BIOINFORMATICS REVIEW

Vol. 23 no. 4 2007, pages 401–407
doi:10.1093/bioinformatics/btl633

Databases and ontologies

Enrichment or depletion of a GO category within a class of genes: which test?

Isabelle Rivals*, Léon Personnaz, Lieng Taing[1] and Marie-Claude Potier[1]

Équipe de Statistique Appliquée and [1]Laboratoire de Neurobiologie et Diversité Cellulaire, École Supérieure de Physique et de Chimie Industrielles (ESPCI), 10 rue Vauquelin, 75005 Paris, France

Received on June 20, 2006; revised and accepted December 11, 2006
Advance Access publication December 20, 2006
Associate Editor: Jonathan Wren

ABSTRACT

Motivation: A number of available program packages determine the significant enrichments and/or depletions of GO categories among a class of genes of interest. Whereas a correct formulation of the problem leads to a single exact null distribution, these GO tools use a large variety of statistical tests whose denominations often do not clarify the underlying *P*-value computations.

Summary: We review the different formulations of the problem and the tests they lead to: the binomial, χ^2, equality of two probabilities, Fisher's exact and hypergeometric tests. We clarify the relationships existing between these tests, in particular the equivalence between the hypergeometric test and Fisher's exact test. We recall that the other tests are valid only for large samples, the test of equality of two probabilities and the χ^2-test being equivalent. We discuss the appropriateness of one- and two-sided *P*-values, as well as some discreteness and conservatism issues.

Contact: isabelle.rivals@espci.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A common problem in functional genomic studies is to detect significant enrichments and/or depletions of Gene Ontology (GO) categories within a class of genes of interest, typically the class of significantly differentially expressed (DE) genes. Many GO processing tools perform this task using various statistical tests refered to as: the binomial test, the χ^2-test, the equality of two probabilities test, Fisher's exact test and the hypergeometric test (see Table 1). The authors of some packages claim the advantages of the test(s) they propose, often seemingly contradicting each other. For example, Zeeberg *et al*. (2003) favor Fisher's exact test: 'Unlike the *Z*-statistics with the hypergeometric distribution, and tests based on it, Fisher's exact test is appropriate even for categories containing a small number of genes', whereas for Martin *et al*. (2004) the hypergeometric test is most appropriate: 'On the average, the hypergeometric distribution seems to be both the most adapted model and the most powerful test'. Moreover, even though the most recent review papers use a number of criteria to exhaustively compare the different existing tools (Khatri and Draghici, 2005), they do not discuss in detail the identity and approximation relationships existing between the different tests. This is precisely the aim of the present paper.

2 PROBLEM STATEMENT

We consider a total population of genes, e.g. the genes expressed in a microarray experiment, and we are interested in the property of a gene to belong to a specific GO category. The aim is to establish whether the class of the DE genes presents an enrichment and/or a depletion of the GO category of interest with respect to the total gene population.

3 CANDIDATE FORMULATIONS

Let $\mathrm{H_0}$ denote the null hypothesis that the property for a gene to belong to the GO category of interest and that to be DE are independent, or equivalently that the DE genes are picked at random from the total gene population. We consider successively the hypergeometric, the comparison of two probabilities, and the 2×2 contingency table formulations of the above problem, and introduce the exact or approximate null distributions they lead to.

Notations (see Table 2): the total number of genes is denoted by n, the total number of genes belonging to the GO category of interest by $\mathrm{n_{+1}}$, the number of DE genes by $\mathrm{n_{1+}}$: n, $\mathrm{n_{+1}}$ and $\mathrm{n_{1+}}$ are hence fixed by the experiment. The number of DE genes belonging to the GO category is denoted by $\mathrm{n_{11}}$.

3.1 Hypergeometric formulation

The hypergeometric formulation is directly derived from the problem statement.

3.1.1 Exact null distribution If $\mathrm{H_0}$ is true, the random variable $\mathrm{N_{11}}$ whose realization[1] is the observed value $\mathrm{n_{11}}$, has a hypergeometric distribution with parameters n, $\mathrm{n_{1+}}$, and $\mathrm{n_{+1}}$, which we denote by $\mathrm{N_{11} \sim Hyper(n, n_{1+}, n_{+1})}$, with:

$$\mathrm{P(N_{11} = x)} = \frac{\binom{\mathrm{n_{+1}}}{\mathrm{x}}\binom{\mathrm{n - n_{+1}}}{\mathrm{n_{1+} - x}}}{\binom{\mathrm{n}}{\mathrm{n_{1+}}}} = \frac{\binom{\mathrm{n_{+1}}}{\mathrm{x}}\binom{\mathrm{n_{+2}}}{\mathrm{n_{12}}}}{\binom{\mathrm{n}}{\mathrm{n_{1+}}}}. \qquad (1)$$

Note that $\mathrm{Hyper(n, n_{1+}, n_{+1}) \equiv Hyper(n, n_{+1}, n_{1+})}$.

