BACTERIAL POPULATION ASSAY VIA K-MER ANALYSIS (EXTENDED ABSTRACT)

D. PAPAMICHAIL AND S. S. SKIENA

Computer Science Department,
SUNY at Stony Brook,
Stony Brook, NY 11794, USA
E-mail: {dimitris|skiena}@cs.sunysb.edu

D. VAN DER LELIE AND S. R. MCCORKLE

Biology Department,
Brookhaven National Laboratory,
Upton, NY 11973-5000, USA
E-mail: {vdlelied|mccorkle}@bnl.gov

Identifying and assaying the relative abundance of members of complex microbial communities is an important problem in ecology. Sandberg et al.[11] investigated the usage of genomic signatures to provide high identification percentages from short sequence samples. In this paper we present an improved naive Bayesian classification method using conditional probabilities, which can be used to classify unsequenced bacterial species, as well as identify and predict the frequency of the dominant species in mixed microbial populations.

1. Introduction

Microorganisms are the largest reservoir of genetic and biochemical diversity on earth. Understanding the structure, functional roles, and diversity of complex communities of microbes is key to using their wide-ranging capabilities. Microorganisms dominate the biosphere, yet most have not been identified or studied. Traditional methods for culturing and characterizing microorganisms limit analysis to those that will grow under laboratory conditions, which represent less than 1% of all microorganisms.

There is currently no effective technology to assay the relative abundance of complex microbial communities. Probe-based methods such as microarrays can only hope to detect species which have already been at least partially sequenced; but these represent a vanishingly small fraction of the millions of microbial species. The *genomic sequence tag* (GST) approach, pioneered by Dunn et al.,[3] promises to make such analysis possible for the first time.

Genomic sequence tags (GSTs) are short (e.g. 21 base) sequence fragments sampled more or less at random from microbial genomes in the given population. Such tags are inexpensive to assay, yet long enough to allow for straightforward species identification against sequence databases. However, such identification techniques cannot hope to iden-

tify non-sequenced species, which will constitute the vast majority of microbes into the foreseeable future.

Hope comes from the intriguing results of Sandberg et al.,[11] who investigated identifying bacterial genomic sequences using k-mer distributions instead of sequence matching. They found that microbial species could be correctly identified with an accuracy of approximately 85% from k-mer distributions from sequence samples as short as 400 bases. In this paper, we build on these observations in several directions:

- *Improved Classification Method* – We give a classification method based on conditional probabilities which performs substantially better than the method of Sandberg et al.[11] when using small amounts of sample sequence. In particular, our conditional probability approach improved species identification accuracy by up to 20% for short sequence segments (35bp) over the naive Bayesian classifier. These results are significant, because the cost of an assay increases linearly with the amount of required sequence.
- *Accurate Recognition Using Fragmented Sequence Data* – We demonstrate that k-mer analysis of short sequence tags is *more* effective than analysis of equivalent amounts of contiguous sequence. These results are fortuitous, because they imply that our results can be readily applied to GST and long SAGE[12] assays. They are also surprising, because (1) fragmentation inherently reduces the information available for k-mer analysis, and (2) individual short tags have a low (between 5-8%) sequence-recognition specificity, as shown in Table 1.
- *Signature Analysis for Unsequenced Species* – Recognizing new and unsequenced species is critical to tagging-based population analysis. Success depends upon the extent to which k-mer distribution is preserved among related strains and higher order classifications (order and genus).

  We demonstrate that k-mer distributions are well-preserved among related strains/species, by demonstrating that bacterial genomes can be clustered into natural groups according to k-mer distribution similarities.

  In the full paper we give accurate methods of identifying the order, genus and species of unsequenced bacteria from short tags. In particular, we show that unsequenced bacterial species can be accurately identified with respect to the 16S ribosomal RNA phylogenetic information on the basis of short tags.
- *Frequency Analysis of Mixed Populations* – We demonstrate that it is possible to identify bacterial species from mixed populations via k-mer distributions using modest amounts of sample sequence. Consider sequence tags collected from a mixture of two equally-represented species: our clustering-based approach proves capable of identifying at least one of two species 95% of the time.

  Further, our methods extend beyond species identification to frequency analysis. By careful analysis of modest amounts of sequence data, we can predict the frequency of the most dominant species in a population – even for unsequenced organisms. Further, our predictions grossly match the actual population over wide range of dominant-species frequencies.

