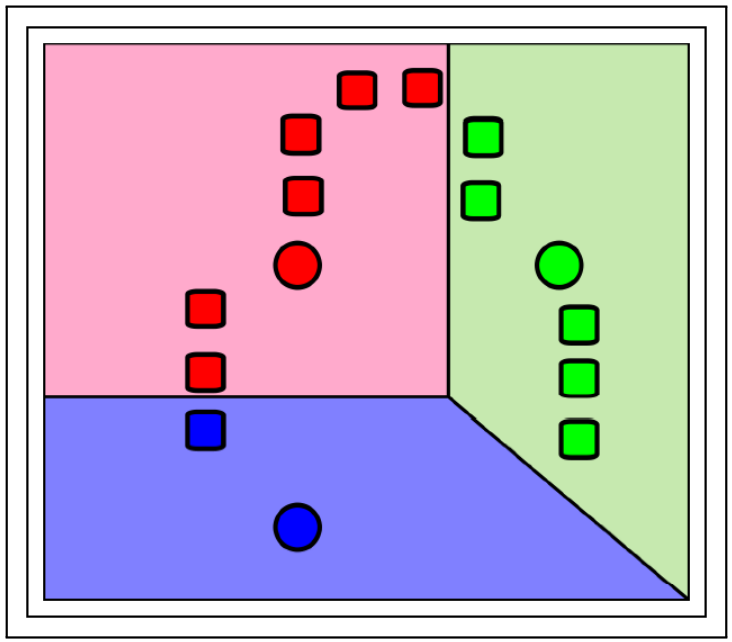
Step 1:



The k initial means (here $k = 3$) are randomly generated within the data domain (shown in color)

k-means clustering

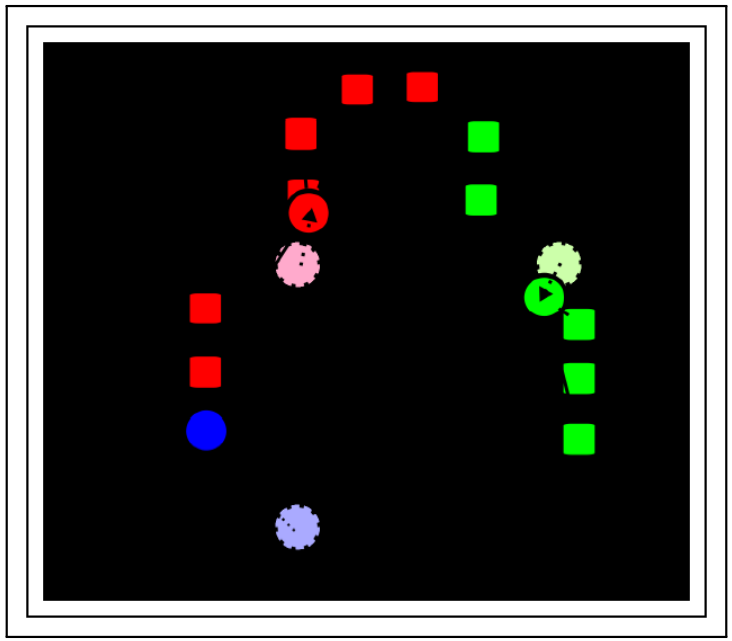
Step 2:



The k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means

k-means clustering

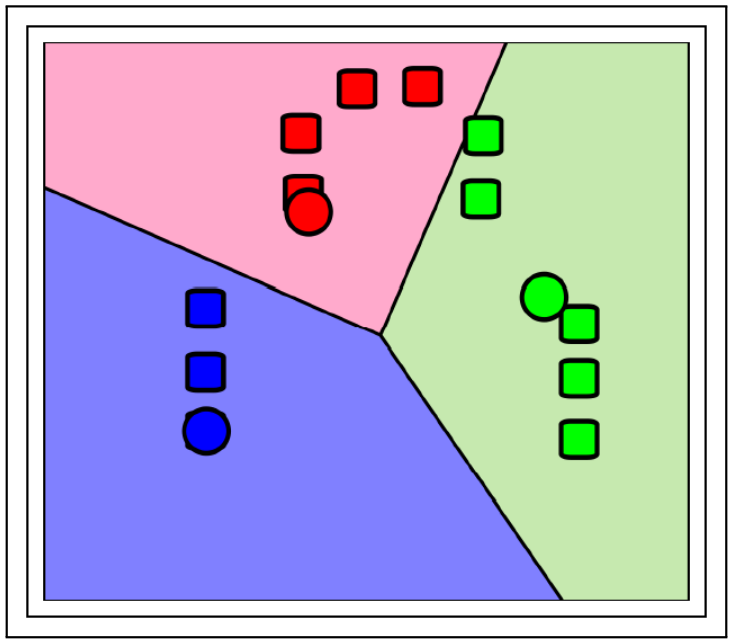
Step 3:



