nxUP - mxUS

Tracks

1. Track UP_i: unmarked points, $i = 1, \ldots, n$
2. Track US_j: unmarked segments, $j = 1, \ldots, m$

Comment:

- This may also be considered as an MP - MS problem. Then MP will consist of the union of the points in UP_i, $i = 1, \ldots, n$, and the mark of a point in the MP segment is i if this point occurs in UP_i. If a point occurs in only one of the UP_i tracks, the mark will be an integer, otherwise the mark will be an integer array. Similarly, MS will consist of the union of the segments in US_j, $j = 1, \ldots, m$. The mark of an MS segment may be a Boolean array with length m indicating whether the segment is part of track US_j or not, $j = 1, \ldots, m$.

Question

For which pairs of tracks (UP_i, US_j) are there more points of track UP_i inside the segments of track US_j than expected by chance?

Comments:

- We assume that segments are fixed.
- For each i, $i = 1, \ldots, n$, we assume the number of points in UP_i are fixed, but that their positions are randomly selected with equal probabilities from the positions in UP, where UP is a new track where the set of points is the union of the points in UP_i, $i = 1, \ldots, n$.
- Significance is determined by means of p-values. Small p-values identify pairs of tracks (UP_i, US_j) where the points in track UP_i is more inside segments of track US_j than expected.

Matrix with counts

Let O_{ij} be the number of points of track UP_i inside the segments of track US_j, $i = 1, \ldots, n$, $j = 1, \ldots, m$. We can make a matrix O of the O_{ij} values as shown in the following table:

Track	US_1	US_2	$\cdots$	US_m	Sum		Track UP
UP_1	O_{11}	O_{12}	$\cdots$	O_{1m}	M_1		N_{UP_1}
UP_2	O_{21}	O_{22}	$\cdots$	O_{2m}	M_2		N_{UP_2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$		$\vdots$
UP_n	O_{n1}	O_{n2}	$\cdots$	O_{nm}	M_n		N_{UP_n}
Sum	N_1	N_2	$\cdots$	N_m	N		N_{UP}

where $M_i = \sum_{j=1}^{m} O_{ij}$, $N_j = \sum_{i=1}^{n} O_{ij}$, and $N = \sum_{j=1}^{m} N_j = \sum_{i=1}^{n} M_i = \sum_{j=1}^{m} \sum_{i=1}^{n} O_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. The last column contains counts for the track UP, where N_{UP_i} is

the number of points in UP_i, $i = 1, \ldots, n$, and $N_{UP} \leq \sum_{i=1}^{n} N_{UP_i}$ is the number of points in UP. Note that $N_{UP} = \sum_{i=1}^{n} N_{UP_i}$ only if there is no overlap between the points in the UP_i tracks. Also, note that if we assume all UP points are in segments of some US_j, then $N \geq N_{UP}$, since points that are inside segments of several US_j are counted several times.

Hypothesis tested

To test whether there are more points of track UP_i inside the segments of track US_j than expected, we need to compare the observed number of points, O_{ij}, to the expected number of points, E_{ij}. This expected value can be based on the assumption that the proportion of points in the US_j segments that are UP_i points is the same as the proportion of points in UP that are UP_i points, i.e.

$$\frac{E_{ij}}{N_j} = \frac{N_{UP_i}}{N_{UP}}.$$

A second possible assumption is that the proportion of points in the US_j segments that are UP_i points is the same as the proportion of points in the average US_j track that are UP_i points, where we define the number of points in the average US_j track to be $\frac{1}{m}\sum_{j=1}^{m} N_i = \frac{N}{m}$ and the number of UP_i points in the average US_j track to be $\frac{1}{m}\sum_{j=1}^{m} O_{ij} = \frac{M_i}{m}$. This means that under this assumption

$$\frac{E_{ij}}{N_j} = \frac{M_i}{N}.$$

We then define the null hypothesis:

H_0: $\frac{O_{ij}}{N_j} = r_i$
where $r_i = \frac{N_{UP_i}}{N_{UP}}$ or $r_i = \frac{M_i}{N}$.

Alternative hypothesis

H_1: $\frac{O_{ij}}{N_j} > r_i$

When $r_i = \frac{N_{UP_i}}{N_{UP}}$ the test is independent of the other US tracks, and it tests whether a particular UP_i track has a higher proportion of points inside the US_j segments than the average UP track. No UP points are counted several times. Note that the test is independent of the other US tracks only if the UP tracks were made without using information about the US tracks.

When $r_i = \frac{M_i}{N}$ the test depends on all $n+m$ tracks, and it tests whether there is a higher proportion of UP_i points inside the US_j segments than inside the segments of the average US track. If the same segment is part of several of the US tracks then the points in these segments are counted several times. Therefore, the tests for the US_j track depends strongly on segments with many points that are part of many of the other US_k tracks, $k = 1, \ldots, m$.

Statistics and rejection of the null hypothesis when $r_i = \frac{N_{UP_i}}{N_{UP}}$

Under the null hypothesis we assume that the proportion of points in the US_j segments that are UP_i points is the same as the proportion of points in UP that are UP_i points, i.e. $\frac{E_{ij}}{N_j} = \frac{N_{UP_i}}{N_{UP}}$. We also assume that the positions of the N_{UP_i} points in UP_i are randomly selected with equal probabilities from the N_{UP} positions in UP. Assuming that the points are

selected without replacement, we observe that O_{ij} follows a hypergeometric distribution with parameters N_{UP}, N_j and N_{UP_i} as follows: i) We have an urn with N_{UP} red and white balls, where the N_j red balls represent positions in UP within segments of US_j, while the $N_{UP} - N_j$ white balls represent positions outside segments of US_j; ii) We draw N_{UP_i} balls/positions without replacement from the urn; iii) O_{ij}, the number of these N_{UP_i} positions that are within segments of US_j, i.e. the number of red balls drawn, follows a hypergeometric distribution.

