Comparing module content

We implemented a strategy for the comparison of (possibly overlapping) module compositions, suggested in [1]. Similarity was defined in terms of the symmetrical Jaccard distance measure [2], although the authors claim the results are quite insensitive to the exact choice of the similarity measure.

Jaccard similarity

The Jaccard similarity measure is defined as [2]

$$J = \frac{TP}{TP + FN + FP}. \tag{1}$$

In this equation, TP stands for True Positives, FP for False Positives, and FN for False Negatives, all defined with respect to two binary matrices that represent pairwise occurrences for two module partitions. When focussing on genes for example, a True Positive would indicate a pair of genes (one entry in the gene x gene co-occurrence matrix) that occurs together in at least one module, in both partitions.

Approximating the randomization process

The original implementation in [1] was based on the empirical distribution of Jaccard values, obtained by repeatedly randomizing the module composition. This distribution can then be used to attach a significance score to the Jaccard similarity between results obtained with two parameter settings, either from the same algorithm, or across algorithms. For more details on this procedure, we refer to the original paper [1].

An obvious disadvantage of this strategy is the need for many randomizations in order to get reasonable estimates for the mean and standard deviation on the distribution of the Jaccard coefficients. In the case of sparse matrices, it is unlikely that any TP occur upon randomization, making (almost) all values equal to zero. This makes it practically impossible to work with empirical p-values for such sparse matrices. Therefore, an analytical derivation for the distribution of Jaccard similarities between randomized compositions is useful to reduce computational cost. It also allows to produce smoother sensitivity analysis plots.

Define n as the total number of entries in the co-occurrence matrix For instance, n could be the number of genes squared in a gene x gene matrix. Assume that p_1 and p_2 represent the densities of ones in the binary co-occurrence matrices for partition one and two respectively. Obviously, TP, FP and FN are related, because

$$TP + FP = n\,p_1 \tag{2}$$
$$TP + FN = n\,p_2. \tag{3}$$

Using this to rewrite (1) yields

$$y = \frac{TP}{TP + FN + FP} = \frac{TP}{(TP + FN) + (TP + FP) - TP} = \frac{TP}{n\,(p_1 + p_2) - TP}. \tag{4}$$

The latter equation is of the form $\frac{x}{a-x}$ with x a random variable indicating the probability of observing a specific number of TP.

Let us assume we are looking at two matrices with given densities that are uncorrelated. In this case, the number of TP is a random binomial variable. The probability of observing exactly k TP is given by $Bin(p_1\,p_2, n)$, the probability of having k successes in n trials with success rate $p_1\,p_2$. If the number of TP x is much smaller than a, $ay \approx x$ is binomial too:

$$p_{ay}(s) = Bin(s, n, p_1\,p_2) \tag{5}$$

In the plots of parameter sensitivity and module comparison, we expressed significance as the number of standard deviations from the mean of this distribution.

1. Shakhnovich BE, Reddy TE, Galinsky K, Mellor J, Delisi C: **Comparisons of predicted genetic modules: identification of co-expressed genes through module gene flow.** *Genome Inform Ser Workshop Genome Inform.* 2004, **15(1)**:221-8

2. Jaccard P: **tude comparative de la distribution florale dans une portion des Alpes et des Jura.** *Bull Soc Vaudoise Sci Nat* 1901, **37**:547-579.