The centroid of each of the k clusters becomes the new mean

k-means clustering

Step 4:



Steps 2 and 3 are repeated until convergence has been reached

k-means clustering

Criterion function for the k-means clustering can be of the following types:

- **Sum-of-squared error:** $\sum_{i=1}^{k} \sum_{x \in d_i} ||x - \mu_i||^2$
- **Related minimum variance:** $\frac{1}{2} \sum_{i=1}^{k} n_i \bar{s}_i$, where $\bar{s}_i = \frac{1}{n^2} \sum_{x \in d_i} \sum_{x' \in d_i} ||x - x'||^2$
- **Scattering property:** $S_W + S_B$, where S_W and S_B denote within-cluster scatter matrix and the between-cluster scatter matrix, respectively

Note: The disadvantages of k-means algorithm are (i) choosing the value of k is difficult, (ii) It does not work well with global clusters, (iii) different initial partitions can result in different final clusters, and (iv) it does not work well with clusters of different sizes and different densities.

Hierarchical clustering

Hierarchical clustering seeks to build a hierarchy of clusters. The strategies for hierarchical clustering generally fall into the following two types:

- **Agglomerative:** This is a bottom-up approach where each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a top-down approach where all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

The linkage criterion is a function of the pairwise distances that determines the distance between two sets of samples (say A and B) to be combined. The alternatives are (i) single linkage ($\min_{a \in A, b \in B} d(a, b)$), (ii) complete linkage ($\max_{a \in A, b \in B} d(a, b)$), and (iii) average linkage ($\frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$).

Hierarchical clustering

Let us illustrate this step by step with an example.

Step 1:

samples	A	B	C	D	E	F	G
A	0	0.5000	0.4286	1.0000	0.2500	0.6250	0.3750
B	0.5000	0	0.7143	0.8333	0.6667	0.2000	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.6667	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8000	0.8571
E	0.2500	0.6667	0.4286	1.0000	0	0.7778	0.3750
F	0.6250	0.2000	0.6667	0.8000	0.7778	0	0.7500
G	0.3750	0.7778	0.3333	0.8571	0.3750	0.7500	0

Hierarchical clustering

Step 2:

samples	A	(B,F)	C	D	E	G
A	0	0.6250	0.4286	1.0000	0.2500	0.3750
(B,F)	0.6250	0	0.7143	0.8333	0.7778	0.7778
C	0.4286	0.7143	0	1.0000	0.4286	0.3333
D	1.0000	0.8333	1.0000	0	1.0000	0.8571
E	0.2500	0.7778	0.4286	1.0000	0	0.3750
G	0.3750	0.7778	0.3333	0.8571	0.3750	0

Hierarchical clustering

Step 3:

samples	(A,E)	(B,F)	C	D	G
(A,E)	0	0.7778	0.4286	1.0000	0.3750
(B,F)	0.7778	0	0.7143	0.8333	0.7778
C	0.4286	0.7143	0	1.0000	0.3333
D	1.0000	0.8333	1.0000	0	0.8571
G	0.3750	0.7778	0.3333	0.8571	0

Hierarchical clustering

Step 4:

samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0	0.7778	0.4286	1.0000
(B,F)	0.7778	0	0.7778	0.8333
(C,G)	0.4286	0.7778	0	1.0000
D	1.0000	0.8333	1.0000	0

Hierarchical clustering

Step 5:

