Enrichment Analysis Methods

Botong Qu

1. Test of Significance

In this part, I want to review some basic knowledge of the test of significance, and the reason of using these statistical methods to analyze the data.

In the Annotation database, we have lots of genes (or QTL or germplasm) and their corresponding annotation information. This is the background data, which is the associations between genes and the ontology terms. When the biologists found an interesting gene list, they may want to know the big picture of these genes, i.e. the relationship and the hierarchy of the ontology terms associated to these genes. Or they may want to know what feature (Trait ontology, environment ontology) or underlying biological process (GO) are related to this set of genes. So we need to find out the enriched ontology terms based on the selected gene list. This analysis procedure is called enrichment analysis. The enrichment means that a class of genes are over-represented in a large set of genes (selected gene list).

The problem is every gene could be associated to more than one ontology terms, most of the ontology terms related to the gene list could not represent the main function of this gene list. So finding the ontology term which is enriched by the interesting set of gene list would become unintuitive. The key is finding the ontology terms, associated to this set of genes, which are not selected by chance. The founded ontology term should be **overrepresented** by the gene list.

In statistics, when we want to make sure something happen caused by real instead of by chance. We need to do the **Test of Significance**. **Null hypothesis (H_0)** expresses the idea that an observed difference is due to chance, the **alternative hypothesis (H_a)** is the other side of the statement, which the difference is real. By using some statistic method, we could calculate a **p-value,** which represent the observed significance level. The lower the p-value is, the stronger evidence we have to reject the null hypothesis. When p-value less than 5%, the result is “statistically significant”, if the p-value less than 1%, the result is “highly significant”.

2. Urn Problem

In probability and statistics, an urn problem is an idealized mental exercise in which some objects of real interest (such as atoms, people, cars, etc.) are represented as colored balls in an urn or other container. One pretends to remove one or more balls from the urn; the goal is to determine the probability of drawing one color or another, or some other properties.

If we have an urn, and have n balls in the urn, some of them are red and some of them are green. When drawn a number of balls from the urn, calculation of the probability of the arrangement is the Urn Problem. If the balls are not returned to the urn once extracted. Hence, the number of total marbles in the urn decreases. This is referred to as "drawing without replacement". To solve drawing without replacement urn problem, we could use hypergeometric distribution.

We represent the cells by the letters a, b, c and d, call the totals across rows and columns **marginal totals**, and represent the **grand total** by n. So the table now looks like this:

	drawn	Not drawn	Row Total
red	a	b	a + b
green	c	d	c + d
Column Total	a + c	b + d	a + b + c + d (=n)

3. Hypergeometric test

Hypergeometric distribution is calculated by following equation:

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

Where $\binom{n}{k}$ is the binomial coefficient and the symbol "!" indicates the factorial operator. $\binom{n}{k}$ is as same as C(n, k), which represent the combination of n things taken k at a time without repetition.

4. Fisher's exact test

The fisher's exact test is useful for categorical data that result from classifying objects in two different ways; it is used to examine the significance of the association (contingency) between the two kinds of classification. Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution.

Example:

A sample of teenagers might be divided into male and female on the one hand, and those that are and are not currently dieting on the other. We hypothesize, for example, that the proportion of dieting individuals is higher among the women than among the men, and we want to test whether any difference of proportions that we observe is significant. The data might look like this:

	Men	Women	Row total
Dieting	1	9	10
Non-dieting	11	3	14
Column total	12	12	24

$$p = \frac{\binom{10}{1}\binom{14}{11}}{\binom{24}{12}} = \frac{10!\,14!\,12!\,12!}{1!\,9!\,11!\,3!\,24!} \approx 0.001346076$$

Above formula gives the exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that men and women are equally likely to be dieters. There is another more extreme arrangement which when the men in dieting is 0. In this case, that means women are more on diet. The probability of that is as following equation.

$$p = \frac{\binom{10}{0}\binom{14}{12}}{\binom{24}{12}} \approx 0.000033652$$

In order to calculate the significance of the observed data, i.e. the total probability of observing data **as extreme or more extreme** if the null hypothesis is true, we have to calculate the values of p for both these two arrangement, and add them together. This gives a one-tailed test, with p approximately 0.001346076 + 0.000033652 = 0.001379728.