This paper is organized as following. Genomic sequence tag methods and previous work on bacterial population assays are discussed in Section 1.1. In Section 2, we extend the work of Sandberg et al.[11] on k-mer recognition of contiguous sequence fragments. In Section 3, we generalize this work to short sequence tags. We consider the clustering and recognition of sequenced species with the respect to k-mer distribution and phylogenetic classifications in Section 4. Finally, we consider the problem of deconvolving tags from mixed species populations in Section 5.

1.1. *Previous Work*

Genomic Sequence Tags (GSTs) are short (21 base) fragments, product of a method for identifying and quantitatively analyzing genomic DNAs without a priori knowledge of the genome. The DNA is initially fragmented with a type II restriction enzyme. An olinucleotide adaptor containing a recognition site for M*me*I, a type IIS restriction enzyme, is then used to release 21-bp tags from fixed positions in the DNA relative to the sites recognized by the fragmenting enzyme. These tags are PCR-amplified, purified, concatenated and sequenced, to create a high-resolution GST sequence profile of the genomic DNA.

The GST approach has proven efficient in providing quantitative information for samples of different microbe sequences, even from non-sequenced genomes. Tags that appear in a sample with significantly different frequencies presumably come from organisms occurring with different frequencies in the population. Difficulty arises when specific organisms appear with similar frequency in the sample, or when tags appear with more than singular multiplicity.

This approach for characterizing prokaryotic or eukaryotic genomes is similar to long serial analysis of gene expression (long SAGE[12]) in that it produces large numbers of positionally defined 21-bp tag sequences that can be used to examine intra-specific genomic variation and, if genome information is available, provide immediate species identity. Other methods of large-scale scanning of microbial genomes on a quantitative and qualitative basis include the *Not*I passporting[14] and the restriction site tagged (RST) microarrays,[15] as well as the original SAGE procedure,[13] which produces positionally defined short tags of 13 to 14 bp with an increased throughput.

Genomic signatures based on compositions of nucleotides have been proven useful in identifying the origin of small sequences.[6] Frequencies of short sequence motifs – down to the level of dinucleotides – have shown great potential in providing a way of distinguishing different genuses in a coarse level,[4] but also differentiate between strains of the same species in eubacterial organisms.[7] Genomic signatures have been used for identification/detection of pathogenicity islands,[7] while differences in the use of mutually symmetric and complementary triplets distinguish between coding and non-coding genomic sequences.[10] Bacterial phage genome signatures are strongly correlated with the nature of the host and the extent to which the phage uses the host-cell machinery.[1] Intragenomically, the dinucleotide relative abundance varies little between 50 kilobase or longer windows on a given genome,[2] but is stable even in windows ranging in size from 50 kilobases down to 125 bases.[5]

Table 1. Average Origin Identification accuracy of 1000 randomly drawn 20-mers for varying k-mer size

	3-mer	4-mer	5-mer	6-mer	7-mer	8-mer
Recognition percentage	5.58%	6.03%	6.51%	6.68%	7.09%	8.16%

2. Identifying the Origin of Contiguous Sequence

Sandberg et al.[11] developed a naive Bayesian classifier to investigate the possibility of predicting the genome of origin for a specific genomic sequence. They found that sequences as short as 400 bases could be correctly classified with an accuracy of approximately 85%. The classifier was applied to 25 fully sequenced genomes, all of which came from unrelated species. The samples in all experiments originated from the same set of organisms.

The Sandberg et al. classifier calculates the probability of finding a sequence S of length N in a genome G_i as the product of the $N-(k-1)$ probabilities of finding each of the $N-(k-1)$ k-mers (motifs of length k, $k \leq N$) that constitute S in G_i. This is a valid measure of relating a sequence with a genome which can effectively be used as a rating, although it does not represent a correctly defined probability.

We propose a different method for classifying sequences. Instead of using the absolute probability of a k-mer being drawn from a genome G_i, we calculate the conditional probability of the last character of a k-mer appearing after the $k-1$ preceding characters of the k-mer. This conditional probability takes into consideration the dependence of the overlapping k-mers in a sequence, recognizing that the first $k-1$ characters have already appeared as a suffix of the previous k-mer, so it is the last character of the k-mer that will provide new information. This modification overcomes the k-mer independency assumptions and does not increase the order of needed computation. Further information can be found in the context of statistical natural language processing.[8]

We say that a bacterial genome is identified when the Bayesian/conditional probability, calculated as the product of the individual k-mer statistical probabilities, is the highest among the 104 probabilities calculated for all the genomes.

