KEYWORDS: serial analysis of gene expression, SAGE, mRNA expression

Special section on techniques:

What a Wise SAGE Once Said about Gene Expression...

Jina Song

Pathology, University of British Columbia

Submitted April 2003

Introduction

High throughput analysis of differential gene expression can be applied to many areas in molecular cell biology such as differentiation, development, physiology and pharmacology. The human genome has been mapped recently and in order to analyze such large sized data, novel technologies for identifying new genes and proteins in a tissue and cell specific manner are desperately required[1]. In recent years, a variety of techniques such as DNA microarrays and serial analysis of gene expression (SAGE) have been developed to simultaneously test the expression of thousands of gene and to automatically identify the genes of interest. On top of identifying gene expression, microarrays and SAGE are also very powerful tools in biotechnology including identification of molecular markers for various disease processes, potential drug targets and pharmacogenomics[2].

Genomics can be roughly divided into five different categories: (1) **Structural genomics** is the study of the structure of individual genes in a genome. The sequencing and mapping of genes are a typical structural genomics study. (2) **Bioinformatics** involves getting information from sequenced data. According to a book by Andreas D. Baxevanis[3], bioinformatics is called "everything from the magic bullet that will cure all infirmities known to man to a brute-force powertool for dismantling sequence data to simply a sexy way to do science." (3) **Functional genomics** is the study of gene function, expression and regulation. The other two categories are (4) **comparative genomics** and (5) **evolutionary genomics** where various data obtained could be studied further in order to conduct comparative or evolutionary studies. As mentioned earlier, DNA microarrays and SAGE are useful tools in simultaneous and quantitative monitoring of the expression levels of numerous genes, thus they are used to simplify functional genomics research.

The genomes of eukaryotic organisms are massive and contain an enormous number of genes. These genes then encode proteins. Even though it is widely believed that the amount of protein produced is directly dependent on the amount of mRNA that encodes it, there are numerous cases where the correlation between the mRNA concentration and protein concentrations does not exist[4]. In this paper, the focus will be on a high throughput analysis technique called serial analysis of gene expression (SAGE) and its data analysis and also how to attempt to confirm the SAGE results by using protein detection method.

What is SAGE?

SAGE is a technique that allows a rapid, detailed analysis of thousands of transcripts. This method uses a unique sequence tag of 13 or more bases generated from each transcript in a cell or tissue of interest. A tag, therefore, is defined as a unique small sequence that is characteristic/diagnostic of a specific message[5]. These sequence tags are then ligated in a defined series of steps. This ligated sequence then represents short, unique tags for genes[2]. There are two principles that SAGE is developed upon: (1) a short nucleotide sequence tag contains unique information about a transcript (provided that it is isolated from a defined position within the transcript) and (2) concatenated short sequence tags can be cloned to allow the efficient analysis of transcripts in a serial manner[5].

SAGE is performed to analyze mRNA expression (Figure 1). The mRNA of eukaryotics have polyadenylate (poly A) tail in the 3' end of strand. Therefore, oligo(dT) primer can be made to bind to all mRNAs in a cell or a tissue of interest. By using reverse transcriptase, double stranded cDNA is synthesized from the oligo(dT) primers for all expressed mRNA. The primers are biotinated so they could be bound to streptavidin beads. The cDNA made is then cleaved with a restriction endonuclease, which is expected to cleave most transcripts at least once. The most commonly used restriction endonuclease is a former, which has a cleaving site with a four base pairs. Thus, cleaving site would vary among different transcripts but identical for the same sequence. Binding to streptavidin beads then isolates the most 3' end position of the cleaved cDNA. The importance here is that each transcript has a unique cleavage site of a former cutter located closest to the poly A tail. Since the former cutters (restriction endonuclease) are used to anchor a specific sequence to beads, they are referred to as anchoring enzyme (AE). The cDNA is then divided in half and ligated to one of two linkers, A and B. These linkers contain a type IIS restriction site. Type IIS restriction endonucleases (tagging enzyme) cut at a defined distance up to 20 base pairs away from their recognition sites. The cleavage by type IIS restrictive endonuclease generates blunt ends about 20 base pairs away from the recognition site of linker A and B. The cleaved cDNA can then be ligated and amplified with primers specific to each linker A and B. The amplification is carried out by PCR[6].

