BMC Bioinformatics

BioMed Central

Research article

Basic properties and information theory of Audic-Claverie statistic for analyzing cDNA arrays

Peter Tiňo

Address: School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, UK

Email: Peter Tiňo - P.Tino@cs.bham.ac.uk

Published: 23 September 2009

BMC Bioinformatics 2009, **10**:310 doi:10.1186/1471-2105-10-310

Received: 25 March 2009
Accepted: 23 September 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/310

Abstract

Background: The Audic-Claverie method [1] has been and still continues to be a popular approach for detection of differentially expressed genes in the SAGE framework. The method is based on the assumption that under the null hypothesis tag counts of the same gene in two libraries come from the same but unknown Poisson distribution. The problem is that each SAGE library represents only a single measurement. We ask: Given that the tag count samples from SAGE libraries are extremely limited, how useful actually is the Audic-Claverie methodology? We rigorously analyze the A-C statistic that forms a backbone of the methodology and represents our knowledge of the underlying tag generating process based on one observation.

Results: We show that the A-C statistic and the underlying Poisson distribution of the tag counts share the same mode structure. Moreover, the K-L divergence from the true unknown Poisson distribution to the A-C statistic is minimized when the A-C statistic is conditioned on the mode of the Poisson distribution. Most importantly, the expectation of this K-L divergence never exceeds 1/2 bit.

Conclusion: A rigorous underpinning of the Audic-Claverie methodology has been missing. Our results constitute a rigorous argument supporting the use of Audic-Claverie method even though the SAGE libraries represent very sparse samples.

Background

It is of utmost importance for biologists to be able to analyze patterns of expression levels of selected genes in different tissues possibly obtained under different conditions or treatment regimes. Even subtle changes in gene expression levels can be indicators of biologically crucial processes such as cell differentiation and cell specialization [2]. Measurement of gene expression levels can be performed either via hybridization to microarrays, or by counting gene tags (signatures) using e.g. Serial Analysis of Gene Expression (SAGE) [3] or Massively Parallel Signature Sequencing (MPSS) [4] methodologies. The SAGE procedure results in a library of short sequence tags, each representing an expressed gene. The key assumption is that every mRNA copy in the tissue has the same chance of ending up as a tag in the library. Selecting a specific tag from the pool of transcripts can be approximately considered as sampling with replacement. The key step in many SAGE studies is identification of "interesting" genes, typically those that are differentially expressed under different conditions/treatments. This is done by comparing the number of specific tags found in the two SAGE libraries corresponding to different conditions or treatments. Several statistical tests have been suggested for identifying differentially expressed genes through comparing such digital expression profiles, e.g. [1,2,5,6].

Audic and Claverie [1] were among the first to systematically study the influence of random fluctuations and sampling size on the reliability of digital expression profile data. Typically, cDNA libraries contain a large number of different expressed genes and observing a given cDNA qualifies as a rare event [1]. For a transcript representing a small fraction of the library and a large number N of clones, the probability of observing x tags of the same gene will be well-approximated by the Poisson distribution parametrized by $\lambda \geq 0$.

$$P(X = x \mid \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}. \quad (1)$$

The unknown parameter λ signifies the number of transcripts of the given type (tag) per N clones in the cDNA library. When comparing two libraries, it is assumed that under the null hypothesis of not differentially expressed genes the tag count x in one library comes from the same underlying Poisson distribution $P(\cdot|\lambda)$ as the tag count y in the other library. However, each SAGE library represents a single measurement only. From a purely statistical standpoint resolving this issue is potentially quite problematic. One can be excused for being rather skeptical about how much can actually be learned about the underlying unknown Poisson distribution from a single observation.

The key instrument of the Audic-Claverie approach is a distribution $\tilde{P}(y|x)$ over tag counts y in one library informed by the tag count x in the other library, under the null hypothesis that the tag counts are generated from the same but unknown Poisson distribution. $\tilde{P}(y|x)$ is obtained by Bayesian averaging (infinite mixture) of all possible Poisson distributions $P(y|\lambda')$ with mixing proportions equal to the posteriors $p(\lambda'|x)$ under the flat prior over λ. When the two libraries are of the same size, we obtain [1]:

$$\tilde{P}(y \mid x) = \frac{1}{2^{x+y+1}} \frac{(x+y)!}{x!\,y!}, \quad (2)$$

$$= \frac{1}{2^{x+y+1}} \binom{x+y}{x}. \quad (3)$$

We will refer to $\tilde{P}(y|x)$ as *Audic-Claverie statistic* (A-C statistic) based on counts x and y. Note that $\tilde{P}(y|x)$ is symmetric, i.e. for $x, y \geq 0$, $\tilde{P}(y|x) = \tilde{P}(x|y)$. Audic and Claverie [1] point out that this is a desirable property, since if the counts x, y are related to two libraries of the same size, they should be interchangeable when analyzing whether they come from the same underlying process or not. The A-C statistic $\tilde{P}(y|x)$ can be used e.g. for principled inferences, construction of confidence intervals, statistical testing etc. For further details regarding the derivation and mathematical treatment of the A-C statistic see [1].

