Evaluating statistical significance of pathways and network in MetaCore™

Scoring and prioritization of networks/pathways according to the relevance to input data In most cases high-throughput experiments result in lists of genes or proteins of interest. These lists could for example be genes differentially expressed between two conditions, or proteins identified in a sample. The datasets usually contain anywhere between few dozens and few thousand genes/proteins. It his paper we address the issue of how different networks and pathway modules in MetaCore can be prioritized based on their statistical significance with respect to such experimental datasets. Significance is evaluated based on the size of the intersection between user's dataset and set of genes/proteins corresponding to a network module/pathway in question. This problem can be cast as selection without replacement and the probability to randomly obtain intersection of certain size between user's set and a network/pathway follows hypergeometric distribution. Let us consider a set of size ***N***, representing all nodes in MetaCore database of interactions. When considering user's set of genes (***I***) a subset ***R*** of these nodes would become "marked" because they correspond to user's data. There two important things that have to be noted. First is that generally not all the genes from I could be associated with nodes in MetaCore network of interactions. Second, Some genes/proteins may correspond to multiple nodes and some nodes may correspond to multiple genes/proteins. Such is the case with nodes representing protein complexes for example. Next, consider a network module containing ***n*** nodes. This module could either be a pre-built one, like pathway map or be built by one of the algorithms in MetaCore. More generally, ***n*** could be any subset of nodes selected based on a common property: Gene Ontology category, set of nodes related to a certain disease, etc. Invariably there will be some number ***r*** of marked nodes among the ***n*** nodes of the module. The probability of a subset of size ***n*** to include ***r*** marked ones provided that ***n*** and ***R*** are unrelated (null-hypothesis) follows the hypergeometric distribution

$$P(r,n,R,N)=\frac{C_R^r\cdot C_{N-R}^{n-r}}{C_N^n}=\frac{C_n^r\cdot C_{N-n}^{R-r}}{C_N^R}=\frac{R!\cdot(N-R)!}{N!}\cdot\frac{n!\cdot(N-n)!}{r!\cdot(R-r)!}\cdot\frac{1}{(n-r)!\cdot(N-R-n+r)!}$$

The mean of this distribution is equal to:

$$\mu=\sum_{r=0}^{n}r\cdot P(r,n,R,N)=\frac{n\cdot R}{N}=n\cdot q\,,$$

Where $q=R/N$ defines the ratio of marked objects.
The dispersion of this distribution is described as:

$$\sigma^2=\sum_{r=0}^{n}r^2\cdot P(r,n,R,N)-\mu^2=\frac{n\cdot R\cdot(N-n)\cdot(N-R)}{N^2\cdot(N-1)}=n\cdot q\cdot(1-q)\cdot\left(1-\frac{n-1}{N-1}\right).$$

.

It is essential that these equations are invariant in terms of exchange of ***n*** for ***R*** which means that the "subset" and "marked" are equivalent and symmetrical sets. Importantly, in the cases of

$r > n$, $r > R$ or $r < R+n-N$, $P(r,n,R,N) = 0$

We also use the z-score for comparison and prioritization of some of the networks in MetaCore™

$$z-score = \frac{r - n\frac{R}{N}}{\sqrt{n\left(\frac{R}{N}\right)\left(1-\frac{R}{N}\right)\left(1-\frac{n-1}{N-1}\right)}} = \frac{r-\mu}{\sigma}$$

Where μ and σ are the mean and dispersion of the hypergeometric distribution. *Z*-score essentially represents a measure of relative deviation of ***r*** from its expected mean value.

