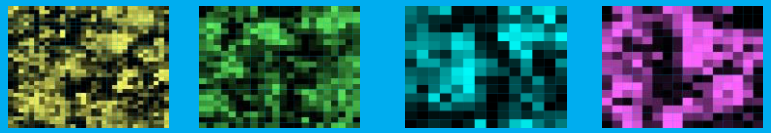
Gene Set Enrichment Analysis (GSEA) Part II



Network Analysis in Systems Biology

Neil Clark, PhD

Postdoctoral Fellow, Ma’ayan Lab

Department of Pharmacology and Systems Therapeutics

Icahn School of Medicine at Mount Sinai, New York, NY 10029

Mount Sinai

A simple example: Ranking the data

- Suppose we have two classes of data, one may be from (A) muscle biopsies of patients not having diabetes, and (B) with diabetes.
- Take the simple data set:

Class A:

Gene	Sample 1	Sample 2
A	1.0	1.0
B	1.2	1.1
C	0.5	0.6
D	0.7	0.4
E	0.2	0.4

Class B:

Gene	Sample 3	Sample 4
A	2.0	1.9
B	0.6	0.7
C	0.1	0.2
D	2.2	2.0
E	0.3	0.2

- Now rank the genes according to descending differential expression across the classes:

	Class A		Class B	
	Sample 1	Sample 2	Sample 3	Sample 4
D	0.7	0.4	2.2	2.0
A	1.0	1.0	2.0	1.9
E	0.2	0.4	0.3	0.2
C	0.5	0.6	0.1	0.2
B	1.2	1.1	0.6	0.7

A simple example: Generating the running sum

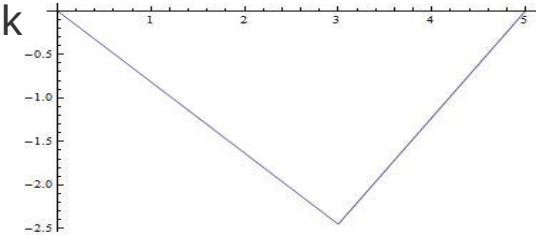
- We would like to test whether the set of genes {B, C}, which we know belongs to a particular biological category, play a significant role in the difference between the two classes of data.
- We next move down the list of genes shown in the previous table and record a running sum. If we encounter a gene which is not in our test set then we add: $-\sqrt{\frac{G}{N-G}}$

  If we encounter a gene which is in our set then we add: $\sqrt{\frac{N-G}{G}}$

  Where N is the number of genes in the microarray, and G is the number of genes in our test set.

A simple example: The supremum of the running sum

- If we calculate this running sum for the ranked data shown in the last table we obtain a walk

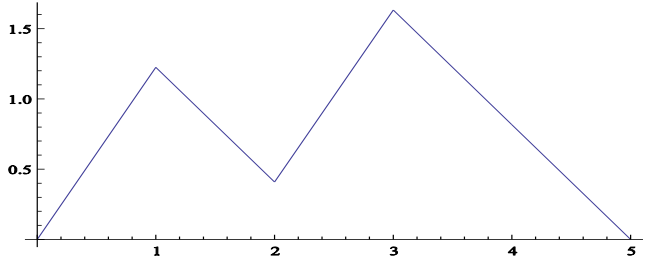


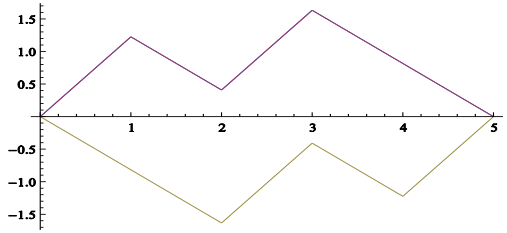
- Note the similarity to the Brownian Bridge.
- The supremum of this walk gives a quantification of the significance of the gene set. If it is significantly positive then the genes are significantly positively differentially expressed.
- If the supremum is significantly negative then the gene set is significantly negatively differentially expressed.
- The significance of the supremum is rated against similar random walks which have been constructed from the data by randomly permuting the class labels.