In order to compare the two methods with respect to the original study of Sandberg et al., we reproduced the original experiment conditions using 25 eubacteria and archea species whose completely sequenced genomes were available before September 2001. Random pieces of different sizes were drawn from each of the 25 microbe sequences and k-mer distributions used in calculating the probabilities for varying values of k.

Figure 1 compares the results of the naive Bayesian classifier method and the conditional probability method. We use whole genomic sequences to create the k-mer statistics and also draw random sequences from the same genomes. For each point in the graphs, all 25 microbe sequences are sampled and 10 samples are drawn in random. The classification accuracy is then averaged over the 250 cases.

Figure 1 shows that our conditional probability method performs consistently better, with up to 20% improvement in short sequences of 35 bases. Using the conditional probability method we can now identify short sequences of 400 bases with more than 90% accuracy using 8-mer frequency distributions.

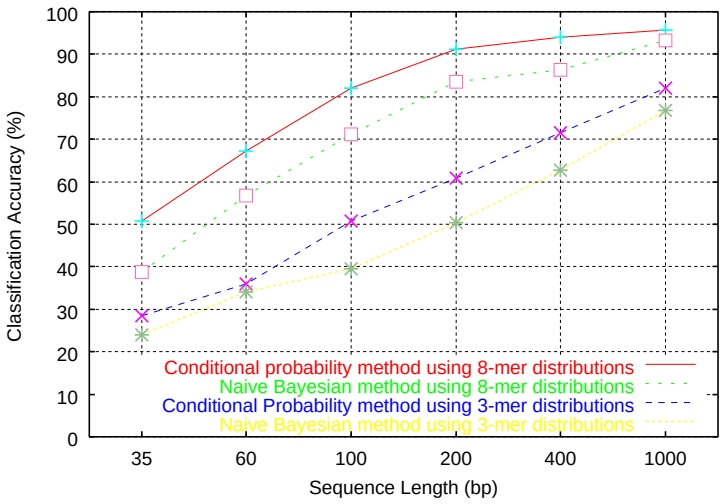


Figure 1. Comparison of Naive Bayesian and Conditional Probability Classifiers.

The probabilities in both methods are calculated by multiplying overlapping k-mer probabilities. One must be careful when handling k-mers that do not appear in specific distributions, since the frequency appears as 0. Since we want to be able to classify sequences from unknown bacteria, we must be able to handle k-mers that do not appear in some or all of the available genomes. For that reason, we discount the probabilities of finding a k-mer by assigning a small portion of the probability space to events that have not been encountered. We use Lidstone's Law[8] for discounting:

$$P(w) = \frac{C(w) + \lambda}{N + B\lambda}$$

where P is the assigned probability, w is a training instance, $C(w)$ is the training instance frequency, N is the number of training instances, B is the number of bin training instances are divided into and λ is a constant.

2.1. *Correcting for Repeated Strains*

Sandberg et al.[11] experimented on the 28 different archea and eubacteria organism genomic sequences available on May 2000. In September 2003, when we started our experiments, 104 full genome sequences were available from NCBI.

Although complete genome sequences are rapidly becoming available, the species diversity of available genomes is increasing at a slower rate because of research biases. Attention is concentrated on human pathogenic microbes, which results in different sequenced strains of similar species.

The frequency profiles of short oligonucleotides (k-mers) of certain length for different microbes, although providing enough specificity for distinguishing different species, becomes less effective for intra-species variation. Sandberg et al.[11] dealt with the problem of

reduced specificity by merging multiple strains of the same species in classes, resulting in 25 different classes, out of 28 available microbial sequences.

The 104 available bacterial genomes we studied included several resequenced strains. To eliminate this bias, we grouped bacteria into clusters based on correlation of the k-mer frequency distributions. We found that partitioning into 80 classes satisfied both a close proximity in distribution correlation difference while retaining biological significance, and so will use these classes in subsequent sections of this paper.

3. Dealing with Fragmented Sequences

The genomic sequence tag (GST) method results in fragments of approximately 20 bases extracted from specific locations in a genome, relative to restriction sites. Using short tags has the advantage of avoiding oversampling from repetitive or non-representative (in a genomic signature sense) regions, but individually have low specificity, inadequate of discriminating species, as seen in Table 1.