As result of PCR amplification, concatenated tag sequences are produced in larger quantify and each tag sequence can provide information about gene expression in a serial manner. As shown in Figure 1, two tags are joined tail to tail to form a ditag and each ditag is separated from the other by anchoring enzyme sites. The question here is why not just use the sequence of each tag punctuated by anchoring enzyme sites? It is because the use of PCR for amplification can easily lead to making mistakes. For instance, one tag could be amplified more than once during one cycle of PCR. Because the probability of any two tags being coupled to form a ditag is very small, if the PCR result contains repeated ditags, the result could be excluded from data analysis. Therefore, ditags eliminate PCR artifacts to a certain extent. The amplified products are then cleaved by an anchoring enzyme and separated by polyacrylamide gel electrophoresis (PAGE). The separated ditags are then cloned into a plamid vector for sequencing[7].

After the sequences are obtained, they are compared to different genome databases in order to identify the tags. Even though ditag sequence analysis using SAGE could potentially identify all the unique mRNA and their copy numbers isolated, the interpretation of SAGE data could be crucial in obtaining valuable results. In a typical SAGE experiment, there are at least two samples. The aim is to identify genes of interest by comparing the number of specific tags found in two different SAGE libraries[8]. The usual question asked when using SAGE is whether one sample has a significant change in gene expression relative to the other sample. One useful application is to investigate expression differences between normal and diseased samples or between samples with and without drug treatment.

Statistical significance of SAGE

In last few years, several methods have been developed to determine the statistical significance of gene expression difference in SAGE experiments. Zhang *et al.* use an approach to determine the probability of obtaining the observed difference[9]. This method is part of the SAGE software package SAGE300. Because this method entails large number of simulations, it is not suitable for fast and interactive applications.

The Fisher's Exact test has been proposed by the Cancer Genome Anatomy Project for comparison of specific tags between SAGE libraries[10]. This test is based on recognizing the data from tags by using a 2x2 contingency table. The rows of this table contain specific and other tags and the columns contain library 1 and library 2. The Fisher's Exact test calculates the pooled probability of obtaining the tables with a more extreme difference with the row and column totals. The chi-squared test is also based on a 2x2 table and this test is based on a Z-statistic to text the equality of two proportions in two experimental conditions. The test Z-statistic is calculated as the observed difference between proportions of specific tags in both libraries divided by the standard error of this difference when the null hypothesis is true. This test is based on an assumption that the probability of the resulting tag counts follows a normal binomial distribution. The Fisher's exact test is appropriate over the chi-squared test when the sample size is small but it is generally more computationally expensive[11].

There are many other statistical tests that are useful in dealing with SAGE data. Each test has its advantages and disadvantages over the others. It seems like there are no rules on what test should be used for what kind of SAGE data. Since SAGE is a fairly new technology, there is a lot to be worked out for its method of data analysis.

Problems with SAGE

There are several problems that exist when using the SAGE procedure. The length of a gene tag is extremely short (13or 14 base pairs). If the tag is derived from an unknown gene, it is difficult to analyze with such a short sequence[7]. However, this disadvantage could also be considered an advantage since the isolation of the unknown gene is often the ultimate goal for most analysis using the SAGE procedure. Thus,

