Deposited research article

Improving SAGE di-tag processing

Jacques Colinge and Georg Feger

Address: Serono Pharmaceutical Research Institute, Ch. des Aulx 14, CH-1228 Plan-les-Ouates, Switzerland.

Correspondence: Georg Feger. E-mail: Georg.Feger@serono.com

Posted: 22 February 2001

Genome ***Biology*** 2001, **2(3)**:preprint0002.1–0002.10

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2001/2/3/preprint/0002

Received: 13 February 2001

This is the first version of this article to be made available publicly. This article has been submitted to *Genome Biology* for peer review.

Improving SAGE di-tag processing

Jacques Colinge and Georg Feger

Address: Serono Pharmaceutical Research Institute, Ch. des Aulx 14, CH-1228 Plan-les-Ouates, Switzerland.

Correspondence: Georg Feger, E-mail: Georg.Feger@serono.com

Abstract

Background: SAGE is a genome-wide method for obtaining gene expression profiles. It generates tags of 10 nucleotides in length, which are assumed to determine the corresponding gene transcript. In practice however, this is not always sufficient for uniquely identifying a gene.

Results: We propose an improved processing of SAGE sequences that allows us to obtain one extra base for reasonably abundant tags. This method includes a statistical test for controlling the relevance of extra base predictions.

Conclusions: The improved SAGE sequence processing we present reduces the uncertainty in SAGE tag to gene mapping and can be applied to any SAGE library.

Background

Serial Analysis of Gene Expression (SAGE) is a method for measuring the relative abundance of gene transcripts in different mRNA samples. It identifies a short mRNA tag from each individual transcript and concatenates them into long DNA molecules, which are then sequenced. By counting these tags one can estimate, for example, the expression of genes in a cell [1]. SAGE popularity is growing fast and many public data are accessible from the Internet [2].

Processing of SAGE sequences is described in [1], [2] and [3]. The usual length of SAGE tag is 10 bases. In practice, this length is not sufficient to uniquely identify each gene: several genes share the same tag. The SAGE method uses tag-pairs to avoid bias by PCR amplification. As pointed out in [2], the observed length of the di-tags is not constant, it varies between 20 and 26 (see Table 1) due to a certain flexibility in the enzyme used. The usual processing of SAGE sequences does not take advantage of these longer di-tags. Here we present a new method to predict an 11^{th} base for sufficiently abundant tags, hence increasing precision in gene identification (the number of possible genes to which a tag is mapped is divided by 4 on average).

Results and Discussion

We use di-tags of sufficient length to compute the frequencies of the four possible extra bases (A, C, G and T) for every tag. Then we use contingency tables and hypothesis testing [4] to determine relevant extra bases. The null hypothesis we apply is that every possible extra base has the same probability to be sequenced.

In the publicly available data set [5], we used as an example a SAGE library made for Homo sapiens normal white matter [6]. We used the di-tag list of this SAGE library to exemplify the

usefulness of our method. [6] contains 51640 di-tags of length between 20 and 26 bases. We rejected 3856 suspect repeated di-tags (see [2] and [3]). From the remaining 47784 di-tags we extracted 32668 different tags. The number of tags for which we could predict an 11th base by applying our method is given in Table 2.

For illustration purpose, we identified these tags by extracting SAGE tags of UniGene [7] clusters (build 108). We only considered tags at the end of the UniGene sequences, i.e. we consider UniGene sequences as 5' oriented. Other identification strategies are possible, see for instance [2]. An example of a tag is CAAGCATCCC, observed 1510 times with 5 extra As (the base A was observed at the 11th position in the di-tag), 1426 extra Cs, 13 extra Gs and 13 extra Ts. We uniquely identified this tag in UniGene as Hs.250444 *small inducible cytokine A7 (monocyte chemotactic protein 3)*. The 11th base found in the UniGene cluster sequence matches with the dominant extra C we mention above. According to the method we propose (see Materials and Methods), the prediction of C as an extra base is relevant at the 99.9% level.

An example of a tag shared by two genes, one of which is apparently not expressed, is provided by GGGCTGGGGT, observed 86 times with 5 extra As and 80 extra Cs. GGGCTGGGGT**A** is identified in UniGene [7] as Hs.90436 *sperm acrosomal protein (SPAG7)* and GGGCTGGGGT**C** as Hs.183698 *ribosomal protein L29 (RPL29).* According to the extra bases observed, it seems that only RPL29 is expressed (99.9% relevant, SPAG7 is possibly weakly expressed).