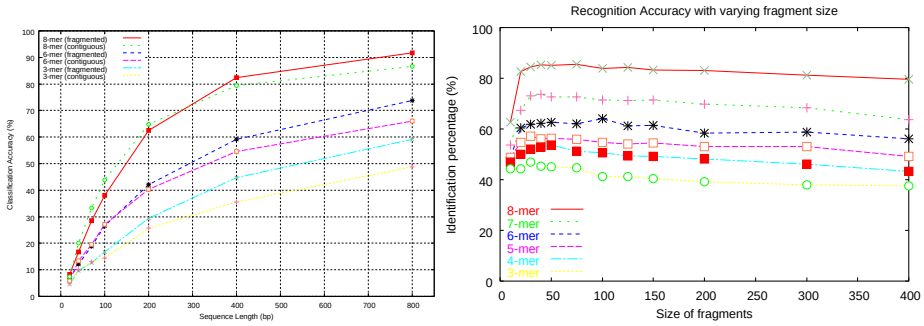
For a fixed size sample of sequence, fragmented sequences give a reduced amount of k-mers over unfragmented sequences. For example, a sequence of 400 bases can yield 396 5-mers if in one contiguous piece, but only 320 5-mers if the sequence is fragmented into 20 pieces of size 20. Still, for the same sequence length, our methods prove better at identifying fragmented sequences than contiguous sequences. Our results appear in Fig. 2(a). Here the contiguous and fragmented sequence experiment results are presented for 3-mer, 6-mer and 8-mer distributions.

To see how the tag size affects the recognition accuracy, we conducted an experiment where we kept the amount of available sequence constant at 400bp and varied the tag size. The results are shown in Fig. 2(b). We observe that the optimal tag size varies with the size of the k-mers used to analyze the data. For distributions of trinucleotide frequencies, the tag length where identification accuracy is maximized is around 30bp, where the optimal tag size is around 75bp for 8-mer frequency distributions. These experiments were performed on all 104 bacteria, with random sampling of 400bp in tags of varying size, where each data point represents 20 averaged repeats.

There are two reasons behind this surprising result. First, although the number of k-mers is reduced when using fragmented pieces, the size of the largest independent set of non-overlapping k-mers is not significantly smaller. With fragmented pieces we get at least one new non-overlapping k-mer every time we have a new piece. Second, by sampling from different locations of the genome we decrease the chance that the samples were drawn from an area not representative of the frequency distribution for the specific bacteria.

4. Phylogenetic Classification from k-mer Distributions

Estimates on the number of distinct bacterial species go into the millions, which makes it unlikely an observed species will correspond to a sequenced organism. In general, we are interested in obtaining coarser identification than distinct species. Thus we seek to identify which general class of bacteria our prediction indicates as the origin of a sequence.



(a) Comparison of classification accuracy of contiguous and fragmented sequences

(b) Classification accuracy as a function of fragment/tag size

Figure 2. Classification accuracy for single bacterial targets

In this section, we will show that sequenced bacteria can be identified with even greater accuracy with respect to phylogenetic categorization.

We use a procaryotic phylogenetic listing of small subunit 16S rRNA found at the Ribosomal Database Project Website.[9] Each bacterial species has a unique index number, consisting of a series of numbers separated by '.', each indicating a different genealogic attribute (kingdom, order, genus, species). We consider each of these numbers as branching points in our inferred tree.

All experiments in this section, involving identifying bacteria based on 16S ribosomal criteria, were averaged over 100 repeats, where fifty 20-mer tags (1000 bp) were randomly selected from each sampling bacteria.

4.1. *Identifying bacteria with known k-mer statistical distributions*

In this section, we analyze how often the top-scoring bacterium of our classifier happens to match the order, genus, and species of the closest appropriate species in the 16S rRNA database. We measure distance to our sampling bacteria using the inferred subtree (which now contains only our 104 fully sequenced genomes). Closest to our sampling bacteria is considered the species of the inferred subtree with the minimum distance in the number of node traversals (hops) needed to reach the former in the 16S rRNA phylogenetic tree.