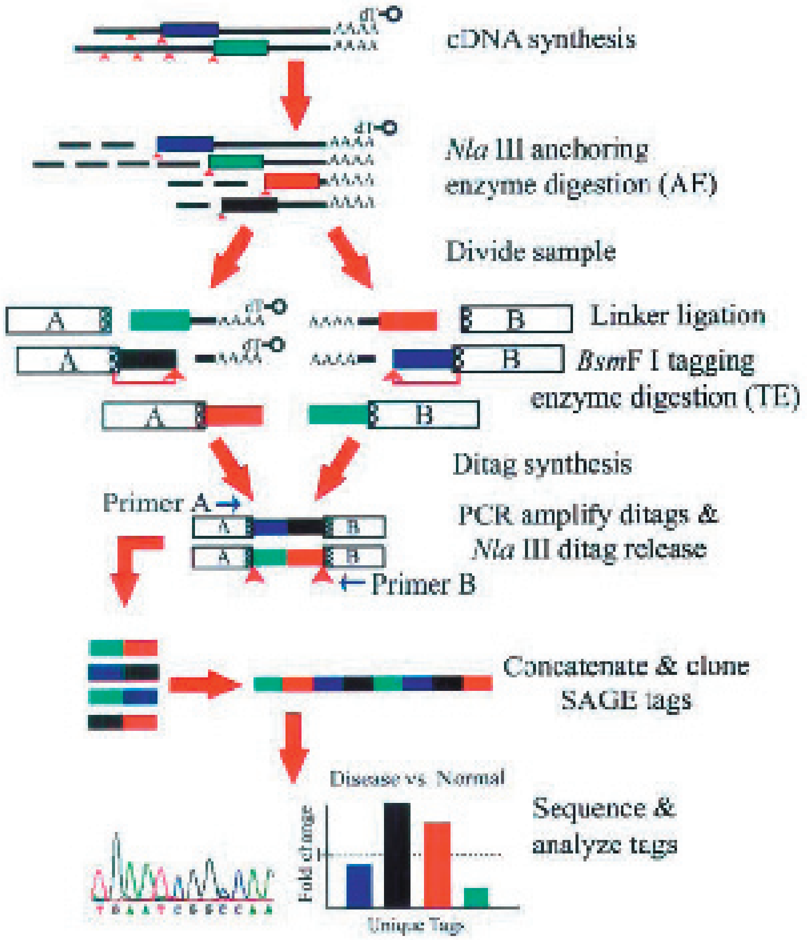

Figure 1: a schematic diagram on how SAGE is performed (Circularion Research 91:565-569

SAGE could also be used as a "gene finding method". In cancer research, for example, the most attractive feature of SAGE is its ability to evaluate the expression pattern of thousands of genes in a quantitative manner without prior sequence information. This gene-finding feature has been exploited in three applications: (1) to define transcriptomes, (2) to analyze differences between the gene expression patterns of cancer cells and their normal counterparts and (3) to identify downstream targets of oncogenes and tumor suppressor genes[12].

Another downfall of the SAGE technique is that typeIIS restriction enzyme (mainly *Bsm*FI) does not always yield same length fragments. *Bsm*FI should yield exact 14 base pair tags but depending on the temperature that the experiment is carried out, the length of fragments produced varies. Since two tags are ligated tail to tail, it is hard to make sure that each tag is 14 base pairs long in a ditag of 28 base pairs. The ditag could be composed of a 12 base pair tag and a 16 base pair tag or a 13 base pair tag and a 15 base pair tag or a 14 base pair tag and a 14 base pair tag. In order to minimize this problem, the temperature of experiment should be kept constant preferably at 65 degrees[7].

The SAGE technique could yield another critical problem when enzymes do not recognize some mRNA. Depending on anchoring enzyme and tagging enzyme used, some fraction of mRNA species could be lost. Even though recognition sites for four base cutters are present every 256 base pairs, some species might have transcripts that do not contain the sequence. In order to avoid this problem, two different combinations of anchoring enzyme and tagging enzyme could be used and investigating the gene expression correlation between the two data generated by these two sets of AE and TE would allow us to compare the generated data. This could require twice the work, however, the comparison between the products could provide powerful information on some genes that do not get represented by a certain combination of enzymes[7].