The special situation of several expressed genes sharing the same 10-base tag is illustrated by GTGAAACCCC, observed 422 times with 161 extra As, 22 extra Cs, 202 extra Gs, and 12 extra Ts. According to the null hypothesis (equiprobability of every extra base), both A and G are relevant at a higher probability than 95% (99.9% in this case). We can estimate a count of

161/(161+202).422=187 for GTGAAACCCCC**A** and 202/(161+202).422=235 for GTGAAACCCCC**G**. We subsequently found that this tag is shared by many UniGene[7] clusters: 49 clusters with extra A, 7 cluster with extra C and 54 clusters with extra G.

[2] proposes the assignment of a score to each identification, in order to characterize its reliability. If a tag comes with a predicted extra base, the latter should be checked with the database sequence and the result included in the score computation.

The complex situation of tag GTGAAACCCC above suggests a possible extension of our method. We test the relevance of predicted extra bases by comparing (hypothesis testing) the observed frequencies with the hypothetical situation of equiprobability. Another possible null hypothesis would be (1) to chose a method for identifying tags, as we did with UniGene [7], and (2) to estimate the relevance of the possible extra bases according to this new null hypothesis. Returning to the example of tag GTGAAACCCC, none of the four possible extra bases significantly departs from the distribution obtained from UniGene. This implies that no extra base can be selected reliably and, consequently, every possible extra base should be considered. We cannot obtain any simplification of the data in that case, contrary to what we found with the equiprobability null hypothesis.

We do not apply the latter extension of the method in practice for two reasons: first, this extension is dependant on the method for identifying tags and, second, considering the difficulty in analyzing SAGE data, we prefer to concentrate on dominantly abundant extra bases for the sake of simplicity.

We presented a method that allows for the prediction of one extra base for sufficiently abundant tags (at least 7 occurrences). The method applies to every SAGE library, without any special

preparation. The predictions may be controlled in terms of relevance by using appropriate hypothesis testing techniques. The longer tags permit a better identification of expressed genes.

Materials and Methods

We assume that, in the case of di-tags of length 20 or more, the first 10 bases belong to the first tag and the last 10 bases belong to the second tag. Since the tags are linked into di-tags randomly, the extra available bases, in the middle of a di-tag of length 22 or more (see Figure 1), belong to each tag with a probability that is symmetrical. Accordingly, we propose a new di-tag processing.

Algorithm

Let c(t) denotes the counter associated with a tag t. Let A(t), C(t), G(t), T(t) denote the counters associated with each 4 possible extra base of tag t. We denote by R(s) the operation to take the complementary reverse of s (read s in reverse order and exchange letters: 'A' with 'T', 'C' with 'G').

1. For each di-tag d of length k:

   Take 10 bases at each end of d in order to obtain the two tags t_1 and t_2 it contains. Namely, we have t_1=d[1..10] and t_2=R(d[k-10..l]). Increment the counters c(t_1) and c(t_2). If k • 22, then extract one extra base for each tag: b_1=d[11] and b_2=R(d[k-11]). These extra bases are used to increment counters A, C, G, T: If b_1='A' then increment A(t_1), if b_1='C' then increment C(t_1), etc. The same for b_2.
2. We chose a degree of relevance, typically 95% or 99%. Then, for each different tag t, which has at least one of its extra base counter different from 0, we test whether each possible extra base is relevant (it is possible that more than one extra base is relevant). This is achieved by using contingency tables and hypothesis testing [4].

Hypothesis testing

We describe in detail a possible method for implementing Step 2. We apply hypothesis testing to decide whether an extra base is relevant or not. Namely we use contingency table methods [4]. Let us denote by D the counter of an extra base to test. Our null hypothesis is that every possible extra base has the same probability to be sequenced. We denote by Q the sum of the other counters. If D+Q is not a multiple of 4, we add 1, 2, or 3 to Q in order to have N=D+Q a multiple of 4. The null hypothesis is equivalent to test whether D is significantly different from N/4. Since we are interested in extra bases that are in excess from N/4, we consider as non-relevant extra bases with D

- N/4. The situation is summarized in a contingency table (see Table 3).

Chi-squared statistics allows estimation of the significance of the departure from the null hypothesis. This can be done, for instance, by using the Chi-squared distribution with 1 degree of freedom or Fisher's exact test as soon as $D < 5$, see [4].

In our algorithm we only consider di-tags of a length of at least 22 for extra base prediction. We do not use 21-base long di-tags for extra base prediction because (1) the distribution of di-tag lengths (Table 1) shows that there are enough 22-base long di-tags, and (2) this would generate too many wrong 11^{th} base counts, hence making the application of hypothesis testing more difficult.