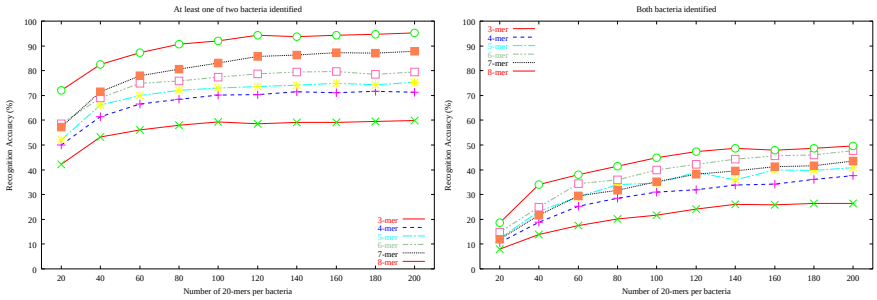
The bacterial samples were identified in the correct order with 99.98% accuracy, in the correct genus with 99.95% accuracy and in the correct species category with 99.83% accuracy. The exact strain of origin was identified correctly 99.42% of the time.

The higher than 99% positive identification exceeds even the classification accuracy using the statistically derived clustering tree by approximately 3%, for similar group sizes. For classification in the corresponding order, 98% of all bacteria were correctly classified 100% of the time, where the percentages for perfect identification in the genus and species categories were 97% and 87% respectively.

5. Identifying bacteria from mixed samples

More than just identify the members of a complex microbial community, we seek to assay their relative population frequency. We have shown that individual 20-mers identify the correct species only 8% of the time, using 8-mer frequencies, thus identifying the relative frequencies of bacteria in a mixed sample is a difficult task. An easier problem is the identification of a subset of species in the sample, especially the single most populous member of the sample.

For this purpose we constructed 20-mer tag data sets where half were derived from one bacteria and half from another. To identify the appropriate species, we cluster the 20-mers according to k-mer similarity, as follows: First we create for each 20-mer a vector of size 104, each position containing the conditional probability of the 20-mer being originated from the corresponding known bacteria genome. Then we cluster the 20-mers using *k-means* clustering into two clusters, according to the Euclidean distance of their corresponding vectors. We then classify the 20-mers of the two clusters separately, which gives us two candidate bacteria.



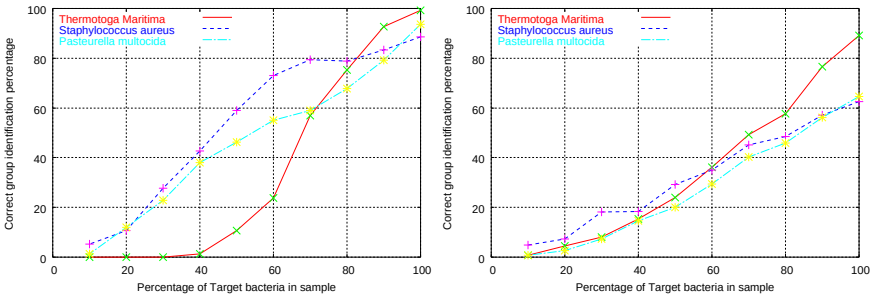
(a) Recognition rates of ≥ 1 bacteria (b) Recognition rates of both bacteria

Figure 3. Recognition accuracy of pairs of equi-probable bacteria, averaged over 500 different bacteria genome pairs.

Figure 3 shows that we can identify both bacteria 50% of the time and one of the two 95% of the time, provided we sample a sufficient number of 20-mers from each bacteria.

As a second experiment, we created samples where a specific percentage p is taken from a primary bacteria that we want to identify, where the rest of the sample is populated with 20-mers from randomly selected bacteria genomes. Then we try to identify the specified bacteria by creating a number of clusters and counting the total percentage of the identified clusters that matches the primary sampled bacteria.

Results for three different bacterial strains (*Thermotoga maritima*, *Pasteurella multocida* and *Staphylococcus aureus subsp. aureus Mu50*) are provided in Fig. 4. These three bacteria were selected as random choices of a hard, medium and easy-to- recognize bacteria strains by their k-mer distribution frequencies. *T. maritima* is pretty distant to other



(a) Using 2 k-means cluster

(b) Using 8 k-means clusters

Figure 4. Identifying bacteria from mixed sample containing percentage p of target bacteria, using 8-mer frequency distributions and variable cluster numbers

bacteria found in our database, *P. multocida* frequency distribution resembles few other in our database and *S. aureus* has another three relative strains present, which are divided in two groups according to an 80-group clustering of the available genomic sequences.

In Fig. 4, we can observe that recognition accuracy when a specific bacteria is comprising more than half of the sequence material in our sample is significant, especially when compared with an expected recognition percentage of 1.25% of a totally random sample. As expected, the recognition rates for *T. maritima* drop significantly faster than of the other representative samples, since having related strains in the database gives a larger space for recognition and *T. maritima* has a pretty distant k-mer frequency distribution. All three bacteria have a higher than 50% recognition percentage when they comprise more than 70% of the sample.