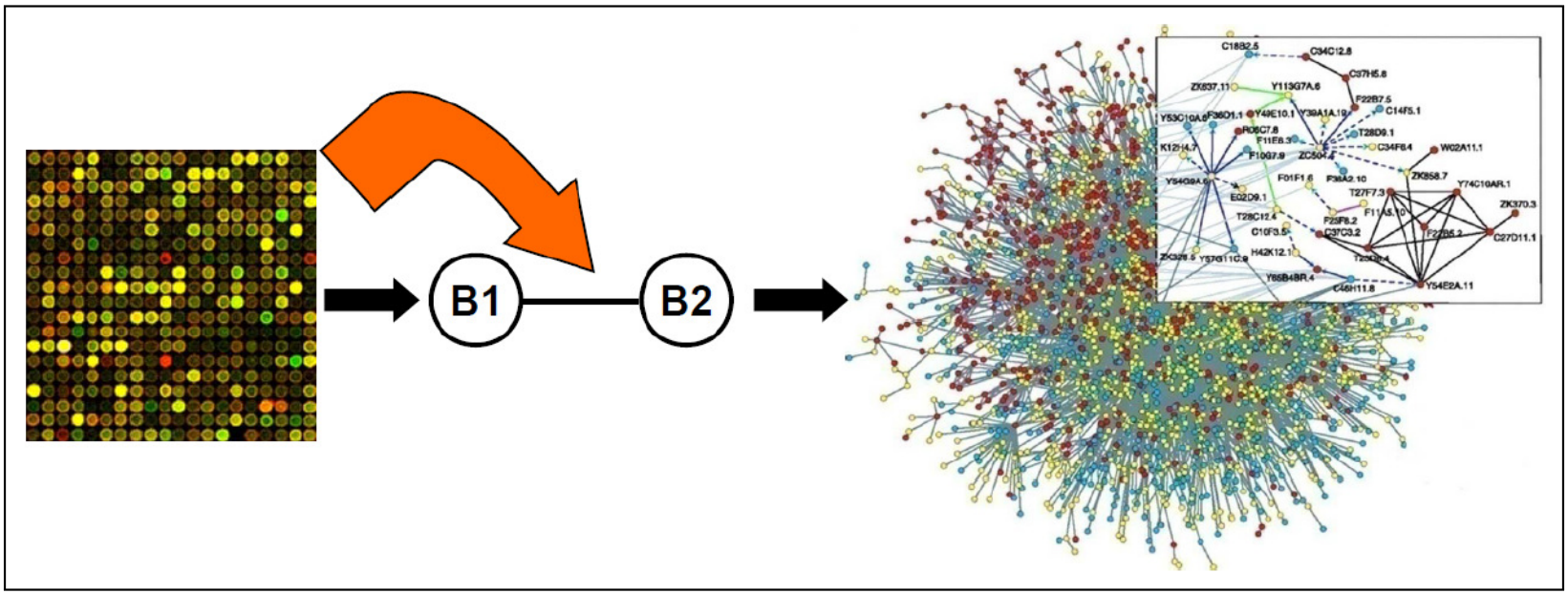
samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0	0.7778	1.0000
(B,F)	0.7778	0	0.8333
D	1.0000	0.8333	0

Step 6:

samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0	1.0000
D	1.0000	0

What is a co-expression network?

Co-expression networks, which provide a global pattern of gene expressions, are built from expression profiles to explore interesting substructures. A co-expression network reflects the global co-expressibility/repressibility between biomolecules. The nodes of a co-expression network are biomolecules and the edges denote the degree of co-expression.



From co-expression to co-functionality and co-regulation

We often infer functional coherence from the modules of a co-expression network. But that may not be always true.