[1] Random variables and their realizations are denoted respectively by uppercase and lowercase letters.

*To whom correspondence should be addressed.

Table 1. Reviewed GO processing tools

GO tool	Statistical tests	Reference
BINGO	Hypergeometric	Maere *et al*., 2005
CLENCH	Hypergeometric, binomial, χ^2	Shah and Fedorov, 2004
DAVID	Fisher	Dennis *et al*., 2003
EASEonline	Fisher	Hosack *et al*., 2003
eGOn	Fisher	http://www.genetools.microarray.ntnu.no/help/help_egon.php?egon=1#intro
FatiGO	Fisher	Al-Sharour *et al*., 2004
FuncAssociate	Fisher	http://llama.med.harvard.edu/cgi/func/funcassociate
FunSpec	Hypergeometric	Robinson *et al*., 2002
GeneMerge	Hypergeometric	Castillo-Davis and Hartl, 2003
GFINDer	Hypergeometric, binomial, Fisher[a]	Masseroli *et al*., 2004
GoMiner	Fisher	Zeeberg *et al*., 2003
GOstat	χ^2, Fisher	Beißbarth and Speed, 2004
GoSurfer	χ^2	Zhong *et al*., 2004
GO TermFinder (CPAN)	Hypergeometric	Boyle *et al*., 2004
GO TermFinder (SGD)	Binomial	http://www.yeastgenome.org/help/goTermFinder.html
GOTM	Hypergeometric	Zhang *et al*., 2004
GOToolBox	Hypergeometric, binomial, Fisher	Martin *et al*., 2004
L2L	Binomial	Newman and Weiner, 2005
NetAffx GO Mining Tool	χ^2	Cheng *et al*., 2004
Onto-Express	Binomial, χ^2, Fisher	Khatri *et al*. 2002; Draghici *et al*., 2003
Ontology Traverser	Hypergeometric	Young *et al*., 2005
STEM	Hypergeometric	Ernst *et al*., 2005
THEA	Hypergeometric, binomial	Pasquier *et al*., 2004

[a] The website now proposes 3 additional tests, but they are not documented.

Table 2. Classification of the genes expressed in a microarray experiment

	Category 1 ($\in$GO category)	Category 2 ($\notin$GO category)	Total
Class 1 (DE)	n_{11}	n_{12}	n_{1+}
Class 2 (not DE)	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

3.1.2 Approximate null distribution For a large sample, N_{11} has approximately a binomial distribution with parameters n_{1+} and n_{+1}/n: $N_{11} \sim Bi(n_{1+}, n_{+1}/n)$. Note that if $n_{1+}\ n_{+1}/n$ is also large, the binomial approximation can further be approximated by a Gaussian distribution.

3.2 Comparison of two probabilities formulation

In a second formulation, we consider two samples, that of the DE genes of size n_{1+}, among which n_{11} genes belonging to the GO category of interest, and that of the not DE genes of size n_{2+}, among which n_{21} genes belonging to the GO category. The proportions of genes belonging to the GO category in the two samples are thus $f_1 = n_{11}/n_{1+}$ (DE genes) and $f_2 = n_{21}/n_{2+}$ (not DE genes). Let p_1 and p_2 denote the probabilities to belong to the GO category in the two samples; then $N_{11} \sim Bi(n_{1+}, p_1)$ and $N_{21} \sim Bi(n_{2+}, p_2)$. In this formulation, the null hypothesis H_0 is the equality of the two probabilities $p_1 = p_2 = p$, i.e. there is neither enrichment nor depletion in the sample of DE genes with respect to that of the not DE genes.

3.2.1 Approximate null distribution The case of large samples arises frequently. Then, the binomial distributions can be approximated with Gaussian distributions. Under H_0, n_{1+} and n_{2+} being large, the probability p can be correctly estimated with $f = (n_{11} + n_{21})/(n_{1+} + n_{2+}) = n_{+1}/n$, leading to the approximately normally distributed variable:

$$Z = \frac{F_1 - F_2}{\sqrt{F(1-F)}\sqrt{\dfrac{1}{n_{1+}} + \dfrac{1}{n_{2+}}}} \sim N(0,1). \qquad (2)$$

This distribution is approximate for two reasons: (1) the replacement of the binomial distributions by Gaussian distributions holds only for large samples (both n_{1+} and n_{2+} must be large), and (2) it has not been taken into account that, according to our problem statement, the sum $N_{11} + N_{21}$, the total number of genes belonging to the GO category, is fixed and equal to n_{+1}.