Even though the SAGE technique is based on a principle that the short sequence tags around 13 base pairs from mRNA represent a unique sequence, there are instances in which multiple genes share the same tag. Also, same genes could yield multiple tags if the gene has alternate poly A sites. This problem could be decreased if longer sequence tags are generated[13]. In the original method as outlined in Figure 1, *Nla*III is used as an anchoring enzyme and *Bsm*FI is used as a tagging enzyme. Ryo *et al.* used *Rsa*I as an anchoring enzyme with *Bsm*FI as a tagging enzyme in order to generate 18 base pair long tags. By using the elongated tags, more of transcripts are represented[13].

Another serious disadvantage is SAGE (also with DNA microarrays) is that mRNA level and protein expression do not always correlate[4]. Also, even if protein is synthesized, protein functionality depends on post-translational modifications of the precursor protein such as phosphorylation, sulphation, glycosylation and hydroxylation. In parallel to either DNA microarray or SAGE, high-throughput protein analysis should be performed in order to ensure that the presence of mRNA does indeed result in protein synthesis[14].

Microarray assays using nucleic acid-nucleic acid interactions are well established. Just recently, protein microarrays have become popular. Both DNA and protein arrays are small flat surfaces that allow the simultaneous analysis of thousands of molecules within a single experiment. The development and applications of DNA microarray technology began to expand during the late 1990's. A variety of DNA microarray and DNA chip devices and systems have been developed and commercialized[15]. Compared to nucleic acids, proteins are more diverse and complex, however, if a novel technique for high-throughput protein array could be developed, it would, in my opinion, provide way more valuable information on gene expression.

To generate high-density protein arrays, robots that transfer protein expressed onto polyvinylidene difluoride (PVDF) membrane are used. Compared to DNA arrays in which the membrane can be reused at least 20 times, protein arrayed on PVDF membrane could only be used once. There are different classes of capture molecules for protein microarrays[17]. In Figure 2, (a) and (b) show antigen-antibody interaction and sandwich immunoassay methods, respectively. In (c), specific protein probes are bound to interact with specific sample proteins. In (d), synthetic molecules referred to as aptamers are used to capture sample proteins. The aptamers can be nucleotides, ribonucleotides or peptides. Interaction -between enzyme-substrate can be used to trap certain enzyme-substrate complexes as shown in (e). Lastly, synthetic low molecular mass compounds can be immobilized as capture molecules as shown in (f) for a receptor-ligand interaction[14].

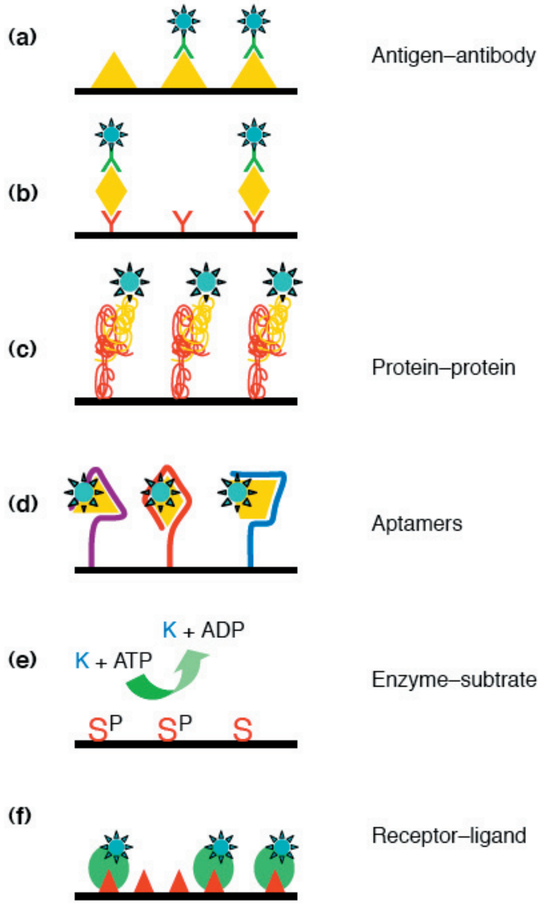

Figure 2: Trends in Biotechnology 20(4):160-166

Conclusion

Different approaches of protein microarrays mentioned above show that protein microarray technology is already a useful tool to study different kinds of protein interactions. However, since this technology is relatively new, further developments are required to increase the usefulness of the information you acquire from this method. Most importantly, a method for high-throughput generation of protein targets and ligands is needed in order to extend the number of application of protein microarrays dramatically. Also further developments and optimization of array production and assay performance will strengthen this technology.

