Comparative Enumeration Gene Expression

Leonardo Varuzza* Carlos Alberto de Bragana Pereira*

19 June 2008

Abstract

This paper is about differential gene expression measured by transcript counting methods such as SAGE or MPSS. It introduces two significance tests for detection of differential expressed tags: frequentist and Bayesian. Under the frequentist view, it is proposed a test that computes the critical level as a function of each tag total frequency. Under the Bayesian view the Full Bayesian Significance Test is used considering the logistic normal distribution. The two proposed significance levels, the frequentist and the Bayesian, are compared for a data set with four libraries. The linking function between them is a Beta distribution function with mean 0.39 and standard deviation 0.30.

1 Introduction

Gene expression is an important measure that helps to elucidate the genes functions in the biological processes. Specially in complex diseases, like cancer or Alzheimers, the knowledge of transcriptopme can helps in diagnose, prognostic and treatment. Discovering a gene that differentiates its expression between case and control is one of the molecular genetics main objectives. In the last twenty years, several methods were developed to measure the expression level of multiple thousands – genes at once. Many of those methods are based on transcript enumeration, such as SAGE[15] and MPSS[4].

The present work presents two new exact significance tests for tag differential expression detection in transcript counting data. It introduces, for the frequentist test, a critical level that is a function of the tag expression power, the total tag frequency among all studied libraries.

There are several statistical tests with the same objective: comparing gene expression. Most of then rely on asymptotic considerations. However, since there are drastic differences in expression power among tags, the appropriateness level in using these asymptotic methods varies between tags. There is also the use, in same data set, of different methods: some exact and some asymptotic. In those cases there are distinct criteria to judge a tag quality for differentiation. We notice that here we establish methods that judges all tags using the

*Bioinformatics Nucleus - University of São Paulo

same criterion. Moreover, we do not need to make any use of asymptotic considerations. In fact, we use the original definition of tail area to calculate our *p-value* and *e-value*: **p** for probability and **e** for evidence. The *p-value* is for the frequentist significance test, and the e-value for the Bayesian significance test.

2 Frequentist Significance Test

The main problem addressed by this paper is the detection of tags that are differentially expressed among k (> 1) libraries known to be collected from individuals under different conditions. The total number of distinct tags is represented by M and the total number of tags in library j ($j = 1, 2, \cdots, k$) is represented by N_j. The frequency of the i-th tag in the j-th library is denoted by X_{ij}. Hence, $Nj = X_{1j} + X_{2j} + ... + X_{Mj}$. By power of the i-th tag we mean the total frequency of the i-th tag on all k libraries, that is, $Y_i = X_{i1} + X_{i2} + \cdots + X_{ik}$. The basic statistical model can be stated as:

1. For each j, the random vector $\mathbf{X}_{\bullet j} = (X_{1j}; X_{2j}; \cdots; X_{Mj})$ is multinomial distributed with parameters N_j and $\mathbf{P}_j = (p_{1j}; p_{2j}; \cdots; p_{Mj})$.

2. The random vectors $\mathbf{X}_{\bullet 1}, \mathbf{X}_{\bullet 2}, \cdots$, and $\mathbf{X}_{\bullet k}$ are mutually statistical independent. That is, we consider that the libraries are independently collected.

3. Since low frequency models have been considered in the literature for problems like the ones presented here, the tag frequency X_{ij} is approximated by a Poisson distribution[2]. That is, considering each tag j alone, it is not absurd to use the approximation of the binomial to the Poisson distribution. In other words we have that X_{ij} is approximately distributed as Poisson with mean $N_j p_{ij}$.

4. Finally, since the libraries are collected independently, the full model considered here is as follows:

$$\Pr\{X_{i1} = x_{i1}, \cdots, X_{ik} = x_{ik} | p_{i1}, \cdots, p_{ik}\} = \frac{(N_1 p_{i1})^{x_{i1}} \cdots (N_k p_{ik})^{x_{ik}}}{x_{i1}! \cdots x_{ik}!} \exp(N_1 p_{i1} + \cdots + N_k p_{ik}) \quad (1)$$