3.2.2 Exact null distribution Without approximating the binomial distribution, and taking into account that $N_{11} + N_{21} = n_{+1}$, we naturally obtain $N_{11} \sim Hyper(n, n_{1+}, n_{+1})$ (see (Fisher, 1935; Lehman, 1986) for the complete computation with the binomial distribution conditionally on $N_{11} + N_{21} = n_{+1}$). Hence, the exact distribution of N_{11} under H_0 is as before the hypergeometric distribution.

3.3 Contingency table formulation

A third formulation is based on Table 2 seen as a 2 × 2 contingency table. Let again H_0 denote the hypothesis that the property to belong to the GO category of interest and that to be DE are independent.

3.3.1 Approximate null distribution The case of a large sample is frequently considered where, if H_0 is true, the following variable is asymptotically χ^2 distributed with one degree of freedom (Mood *et al*., 1974):

$$D^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(N_{ij}-\frac{N_{i+}N_{+j}}{n}\right)^2}{\frac{N_{i+}N_{+j}}{n}} \sim \chi^2(1). \tag{3}$$

Note that d^2 is the square of the realization z of the normal variable Z given by Equation (2):

$$d^2 = \frac{n(n_{11}n_{22}-n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}} = z^2. \tag{4}$$

3.3.2 Exact null distribution Whatever the sample size, Fisher's formula gives the probability of the observed configuration of the contingency table under H_0 (Fisher, 1935; Mood et *al*., 1974; Agresti, 2002):

$$P(\{N_{ij} = n_{ij}\}) = \frac{n_{1+}!n_{2+}!n_{+1}!n_{+2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}. \tag{5}$$

It is easy to show that $N_{11} \sim \text{Hyper}(n, n_{1+}, n_{+1})$:

$$\begin{aligned} P(N_{11} = x \,|\, N_{1+} = n_{1+}, N_{+1} = n_{+1}) &= \frac{\frac{n_{+1}!}{x!n_{21}!}\,\frac{n_{+2}!}{n_{12}!n_{22}!}}{\frac{n!}{n_{1+}!n_{2+}!}} \\ &= \frac{\binom{n_{+1}}{x}\binom{n_{+2}}{n_{12}}}{\binom{n}{n_{1+}}}. \end{aligned} \tag{6}$$

As expected, the exact distribution of N_{11} under H_0 is again the hypergeometric distribution, see Equation (1).

3.4 Summary

Under H_0, i.e. assuming the independence of the property to belong to the GO category of interest and of the property to be DE, or equivalently assuming $p_1 = p_2$ where p_1 is the probability of the DE genes to belong to the GO category, and p_2 the probability of the not DE genes to belong to the GO category, the exact distribution of N_{11} is the hypergeometric distribution $N_{11} \sim \text{Hyper}(n, n_{1+}, n_{+1})$ which, if n is large, can be approximated with the binomial distribution $\text{Bi}(n_{1+}, n_{+1}/n)$. If the two samples are large, it is also possible to exhibit an approximately normal variable Z or its square $D^2 = Z^2$, the latter being hence approximately χ^2 distributed with one degree of freedom.

4 TESTS AND *P*-VALUES

Generally, when performing the test of a null hypothesis H_0 against some alternative hypothesis H_a, one disposes of a realization x of a random variable X with known distribution under H_0, the null distribution. One chooses *a priori* a probability α of type I error (the error to reject H_0 whereas it is true) that must not be exceeded, also called significance level, the decision to reject H_0 being taken when x falls in the critical region. In this context, the *P*-value is the minimum significance level for which H_0 would be rejected, or equivalently, it is the probability, under H_0, of the minimal critical region containing x.

The choice of a critical region in order to maximize the power of the test, and hence the choice of the corresponding *P*-value, depends on the alternative hypothesis H_a, which may be 'enrichment' ($p_1 > p_2$, one-sided test, critical region right), 'depletion' ($p_1 < p_2$ 'one-sided' test, critical region left), or 'enrichment or depletion' ($p_1 \neq p_2$, two-sided test, critical region left and right). Enrichment, depletion and enrichment or depletion are later denoted by E, D, and E/D, respectively.

