BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 9 2001
Pages 840–842

Detecting the impact of sequencing errors on SAGE data

Jacques Colinge[1,2] and Georg Feger[1,*]

[1]Serono Pharmaceutical Research Institute, Ch. des Aulx 14, CH-1228 Plan-les-Ouates, Switzerland

Received on March 13, 2001; revised on May 30, 2001; accepted on June 7, 2001

ABSTRACT

Summary: SAGE data are obtained by sequencing short DNA tags. Due to the mistakes in DNA sequencing, SAGE data contain errors. We propose a new approach to identify tags whose abundance is biased by sequencing errors. This approach is based on a concept of neighbourhood: abundant tags can contaminate tags whose sequence is very close. The application of our approach reveals that moderately abundant tags can be generated by sequencing errors uniquely. It also allows for detecting correct rare tags.

Availability: Software is available only to non-profit entities and for non-commercial purposes upon request.

Contact: Georg.Feger@serono.com

Serial Analysis of Gene Expression (SAGE) is a method for estimating the abundance of gene transcripts (mRNA) (Velculescu *et al.*, 1995). SAGE is based on the isolation of short sequence tags characteristic of each individual transcript. The sequence tags are then concatenated into long DNA molecules thus facilitating their identification by conventional DNA sequencing. By counting these tags one can estimate the expression level of transcripts in a cell. However, due to the inaccuracy of the DNA sequencing process, SAGE data contain errors: moderately abundant or rare tags may have their observed abundance substantially modified. Moreover, non-existing tags can be created. We describe two methods that detect those tags whose abundance has a large probability to be strongly biased by sequencing errors. Stollberg *et al.* (2000) studied the macroscopic impact (library-wide) of sequencing errors in details. We focus on techniques to compute information at the tag level. Our approach is based on the following idea: an abundant tag, or the combination of several tags, can significantly modify the observed abundance of tags with close sequences because of sequencing errors.

Let L_T be the set of sequenced SAGE tags (library). For the sake of clarity, we consider a simple definition of the set of tags $q \in L_T$ that can generate a given tag $t \in L_T$ because of sequencing errors. We name t neighbourhood the set of such tags q. We assume that sequencing errors are independent for each base and the error rate is 1%. Then the probability to have two or more errors in a 10-base tag is $1 - 0.99^{10} - 10 \times 0.01 \times 0.99^9 = 0.43\%$. The probability to have 1 error or more is $1 - 0.99^{10} = 9.56\% \gg 0.43\%$. Hence we define the set $N(t)$ (t neighbourhood) as the tags $q \in L_T$ that are at an edit distance equal to 1 from t (Gusfield, 1997). The probability to have one error only is $10 \times 0.01 \times 0.99^9 = 9.13\%$. Method 1 is the following: (a) For each tag $t \in L_T$, we list every tag $q \in N(t)$. (b) We compute the average contribution $v(t)$ of tags $q \in N(t)$ to the observed abundance of t. We have $v(t) = \sum_{q \in N(t)} \frac{0.0913 \times \text{count}(q)}{\#N(q)}$, where count($q$) is the observed abundance of q and # denotes the cardinality. The factor $0.0913/\#N(q)$ is the average contribution of q to each of its neighbours. (c) Let $C \in \mathbb{R}_*^+$ be a chosen cut-off. Mark as suspect every tag t such that $\frac{v(t)}{\text{count}(t)} \geqslant C$. We further study the status of tags $t \in L_T$ with no neighbour, i.e. $N(t) = \emptyset$. Let us estimate the probability that such a tag has been generated by sequencing errors only. There must exist $B \geqslant \text{count}(t)$ tags in the neighbourhood of t before sequencing, all erroneously sequenced ($N(t) = \emptyset$) with probability 0.0956^B. Therefore, t is correct with probability at least $1 - 0.0956^{\text{count}(t)}$. Accordingly, a tag with count = 1 and no neighbour is correct with probability 90.44%.

In practice, we use more realistic models of sequencing errors. For example, one can consider the number of ways $q \in N(t)$ can be erroneously sequenced as t to weight the contributions of q. For instance, ACGTT can be transformed into ACGGT either by substituting a G for the first T or by inserting a G before or after the existing G. A more refined model could make use of phred scores (Ewing *et al.*, 1998): for each copy of a tag t, add specific tags q into $N(t)$ and weight their probable contribution to t based on t phred score (parts of t sequence that

*To whom all correspondence should be addressed.

[2]Present address: GeneProt Inc., Pré-de-la-Fontaine 2, CH-1217 Meyrin, Switzerland.

Table 1. A simple example with four tags. Tags 2 and 3 are at edit distance 1 from tag 1. $s = 0.0913$ is the probability to have one error and $r = 1-s$. We observe that tag 2 is probably entirely due to sequencing errors, while tag 3 has its real abundance probably quite different from its observed abundance

Tag	y	A				x
ATTACATGCG	100	r	s	s	0	109.5
CTTACATGCG	5	s/2	r	0	0	0.0
ATTACATTGC	10	s/2	0	r	0	5.5
TCAGGTCGAT	1	0	0	0	1	1.0
Total	116					116.0

are very reliable exclude certain neighbours). See also Margulies (2000). More generally, by introducing $P(t|q)$, $t, q \in L_T$, the probability that q is erroneously sequenced as t, we have $v(t) = \sum_{q \in L_T, q \neq t} P(t|q)\text{count}(q)$. So stated, Method 1 is a generalization of a method applied by Velculescu *et al.* (1999, supplementary material): more complete treatment of tags with count = 1 and arbitrary error model.