2.1 Partial Likelihood

One can now write equation (1) using an alternative parametrization. Let the new parameters be $\theta_i = N_1 p_{i1} + \cdots + N_k p_{ik}$, $\pi_{ij} = N_j p_{ij}/\theta_i$ and $\Pi_i = (\pi_{i1}; \cdots; \pi_{ik})$. Note also that, taking $y_i = x_{i1} + \cdots + x_{ik}$, the following two events are equivalent:

$$\{X_{i1} = x_{i1}, \cdots, X_{ik} = x_{ik}\} \equiv \{X_{i1} = x_{i1}, \cdots, X_{ik} = x_{ik}; Y_i = y_i\}.$$

Therefore, the alternative statistical model is then written as:

$$\Pr\{X_{i1} = x_{i1}, \cdots, X_{ik} = x_{ik}, Y_i = y_i | \Pi_i, \theta_i\} = \frac{y!}{x_{i1}! \cdots x_{ik}!} (\pi_{i1})^{x_{i1}} \cdots (\pi_{ik})^{x_{ik}} \frac{\theta_i^y e^{-y}}{y!} \quad (2)$$

With this new parametrization, the full likelihood is a product of a multinomial by a Poisson probability functions. Also, it is important to note that the new parameters, Π_i and θ_i, are of independent variation; i.e., the value of one carries no information about the value of the other. Following Basu[3] and Cox[5], to perform inference about Π_i (θ_i) one only has to consider as the likelihood the multinomial (Poisson) factor of equation (2). With this new parametrization the null hypothesis of interest – tag i have the same expression in all libraries – is reduced to a simple hypothesis as following:

$$H_0' : \Pi_i = (\pi_{i1}, \cdots, \pi_{ik}) = \left(\frac{N_1}{N}, \cdots, \frac{N_k}{N}\right) \quad (3)$$

Recall that for the full likelihood models, the original hypothesis is, equivalently, as follows:

$$H_0 : \Pi_i = (p_{i1}, \cdots, p_{ik}) = (p_i, \cdots, p_i) \quad (4)$$

This approach of partial likelihood introduced by Cox[5] simplifies considerably the problem of comparing the expression of the j-th tag in all k libraries. Hence, under the null hypothesis the likelihood is simple a multinomial probability function evaluated for Π_0. In symbols, letting $Y_i = X_{i1} + \cdots + X_{ik}$ and $X_{i\bullet} = (X_{i1}, \cdots, X_{ik})$, the distribution under the null and alternative hypotheses are

$$H_0' : \Pr\{X_{i\bullet} = (x_1, \cdots, x_k) | Y_i = y_i; \Pi_i = \Pi_0\} = \frac{y_i!}{N^{y_i}} \prod_{j=1}^{k} \frac{N_j^{x_j}}{x_j!} \quad (5)$$

and

$$H_1' : \Pr\{X_{i\bullet} = (x_1, \cdots, x_k) | Y_i = y_i; \Pi_i\} = y_i! \prod_{j=1}^{k} \frac{\pi_{ij}^{x_j}}{x_j!} \quad (6)$$

2.2 Significance Test: *p-value*

According to Cox[6] and Kempthorne[10], a significance test is a method that measures the consistency of the data with the null hypothesis. The common index used to perform this task is the well known *p-value*. We refer to Kempthorne and Folks[11] for important discussions on the evaluation of *p-values*. For an experiment that observes the value of a random vector $\mathbf{X}$, let $T(X) = T$ be a statistic that small values of it cast doubt about H_0. If for an observation x the value of $T(x) = t$, the *p-value* associated to the observation x is the value of the probability, under H_0, of the event $\{T \leq t\}$; that is, $p = \Pr\{T \leq t | H_0\}$. The

consequence of this definition is that the random variable T must be a function that produces an order in the sample space. This ordered sample space indicates that sample points with low (high) order favor the alternative (null) hypothesis. The difficult is that most of the sample spaces have dimension higher than one, as in the case of the present paper.

It seems that the really appropriate candidate for ordering the sample space relative to the null hypothesis is the likelihood ratio. Recall that the likelihood ratio for an observation x is the maximum of the likelihood under H_0, the null hypothesis, divided by the maximum under H_1, the alternative hypothesis. Clearly, small (large) values of the likelihood ratio favor the alternative (null) hypothesis.

If R is the likelihood statistics and $R(x) = r$ is the value of that statistic at the observation x, the *p-value* associated to x in relation to H_0 is $\Pr\{R \leq r|H_0\}$. Since a sample point w, such that $R(w) < R(x)$, favors H_1 more than x does, it should not be difficult to prove that $\Pr\{R \leq r|H_0\} < \Pr\{R \leq r|H)1\}$. This is a desired property of a *p-value*.