P-value and evaluation of statistical significance of networks In MetaCore user often employs algorithms like “Analyze networks” and similar to build a set of network modules associated with input data (e.g. a list of genes). The resulting networks have to be evaluated for whether algorithm have succeeded in creating modules that have higher than random saturation with the genes of interest. To address this question MetaCore calculates p-values for networks generated by such algorithms. P-value calculations are based on hypergeometric distribution discussed above. MetaCore also uses p-value calculation to evaluate network’s relevance to Gene Ontology biological processes classification. Following discussion in the previous section let us consider set of all nodes in MetaCore database of interactions representing the “Global network” of size ***N***. Respectively a set of nodes corresponding to user’s list is ***R*** and set of nodes in the network under evaluation is ***n***. Again, consider that r nodes in n turned out to be “marked” by association with user’s list. Note that ***R*** and ***r*** could also represent nodes marked by their association with some other list, such as “process” category from Gene Ontology. For the evaluation of statistical significance we consider the null-hypothesis which states that the subsets ***R*** and ***n*** are independent and, therefore, the size of their intersection follows the hypergeometric distribution. The alternative hypothesis states that there is positive correlation between the subsets. Based on these assumptions, we can calculate a p-value as the probability that intersection of two randomly selected sub-sets of N would have having the size of ***n*** or larger:

$$pVal(r,n,R,N) = \sum_{i=\max(r,R+n-N)}^{\min(n,R)} P(i,n,R,N) = \frac{R!\cdot n!\cdot(N-R)!\cdot(N-n)!}{N!} \sum_{i=\max(r,R+n-N)}^{\min(n,R)} \frac{1}{i!\cdot(R-i)!\cdot(n-i)!\cdot(N-R-n+i)!}$$

Options for p-value calculation

In MetaCore there are several option as to what nodes are considered as an entire set from which selections are made. The differences are represented on figure 1.