Acknowledgements

The authors would like to acknowledge Mark Ibberson for reading an early version of this paper. We also thank Massimo de Francesco for his help and his support.

References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial Analysis of Gene Expression.** *Science* 1995, **270:**484-487.
2. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: A Public Gene Expression Resource.** *Gen. Res.* 1998, **8:**175-185.
3. Margulies EH, Innis JW: **eSAGE: Managing and Analyzing Data Generated with Serial Analysis of Gene Expression (SAGE).** *Bioinformatics* 2000, **16(7):**650-651.
4. Everitt BS: **The Analysis of Contingency Tables.** London: Chapman and Hall, 1977.
5. **CGAP (Cancer Genome Anatomy Project)** [http://www.ncbi.nlm.nih.gov/CGAP/].
6. **SAGE library for human normal white matter** [ftp://ncbi.nlm.nih.gov/pub/sage/extr/SAGEBB542whitematter].
7. Schuler GD, Boguski MS, Stewart EA, Stein LD, Gyapay G, Rice K, White RE, Rodriguez-Tome P, Aggarwal A, Bajorek E, Bentolila S, Birren BB, Butler A, Castle AB, Chiannilkulchai N, Chu A, Clee C, Cowles S, Day PJ, Dibling T, Drouot N, Dunham I, Duprat S, East C, Hudson TJ, et al.: **A Gene Map of the Human Genome.** *Science* 1996, **274(5287):**540-546.

Table 1

Di-tag length distribution

Di-tag length	Number detected	Percentage
20	233	0.5%
21	2524	5.3%
22	25502	53.3%
23	17052	35.7%
24	2151	4.5%
25	129	0.3%
26	191	0.4%

Example of di-tag length distribution obtained for a human white matter SAGE library [5]. Distributions obtained for other libraries follow the same pattern.

Table 2

Number of 11^{th} base prediction

Relevance	Number of predictions	Average count	Median count
95.0%	1700 (432)	28.8 (8.5)	25 (7)
99.0%	1268 (488)	35.7 (10.7)	20 (13)
99.9%	780	51.4	22

Number of 11^{th} base predictions for human normal white matter SAGE library [5]. Statistics about tag abundance for each relevance degree are given both as the average and median counts. Statistics for a specific relevance degree only are in parentheses.

Table 3

Contingency table			
	To test	Others	Total
Observed counts	D	Q	N=D+Q
Null hypothesis	N/4	3N/4	N

Contingency table for testing the relevance of a possible extra base.

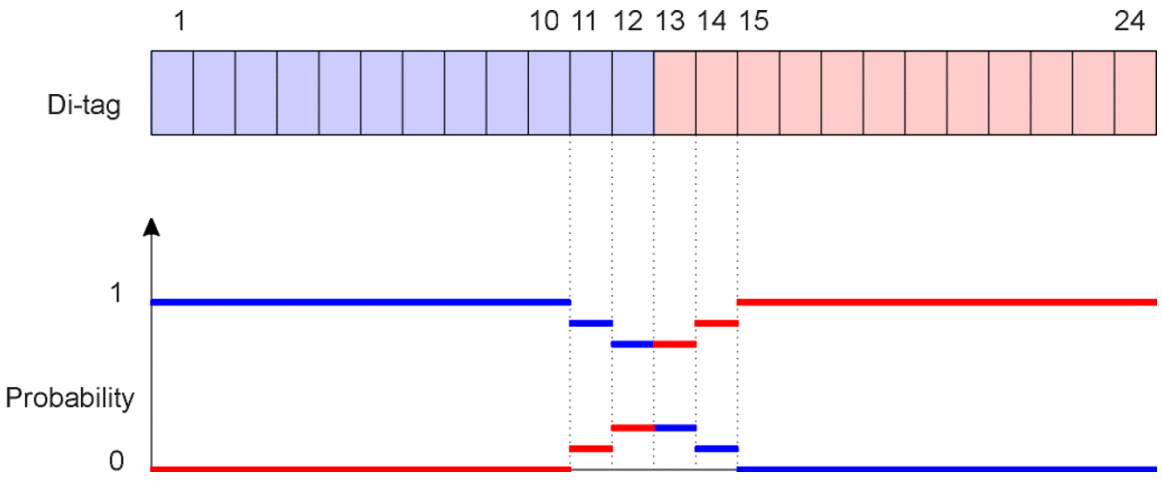

Figure 1: A di-tag of length 24. The two 10-base tags are made of the bases 1 to 10 and 15 to 24. The bottom part of the figure shows an idealization of the probability that each base belongs to a specific tag (blue line for the first tag, red line for the second tag)