Comparing the results of clustering in 2 or 8 groups, we can see that 2-group clustering performs generally better, which is expected considering we are seeking to identify only one bacterial strain. The difference, though, diminishes (or even reverses, in the case of distant bacteria like *T. maritima*) when the bacteria comprises a smaller percentage of our sample. This can be explained by the fact that the specificity of the existing 20-mers of our target bacteria in the sample is absorbed by the noise of the other 20-mers in a larger group, where the target 20-mers could actually form smaller easier to identify groups (given enough 20-mers).

All majority-identifying experiments were performed 100 times for each bacteria to create data points in our graph, for 100 20-mers drawn randomly from the target and random other bacteria in our frequency distribution database, and averaged.

6. Conclusions

Through computational experiments, we have demonstrated that the analysis of short DNA sequence reads or tags can be used to determine the composition of complex microbial communities. Such methods hold particular promise as inexpensive, high-throughput meth-

ods of producing short sequence reads become available. Unlike microarray-based techniques for population analysis, our approach appears capable of recognizing previously unsequenced species. We are now applying these techniques to the analysis of actual sequence data from samples of the poplar rhizosphere grown under different environmental conditions.

References

1. B.E. Blaisdell, A.M. Campbell, and S. Karlin. Similarities and dissimilarities of phage genomes. *Proc. Natl. Acad. Sci.*, 93:5854–5859, 1996.
2. P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol.*, 16(10):1391–9, 1999.
3. J.J. Dunn, S.R. McCorkle, L.A. Praissman, G. Hind, D. van der Lelie, W.F. Bahou, D.V. Gnatenko, and M.K. Krause. Genomic Signature Tags (GSTs): A System for Profiling Genomic DNA. *Genome Research*, 12:1756–1765, 2002.
4. S.V. Edwards, B. Fertil, A. Giron, and P.J. Deschavanne. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst Biol.*, 51(4):599–613, 2002.
5. R.W. Jernigan and R.H. Baran. Pervasive properties of the genomic signature. *BMC Genomics*, 3:23, 2002.
6. S. Karlin. Global dinucleotide signatures and analysis of genomic heterogeneity. *Current Opinion in Microbiology*, 1:598–610, 1998.
7. S. Karlin, A.M. Campbell, and J. Mrazek. Comparative DNA analysis across diverse genomes. *Annu. Rev. Genet.*, 32:185–225, 1998.
8. C.D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA, 2003.
9. Michigan State University. Ribosomal Database Project (RDP). http://rdp.cme.msu.edu/download/SSU_rRNA/SSU_Prok.phylo.
10. C. Nikolaou and Y. Almirantis. Mutually symmetric and complementary triplets: differences in their use distinguish systematically between coding and non-coding genomic sequences. *Journal of Theoretical Biology*, 223(4):477–487, 2003.
11. R. Sandberg, C. Winberg, C. Branden, A. Kaske, I. Ernberg, and J. Coster. Capturing Whole-Genome Characteristics in Short Sequences Using a Naive Bayesian Classifier. *Genome Research*, 11:1404–1409, 2001.
12. V.E. Velculescu. Using SAGE to explore the genome. *Proceedings from SAGE 2001: Frontiers in transcritome exploration*, page 15, 2001.
13. V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler. Serial analysis of gene exression. *Science*, 270:484–487, 1995.
14. V. Zabarovska, A.S. Kutsenko, L. Petrenko, G. Kilosanidze, O. Ljungqvist, E. Norin, T. Midtvedt, G. Winberg, R. Mollby, V.I. Kashuba, I. Ernberg, and E.R. Zabarovsky. NotI passporting to identify species composition of complex microbial systems. *Nucleic Acids Res.*, 31(2):E5–5, 2002.
15. E.R. Zabarovsky, L. Petrenko, A. Protopopov, O. Vorontsova, A.S. Kutsenko, Y. Zhao, G. Kilosanidze, V. Zabarovska, E. Rakhmanaliev, B. Pettersson, V.I. Kashuba, O. Ljungqvist, E. Norin, T. Midtvedt, R. Mollby, G. Winberg, and I. Ernberg. Restriction site tagged (RST) microarrays: a novel technique to study the species composition of complex microbial systems. *Nucleic Acids Res.*, 31(16):e95, 2003.