5. Chi-squared test

The chi-square test provides a method for testing the association between the row and column variables in a two-way table. The null hypothesis H_0 assumes that there is no association between the variables (in other words, one variable does not vary according to the other variable), while the alternative hypothesis H_a claims that some association does exist. The alternative hypothesis does not specify the type of association, so close attention to the data is required to interpret the information provided by the test.

The chi-square test is based on a test statistic that measures the **divergence of the observed data from the values that would be expected** under the null hypothesis of no association. This requires calculation of the expected values based on the data.

$$\chi^2 = \sum \frac{(observed - expected\,)^2}{expected}$$

The **expected value** for each cell in a two-way table is equal to

$$\frac{(row\ total * column\ total\,)}{n}$$

Where n is the total number of observations included in the table.

The **degree of freedom** is calculated by

(row number-1)*(column number -1).

Based on the χ^2 value and the degree of freedom, we could search the p-value from below table, normally, we could use computer to calculate the p- value of the corresponding χ^2.

Degrees of freedom (df)	χ2 value										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27

4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Example:

We use the same example as above to illustrate the chi-squared test.

Observed table

	Men	Women	Row total
Dieting	1	9	10
Non-dieting	11	3	14
Column total	12	12	24

Based on above statement, Calculation of the expected values as following:

Men on diet: (10/24)*12 = 5

Men not on diet: 12 – 5 = 7

Women on diet: 10-5=5

Women not on diet: 12-5 = 7

So the expected table as following:

Expected table

	Men	Women	Row total
Dieting	5	5	10
Non-dieting	7	7	14
Column total	12	12	24

The degree of the freedom is (2-1)*(2-1) = 1.

So the

$$\chi^2 = \frac{(1-5)^2}{5} + \frac{(9-5)^2}{5} + \frac{(11-7)^2}{7} + \frac{(3-7)^2}{7} = 3.2 + 3.2 + 2.28 + 2.28 \approx 10.9$$

From the table, we could found that when χ^2 is 10.9, the p-value is smaller than 0.001, so we have enough evidence to reject the null hypothesis. So there is an association between gender and dieting.

6. Gene Enrichment Analysis model

As same as the urn problem, the gene enrichment analysis is also a two way table, and we need to find one sample distribution whether or not caused by chance, so a significance test is required.

	Gene list (interesting)	Not in Gene list (not interesting)	
term A annotated	m	k-m	k
term A un-annotated	n-m	(N-n)-(k-m)	N-k
	n	N-n	N

- Term A: An ontology term, like GO: 0000001 or GO: 0000089. In our system, this term A is not limited to other Ontology databases such as TO, EO.
- N is the number of Genes (object term) in the database, that is the quantity of genes of background.
- k is the number of genes annotated to Term A in the whole database.
- n is the number of selected genes, or the quantity of genes in the interesting gene list.
- m is the number of genes which are annotated to term A in the selected gene list.
- H_0: The selected gene list is not enriched by Term A, that is there is no association between selected gene list with term A. Or we can say that the m annotated genes in selected gene list is chance variation, which means we get this m on coincidence.
- H_a: There is an association between the gene list and the Term A. Or we can say that gene list is enriched by term A.
- P-value: if the p-value calculated by following three methods is lower than 0.05, we should reject the Null Hypothesis, which means that we found the association between gene list and the term A. The p-value could also be set to 0.01 which means that we will only find more strong evidence to reject the Null Hypothesis.