Even though there have been further developments in comparison techniques for cDNA libraries (e.g. while Audic and Claverie [1] only deal with two libraries, Stekel et al. [7] suggest an approach to compare gene expressions across multiple cDNA libraries; for links to further approaches see [2]), the Audic-Claverie method has been and still continues to be a popular approach in current biological research, e.g. [8-17], with 427 citations (based on ISI Web of Knowledge), over 100 citations in the past 3 years. Given the widespread use of the Audic-Claverie method, it is somewhat surprising that a rigorous underpinning of the methodology has not yet been fully developed. Audic and Claverie did demonstrate the desirable behavior of their method through Monte Carlo simulations randomly sampling tags based on two experimentally obtained sequence tag distributions [1]. The rate of false alarm, e.g. how often random fluctuations in tag counts are interpreted as significant differences, was small for genes associated with small tag counts and increased for higher tag counts, but never exceeded the significance level of the test. Of course, one may argue that false alarm rate (false positives) is only one side of the story and ideally one would like to minimize both the false positive and false negative rates. The false negative rate quantifies how often significant differences get interpreted as just random fluctuations. However, of equal importance is the issue of why the Audic-Claverie approach seems to be well-behaved, e.g. when compared to an approach based on Ricker's confidence intervals (see [1]). In this contribution, we provide rigorous arguments as to why the Audic-Claverie method can be expected to work well, even though from the purely statistical standpoint one could be excused for being skeptical. We start by assuming that for a given gene there is a hidden (unobserved) underlying Poisson distribution generating the tag counts. We then go beyond simple Monte-Carlo-style verification by rigorously studying *how much* and *in what form* can be actually learned about the distribution in the Audic-Claverie framework, given a single observation provided by a SAGE library. In particular, we ask:

1. How natural is the A-C statistic's representation of the underlying unknown Poisson distribution governing the tag counts?

2. Given that the observed tag count sample is very limited, how well can the Audic-Claverie approach

work, i.e. how well does the A-C statistic capture the underlying Poisson distribution?

Methods

Basic properties of the A-C statistic

In this section we answer the first question posed above. It turns out that the A-C statistic and the underlying Poisson distribution are quite similar in their nature: for any (integer) mean tag count $\lambda \geq 1$, the Poisson distribution $P(\cdot|\lambda)$ has two neighboring modes located at λ and λ - 1, with $P(\lambda|\lambda) = P(\lambda - 1|\lambda)$. When it comes to the observed tag counts, given a count $x \geq 1$, the A-C statistic $\tilde{P}(\gamma|x)$ has two neighboring modes, one located at $\gamma = x$, the other at $\gamma = x - 1$, with $\tilde{P}(x|x) = \tilde{P}(x - 1|x)$. As in Poisson distribution, the values of $\tilde{P}(\gamma|x)$ decrease as one moves away from the modes in both directions.

Theorem 1 *Let x, γ and d be integers with ranges specified below. It holds:*

1. $\tilde{P}(x|x) > \tilde{P}(x + d|x)$ *for any* $x \geq 0$ *and* $d \geq 1$.

2. *For* $x \geq 1$, $\tilde{P}(x|x) = \tilde{P}(x - 1|x)$.

3. $\tilde{P}(x|x) > \tilde{P}(x - d|x)$ *for any* $x \geq 2$ *and* $2 \leq d \leq x$.

Proof:

1. We have

$$\tilde{P}(x+d \mid x) = \frac{1}{2^{2x+d+1}} \frac{(2x+d)!}{x!(x+d)!} = \frac{1}{4^x 2^{d+1}} \frac{\prod_{i=x+1}^{2x+d} i}{(x+d)!}.$$

In particular,

$$\tilde{P}(x \mid x) = \frac{1}{2 \cdot 4^x} \frac{\prod_{i=x+1}^{2x} i}{x!}.$$

Hence,

$$\frac{\tilde{P}(x+x)}{\tilde{P}(x+d|x)} = 2^d \frac{\prod_{j=x+1}^{x+d} j}{\prod_{k=2x+1}^{2x+d} k} = 2^d \prod_{j=1}^{d} \frac{x+j}{2x+j}.$$

Now, for $x \geq 0$, we have

$$\frac{x+j}{2x+j} > \frac{1}{2}.$$

This can be easily seen, as for $j \geq 1$, $2(x + j) > 2x + j$. It follows that

$$\frac{\tilde{P}(x|x)}{\tilde{P}(x+d|x)} > 2^d \prod_{j=1}^{d} \frac{1}{2} = 2^d \frac{1}{2^d} = 1.$$

2. and 3) For $d \leq x$,

$$\tilde{P}(x-d \mid x) = \frac{1}{2^{2x-d+1}} \frac{(2x-d)!}{x!(x-d)!} = \frac{1}{4^x 2^{-d+1}} \frac{\prod_{i=x+1}^{2x-d} i}{(x-d)!}.$$

Hence,

$$\frac{\tilde{P}(x|x)}{\tilde{P}(x-d|x)} = \frac{1}{2^d} \frac{\prod_{i=x+1}^{2x} i}{\prod_{i=x+1}^{2x-d} i} \frac{(x-d)!}{x!} = \frac{1}{2^d} \frac{\prod_{i=2x-d+1}^{2x} i}{\prod_{j=x-d+1}^{x} j} = \frac{1}{2^d} \prod_{j=1}^{d} \frac{2x-d+j}{x-d+j}. \quad (4)$$

If $d = 1$,

$$\frac{\tilde{P}(x|x)}{\tilde{P}(x-1|x)} = \frac{1}{2} \frac{2x}{x} = 1.$$

When $2 \leq d \leq x$, we have for all j such that $1 \leq j \leq d$ - 1,

$$\frac{2x-d+j}{x-d+j} > 2.$$

This follows from $2(x - d + j) < 2x - d + j$, which can be easily verified, since for $j \in \{1, 2, \ldots, d - 1\}$, we have $(j - d) > 2 \cdot (j - d)$.