4.1 One-sided tests

The one-sided *P*-value is defined as:

$$\begin{cases} \text{if } H_a = E, p_{\text{one}}(n_{11}) = P(N_{11} \geq n_{11}) \\ \text{if } H_a = D, p_{\text{one}}(n_{11}) = P(N_{11} \leq n_{11}) \end{cases}. \tag{7}$$

If the case of a discrete distribution, like the exact hypergeometric distribution or the approximate binomial distribution, it is not possible to guaranty any value of the significance level with the rule 'reject H_0 if $p_{\text{one}}(n_{11}) \leq \alpha$'. Due to the discreteness, the actual significance level (or size of the test) is generally smaller than the nominal (desired) significance level α, which results in a loss of power.

To minimize this loss, a good remedy is the use of mid-*P*-values (Agresti and Min, 2001; Agresti, 2002). The one-sided mid-*P*-value, which we denote by π_{one}, is defined as:

$$\begin{cases} \text{if } H_a = E, \pi_{\text{one}}(n_{11}) = P(N_{11} > n_{11}) + \frac{1}{2}P(N_{11} = n_{11}) \\ \text{if } H_a = D, \pi_{\text{one}}(n_{11}) = P(N_{11} < n_{11}) + \frac{1}{2}P(N_{11} = n_{11}) \end{cases}. \tag{8}$$

It must be noted that the actual significance level, i.e. the actual probability of type I error, is no longer guaranteed to be smaller than the nominal significance level. However, it is rarely much greater (Agresti, 2002).

Another remedy is randomization, with which any desired significance level can be achieved. However in practice, randomization having nothing to do with the data does not make much sense (Lehmann, 1986; Agresti, 2002).

If the approximately normal variable Z is considered, we have:

$$\begin{cases} \text{if } H_a = E, p_{\text{one}}(z) = P(Z > z) \\ \text{if } H_a = D, p_{\text{one}}(z) = P(Z < z) \end{cases}. \tag{9}$$

If the approximately χ^2 distributed variable D^2 is used, a one-sided test cannot be performed, since both enrichment (large observed n_{11}) and depletion (small observed n_{11}) lead to a large value of D^2, i.e. there is a single critical region.

4.2 Two-sided tests

In the case of a two-sided test i.e. $H_a = E/D$, and of a discrete null distribution, there are several popular definitions of the *P*-value, see (Agresti, 1992, 2002). A first approach defines the two-sided *P*-value as twice the one-sided *P*-value:

$$p_{\text{two}}^{\text{doubling}}(n_{11}) = 2 \times \min[P(N_{11} \geq n_{11}), P(N_{11} \leq n_{11})]. \tag{10}$$

Yates and Fisher himself were in favor of this 'doubling' approach (Yates, 1984). A second approach, which after Gibbons and Pratt (1975) we name the 'minimum-likelihood' approach, defines the *P*-value as the sum of the probabilities of the values of N_{11} that are smaller or equal to that of the observed value n_{11}, as recommended for example in (Mehta and Patel, 1998):

$$p_{\mathrm{two}}^{\mathrm{min\ lik}}(n_{11}) = \sum_{P(N_{11}=m)\leq P(n_{11})} P(N_{11}=m). \tag{11}$$

The minimum-likelihood approach is the only one we have encountered in the GO tools of Table 1. A third approach defines the *P*-value as the sum of the probabilities of the values of N_{11} that are at least as or more extreme (with respect to the mathematical expectation of N_{11}) than the observed one (Gibbons and Pratt, 1975; Yates, 1984; Agresti, 2002). A fourth approach defines the two-sided *P*-value as $\min[P(N_{11} \geq n_{11}), P(N_{11} \leq n_{11})]$ plus an attainable probability in the other tail that is as close as possible to, but not greater than, that one-tailed probability (Agresti, 2002).

These definitions lead to equal *P*-values in the case of symmetric distributions, i.e. when $n_{1+} = n_{2+}$; else, they possibly lead to different *P*-values and corresponding test results, each of them having advantages and disadvantages, due to the discreteness and skewness of the hypergeometric distribution. The problem is also that these *P*-values do not correspond to any well-defined two-sided test. This issue is discussed for example in (Dunne *et al*., 1996), where a two-sided *P*-value based on an uniformly most powerful unbiased test is proposed. However, this *P*-value is obtained with an iterative procedure, which makes this approach inadequate for the screening of hundreds of different GO categories.

