The Second International Symposium on Optimization and Systems Biology (OSB'08)
Lijiang, China, October 31– November 3, 2008

Predicting Non-protein-coding RNA Genes in Escherichia Coli Using SVM with Signature Descriptor

Hong-Wei Liu[1,2,*]

[1]School of Information, Beijing Wuzi University, Beijing 101149, China
[2]Institute of Applied Mathematics Academy of
Mathematics and Systems Science, CAS, Beijing 100080, China

Abstract Non-protein-coding RNA (ncRNA) genes are known to play significant roles. Along with transfer RNAs, ribosomal RNAs and mRNAs, ncRNAs contribute to gene splicing, nucleotide modification, protein transport and regulation of gene expression. Several methods exist for predicting ncRNA genes in Escherichia coli (E.coli). In this paper, we describe a very general, high-throughput method for predicting ncRNA genes in E.coli. The method predicts more than two hundred intergenic regions to contain ncRNA genes, and over half of these overlap with previous tested candidates. Our results indicate that the number of ncRNA genes in E.coli is larger than what has previously been estimated.

Keywords Escherichia coli; non-protein-coding RNA; SVM; signature descriptor

1 Introduction

Cellular RNAs that do not function as messenger RNAs (mRNAs), transfer RNAs (tRNAs) or ribosomal RNAs (rRNAs) comprise a diverse class of molecules that are commonly referred to as non-protein-coding RNAs (ncRNAs) whose function lies in the RNA sequence itself and not as information carriers for protein synthesis[8, 14]. These molecules have been known for quite a while, but their importance was not fully appreciated until recent genome-wide searches discovered thousands of these molecules and their genes in a variety of model organisms. Although long believed to be a minor gene class, in recent years it became increasingly clear, that ncRNAs constitute a large portion of the transcriptional output from the genomes. There is a constantly growing number of novel RNAs that do not encode proteins and do not perform housekeeping functions in the cells. NcRNAs are implicated in a number of cellular processes, but in many cases, it is difficult to precisely determine mechanism of their action[28, 21, 17].

In Escherichia coli, the number of experimentally verified small RNA (sRNA) genes (ncRNA genes excluding rRNA and tRNA,such as snRNAs, snoRNAs, RNAseP RNA, tmRNA in the literature) has increased rapidly. Only 10 sRNA genes were known in 1999 [26], whereas a recent survey listed 55 known sRNA genes [13]. Subsequent RNA

*E-mail address: ryuhowell@163.com.

cloning experiments increased the number of known sRNA genes to 62 [25]. Most of these sRNA genes were identified in seven studies describing systematic searches for new sRNA genes [1, 27, 19, 5, 6, 22, 20]. Together, these seven studies have predicted about 1000 non-redundant sRNA candidates that are yet to be confirmed [13]. Note, however, that only 95 candidates were predicted by more than one study.

We describe a method that uses sequence patterns to predict ncRNA genes in E.coliąfs intergenic regions. Our method is most similar to the methods of [20], we conclude with a brief discussion of the similarities. The main strengths of the method as compared to other methods are as follows. Firstly, our method uses only experimental and sequence information and can therefore be used to study organisms where little is known, it works well with a much larger number of intergenic sequences (negative examples) than known ncRNA sequences (positive examples) [5]. Secondly, we use a local description of intergenic sequences and known ncRNA sequences, which may be more consistent with the actual biology sequences, and uses the local description information of the intergenic sequences and known ncRNA sequences directly as input information, which helps to reduce any potential bias from input feature selection and encoding [5]. Thirdly, it is very robust when it comes to noise in the training data, as for instance intergenic regions that actually are ncRNAs. And lastly, our method does not require physico-chemical information, and do not need to have prior knowledge of ncRNA genes, it does not rely on sequence conservation to predict ncRNA genes.

2 MATERIALS AND METHOD

2.1 Sequence data

Training and testing were performed using Escherichia coli K-12 strain MG1655 cells genome sequence (U00096.2) and its annotations extracted from a 2006 release (release 87) of the EMBLąfs FTP server (http://www.ebi.ac.uk/genomes/bacteria.html). Based on annotations and previous studies [1, 19], we collected a set of 157 experimentally verified ncRNA sequences. These sequences consisted of 86 tRNAs, 22 rRNAs and 49 other sRNA genes. Note that one of these sRNAs was the strain-dependent uptR gene[12]. Based on the positions of known ncRNA genes and protein coding sequences (CDS), we constructed a set of intergenic sequences (INT) by removing all parts of the genome containing ncRNAs and CDSs, along with 100 nt on each side. This resulted in 660 subsequences totaling 130,931 nt and each containing no less than 50 nt. The 50 nt size was chosen because the smallest ncRNA in our dataset was 53 nt (dicF).