Statistics and rejection of the null hypothesis when $r_i = \frac{M_i}{N}$

Under the null hypothesis we assume that the proportion of points in the US_j segments that are UP_i points is the same as the proportion of points in the average US_j track that are UP_i points, i.e. $\frac{E_{ij}}{N_j} = \frac{M_i}{N}$. We also assume that the positions of the N_{UP_i} points in UP_i are randomly selected with equal probabilities from the N_{UP} positions in UP. In this case we cannot choose a hypergeomtric distribution as positions might have been counted several times in N and M_i, not only once as in N_{UP} and N_{UP_i}. Instead of assuming that the points are selected without replacement, we might assume that the points are selected with replacement. Then O_{ij} will follow a Binomial distribution with parameters N_j and $\frac{M_i}{N}$.

Clustering

It might be interesting to find groups of US tracks or groups of UP tracks that are similar. This might be achieved by clustering a matrix S that for each pair of tracks (UP_i, US_j) expresses how different the observed value O_{ij} is from the expected value E_{ij}. A natural score might be the statistics used in the hypotheses tests, but these are not necessarily comparable, and needs to be normalized before clustering the matrix. One possibility is to use $\frac{(Oij-Eij)}{\sqrt{(E_{ij}(1-E_{ij}/N_i)}}$, which is approximately standard normally distributed, as a score in the matrix to be clustered. For the hypothesis test based on the Binomial distribution ($r_i = \frac{M_i}{N}$), the reasoning behind this is that $O_{ij} \sim Binomial(N_i, \frac{M_i}{N})$ might be approximated by $O_{ij} \sim Normal(E_{ij}, E_{ij}(1 - \frac{E_{ij}}{N_i})))$, where $E_{ij} = N_i \cdot \frac{M_i}{N}$ (this is a good approximation if $5 < E_{ij} < N_i - 5$). This means that $\frac{O_{ij}-E_{ij}}{\sqrt{(E_{ij}(1-\frac{E_{ij}}{N_i})))}} \sim Normal(0, 1)$. For the hypothesis test based on the hypergeometric distribution ($r_i = \frac{N_{UP_i}}{N_{UP}}$) we first approximate the hypergeometric distribution with a Binomial distribution: $O_{ij} \sim Binomial(N_i, \frac{N_{UP_i}}{N_{UP}})$. Otherwise the reasoning is as for the case where the hypothesis test based on the Binomial distribution, except that $E_{ij} = N_i \cdot \frac{N_{UP_i}}{N_{UP}}$. The approximation of the Binomial to the Normal becomes better if O_{ij} is substituted by $O_{ij} - 0.5$ (continuity correction).

Example – The disease regulome

The hypothesis testing and clustering described above have been used for a data set where each track with unmarked segments consists of genes that are associated with a given disease, while each track with unmarked points consists of binding sites for a given TF.

Disease input The tracks with unmarked segments are made from gene lists associated with different diseases. These gene lists are obtained from the medical literature, i.e. from a collection of documents found to be relevant. Let N be the number of documents in this

collection and let m be the number of documents that mention disease term d and n the number of documents that mention gene term g. Under the null model that there is no association between the disease term d and gene term g, the number of documents that mention both d and g follows a hypergeometric distribution with parameters N, m and n. We then define the gene list for a given disease to be all genes for which we obtain a p-value less than 0.05.

TF input The tracks with unmarked points are made from DNA locations for TF binding, using a track from UCSC, keeping all predictions in the track.

Gene regions As TFBS acting on a gene are often close to but outside a gene, we have extended gene regions (segments representing each gene in a gene list) to include flanks, set to 5kb in each direction.

Counting When counting binding sites for a given TF acting on a given gene, it is not necessarily that different whether it is one binding site or twenty. We first did a simple count of binding sites across all gene regions associated with a gene. We have now also tried to only count the number of genes having binding sites (i.e. count only one binding site per gene) and we have tried to take the (discretized) logarithm of the count within each gene, then summing the logarithm values across genes. We open up for all options, and have used the summing of logarithm values in our main disease regulome. Mark that the potensial reduction of the number of TFs is done before making the matrix O and the track UP.

Clustering algorithm We have used hierarchical clustering as this gives us information on several levels, both closely related diseases and large groups of diseases with a certain amount of similarity. We have as distance measure between feature vectors (rows/columns of matrix values for TFs/diseases) used the Euclidean distance. As between-cluster distance, we have used the average of all pairwise distances between feature vectors of the respective clusters.

The two null hypotheses The consequence of choosing the second null hypothesis ($r_i = \frac{M_i}{N}$) is that the results for a specific disease will depend on the set of diseases used in the analysis giving results that are differential compared to the disease set. If e.g. only cancer diseases are used, then results will show how a specific cancer differs from general cancers. The first null hypothesis ($r_i = \frac{N_{UP_i}}{N_{UP}}$) will give results for each disease independent of other diseases. We consider the two assumptions / null hypotheses to provide complementary results, and have generated the disease regulome according to either, though we consider $r_i = \frac{M_i}{N}$ as the main result.