Thus, if a single simple and computationally light (see subsection 6.3) procedure were to be recommended, we would advice the doubling approach, against which there is no strong argument, and using the mid-*P*-value, in order to reduce the discreteness and conservatism effects:

$$\begin{aligned}\pi_{\mathrm{two}}^{\mathrm{doubling}}(n_{11}) &= 2 \times \min(P(N_{11} > n_{11}) + \tfrac{1}{2}P(N_{11} = n_{11}),\\ &\quad P(N_{11} < n_{11}) + \tfrac{1}{2}P(N_{11} = n_{11})).\end{aligned} \tag{12}$$

A mid-*P*-value can also be defined for the minimum-likelihood approach, as the sum of the probabilities that are smaller than the probability of the observed value n_{11}, plus half the sum of the probabilities equal to it:

$$\begin{aligned}\pi_{\mathrm{two}}^{\mathrm{min\ lik}}(n_{11}) = &\sum_{P(N_{11}=m)<P(n_{11})} P(N_{11}=m)\\ &+ \frac{1}{2}\sum_{P(N_{11}=m)=P(n_{11})} P(N_{11}=m).\end{aligned} \tag{13}$$

However, we must again emphasize that the actual probability of type I error may exceed the nominal significance level.

If the approximately normal variable Z is considered (a continuous and symmetrically distributed variable), we have:

$$p_{\mathrm{two}}(z) = 2 \times \min[P(Z > z), P(Z < z)]. \tag{14}$$

If the approximately χ^2 distributed variable D^2 is considered, the *P*-value is computed as:

$$p_{\mathrm{two}}(d^2) = P(D^2 > d^2) = p_{\mathrm{two}}(z). \tag{15}$$

4.3 One versus two-sided tests?

Consider a dataset consisting of tissues in a pathological condition and of normal tissues, and a GO category whose genes are directly affected by the condition, i.e. the genes belonging to this GO category are DE (either over- or under-expressed). Such a GO category is likely to be over-represented among the DE genes, i.e. an enrichment is expected. Thus, detecting an enrichment is desirable. On the other hand, consider a GO category such that the normal expression of the corresponding genes is necessary for the condition to develop, i.e. the genes belonging to this GO category are not DE. Such a GO category is likely to be under-represented among the DE genes, i.e. a depletion is expected. Thus, detecting a depletion is also desirable, even if there is a risk to detect the depletion of a GO category corresponding to genes whose normal expression is necessary to the mere survival of the specie.

Thus, both enrichments and depletions of GO categories are potentially of interest. Hence, unless there is a specific reason not to consider enrichment or depletion, the adequate alternative hypothesis is $H_a = E/D$, i.e. two-sided tests are appropriate.

5 SUMMARY AND DISCUSSION

To summarize, there is a single exact null distribution of N_{11}, the hypergeometric distribution, but different exact tests (exact in the sense that they are based on the exact null distribution), one or two-sided, and with several definitions of the *P*-value in the latter case. These tests can equally be called hypergeometric or Fisher's exact tests[2]. Thus, it is not justified to claim, as Masseroli *et al*. (2004) do, that 'the χ^2 and Fisher's exact tests have less power than the hypergeometric and binomial distribution tests'. GFINDER and GOToolBox propose the hypergeometric test and Fisher's exact test as two alternative options: GFINDER indeed provides the same results for the two options (one-sided tests), but strangely enough, GOToolBox gives different results, whereas they should be identical for the same choice of a *P*-value (incorrect results given by some GO tools are detailed in the Appendix, which is available as supplementary data).

The available GO tools often do not explicitly state which *P*-value is computed. For example, BINGO calls the test it performs 'hypergeometric test' (Maere *et al*. 2005), without saying that it is two-sided with the minimum-likelihood approach. According to both references and websites, we could establish that FuncAssociate, GFINDER and THEA provide only one-sided tests in both directions, while FuncSpec, EASEonline, GO Term Finder (CPAN), Term Finder (SGD), GOTM, L2L, Ontology Traverser and STEM only one-sided enrichment tests and that BINGO, DAVID, eGOn, 2004 FatiGo, GeneMerge, GoMiner, GOstat, GoSurfer, NetAffx and Onto-Express provide two-sided tests, the *P*-values being computed according to the minimum-likelihood approach when a discrete distribution is used.

As discussed in section 4.3, two-sided tests are usually most appropriate. Be it with the doubling or the minimum-likelihood approach to the *P*-value, the discreteness and conservatism effects can be efficiently dealt with using mid-*P*-values, a possibility that is not offered by any of the GO tools of Table 1.

[2] As a matter of fact, (Fisher, 1935) describes a one-sided test in the direction of the observed departure of the null hypothesis.