For $j = d$, we have the equality $(2x - d + j)/(x - d + j) = 2$.

Finally, form (4),

$$\frac{\tilde{P}(x|x)}{\tilde{P}(x-d|x)} > \frac{1}{2^d} \prod_{j=1}^{d} 2 = 1$$

Q.E.D

We have shown that after observing a count x, the A-C statistic expects counts $y = x$ and $y = x - 1$ with the highest and equal probability. The other values of count y are, as one would naturally expect, less probable.

As an illustrative example we show in figure 1 plots of both the A-C statistic $\tilde{P}(y|x)$ and the corresponding Poisson distribution $P(y|\lambda)$ at $\lambda = x$ for two values of x, $x = 10$ and $x = 30$. As a result of Bayesian averaging in the A-C statistic, $\tilde{P}(y|x)$ is less peaked at its modes than the Poisson counterpart $P(y|x)$. However, both $\tilde{P}(y|x)$ and $P(y|x)$ have two modes located at x and $x - 1$.

Information theory of the A-C statistic

We now answer, in the framework of information theory, the second question posed in the 'Background' section. Assume that there is some "true" underlying Poisson distribution $P(y|\lambda)$ (1) over possible counts $y \geq 0$ with unknown parameter λ. In the same process, we first generate a count x and then use the A-C statistic $\tilde{P}(y|x)$ (3) to define a distribution over y, given the already observed count x. We ask: How different, in terms of Kullback-Leibler (K-L) divergence, are the two distributions over y? For the A-C statistic to work, one would naturally like $\tilde{P}(y|x)$ to be sufficiently representative of the true unknown distribution $P(y|\lambda)$. In other words, one would expect $P(y|\lambda)$ and $\tilde{P}(y|x)$ to be close, with the smallest "distance" at $\tilde{P}(y|x = \lambda)$ (for λ integer), that is, when count x is exactly equal to the expected tag count under the Poisson distribution $P(y|\lambda)$. In this section we provide a quantitative answer to the above question and show, perhaps surprisingly, that the "statistical distance" between $P(y|\lambda)$ and $\tilde{P}(y|x)$ is not minimized at $x = \lambda$, but it attains minimum at the mode of $P(y|\lambda)$, i.e. when $x = \lambda - 1$.

First, define the K-L divergence from $P(y|\lambda)$ to $\tilde{P}(y|x)$:

$$D(\lambda, x) = D_{KL}[P(y \mid \lambda) \,||\, \tilde{P}(y + x)] = \sum_{y=0}^{\infty} P(y \mid \lambda) \log \frac{P(y|\lambda)}{\tilde{P}(y|x)}. \quad (5)$$

The divergence $D(\lambda, x)$ has a nice information-theoretic interpretation: When the log is base 2, $D(\lambda, x)$ expresses the number of bits of additional information one needs in order to fully specify $\tilde{P}(y|x)$, provided one has a perfect knowledge of $P(y|\lambda)$. The divergence $D(\lambda, x)$ is non-nega-