A simple example: Estimating the significance

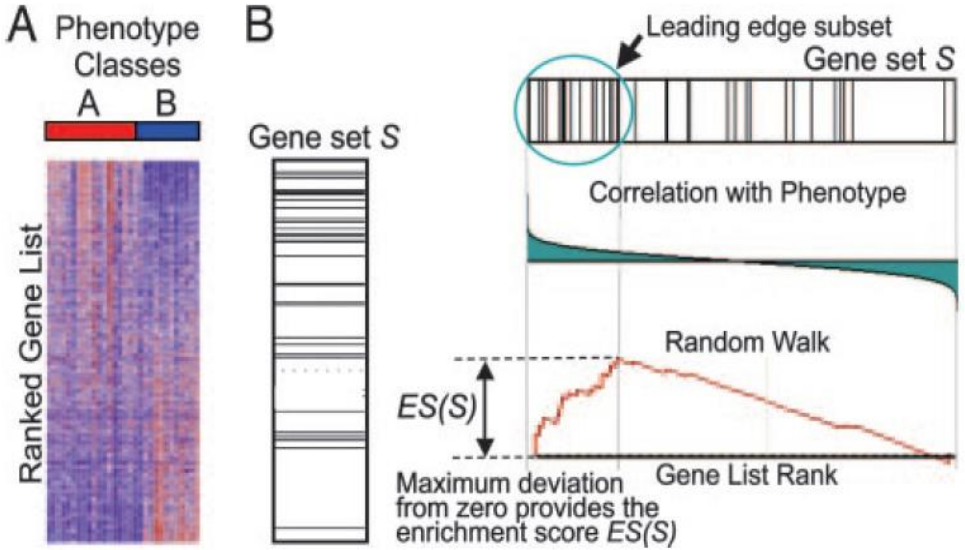
- We permute the class labels of the data – this preserves the correlation structure of the data while randomizing with respect to class
- Repeat the running sum process to produce a new walk.
- Repeat the randomization process many times.
- The number of times that the actual supremum is larger than the random gives an estimate of the significance of the result.
- In the example, the supremum of -2.5 from the data is larger than all the randomized suprema – so we would conclude that the set {B, C} is significant.
- Finally, correct for multiple hypothesis testing

	Class A		Class B	
Gene	Sample 1	Sample 3	Sample 4	Sample 2
A	1.0	2.0	1.9	1.0
B	1.1	0.6	0.7	1.2
C	0.6	0.1	0.2	0.5
D	0.4	2.2	2.0	0.7
E	0.4	0.3	0.2	0.2



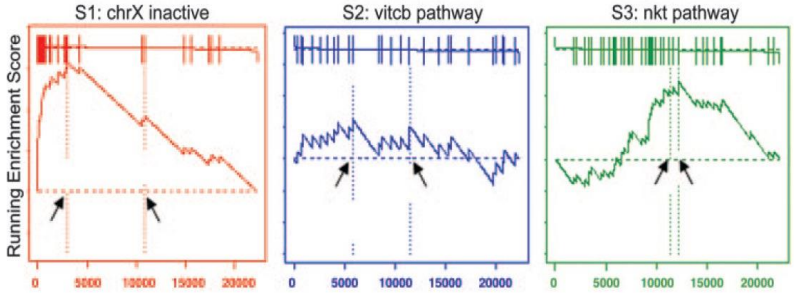


Summary of GSEA



A. Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, **102**, 43 (2005)

An example from the literature



The distribution of three gene sets, from the C2 functional collection, in the list of genes in the male/female lymphoblastoid cell line example ranked by their correlation with gender: S1, a set of chromosome X inactivation genes; S2, a pathway describing vitamin c import into neurons; S3, related to chemokine receptors expressed by T helper cells.

A. Subramanian et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS, **102**, 43 (2005)

An example of GSEA in the literature

- In: “PGC – -responsive genes involved in oxidative phosphorylation are coordinately down-regulated in human diabetes” Mootha, V. K. *et al* 2003 *Nat. Genet*. **34** 267-27 The authors tested expression in muscle samples from 43 age-matched males. 17 had normal glucose tolerance, 8 with impaired glucose tolerance, and 18 had type II diabetes.
- The authors identified a set of genes associated with oxidative phosphorylation to be significant
- Each individual gene in the set was only down-regulated by a small amount, but the down-regulation was well coordinated across the members of the set.
- They could use the biological mechanism of oxidative phosphorylation to investigate the biological mechanism of the disease.

Conclusion

- GSEA is a statistical test which can identify sets of genes, belonging to a particular biological category, which play an important role in distinguishing between two classes of gene expression data.
- The test is particularly sensitive as small changes which are coordinated across the set can be detected.
- The test helps reveal the biological mechanisms responsible for the difference between the two classes because the test set has an *a priori* biological theme.