In order to develop a novel technology to analyze gene and protein expression, fast and accurate methods are required to overcome such large number of samples. As discussed earlier, SAGE technique allows high-throughput analysis of gene expression. However the main disadvantage of this method is that gene expression does not always correlate with protein expression profile. To minimize this problem, protein arrays could be used on parallel with SAGE to look at the expression of proteins. However, since this protein microarray technology is new, further improvement of technology is required. Once this technology is well developed, it will be a very powerful tool to confirm SAGE result in order to confirm the protein expression from genes.

References

1. Denis Soulet, Serge Rivest (1999) Perspective: How to make microarray, serial analysis of gene expression, and proteomic relevant to day-to-day endocrine problems and physiological systems. *Endocrinology* 143(6):1995-2001

2. John P. Carulli, Michael Artinger, Pamela M. Swain, Colleen D. Root, Linda Chee, Craig Tulig, Jennifer Guerin, Mark Osborne, Gary Stein, Jane Lian, Peter Lomedico (1998) High throughput analysis of differential gene expression. *Journal of Cellular Biochemistry Supplements* 30/31:286-296

3. Andreas D. Baxevanis, B.F. Francis Ouellette (Wiley, New York, 1998) Bioinformatics

4. S.P. Gygi (1999) Correlation between protein and mRNA abundance in yeast. *Molecular Cellular Biology* 19:1720-1730

5. Willmar D. Patino, Omar Y. Mian, Paul M. Hwang (2002) Serial analysis of gene expression. *Circulation Research* 91:565-569

6. Victor E. Velculescu, Lin Zhang, Bert Vogelstein, Kenneth W. Kinzler (1995) Serial analysis of gene expression. *Science* 270:484-487

7. Mikio Yamamoto, Toru Wakatsuki, Akiyuki Hada, Akihide Ryo (2001) Use of serial analysis of gene expression (SAGE) technology. *Journal of Immunological Methods* 250:45-66

8. Jam M. Ruijter, Antoine H. C. Van Kampen, and Frank Baas (2002) Statistical evaluation of SAGE libraries: consequences for experimental design. *Physiol. Genomics* 11: 37-44

9. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, and Kinzler KW (1997) Gene expression profiles in normal and cancer cells. *Science* 276: 1268-1272

10. Audic S and Claverie JM (1997) The significance of digital gene expression profiles. *Genome Res* 7:986-995

11. Michael Z. Man, Xuning Wang and Yixin Wang (2000) Power SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics* 16: 953-959

12. Kornelia Polyak, Gregory J. Riggins (2001) Gene discovery using the serial analysis of gene expression technique: implication for cancer research. *Journal of Clinical Oncology* 19(11): 2948-2958

13. A. Ryo, N. Kondoh, T. Wakatsuki, A. Hada, N. Yamamoto, M. Yamamoto (2000) A modified serial analysis of gene expression that generates longer sequence tags by nonpalindromic cohesive linker ligation. *Analytical Biochemistry* 277:160-162

14. Markus F. Templin, Dieter Stoll, Monika Schrenk, Petra C. Traub, Christian F. Vohringer, Thomas O. Joos (2002) Protein microarray technology. *Trends in Biotechnology* 20(4):160-166

15. Patric O. Brown, David Botstein (1999) Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21:33-37

16. David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer, Jeffrey M. Trent (1999) Expression profiling using cDNA microarrays. *Nature Genetics* 21:10-14

17. Dolores J. Cahill (2000) Protein arrays: a high-throughput solution for proteomics research? *Proteomics* 47-51

18. Zhu. H. *et al*. (2000) Analysis of yeast protein kinases using protein chips. *Nature Genetics* 26: 283-289

19. Ge. H. (2000) A universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. *Nucleic Acids Res.* 28,e3