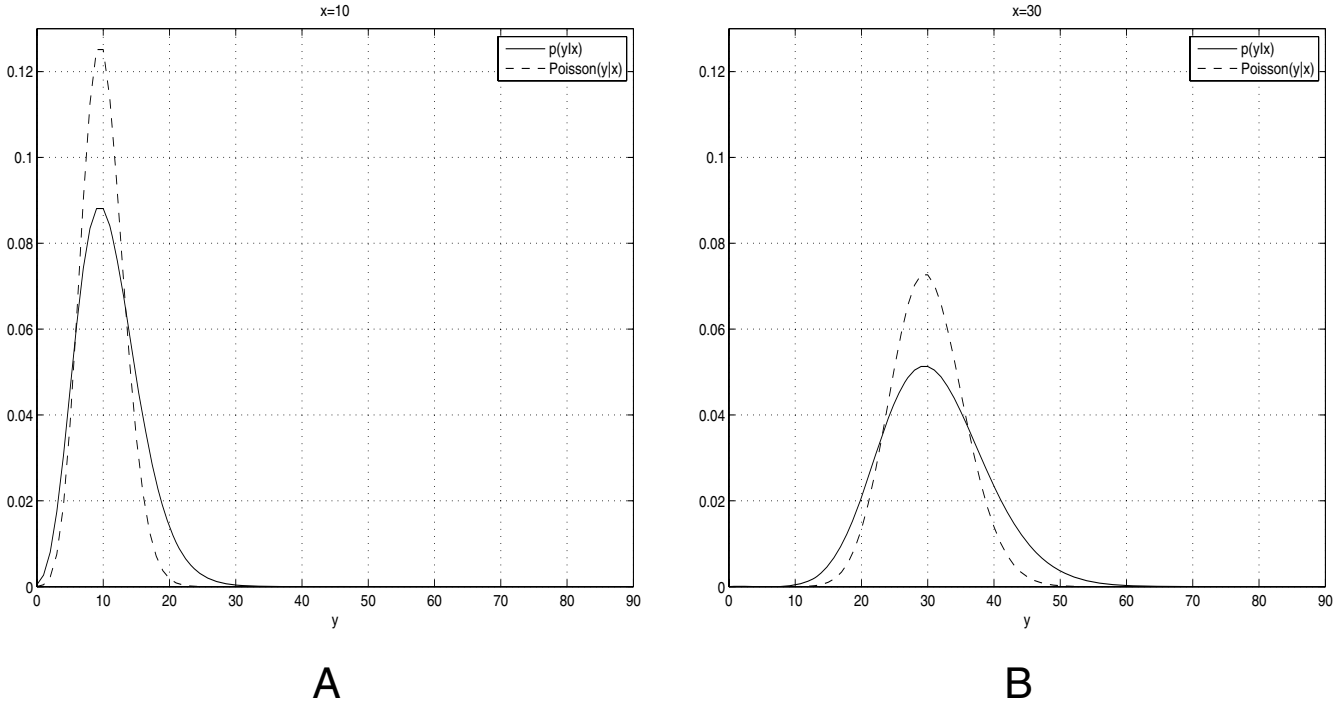

Figure 1

A-C statistic vs. Poisson distribution. Graphs of A-C statistic $\tilde{P}(y|x)$ (solid line) and the corresponding Poisson distribution $P(y|\lambda)$ at $\lambda = x$ (dashed line) for $x = 10$ (A) and $x = 30$ (B).

tive, with $D(\lambda, x) = 0$ if and only if the two distributions $\tilde{P}(y|x)$ and $P(y|\lambda)$ coincide.

Naturally,

$$D(\lambda, x) = -H[P(y \mid \lambda)] - E_{P(y|\lambda)}[\log \tilde{P}(y \mid x)],$$

where

$$\begin{aligned} H[P(y \mid \lambda)] &= -E_{P(y|\lambda)}[\log P(y \mid \lambda)] \\ &= -\sum_{y=0}^{\infty} P(y \mid \lambda) \log P(y \mid \lambda) \end{aligned}$$

is the entropy of $P(y|\lambda)$ and $E_{Q(y)}[f(y)]$ denotes the expectation of the quantity $f(y)$ under the distribution $Q(y)$.

We have

$$\begin{aligned} E_{P(y|\lambda)}[\log \tilde{P}(y \mid x)] = & -\log x! \\ & -(E_{P(y|\lambda)}[y] + x + 1) \log 2 \\ & -E_{P(y|\lambda)}[\log y!] \\ & +E_{P(y|\lambda)}[\log(x + y)!], \end{aligned}$$

and so

$$D(\lambda, x) = -H[P(y \mid \lambda)] + \log x! + (\lambda + x + 1) \log 2 + F(\lambda, 0) - F(\lambda, x), \tag{6}$$

where for each integer $d \geq 0$,

$$\begin{aligned} F(\lambda, d) &= E_{P(y|\lambda)}[\log(y + d)!] \\ &= \sum_{y=0}^{\infty} P(y \mid \lambda) \log(y + d)! \end{aligned} \tag{7}$$

As discussed above, one would intuitively expect $D(\lambda, x)$ to be minimal for $x = \lambda$, as then the conditioning count in the A-C statistic would be the mean of the underlying Poisson distribution. However, the mode of that Poisson distribution, $\lambda - 1$, is surrounded by enough probability mass to yield the following result:

Theorem 2 *For any integer $\lambda \geq 1$, it holds $D(\lambda, \lambda) > D(\lambda, \lambda - 1)$. In other words,*

$$D_{KL}[P(y \mid \lambda) \,||\, \tilde{P}(y \mid \lambda)] > D_{KL}[P(y \mid \lambda) \,||\, \tilde{P}(y \mid \lambda - 1)].$$

Proof: Using (6), we have

$$D(\lambda, \lambda) - D(\lambda, \lambda - 1) = \log \lambda + \log 2 + F(\lambda, \lambda - 1) - F(\lambda, \lambda). \tag{8}$$

Now,

$$F(\lambda, \lambda - 1) - F(\lambda, \lambda) = -E_{P(y|\lambda)}[\log(y + \lambda)]$$

and by Jensen's inequality,

$$\begin{aligned} E_{P(y|\lambda)}[\log(y + \lambda)] &< \log E_{P(y|\lambda)}[y \mid \lambda] \\ &= \log(2\lambda). \end{aligned}$$

By (8), $D(\lambda, \lambda) - D(\lambda, \lambda - 1) = \log(2\lambda) + F(\lambda, \lambda - 1) - F(\lambda, \lambda)$, and since

$$F(\lambda, \lambda - 1) - F(\lambda, \lambda) = -E_{P(y|\lambda)}[\log(y + \lambda)] > -\log(2\lambda),$$

we have $D(\lambda, \lambda) - D(\lambda, \lambda - 1) > 0$, implying $D(\lambda, \lambda) > D(\lambda, \lambda - 1)$.

Q.E.D

We proceed our investigation by asking the following question: Given an underlying Poisson distribution $P(x|\lambda)$, if we repeatedly generated a "representative" count x from $P(x|\lambda)$, what would be the average divergence of the corresponding A-C statistic $\tilde{P}(y|x)$ from the truth $P(y|\lambda)$? In other words, we are interested in the quantity

$$\mathcal{E}(\lambda) = E_{P(x|\lambda)}[D(\lambda, x)]. \tag{9}$$

Lemma 3 *For any $\lambda \geq 0$,*

$$E_{P(x|\lambda)}[F(\lambda, x)] = F(2\lambda, 0).$$

Proof: Employing Malmstén's formula,

$$\log k! = \int_0^{\infty} \left(k - \frac{1 - e^{-kt}}{1 - e^{-t}} \right) \frac{e^{-t}}{t} \, dt, \tag{10}$$

we write

$$\begin{aligned} F(2\lambda, 0) &= E_{P(y|2\lambda)}[\log y!] \\ &= \int_0^{\infty} E_{P(y|2\lambda)}[e^{-yt}] \frac{e^{-t}}{(1 - e^{-t})t} \, dt \\ &\quad + \int_0^{\infty} E_{P(y|2\lambda)}[y] \frac{e^{-t}}{t} \, dt - \int_0^{\infty} \frac{e^{-t}}{(1 - e^{-t})t} \, dt \\ &= \int_0^{\infty} \left(2\lambda - \frac{1 - e^{2\lambda(e^{-t} - 1)}}{1 - e^{-t}} \right) \frac{e^{-t}}{t} \, dt. \end{aligned} \tag{11}$$