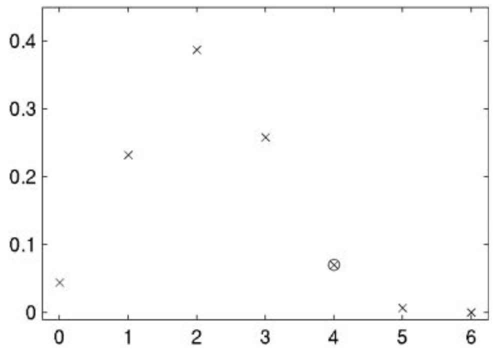

Fig. 1. Hypergeometric null distribution Hyper(20, 7, 6) (crosses). The observed value is $n_{11} = 4$ (circle).

6 NUMERICAL ILLUSTRATIONS

6.1 Small sample

We consider a small sample with $n = 20$, $n_{1+} = 7$, $n_{+1} = 6$ and $n_{11} = 4$, i.e. $f_1 = 0.57$ and $f_2 = 0.15$. The null distribution of $N_{11} \sim$ Hyper(n, n_{1+}, n_{+1}) is shown in Figure 1. The sample being very small, we consider only the tests based on this exact distribution.

6.1.1 One-sided test For illustration purposes, let us first consider a one-sided test (suppose one is interested in enrichments only). The corresponding one-sided *P*-value right equals:

$$\begin{aligned}p_{\text{one}}(4) &= P(N_{11} \geq 4) = P(4) + P(5) + P(6)\\ &= 7.04\times10^{-2} + 7.04\times10^{-3} + 1.81\times10^{-4} = 7.77\times10^{-2}.\end{aligned}$$

The one-sided mid-*P*-value is:

$$\begin{aligned}\pi_{\text{one}}(4) &= P(N_{11} > 4) + P(4)/2\\ &= 7.04\times10^{-3} + 1.81\times10^{-4} + 7.04\times10^{-2}/2 = 4.24\times10^{-2}.\end{aligned}$$

There is a substantial difference between the *P*-value and the mid-*P*-value. With a significance level $\alpha = 5\%$, the mid-*P*-value leads to reject H_0, whereas the *P*-value does not: the use of a mid-*P*-value corresponds to a less conservative test. However, the actual significance level is no longer guaranteed to be smaller than the nominal significance level 5%.

6.1.2 Two-sided tests The two-sided doubling *P*-value equals:

$$\begin{aligned}p_{\text{two}}^{\text{doubling}}(4) &= 2\times\min\left(P(N_{11} \leq 4), P(N_{11} \geq 4)\right) = 2\times P(N_{11} \geq 4)\\ &= 2\times p_{\text{one}}(4) = 1.55\times10^{-1}.\end{aligned}$$

The two-sided doubling mid-*P*-value equals:

$$\begin{aligned}\pi_{\text{two}}^{\text{doubling}}(4) &= 2\times\min\left(P(N_{11} < 4) + P(N_{11} = 4)/2,\right.\\ &\qquad \left.P(N_{11} > 4) + P(N_{11} = 4)/2\right)\\ &= 2\times\left(P(N_{11} > 4) + P(N_{11} = 4)/2\right)\\ &= 2\times\pi_{\text{one}}(4) = 8.49\times10^{-2}.\end{aligned}$$

As for the one-sided test, there is a substantial difference between the two values. Also, with a significance level $\alpha = 5\%$, a two-sided test does not reject H_0.

The two-sided minimum-likelihood *P*-value equals:

$$\begin{aligned}p_{\text{two}}^{\text{min lik}}(4) &= \sum_{P(N_{11}=m)\leq P(4)} P(N_{11} = m) = P(0) + P(4) + P(5) + P(6)\\ &= 4.43\times10^{-2} + 7.04\times10^{-2} + 7.04\times10^{-3} + 1.81\times10^{-4}\\ &= 1.22\times10^{-1}.\end{aligned}$$

The hypergeometric distribution being here asymmetric, the doubling and minimum-likelihood *P*-values are quite different.

The two-sided minimum-likelihood mid-*P*-value equals:

$$\begin{aligned}\pi_{\text{two}}^{\text{min lik}}(4) &= \sum_{P(N_{11}=m)<P(4)} P(N_{11} = m) + P(N_{11} = 4)/2\\ &= P(0) + P(5) + P(6) + P(4)/2 = 8.67\times10^{-2}.\end{aligned}$$

It is always smaller than the *P*-value, and hence corresponds to a less conservative test.