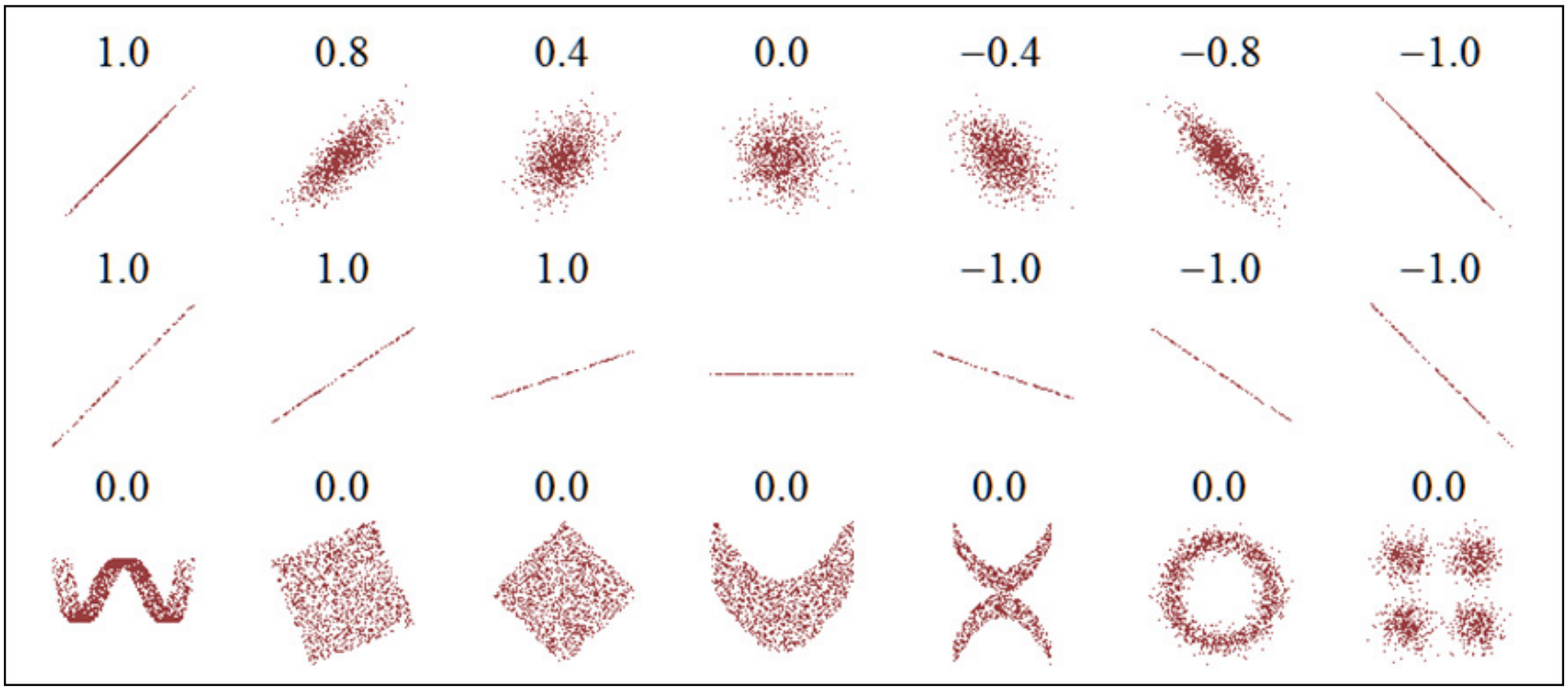
- **Weak inference:** Co-expressed genes are co-functional
- **Strong inference:** Co-expressed proteins are co-functional

Co-expressed biomolecules are expected to be regulated by common regulators. But this is also not necessarily true.

Co-expression might be a compound effect of co-regulation.

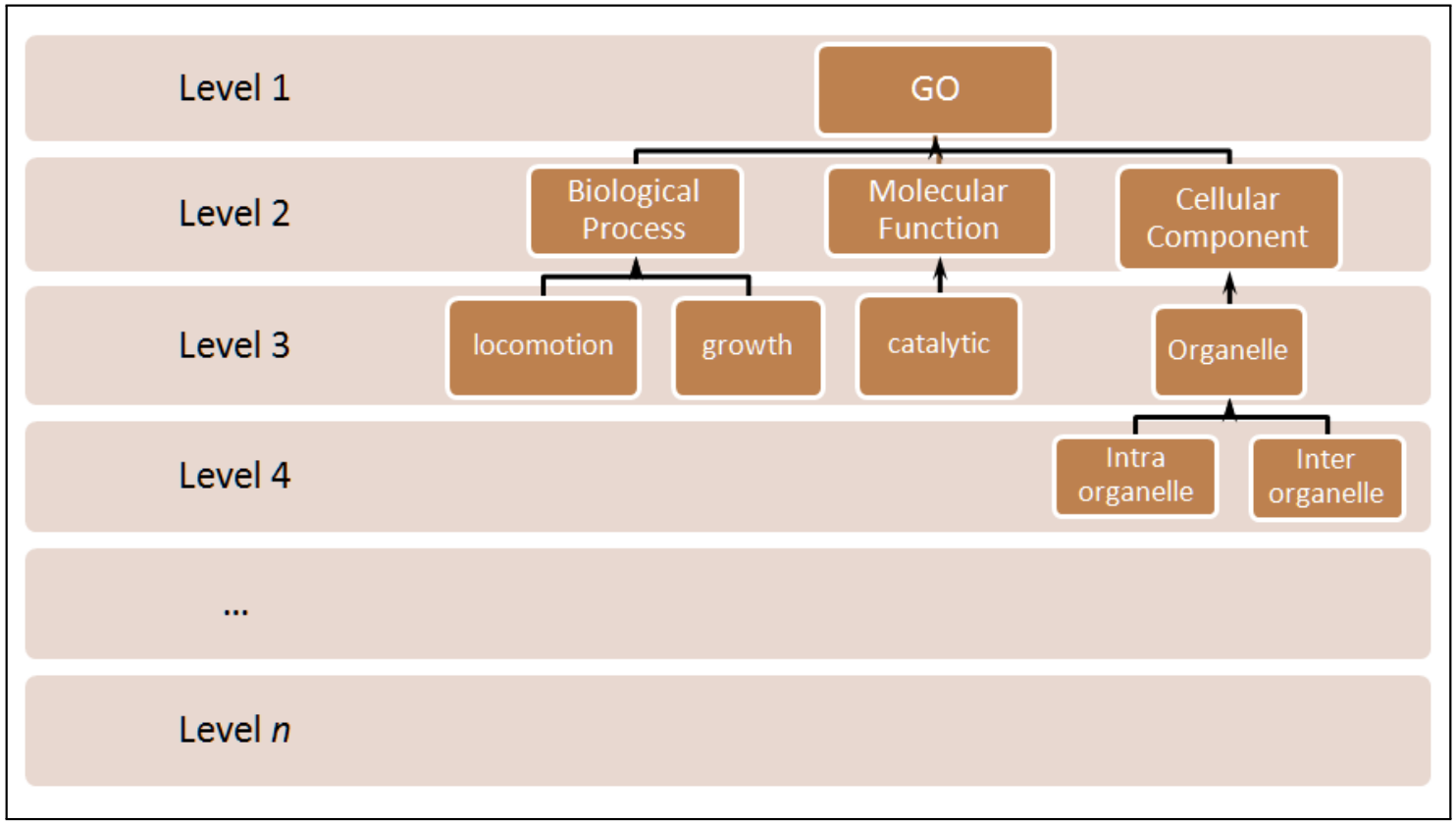
From statistical to biological coherence