The last equality follows from $E_{P(\gamma|2\lambda)}[\gamma] = 2\lambda$ and

$$E_{P(\gamma|2\lambda)}[e^{-\gamma t}] = e^{-2\lambda}\sum_{\gamma=0}^{\infty}\frac{(2\lambda\cdot e^{-t})^{\gamma}}{\gamma!} = e^{-2\lambda}e^{2\lambda e^{-t}}. \quad (12)$$

Let us now evaluate

$$E_{P(x|\lambda)}[F(\lambda, x)] = E_{P(x|\lambda)}[E_{P(\gamma|\lambda)}[\log(x+\gamma)!]].$$

Using Malmstén's formula again, we obtain

$$E_{P(x|\lambda)}[E_{P(x|\lambda)}[\log(x+\gamma)!]] = \int_0^{\infty}\left(2\lambda - \frac{1-E_{P(x|\lambda)}[E_{P(\gamma|\lambda)}[e^{-(x+\gamma)t}]]}{1-e^{-t}}\right)\frac{e^{-t}}{t}\mathrm{d}t. \quad (13)$$

Expansion similar to that in (12) leads to:

$$E_{P(x|\lambda)}[E_{P(\gamma|\lambda)}[e^{-(x+\gamma)t}]] = (E_{P(x|\lambda)}[e^{-xt}])^2 = e^{2\lambda(e^{-t}-1)} \quad (14)$$

Plugging (14) into (13) we obtain (11).

Q.E.D

We will now show that up to terms of order $O(\lambda^{-1})$, the expected divergence of A-C statistic $\tilde{P}(\gamma|x)]$ from the true underlying Poisson distribution $P(\gamma|\lambda)$ is equal to (1/2) log 2.

Theorem 4 *Consider an underlying Poisson distribution* $P(\cdot|\lambda)$ *parametrized by some* $\lambda > 0$. *Then*

$$\mathcal{E}(\lambda) = E_{P(x|\lambda)}[D_{KL}[P(\gamma\mid\lambda)\,||\,\tilde{P}(\gamma+x)]] = \frac{1}{2}\log 2 + O\left(\frac{1}{\lambda}\right)$$

Proof: Since

$$\begin{aligned}\mathcal{E}(\lambda) &= -H[P(\gamma\mid\lambda)] \\ &\quad +E_{P(x|\lambda)}[\log x!] \\ &\quad +(\lambda+1)\log 2 + E_{P(x|\lambda)}[x]\log 2 \\ &\quad +F(\lambda,0) - E_{P(x|\lambda)}[F(\lambda,x)] \\ &= -H[P(\gamma\mid\lambda)] \\ &\quad +(2\lambda+1)\log 2 \\ &\quad +2F(\lambda,0) - E_{P(x|\lambda)}[F(\lambda,x)]\end{aligned} \quad (15)$$

and

$$\begin{aligned}-H[P(\gamma\mid\lambda)] &= E_{P(\gamma|\lambda)}[\log P(\gamma\mid\lambda)] \\ &= -\lambda\log e + E_{P(\gamma|\lambda)}[\gamma]\log\lambda - E_{P(\gamma|\lambda)}[\log\gamma!] \\ &= -\lambda\log e + \lambda\log\lambda + F(\lambda,0)\end{aligned} \quad (16)$$

we have

$$\begin{aligned}\mathcal{E}(\lambda) &= \lambda(\log\lambda - \log e + 2\log 2) + \log 2 \\ &\quad +F(\lambda,0) - E_{P(x|\lambda)}[F(\lambda,x)].\end{aligned}$$

By lemma 3,

$$\begin{aligned}\mathcal{E}(\lambda) &= \lambda(\log\lambda - \log e + 2\log 2) + \log 2 \\ &\quad +F(\lambda,0) - E(2\lambda,0).\end{aligned} \quad (17)$$

We next approximate the terms $F(\lambda, 0)$ and $F(2\lambda, 0)$. To that end, note that the entropy $H[P(\gamma|\lambda)]$ can be approximated as [18]

$$H[P(\gamma\mid\lambda)] = \lambda(\log e - \log\lambda) + F(\lambda,0) = \frac{1}{2}\log(2\pi e\lambda) + O(\lambda^{-1}).$$

Hence,

$$F(\lambda,0) = \lambda(\log\lambda - \log e) + \frac{1}{2}\log(2\pi e\lambda) + O(\lambda^{-1}). \quad (18)$$

By the same token

$$F(2\lambda,0) = 2\lambda(\log 2\lambda - \log e) + \frac{1}{2}\log(4\pi e\lambda) + O(\lambda^{-1}). \quad (19)$$

Plugging (18) and (19) into (17) we obtain

$$\mathcal{E}(\lambda) = \frac{1}{2}\log 2 + O\left(\frac{1}{\lambda}\right).$$

Q.E.D

In fact, one can obtain a more precise characterization of the expected divergence $\mathcal{E}(\lambda)$ by using a higher order entropy expansion (for log base 2):

$$\begin{aligned}H[P(\gamma\mid\lambda)] &= \lambda(\log_2 e - \log_2\lambda) + F(\lambda,0) \\ &= \frac{1}{2}\log_2(2\pi e\lambda) - \frac{1}{12\lambda} - \frac{1}{24\lambda^2} + O(\lambda^{-3}).\end{aligned}$$