2.2 Support vector machine(SVM)

SVMs are classifiers that are described thoroughly by their inventor [23], and SVMs are very adaptable and have been applied successfully to a wide variety of problems. Recently, there has been interest in the application of SVMs to biological problems such as classification of gene expression data[11], homology detection [16] and prediction of protein-protein interaction [3], as well as many additional problems. Unlike many traditional methods which implement the Empirical Risk Minimization Principle which aims at minimizing the training error, SVM implements the Structural Risk Minimization Principle which seeks to minimize an upper bound of the generalization error, which eventually results in better generalization of SVM than that of traditional techniques.

To describe an SVM precisely, suppose the data are given as pairs $\{(x_i, y_i)\} \subset R^n \times \{\pm 1\}$, and the classifiers created by SVM algorithm are sequence patterns that can only give binary answers. In other words, given a sequence, each pattern answers either ąőyesąŕ(1) orąőnoąŕ(-1), as to whether the pattern matches parts of the sequence or not. Using this notation an SVM assumes the form $f(x) = \sum_i \alpha_i y_i k(x_i, x) + b$, where $f : R^n \rightarrow R$ is a decision function (x belongs to class 1 if $f(x)$ is greater than some threshold t , or to class -1 otherwise), $k : R^n \times R^n \rightarrow R$ is a kernel function, otherwise known as a dot product in some vector space , and the constants b and α_i are obtained by solving a quadratic programming problem (for details see [4]). The threshold t is typically 0, although it may be varied to obtain classifiers that are more or less accurate on positive predictions.

We use SVMs create classifiers that predict whether or not a sequence is an ncRNA gene. SVMs in the context takes as input a set of positive and negative sequences and creates a classifier that predicts whether or not an unknown sequence belongs to the positive set. Here, the positive and negative sequences are the ncRNAs and INT sequences described in the previous section. Thus, the classifier created by SVMs can predict whether or not a given sequence comes from a ncRNA.

2.3 Signature

One of the main challenges in using SVMs for the prediction of ncRNA in genome sequence is a suitable encoding of the genome sequences information in some vector space. In our case, we have the problem of representing variable length genome sequences as vectors containing the necessary information to be distinguished.

Our solution to this problem is to use the signature molecular descriptor[24, 9, 10, 7, 18]. Signature is based on the molecular graph of a molecule, where the vertices denote atoms in the molecule, and the edges correspond to the bonds between atoms. In this context, a molecule is characterized by a set of canonical subgraphs, each rooted on a different vertex with a predefined level of branching referred to as the height h. The branching of a vertex is an extended-degree sequence that describes the local neighborhood, up to a distance h away from the root. A height 0 signature consists of a count of the number of each of the ribonucleoside types present in the strand. A height 1 signature counts each of the possible tri-mers present in the peptide. A height 2 signature counts each of the possible five-mers present in the sequence, and so forth. In fact, signature was originally developed to describe molecules in cheminformatics. Recently, however, signature has also been used successfully in applications to HIV protease-1 peptide prediction [9] and inverse design of LFA-1/ICAM-1 peptide [7]. Thus, signature is information rich, and, in particular, enables the solution of inverse problems.The choice of the signature height depends on the specific problem. In our experience the best heights are usually 0, 1, or 2. For the prediction of ncRNA problem, we found height 1 to provide the best test set accuracy, and therefore consider only height 1 signatures in this paper and formulate signature as a function s : variable length genome sequencesąú F defined by $s(A) = \sum_i \sigma_i z_i$, where A is a genome sequence, z_i is a basis vector in the signature space $F \cong R^N$ and σ_i is the number of occurrences of z_i in A.

As an example, consider the seven-letter genome sequence ATCGGCG. All height 1 signatures are based on trimers and there are five trimers in this sequences: ATC, TCG, CGG, GGC and GCG. Each signature consists of a root (the middle letter) and its two neighbors, ordered alphabetically. Thus, the signatures corresponding to the

four trimers are T(AC), C(GT), G(CG), G(CG) and C(GG), so that s(ATCGGCG) = T(AC)+C(GT)+2G(CG)+C(GG). Notice that CGG and GGC generate the same signature (due to symmetry) and therefore contribute two occurrences to the sum $s(A) = \sum_i \sigma_i z_i$.