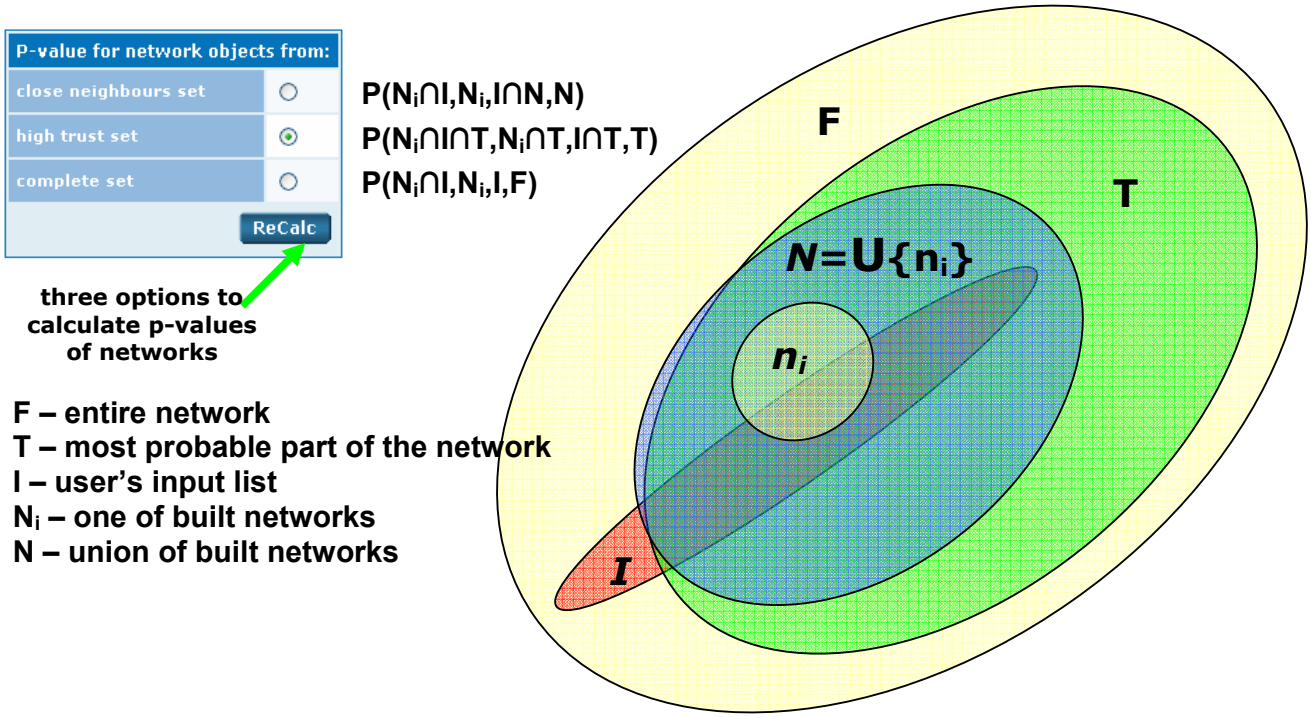

Figure 1 Different options for p-value calculation for network scoring

Definitions:

F is the entire set of network nodes corresponding to genes and protein in MetaCore. This set includes all nodes - those annotated with direct physical interactions ("most probable network") as well as nodes that have only low trust interactions such as influence on expression.
T is the set of network objects that form "the most probable part of the network – network formed by nodes annotated with direct physical interactions
I – network nodes corresponding to user's input list
$\mathbf{N_i}$ – is one of the subsets being evaluated for its significance (for example a Gene Ontology category, or one of the generated networks: index *i* represents the fact that there may be many be such subsets).
N is the union of all the subsets tied to the particular characteristic or category (e.g., objects that can be associated with at least one GO processes or one of the built networks). Note that depending on options selected by user in network building algorithm, **N** may include nodes outside of the "most probable" set.

Options in ANALYZE Network algorithm:

CLOSE NEIGHBORS SET – estimates the significance of the overlap of the input gene list with each of the generated sub-networks, based on the intersection of the input list with the union of ALL generated sub-networks (**N**).

HIGH TRUST SET – estimates the significance of the overlap of the input gene list with each of the generated sub-networks, based on the intersection of the input list with the "most probable" part of the network (**T**).

COMPLETE SET - estimates the significance of the overlap of the input gene list with each of the generated sub-networks, based on the intersection of the input list with the whole network (**F**).

Where it is

1. Distribution histograms (for all files in the Active data window)
    a. GO processes (hierarchy): each process ($\mathbf{N_i}$) is compared to each experiment (**I**)
    b. Diseases (hierarchy): each disease-associated set of genes ($\mathbf{N_i}$) is compared to each experiment (**I**)
    c. GeneGO processes Networks (hierarchy): each GeneGo process network ($\mathbf{N_i}$) is compared to each experiment (**I**)
    d. Diseases (networks): each disease network ($\mathbf{N_i}$) is compared to each experiment (**I**)
    e. Metabolic Pathways (networks): each network ($\mathbf{N_i}$) is compared to each experiment (**I**)
    f. Maps: each map ($\mathbf{N_i}$) is compared to each experiment (**I**)
2. List of most significant (for activated experiments)
    a. GO processes (hierarchy): each process ($\mathbf{N_i}$) is compared to the union of all experiments.
    b. Diseases (hierarchy): each disease ($\mathbf{N_i}$) is compared to the union of all experiments.
    c. Maps: each map ($\mathbf{N_i}$) is compared to the union of all experiments.
3. In the network list (after analyze, etc.):
    a. Each network ($\mathbf{N_i}$) is compared t the original list (I)
    b. Top 5 GO processes for each network: each GO process is compared to each network ($\mathbf{N_i}$)
4. On the network:
    a. Tissues: each tissue (from ABI list) (n_i) is compared to set of nodes on the network
    b. Localizations: each localization (from the original list) (n_i) is compared to set of nodes on the network
    c. GO processes: each process (n_i) is compared to set of nodes on the network
    d. Diseases: each disease (n_i) is compared to set of nodes on the network
5. Compare Experiments:
   Histogram Distributions (same principles)