Similarity between two observations can be of various types. So, they are statistically modeled in different ways.



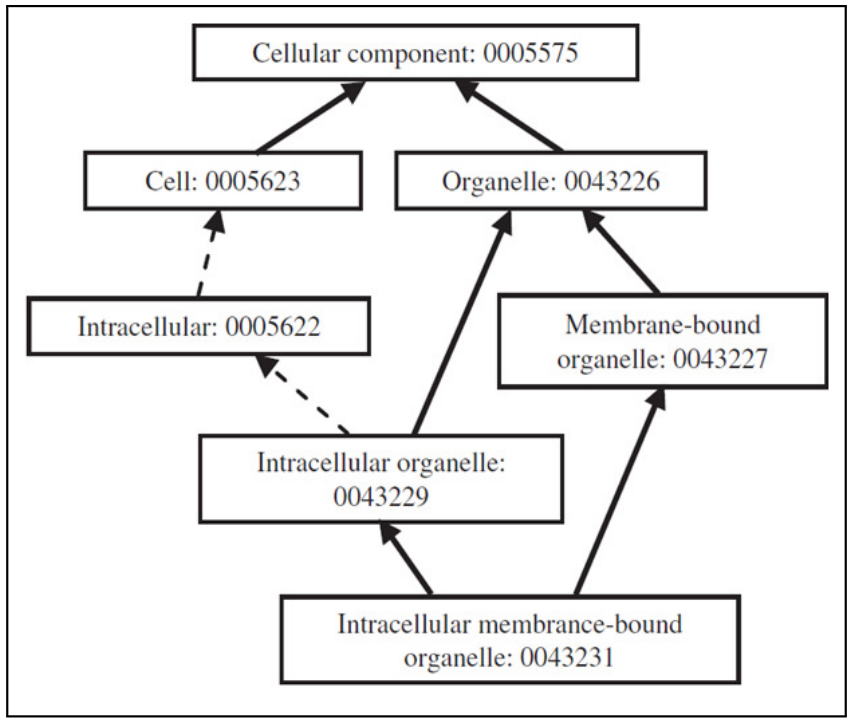
From statistical to biological coherence

Statistical coherence can sometimes interpreted as biological coherence. Gene ontology is a tool that helps to explore biological coherence.



From statistical to biological coherence

A closer view of gene ontology provides much complicated relations as follows.



Relations in GO are of three types – 'is a' (subtype), 'part of' (part-whole) and 'regulates' (have effect on).

Hands-on

1. Open The Cancer Genome Atlas (TCGA) portal and do the following steps.
   i) Access TCGA data from `https://portal.gdc.cancer.gov`.
   ii) Choose a cancer type by the primary site, say "Brain".
   iii) Explore the different tables present and try to download some expression data.