The use of likelihood ratios for computing *p-values* was discussed by Neyman-Pearson[12], Pereira and Wechsler[13], and Dempster[9]. The next section presents the steps to compute the *p-value* described here. We end this section with the likelihood function used in this paper. Let a multinomial sample point for the i-th tag be represented by $\mathbf{W} = (w_1, \cdots, w_k)$, for which w_is are non negative integers and $w_1 + \cdots + w_k = y_i$. Taking $N = N_1 + \cdots + N_k$, the likelihood associated to $\mathbf{W}$ is as follows:

$$R_i(\mathbf{W}) = \left(\frac{y_i}{N}\right)^{y_i} \prod_{j=1}^{k} \left(\frac{N_j}{w_j}\right)^{w_j} \tag{7}$$

Define now the tail set, T_i, of extremer frequencies than $\mathbf{x}_i$. This set can be written as

$$T_i = \{\mathbf{w} = (w_1, \cdots, w_k)|w_1 + \cdots + w_k = y_i \vee R(\mathbf{w}) \leq R(\mathbf{x}_i)\}.$$

The *p-value* of the tag i, pv_i, that provides the significance test for H_0 when the observation is $\mathbf{x}_i$, is evaluated as follows:

$$pv_i = \sum_{w \in T_i} \frac{y_i!}{N^{y_i}} \prod_{j=1}^{k} \frac{N_j^{w_j}}{w_j!} \tag{8}$$

For the computation of pv_i the algorithm based in Monte Carlo Method is in figure 1.

2.3 Critical significance level

As mentioned before, the tag abundances can be very different among the tags. Considering the same significance level would be unfair for the tags with low frequencies. Following the recommendations of DeGroot[8], we use the decision

```
P-VALOR(X, N, runs)
 1  t ← R(X, N)
 2  y ← ∑ X_i
 3  p ← N/∑ N_i
 4  c ← 0
 5  for i ← 1 to runs
 6  do W ← Random vector with Multi(y, p)
 7     if R(W, N) ≤ t
 8        then c ← c + 1
 9  return c/runs
```

Figure 1: Algorithm for p-value computation.

theory optimum procedure that minimizes the risk function $a\alpha + b\beta$, a linear combination of the two kinds of errors: α and β are, respectively, the first and the second kind of errors. After a long discussion with molecular scientists our choice for the scalars of the combination are a=4 and b =1. With this choice the minimization of α is stronger than that of β. We believe that the first kind of error, deciding in favor of differentially expressed when it is not, is more dangerous than the second kind of error, deciding against differentially expressed when it is.

The value of α is simply the value of the probability of the critical region using the parameter value defined by the null hypothesis. However, the computation of β is not so simple since the alternative hypothesis H_1 is composed, not a single point hypothesis. To solve the problem of defining the appropriate β, we consider the average of all possible single alternative hypotheses inside the set that defines the alternative hypothesis. To perform this computation we use a uniform prior for the parameter and consider the predictive distribution for this prior choice. Fortunately, it happens that this predictive distribution is a uniform discrete distribution in the sample space. Hence, a constant that is equal to the inverse of the number of points of the sample space. The mentioned average of β is then the number of points inside the acceptance region divided by this constant.

To choose the critical level we consider all possible critical regions. The critical level is then the value of α for the critical region that gives the smallest value of $4\alpha + \beta$.

To establish the function of y that gives the approximate α, we consider the pairs $\{L = Log(y); Log(\alpha)\}$ and use the least squares method piecewisely in two difference regions of y values: [1;50] and [51;10,000] . Here the αs are the ones obtained as above. Figure 1 illustrates the adjusted functions for the cases of 2, 3, 4, 5 and 6 libraries. The low region adjusts a second degree polynomial $a_k L^2 + b_k L + c_k$ and the high region adjusts a line $u_k + v_k L$. Table 1 presents the coefficient values for those functions. Figure 2 shows their adjustment to the calculated points.

	a_k	b_k	c_k	u_k	v_k
k=2	0.00957978	-0.463118	-2.76474	-2.37781	-0.530119
k=3	-0.304365	1.18976	-4.60784	-0.713611	-0.968513
k=4	-0.931159	5.00318	-10.1863	0.385118	-1.28105
k=5	-0.685327	3.39467	-7.59502	1.47602	-1.57657
k=6	-0.914225	4.84175	-9.81444	1.93518	-1.70783

Table 1: Coefficients values of fitted critical level functions.