6.2 Large sample

We now consider a sample whose size is analogous to that of samples encountered when testing enrichment of GO categories among DE genes on dedicated microarrays. We have $n = 800$, $n_{1+} = 40$, $n_{+1} = 100$, and observe $n_{11} = 10$, i.e. $f_1 = 0.25$ and $f_2 = 0.12$. The alternative hypothesis is $H_a = E/D$ (two-sided test).

- The exact two-sided doubling *P*-value obtained with the hypergeometric distribution is $p_{\text{two}}^{\text{doubling}}(n_{11}) = 3.95\times10^{-2}$, and the two-sided mid-*P*-value is $\pi_{\text{two}}^{\text{doubling}}(n_{11}) = 2.66\times10^{-2}$. With the minimum-likelihood approach, $p_{\text{two}}^{\text{min lik}}(n_{11}) = 2.39\times10^{-2}$, and the two-sided mid-*P*-value is $\pi_{\text{two}}^{\text{min lik}}(n_{11}) = 1.74\times10^{-2}$. Note that, the null distribution being asymmetric, there is a noticeable difference between the two approaches, and, though the sample is quite large, between the *P*-values and the corresponding mid-*P*-values.
- The approximate binomial test leads to a doubling *P*-value of 4.54×10^{-2}, and to a doubling mid-*P*-value of 3.11×10^{-2}, to a minimum-likelihood *P*-value of 2.75×10^{-2}, and to a minimum-likelihood mid-*P*-value of 2.03×10^{-2}. Note that though the sample is not small, there is quite a difference with the exact distribution.
- The approximate test of equality of two probabilities leads to the value of an approximately normal statistic $z = 2.45$, and to a two-sided *P*-value of $p_{\text{two}}(z) = 1.42\times10^{-2}$. This value is even less accurate than that obtained with the binomial approximation, because the DE sample is too small ($n_{1+} = 40$).
- The χ^2-test indeed leads to a statistic value $d^2 = 6.015 = z^2$, and hence to the same two-sided *P*-value.

In the case of larger samples, obtained with mouse or human pangenomic microrrays, typically with n of the order of 25 000:

- The approximate binomial test leads to (mid-) *P*-values that are very close to those of the exact hypergeometric test. However, with todays computing means, there is no decisive advantage in performing this approximation (see next section).
- The approximate test of equality of two probabilities becomes closer to the exact one only if the number of DE genes is large,

which is not necessarily the case. There is thus no reason to use this test.

- This is hence also true for the equivalent χ^2 test.

6.3 Implementation with R and computational issues

All the exact tests can be implemented 'by hand' with the hypergeometric cumulative distribution function 'phyper' and the distribution function 'dhyper', and the binomial approximations with 'pbinom' and 'dbinom'[3].

The default implementation of the exact test with R provides the two-sided minimum-likelihood P-value. The corresponding instruction is 'fisher.test(c)', where the matrix c is the 2 × 2 contingency table [n_{11} n_{12}; n_{21} n_{22}]. The one-sided enrichment test is obtained with 'fisher.test(c, alternative = ''greater'')', the one-sided depletion test with 'fisher.test(c, alternative = ''less'')'.

In order to evaluate the computation time of the two-sided tests, let us consider the case of a microarray with n = 25 000 genes, n_{1+} = 1000 DE genes, and 500 different GO categories. We take n_{+1} uniformly distributed in [0,n], and n_{11} uniformly distributed in [max(0, n_{+1}+n_{1+}–n), min(n_{1+}, n_{+1})]. With R 2.1.0 running under Mac OS X on a 2 GHz two processor Macintosh (PowerPC 970 2.2), we obtain the following total elapsed times (mean and standard error on 20 runs) for the doubling approach:

- hypergeometric doubling P-values, computed with the functions 'dhyper' and 'phyper': 0.17 ± 0.02 s, and 0.20 ± 0.02s for the mid-P-values.
- binomial doubling P-values, computed with the functions 'dbinom' and 'pbinom': 0.16 ± 0.02s, and 0.19 ± 0.02s for the mid-P-values.

Hence, the gain in time obtained by using the binomial approximation to the hypergeometric distribution is negligible.

For the minimum-likelihood approach, the R function 'fisher. test', (which does not only compute a P-value) is much slower than a computation 'by hand':

- hypergeometric minimum-likelihood P-values, computed with the function 'fisher.test': 17.15 ± 0.21 s.
- hypergeometric minimum-likelihood P-values, computed with the functions 'dhyper' and 'phyper': 1.83 ± 0.04s and 2.10 ± 0.05s for the mid-P-values.