Fisher's exact test

Use the formula above to calculate the P(m<X<k), X is the number of genes in the gene list is annotated to ontology tern A.

$$P(m) = C(k,m)*C(N-k, n-m) / C(N, n)$$

$$\text{p-value} = P(m) + P(m+1) + P(m+2) +\ldots + P(k)$$

Hypergeometric test

The fisher's exact test is just using the Hypergeometric distribution to calculate the p-value. From my understanding, the so called hypergeometric test is as same as fisher's exact test. I used the enrichment test tool in **AgriGo** to test this assumption, found that the visualization results of using fisher's exact and hypergeometric are same, even the calculated p-values are same for each ontology terms.

Chi-squared test

With **large samples**, a chi-squared test can be used in this situation. However, the significance value it provides is only an approximation, because the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution. The usual rule of thumb for deciding whether the chi-squared approximation is good enough is that the chi-squared test is not suitable when the expected values in any of the cells of a contingency table are below 5, or below 10 when there is only one degree of freedom.

The Degree of freedom is (2-1)*(2-1), and the expected value as following table

	Gene list (interesting)	Not in Gene list (not interesting)	
term A annotated	N*k/N	k-N*k/N	k
term A un-annotated	n-N*k/N	N-k-n+N*k/N	N-k
	n	N-n	N

Then calculate the chi-squared value by using following equation:

$$\chi^2 = \frac{(m - \mathrm{N} * \mathrm{k}/\mathrm{N})^2}{\mathrm{N} * \mathrm{k}/\mathrm{N}} + \frac{(\mathrm{k} - \mathrm{m} - \mathrm{k} + \mathrm{N} * \mathrm{k}/\mathrm{N})^2}{\mathrm{k} - \mathrm{N} * \mathrm{k}/\mathrm{N}} + \frac{(\mathrm{n} - \mathrm{m} - \mathrm{n} + \mathrm{N} * \mathrm{k}/\mathrm{N})^2}{\mathrm{n} - \mathrm{N} * \mathrm{k}/\mathrm{N}} + \frac{(\mathrm{N} - \mathrm{n} - \mathrm{k} + \mathrm{m} - \mathrm{N} + \mathrm{k} + \mathrm{n} - \mathrm{N} * \mathrm{k}/\mathrm{N}))^2}{\mathrm{N} - \mathrm{k} - \mathrm{n} + \mathrm{N} * \mathrm{k}/\mathrm{N}}$$

$$= \frac{(m - \mathrm{N} * \mathrm{k}/\mathrm{N})^2}{\mathrm{N} * \mathrm{k}/\mathrm{N}} + \frac{(-\mathrm{m} + \mathrm{N} * \mathrm{k}/\mathrm{N})^2}{\mathrm{k} - \mathrm{N} * \mathrm{k}/\mathrm{N}} + \frac{(-\mathrm{m} + \mathrm{N} * \mathrm{k}/\mathrm{N})^2}{\mathrm{n} - \mathrm{N} * \mathrm{k}/\mathrm{N}} + \frac{(\mathrm{m} - \mathrm{N} * \mathrm{k}/\mathrm{N}))^2}{\mathrm{N} - \mathrm{k} - \mathrm{n} + \mathrm{N} * \mathrm{k}/\mathrm{N}}$$

Than find the p-value for the chi-squared value.

7. Gene Enrichment analysis steps

1. Analysis the background database to create a reference file, calculation of the **N** (Gene number in background data), and for each ontology term, calculation the **k** (the associated gene number).
2. A list of interesting genes, calculate the **n** (gene number in the list).
3. Find all associated ontology terms in one ontology database (GO, TO, EO, etc.).
4. For each ontology terms, find the number of associated genes in the gene list --- **m**.
5. Using the above formula to calculate the **p-value**.
6. Find the ontology terms whose corresponding p-value below 0.05 (0.01 if the user select). These ontology terms are the enriched ontologies.
7. Visualization the hierarchy among the enriched ontology terms.

8. Related Tools

All these tools is selected based on first impression of Amigo, we may change to better ones if we find some other choices.

- Using **JavaScript** to query from the **SOLR** database, the reply data is in format of **JSON**.
- Using **Jstat** to calculate all statistic values with using JS.
- Using **Graphviz** to visualize the graphics.