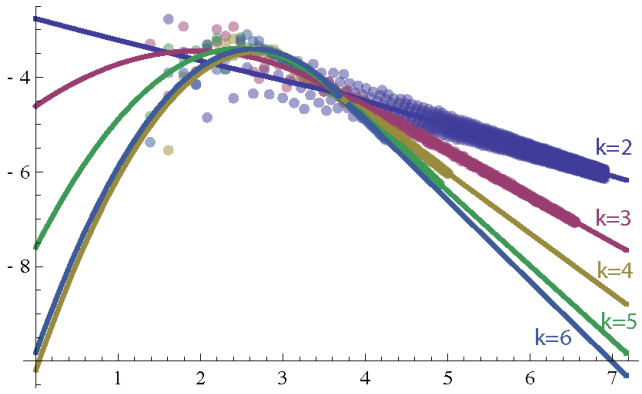

Figure 2: Dilog graphic with the simulated values of the critical level, and the fitted functions.

We also define a score for the tags with $pv < \alpha$: $S = 10\left(1 - \frac{pv}{\alpha}\right)$, which is a practical device to order the tags by the ones which the *p-value* are more distant from the critical level.

3 Bayesian Significance Test: e-value

The Bayesian test is defined on the original libraries parameters $\mathbf{P}_1, \cdots$, and $\mathbf{P}_k$. As a prior for these parameters we consider independent and identically distributed Dirichlet with meta parameter $\alpha_1, \cdots, \alpha_M$. Consequently, the posterior distribution for each independent parameter $\mathbf{P}_j$ is Dirichlet with meta parameters $x_{1j}+\alpha_1, \cdots, x_{Mj}+\alpha_M$[1].For this model the posterior marginal density for the parameter p_{ij} is $Beta(x_{ij}+\alpha_i, N_j+S-x_{ij}-\alpha_i)$, with $S = \alpha_1+\cdots+\alpha_n$, and, from independency, the joint probability of all libraries for one tag is a product of these Beta densities.

The FBST, Full Bayesian Significance Test, was introduced in Pereira & Stern[7] and with a long review in Pereira, Strern & Wechsler[14]. To perform this test one needs two numerical procedures: optimization and integration. The objective is to obtain an alternative index to *p-values*, namely *e-values*: **p** for probability and **e** for evidence. Both indexes are numbers between zero and one, since they are probabilities; the *p-value* on the sample space and the

e-value on the parameter space.

Due to the difficulties caused by the low values of $x_{ij} + \alpha_i$, compared with the high values of $N_j + S - x_{ij} - \alpha_i$, we consider the transformation of the betas to the logistic-normal distributions[1]. The problem becomes more tractable since we have to integrate regions of a joint distribution of independent normal variables. In addition, the parameters of the normal distributions involved avoid numerical approximation problems that would appear when using directly the beta distributions. We recall here that the mean and the variance of the normal distribution obtained after the logistic transformation, $\zeta_{ij} = \log[p_{ij}/(1 - p_{ij})]$, are the digamma for the mean and the trigamma for the variance. In other words, the mean and the variance are, respectively:

$$\mu_{ij} = \Psi(x_{ij} + \alpha_i) + \Psi(N_j - S - x_{ij} - \alpha_i) \tag{9}$$

$$\sigma^2_{ij} = \Psi(x_{ij} + \alpha_i) - \Psi(N_j - S - x_{ij} - \alpha_i) \tag{10}$$

Notice that the null hypothesis $\zeta_{i1} = \cdots = \zeta_{ik}$ is equivalent to the original hypothesis $p_{i1} = \cdots = p_{ij}$. Hence, the test is performed in this normal density replacing the work with the beta density.

4 *p-value* versus *e-value*

In order to illustrate the results of the two tests we calculate both *p-values* and *e-values* for all tags that appears in the data set of Alzheimer Disease (GSE6677 of GEO) with four libraries.

To obtain the relationship between *p-values* and *e-values* we consider *p-value* intervals of length 0.04. In each of those intervals we calculate the weighted average of the *e-values*. The weighing system was defined by the y values. We them obtain the pairs (p,e) where p is the center of the interval and e the weighted average. A Beta distribution function was adjusted to those pairs as a function of p-values. The best fit was the Beta distribution with parameters a=0.66 e b=1.036, corresponding to a beta with mean 0.39 and standard deviation 0.30. The values of the two significance levels and the fitted function are illustrated by Figure 3.