Signature has been a useful descriptor in the past, and has the important practical advantage for us in that it provides a vector representation of a genome sequence. We exploit this fact to develop a signature-based SVM for use in the prediction of ncRNA problem.

2.4 Implement

We use the signature method to deal with the induced 660 INTs and 157 ncRNAs and get the data_file satisfied the following format:

$$< line >=< target >< feature >:< value > ... < feature >:< value >$$

where the first entry$< target >= [+1|-1]$ gives the class labels(ncRNA or INT), $< feature >= [integer]$ denotes the basis vector's index in the basis vector set and $< value >= [float]$ denotes the basis vector's weight satisfied that the summation of the weight's square in the same sequence is equal to 1. Note that the target value and each of the feature/value pairs are separated by a space character and feature/value pairs MUST be ordered by increasing feature number and features with value zero can be skipped. Note again that indices start at 1.

SVMs have several advantages over other classifiers though we do not discuss them here. Instead, we refer to Vapnik [23] and Bennett and Campbell[2], among hers. To implement the SVMs in this paper, we used the SVM^{light} algorithm [15] with radial basis kernel based on the induced data_file.

3 RESULTS

When a model is evaluated on a positive and negative set of sequences, four statistics (counts) can be defined: the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). These represent the both predicted and observed, predicted but not observed, neither predicted nor observed, and not predicted but observed, respectively. We used 10-fold cross-validation to train and test our machine learning algorithm and calculated accuracy, precision, sensitivity and specificity. To be precise, we first divided the sets of ncRNA and INT sequences (at random) into 10 roughly equal-sized non-overlapping subsets(i.e., each subset contains 66 INTs and 15 or 16 ncRNAs). We used each subset in turn as a test set, while we trained our method on the union of the remaining 9 subsets. We evaluated the performance of our classifier by computing accuracy (TP+TN)/(TP+FP+TN+FN), precision TP/(TP+FP), sensitivity TP/(TP+FN) and specificity TN/(TN+FP).

In addition to observations about specific classifiers, the accuracy, precision and sensitivity are useful for measuring the behavior of a classifier in general. In particular, the accuracy gives the overall performance of a classifier, the precision gives the percentage of positive predictions that are actually positive and the sensitivity gives the percentage of actual positives that are predicted. By looking at the precision and sensitivity statistics, we can determine if a classifier will identify positives correctly. If a classifier has a high

precision and a low sensitivity, then it is likely to be correct when it makes a positive prediction, although it will make many false negative predictions. Conversely, a classifier with a low precision and a high sensitivity is likely to identify most true positives, even though many of its predictions will be false. In some sense, the first classifier is too conservative while the second is too optimistic.

To estimate the optimal regularization value, we tried several different values and used the one with a average high precision and low sensitivity in the 10 test subsets. For example, when an SVM was trained with radial basis kernel and default regularization value, the average accuracy is 92.776% ,the average precision is 66.375% and the average sensitivity is 79.458%. These optimal models had predicted on average 16 false positive sequences in the test subsets. The algorithm identifies nearly 85% of the sRNAs in the database and predicts 258 intergenic regions to contain ncRNA genes, 145 of these overlap with previous tested candidates, and 113 potential new ncRNA genes of which 53 are confirmed by all these optimal models.

4 Discussion and Conclusion

We have described a novel method that use SVMs, sequence information and experimental data to predict non-protein-coding RNA genes and explored its applicability by analyzing E.coli intergenic regions. An automatic method is used for generating signatures, which depends only on sequence information and does not require us to perform transforming sequence information into physicio-chemical information. Our method also has the advantage of using a principled method (SVMs) to obtain our final classifier by statistical evaluation.

The method predicts more than two hundred intergenic regions to contain ncRNA genes, and over half of these overlap with previous tested candidates. Our results indicate that the number of ncRNA genes in E.coli is larger than what has previously been estimated [29]. Several groups have searched for new ncRNAs in E.coli [25, 1, 27, 19, 6, 22, 20], which have resulted in a list of about 1000 non-redundant and untested candidates [13]. This is because the estimates in the literature were partly based on the number of ncRNA genes predicted by more than one method. We have extended this list by 41.85% (that is, 113/(113+157)), which is a significant increase.

Here we must to claim that more potential new ncRNA genes will be predicted if we do not exploit too conservative attitudes towards dealing with the data set and model selection.