Method 1 considers the inter-tag contributions independently. A natural extension is to consider every contribution simultaneously: a given tag contributes to the observed abundance of other tags and, at the same time, receives contributions also. This is a more realistic model of the dynamics of sequencing errors.

We introduce Method 2, intended to estimate the original tag abundance before sequencing according to such a simultaneous model. A comparison with the observed tag abundance allows for detecting suspect tag abundances. Let N be the number of detected tags and $y \in \mathbb{R}^N$ the corresponding observed abundances. We denote by tag(i) the i^{th} tag whose abundance is y_i. We want to compute $x \in \mathbb{R}^N$, an approximation of the tag abundances before sequencing. We build a linear system $Ax = y$, where A is a $N \times N$ matrix with elements $a_{i,j}$. For each detected tag $t = \text{tag}(j)$ we set the corresponding column of A, i.e. we set elements in A to represent the average contribution of t to its neighbours: $a_{i,j} = P(\text{tag}(i)|\text{tag}(j))$, $i = 1, \ldots, N$. See Table 1 for an example computed with the simple model used in the description of Method 1. The observation concerning the reliability of rare tags is the same as for Method 1. Since $P(t|q)$ is a probability, i.e. $\sum_{q \in S_T} P(q|t) = 1, \forall t \in S_T$, it is straightforward to prove that the total number of tags is conserved. Namely, $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} x_i$. The linear system $Ax = y$ has a unique solution (Gershgorin's Theorem shows that every eigenvalue is different from zero) and is sparse. It is efficiently solved by Lanczos-type methods (Saad, 1996) because it is diagonally dominant. We had very good results with BiCGSTAB algorithm.

From the publicly available data set CGAP (http://www.ncbi.nlm.nih.gov/CGAP/), we used as an example a *Homo sapiens* normal white matter SAGE library (ftp://ncbi.nlm.nih.gov/pub/sage/extr/SAGE_BB542_whitematter/ditags/). After removing repeated di-tags, we found 31 454 different tags (93 748 total tags). By applying Method 2 (with an error model where the number of ways q can be erroneously sequenced as t is taken into account, see above), we estimated the relative error of every tag count. The result is shown in Figure 1. The computation time for BB542-whitematter was 12 s (SGI R10000 processor at 250 MHz).

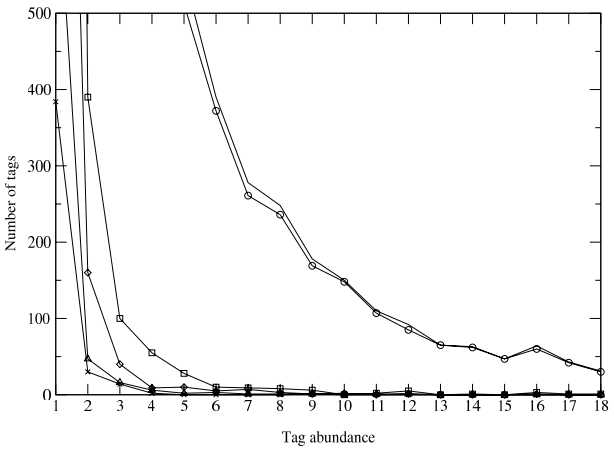

Fig. 1. Number of tags having a given relative difference between observed and estimated abundance. Continuous line: number of tags. Circle: less than 10%. Square: between 10 and 30%. Diamond: between 30 and 60%. Triangle: between 60 and 100%. Cross: more than 100%. Tags having an abundance larger than 18 change by less than 10%.

Coupled with improved gene-tag assignment (Caron *et al.*, 2001), the approach we presented improves confidence in SAGE data analysis and the accuracy of predicted transcriptomes. Our statistical methods are particularly powerful when analyzing large SAGE libraries or pools of several SAGE libraries (Velculescu *et al.*, 1999).

ACKNOWLEDGEMENTS

The authors would like to acknowledge Jonathan Lees, the referees and especially Massimo De Francesco who made very valuable comments.

REFERENCES

Caron,H. *et al.* (2001) The human trancriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.

Ewing,B., Hillier,L., Wendl,M.C. and Green,P. (1998) Base-calling of automated sequencer traces using Phred. *Genet. Res.*, **8**, 175–185.

Gusfield,D. (1997) *Algorithms on String, Trees, and Sequences*. Cambridge University Press, Cambridge.

Lash,A.E., Tolstoshev,C.M., Wagner,L., Schuler,G.D., Strausberg,R.L., Riggins,G.J. and Altschul,S.F. (2000) SAGEmap: a public gene expression resource. *Genet. Res.*, **10**, 1051–1060.

Margulies and Innis (2000) eSAGE: managing and analyzing data generated with Serial Analysis of Gene Expression (SAGE). *Bioinformatics*, **16**, 650–651.

Saad,Y. (1996) *Iterative Methods for Sparse Linear Systems*. PWS Publishing, Boston, MA.

Stollberg,J., Urschitz,J., Urban,Z. and Boyd,C.D. (2000) A quantitative evaluation of SAGE. *Genet. Res.*, **10**, 1241–1248.

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

Velculescu,V.E. *et al.* (1999) Analysis of human transcriptomes. *Nature Genet.*, **23**, 387–388.