5 Final Remarks

This paper presented two new methods to compare transcript counting libraries. There are other alternative ones in the literature that have been widely used. However, when comparing tags by their differentiating expressions we may have been comparing values obtaining from different methods: χ^2 for highly expressed tags, which make use of asymptotic characteristics, and exact tests for low expressed tags. Both of our methods are exact and have the same methodology for all kinds of expressing tags. We based our calculus on the original definition of extremer sample points. We do not make any use of asymptotic results as the

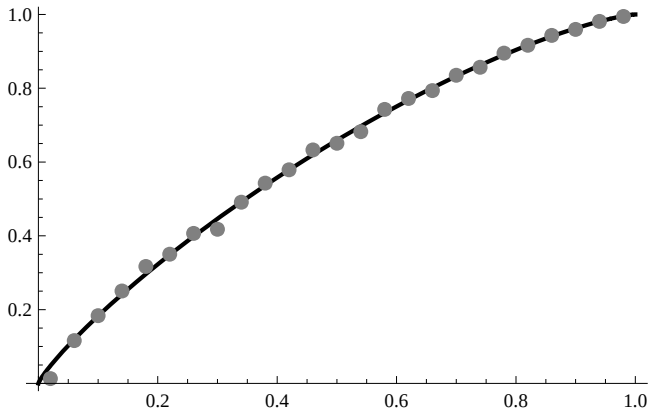

Figure 3: Linking function for *p-value* and *e-value*.

χ^2 test do. Both methods are independent of dimensionality both in parameter and sample spaces.

In the last section we can see that there is no disagreement between the two tests. The differences among their values are due only to the fact that the *p-value* is an integral in the sample space although the *e-value* is an integral in the parameter space. Even with that difference they converge to the same conclusions most of times.

The two methods are implemented in C language and the source code is freely available with GPL license in the internet site `http://code.google.com/p/kempbasu/`

6 Authors' contributions

CAdBP conceived the methods. LV implemented them and performed the analysis. Both authors wrote the paper. Special thanks to Helena Brentani for text revision.

References

[1] J. Aitchison. *The Statistical Annalysis of Compositional Data*, chapter 6, pages 126–128. Chapman and Hall, 2003.

[2] S. Audic and J. M. Claverie. The significance of digital gene expression profiles. *Genome Res*, 7(10):986–95, Oct 1997.

[3] D. Basu. On the elimination of nuisance parameters. *JASA*, Sep 2007.

[4] S. E. Brenner, M. Johnson, J. Bridgham, G. Golda, D. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, et al. Gene expression analysis

by massively parallel signature sequencing (mpss) on microbead arrays. *Nature Biotechnology*, 18:630–634, 2000.

[5] D. Cox. Partial likelihood. *Biometrika*, 62(2):269–276, 1975.

[6] D. Cox. The role of significant test (with discussion). *Scandinavian Journal of Statistics*, 4:49–70, 1977.

[7] C. A. de Bragana Pereira and J. M. Stern. Evidence and credibility: Full bayesian significance test for precise hypotheses. *Entropy*, 1(4):99–110, 1999.

[8] M. DeGroot. *Probability and statistics*. Addison-Wesley Boston, 1986.

[9] A. P. Dempster. The direct use of likelihood for significance testing. *Statistics and Computing*, 7, Nov 1997.

[10] O. Kempthore. Of what use are tests of significance and tests of hypothesis. *Communications in Statistics-Theory and Methods*, 5(8):763–777, 1976.

[11] O. Kempthorne and L. Folks. *Probability, Statistics and Data Analysis*. Ames, Iowa: The Iowa State University Press, 1971.

[12] J. Neyman and E. Pearson. On the use and interpretation of certain test criteria for purposes of statistical inference part I. *Biometrika*, 20(1-2):175–240, 1928.

[13] C. Pereira and S. Wechsler. On the concept of p-value. *Braz. J. Prob. Statist*, 7:159–177, 1993.

[14] C. Pereira, S. Wechsler, and J. Stern. Can a Significance Test be Genuinely Bayesian. *Bayesian Analysis*, 2007.

[15] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270(5235):484–487, Oct 1995.