The computation time is hence an argument in favor of the doubling approach to the two-sided P-value.

7 CONCLUSION

The correct statement of the enrichment and/or depletion testing problem leads to a unique exact null distribution of the number of DE genes belonging to the GO category of interest, given the total gene number and the total number of genes belonging to the GO category. This distribution is the hypergeometric one, whose values are equivalently given by Fisher's formula for a 2 × 2 contingency table. Since both enrichments and depletions of GO categories are potentially of interest, two-sided tests are generally most appropriate. With the doubling or the popular minimum-likelihood definitions of the P-value, a loss of power due to the discreteness of the hypergeometric distribution is efficiently dealt with using mid-P-values, the doubling P-value involving lighter computations than the minimum-likelihood P-value. Finally, since many dedicated microarrays involve small data sets, and given the currently available algorithms and computing means, there is no strong argument in favor of the approximate large sample tests.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by the CNRS and the city of Paris.

Conflict of Interest: none declared.

REFERENCES

Agresti,A. (1992) A survey of exact inference for contingency tables. *Stat. Sci.*, **7**, 131–177.

Agresti,A. and Min,Y. (2001) On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*, **57**, 963–971.

Agresti,A. (2002) *Categorical Data Analysis. 2nd edn.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Agresti,A. (2006) Reducing conservatism of exact small-sample methods of inference for discrete data. *Compstat 2006, 17th Symposium of the IASC,* Rome 28 August—1 September 2006.

Al-Sharour,F. *et al.* (2004) FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Beißbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within & group of genes. *Bioinformatics*, **20**, 1464–1465.

Boyle, E.I. *et al.* (2004) GO: TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.

Castillo-Davis,C.I. and Hartl,D.L. (2003) GeneMerge–post-genomics analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.

Cheng,J. *et al.* (2004) NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.

Dennis,G., Jr *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, R60.

Draghici,S. *et al.* (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.

Dunne,A. *et al.* (1996) Two-sided *P*-values from discrete asymmetric distributions based on uniformly most powerful unbiased tests. *The Statistician*, **45**, 397–405.

eGOn Reference Manual (2004).

Ernst,J. *et al.* (2005) Clustering short time series gene expression data. *Bioinformatics*, **21** (Suppl. 1), i159–i168.

Fisher,R.A. (1935) The logic of inductive inference. *J. Royal Stat. Soc.*, **98**, 39–54.

Gibbons,J.D. and Pratt,J.W. (1975) *P*-values: interpretation and methodology. *Am. Stat.*, **29**, 20–25.

Hosack,D.A. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Khatri,P. *et al.* (2002) Profiling gene expression utilizing onto-express. *Genomics*, **79**, 266–270.

Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Lehman,E.L. (1986) *Testing Statistical Hypotheses. 2nd edn.* Springer-Verlag, New York, LLC.

Maere,S. *et al.* (2005) BiNGO: a Cytoscape plugin to assass overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, **21**, 3448–3449.

Martin,D. *et al.* (2004) GOToolbox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.

Masseroli,M. *et al.* (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysisn and mining. *Nucleic Acids Res.*, **32**, W293–W300.

[3] The code of the R functions can be found at the R project site https://svn.r-project.org/R/trunk/src/nmath/. The best known and most complete software for contingency table methods in general is StatXact (Agresti, 2006).

Mehta,C.R. and Patel,N.R. (1998) Exact inference for categorical data. In: Armitage,P. and Coltin,T. (eds), *Encyclopedia of Biostatistics*, Vol, **2**, Wiley, Chichester, UK, pp. 1411–1422.

Mood,A.M. *et al*. (1974) *Introduction to the Theory of Statistics, 3rd edn*. McGraw-Hill. International Edition.

Newman,J.C. and Weiner,A.M. (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol*., **6**, R8.

Pasquier,C. *et al*. (2004) THEA: ontology-driven analysis of microarray, *Bioinformatics*, **20**, 2636–2643.

Robinson,M.D. *et al*. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.

Shah,N.H. and Fedoroff,N.V. (2004) CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, **20**, 1196–1197.

Yates,F. (1984) Test of significance for 2x2 contingency tables. *J. Royal Stat. Soc. Series A*, **147**, 426–463.

Young,A. *et al*. (2005) Ontology Traverser: an R package for GO analysis. *Bioinformatics*, **21**, 275–276.

Zeeberg,B.R. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.

Zhang,B. *et al*. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of iinteresting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.

Zhong,S. *et al*. (2004) GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in gene ontology space. *Appl. Bioinformatics*, **3**, 261–264.