After expressing $F(\lambda, 0)$ and $F(2\lambda, 0)$ in the style of (18) and (19), respectively, we obtain an expression for the expected divergence measured in bits:

$$\mathcal{E}(\lambda) = \frac{1}{2} - \frac{1}{12\lambda}\left(1-\frac{1}{2}\right) - \frac{1}{24\lambda^2}\left(1-\frac{1}{2^2}\right) + O(\lambda^{-3}). \quad (20)$$

Figure 2 presents values of the expected divergence $\mathcal{E}(\lambda)$ (measured in bits) calculated numerically from the definition (9), as well as their analytical approximation calcu-

lated from (20). As expected, the two curves are in good correspondence, as our approximation is $O(\lambda^{-3})$.

Results of this section suggest that if the true Poisson source $P(\cdot|\lambda)$ is not known, the A-C statistic $\tilde{P}(y|x)$, *based on a single observed tag count realization x* from $P(\cdot|\lambda)$, is on average not further away from the truth $P(y|\lambda)$ than half a bit of additional information. As the mean tag count λ increases, so does the uncertainty in the generating Poisson distribution $P(\cdot|\lambda)$. As a consequence, the average K-L divergence $\varepsilon(\lambda)$ from $P(\cdot|\lambda)$ to the approximating A-C statistic (based on a single realization from $P(\cdot|\lambda)$) gets larger. The average K-L divergence expressed in bits increases with increasing λ from about 0.42 bits to 0.5 bits.

Results and Discussion

The Audic-Claverie method [1] has been and still continues to be a popular approach for detection of differentially expressed genes in the SAGE framework. The method is based on the assumption that under the null hypothesis the tag counts *x, y* in two libraries come from the same but unknown Poisson distribution $P(\cdot|\lambda)$. The problem is that each SAGE library represents only a single measurement. We have rigorously analyzed usefulness of the Audic-Claverie method by investigating the A-C statistic

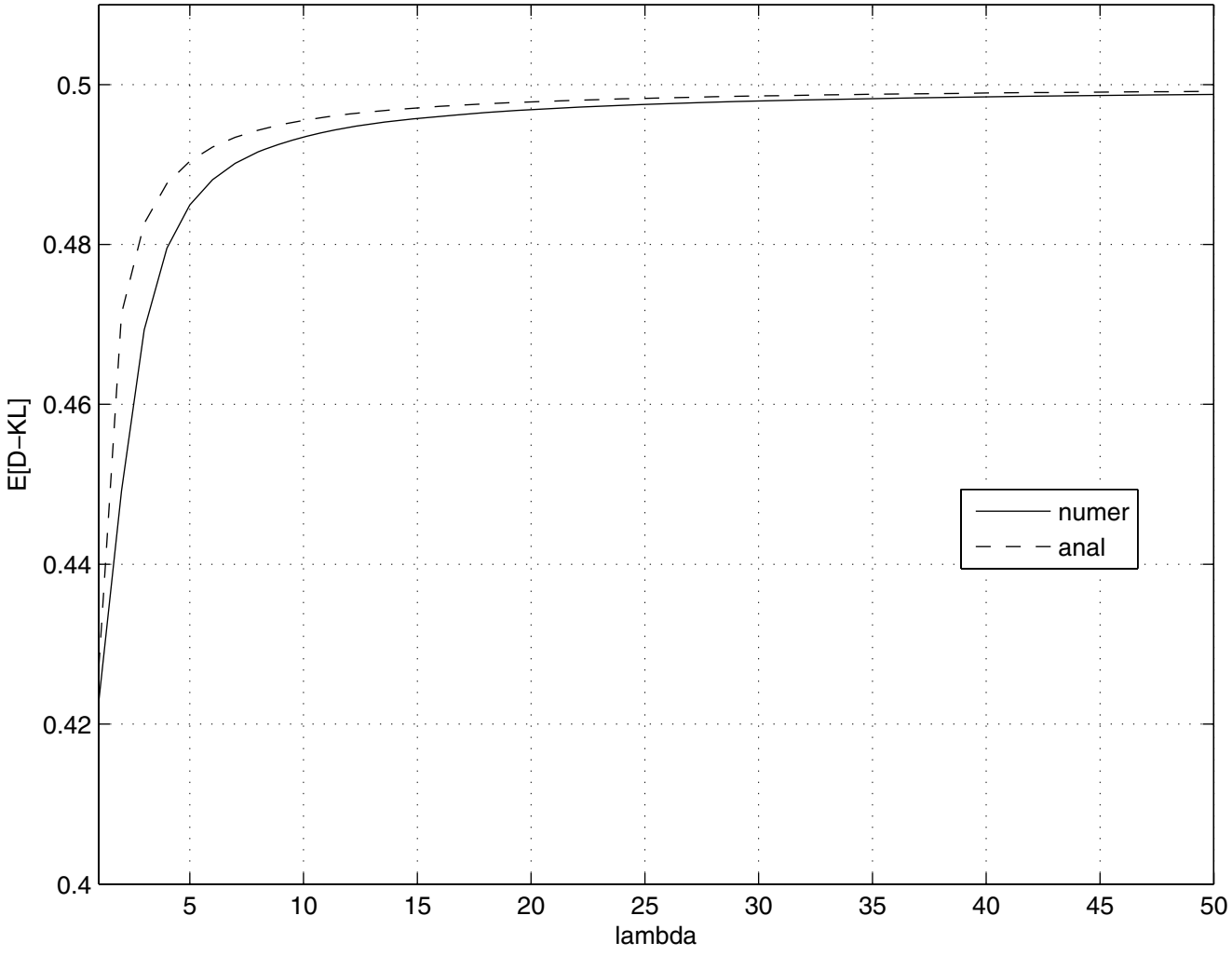

Figure 2
Expected K-L divergence from the underlying Poisson distribution to A-C statistic. Expected K-L divergence $\varepsilon(\lambda)$ (measured in bits) from the true unknown Poisson distribution to the A-C statistic (solid line) and its analytical approximation (20) (dashed line).

$\tilde{P}(y|x)$ that forms a backbone of the method and represents our knowledge of the underlying Poisson distribution $P(\cdot|\lambda)$ based on only one tag count x drawn from it.

It turns out that the Poisson distribution is rather "rigid" in the sense that it is unimodal and parametrized by a single parameter λ representing both its mean and variance. Learning about $P(\cdot|\lambda)$ from a very limited sample (as one is effectively bound to do in the SAGE framework) is much less suspicious than one might naively expect.

We have first shown that the A-C statistic $\tilde{P}(y|x)$, even though not a Poisson distribution itself, naturally captures the distribution of further tag counts y, given a single observation x from the unknown $P(\cdot|\lambda)$. According to Theorem 1, for integer λ, both $\tilde{P}(\cdot|x)$ and $P(\cdot|\lambda)$ have two neighboring modes with decreasing probability values as one moves away from the modes in either direction. In particular, $P(\cdot|\lambda)$ has the modes located at λ and λ - 1, with $P(\lambda|\lambda) = P(\lambda - 1|\lambda)$. Given a tag count $x \geq 1$, $\tilde{P}(y|x)$ has the modes located at x and x - 1, with $\tilde{P}(x|x) = \tilde{P}(x - 1|x)$.

We then analyzed how 'close' is the A-C statistic $\tilde{P}(\cdot|\lambda)$ (in terms of K-L divergence) to the underlying Poisson distribution $P(\cdot|\lambda)$ of tag counts. It turns out that the K-L divergence from $P(y|\lambda)$ to $\tilde{P}(y|x)$ is minimized at the mode of $P(y|\lambda)$, i.e. when $x = \lambda - 1$ (Theorem 2). Most importantly, by Theorem 4, on average, the A-C statistic is never too far from the true underlying distribution. To be precise, up to terms of order $O(\lambda^{-3})$, on average, the A-C statistic is never further away from the truth $P(\cdot|\lambda)$ than half-a-bit of additional information. Hence, the Audic-Claverie method can be expected to work well even though the SAGE libraries represent very sparse samples.

So far the Audic-Claverie methodology for detection of differentially expressed genes has been verified only empirically through a series of specific Monte Carlo simulations [1]. It has not been clear how general the apparently stable simulation findings were. Besides detailed explanations of the nature of A-C statistic capturing the unknown Poisson distribution based on single observation only, we showed that the A-C statistic is *universally* applicable in any situation where inferences about the underlying Poisson distribution must be made based on an extremely sparse sample. Such situations are referred to in machine learning as 'one-shot-learning'. In the Monte Carlo simulations of [1] the false alarm rate was small for genes associated with small tag counts and gradually increased for higher tag counts. The false alarm rate, however, never exceeded the significance level of the test. These findings are consistent with the theoretically calculated divergence function $\varepsilon(\lambda)$ (eq. (20)) illustrated in figure 2. With increasing mean tag count λ, it is more likely that increased counts x will be observed. But as λ increases, so does the uncertainty in the generating Poisson distribution $P(\cdot|\lambda)$. Consequently, the average K-L divergence $\varepsilon(\lambda)$ from $P(\cdot|\lambda)$ to the approximating A-C statistic (based on a single realization x from $P(\cdot|\lambda)$) gets larger. For smaller λ the underlying Poisson distribution is well captured by the A-C statistic and the test that operates on it will be well behaved. As λ grows, the average K-L divergence $\varepsilon(\lambda)$ saturates at 0.5 bits implying that the test based on the A-C statistic will continue to be well behaved even for large values of the mean tag count λ.

The Audic-Claverie method has also been formulated for the case of two cDNA libraries of unequal size. Similar methodologies have been proposed for the case of multiple cDNA libraries (e.g. [7]). Even though developed under the limited assumption of two libraries of the same size, theoretical results obtained in this paper offer deep insights into the workings of the Audic-Claverie approach and provide an information theoretic justification for its use when analyzing expression patterns in cDNA arrays. Of course, when using libraries of unequal size, the A-C statistic will no longer be symmetric, putting more weight on the more populated library. Information theoretic investigation of statistics developed for pattern analysis in the cases of unequal multiple libraries is a matter for our future work.

Conclusion

Detection of differentially expressed genes is a crucial step in any large scale automated analysis of patterns of gene expression data. One of the most popular techniques for identifying genes with statistically different expression in SAGE libraries is the methodology of Audic and Claverie [1]. The methodology relies on learning the underlying Poisson distribution of tag counts from a single observation from it in the form of (A-C statistic). In this paper we rigorously analyzed the A-C statistic. We have shown that under the null hypothesis of not differentially expressed genes:

1. The A-C statistic and the underlying Poisson distribution share the same mode structure.

2. The K-L divergence from the true unknown Poisson distribution to the A-C statistic is minimized when the A-C statistic is conditioned on the mode (not mean) of the Poisson distribution.

3. The expected K-L divergence from the true unknown Poisson distribution to the A-C statistic is never larger than 1/2 bit, irrespective of the mean of the Poisson distribution.

4. The expected K-L divergence from the true unknown Poisson distribution to the A-C statistic can be approximated up to order $O(\lambda^{-3})$ by a simple function of the form $a_0 + a_1\lambda^{-1} + a_2\lambda^{-2}$. For the divergence measured in bits, $a_0 = 1/2$, $a_1 = 1/24$ and $a_2 = 1/32$.

Even though the A-C statistic infers the unknown underlying Poisson distribution based on one count observation only, the Audic-Claverie method should work reasonably well in most cases, since under the null hypothesis, the average divergence from the unknown Poisson distribution to the A-C statistic is guaranteed not to exceed 1/2 bit. This constitutes a rigorous quantitative argument, extending the empirical Monte Carlo studies of [1], that supports the wide spread use of Audic-Claverie method, even though by their very nature, the SAGE libraries represent very sparse samples.

Authors' contributions

I am the sole author of this paper.

Acknowledgements

I would like to thank Hong Yan for introducing me to the problem of cDNA array analysis and Somak Raychaudhury for inspiring me to study estimation of Poisson processes based on extremely limited samples.

References

1. Audic S, Claverie J: **The significance of digital expression profiles.** *Genome Res* 1997, **7:**986-995.
2. Varuzza L, Gruber A, de B Pereira C: **Significance tests for comparing digital gene expression profiles.** *Nature Precedings* 2008. **npre.2008.2002.3**
3. Velculescu V, Zhang L, Vogelstein B, Kinzler K: **Serial analysis of gene expression.** *Science* 1995, **270:**484-487.
4. Brenner S, Johnson M, Bridgham J, Golda G, Loyd D, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, *et al.*: **Gene expression analysis by massively parralel signature sequencing on microbead arrays.** *Nature Biotechnol* 2000, **18:**630-634.
5. Ruijter J, Kampen AV, Baas F: **Statistical evaluation of SAGE libraries: consequences for experimental design.** *Physiol Genomics* 2002, **11(2):**37-44.
6. Ge N, Epstein C: **An empirical Bayesian significance test of cDNA library data.** *Journal of Computational Biology* 2004, **11(6):**1175-1188.
7. Stekel D, Git Y, Falciani F: **The comparison of gene expressiom from multiple cDNA libraries.** *Genome Research* 2000, **10:**2055-2061.
8. Bortoluzzi S, Coppe A, Bisognin A, Pizzi C, Danieli G: **A multistep bioinformatic approach detects putative regulatory elements in gene promoters.** *BMC Bioinformatics* 2005, **6:**121-136.
9. Medina C, Rotter B, Horres R, Udupa S, Besser B, Bellarmino L, Baum M, Matsumura H, Terauchi R, Kahl G, Winter P: **SuperSAGE: the drought stress-responsive transcriptome of chickpea roots.** *BMC Genomics* 2008, **9:**553.
10. Kim H, Baek K, Lee S, Kim J, Lee B, Cho H, Kim W, Choi D, Hur C: **Pepper EST database: comprehensive in silico tool for analyzing the chili pepper (Capsicum annuum) transcriptome.** *BMC Plant Biology* 2008, **8:**101-108.
11. Zhao Y, Li Q, Yao C, Wang Z, Zhou Y, Wang Y, Liu L, Wang Y, Wang L, Qiao Z: **Characterization and quantification of mRNA transcripts in ejaculated spermatozoa of fertile men by serial analysis of gene expression.** *Human Reproduction* 2006, **21(6):**1583-1590.
12. Metta M, Gudavalli R, Gibert J, Schlötterer C: **No Accelerated Rate of Protein Evolution in Male-Biased Drosophila pseudoobscura Genes.** *Genetics* 2006, **174:**411-420.
13. Morin R, O'Connor M, Griffith M, Kuchenbauer F, Delaney A, Prabhu A, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves C, Marra M: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.** *Genome Research* 2008, **18:**610-621.
14. Borecký J, Nogueira F, de Oliveira K, Maia I, Vercesi A, Arruda P: **The plant energy-dissipating mitochondrial systems: depicting the genomic structure and the expression profiles of the gene families of uncoupling protein and alternative oxidase in monocots and dicots.** *Journal of Experimental Botany* 2006, **57(4):**849-864.
15. Lin C, Mueller L, Carthy JM, Crouzillat D, Pétiard V, Tanksley S: **Coffee and tomato share common gene repertoires as revealed by deep sequencing of seed and cherry transcripts.** *Theor Appl Genet* 2005, **112:**114-130.
16. Cervigni G, Paniego N, Pessino S, Selva J, Diaz M, Spangenberg G, Echenique V: **Gene expression in diplosporous and sexual Eragrostis curvula genotypes with differing ploidy levels.** *BMC Plant Biology* 2008, **67:**11-23.
17. Miles J, Blomberg A, Krisher R, Everts R, Sonstegard T, Tassell CV, Zeulke K: **Comparative Transcriptome Analysis of In Vivoand In Vitro-Produced Porcine Blastocysts by Small Amplified RNA-Serial Analysis of Gene Expression (SAR-SAGE).** *Molecular Reproduction and Development* 2008, **75:**976-988.
18. Evans R, Boersma J: **The Entropy of a Poisson Distribution.** *SIAM Review* 1988, **30(2):**314-317.