Acknowledges

This work is partly supported by Chinese National Science Foundations (grant No. 10701080) and Information & Control Research Project on Beijing Wuzi University.

References

[1] Argaman,L., Hershberg,R., Vogel,J., Bejerano,G., Wagner,E.G.H., Margalit,H. and Altuvia,S. (2001) Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. Curr. Biol., 11, 941-950.

[2] Bennett,K.P. andCampbell,C. (2000) Support vectormachines: hype or hallelujah. ACM SIGKDD Explorations, 2, 1-13.

[3] Bock,J. andGough,D. (2001) Predicting proteinÍCprotein interactions from primary structure. Bioinformatics, 17, 455-460.

[4] Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. Knowl. Discov. Data Mining, 2, 121-167.

[5] Carter,R.J., Dubchak,I. and Holbrook,S.R. (2001) A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res., 29, 3928-3938.

[6] Chen,S., Lesnik,E.A., Hall,T.A., Sampath,R., Griffey,R.H., Ecker,D.J. and Blyn,L.B. (2002) A bioinformatics based approach to discover small RNA genes in the Escherichia coli genome. Biosystems, 65, 157-177.

[7] Churchwell,C.J., Rintoul,M.D., Martin,S., Visco,D., Kotu,A., Larson,R.S., Sillerud,L.O., Brown,D.C. and Faulon,J.L. (2004)The signature molecular descriptor. 3. Inverse quantitative structureÍCactivity relationship of ICAM-1 inhibitory peptides. J. Mol. Graph. Model.

[8] Eddy,S.R. (2001) Non-coding RNA genes and the modern RNA world. Nature Rev. Genet., 2, 919-929.

[9] Faulon,J.-L., Churchwell,C. and Visco,D.P.,Jr. (2003a) The signa- ture molecular descriptor. 2. Enumerating molecules from their extended valence sequences. J. Chem. Inf. Comput. Sci., 43, 721-734.

[10] Faulon,J.-L., Visco,D.P.,Jr. and Pophale,R.S. (2003b) The signature molecular descriptor. 1. Extended valence sequences vs. toplo- gical indices in QSAR and QSPR studies. J. Chem. Inf. Comput. Sci., 43, 707-720.

[11] Furey,T., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics, 16, 906-914.

[12] Guigueno,A., Dassa,J., Belin,P. and Boquet,P.L. (2001)Oversynthesis of a new Escherichia coli small RNA suppresses export toxicity of DsbAąś- PhoA unfoldable periplasmic proteins. J. Bacteriol., 183, 1147-1158.

[13] Hershberg,R., Altuvia,S. and Margalit,H. (2003) A survey of small RNA-encoding genes in Escherichia coli. Nucleic Acids Res., 31, 1813-1820.

[14] Huttenhofer,A. and Vogel,J. (2006) Experimental approaches to identify non-coding RNAs. Nucleic Acids Res., 34(2), 635-646.

[15] Joachims,T. (1999) Making large-scale SVM learning practical. In Sch?lkopf,B., Burges,C.J.C. and Smola,A.J. (eds), Advances in Kernel MethodsÍCSupport Vector Learning. MIT Press, Cambridge, MA, pp. 169-184.

[16] Leslie,C., Eskin,E.,Weston,J. and Noble,W. (2003)Mismatch string kernels for SVM protein classification. In Becker,S., Thrun,S. and Obermayer,K. (eds.), Advances in Neural Information Processing Systems. MIT Press, Vol. 15, 1441-1448.

[17] Mattick,J.S. (2004) RNA regulation: a newgenetics? Nature Rev.Genet., 5, 316-323.

[18] Martin,S., Roe,d. and Faulon,J.L. (2005) Predicting proteinÍCprotein interactions using signature products. Bioinformatics, 20, 218-226.

[19] Rivas,E., Klein,R.J., Jones,T.A. and Eddy,S.R. (2001) Computational identification of noncoding RNAs in E. coli by comparative genomics. Curr. Biol., 11, 1369-1373.

[20] sætrom,P., Sneve,R., Kristiansen,K.I., Grünfeld,T., Rognes,T. and Seeberg,E. (2005) Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. Nucleic Acids Res., 33, 3263-3270.

[21] Storz,G. (2002) An expanding universe of noncoding RNAs. Science, 296, 1260-1263.

[22] Tjaden,B., Saxena,R.M., Stolyar,S., Haynor,D.R., Kolker,E. and Rosenow,C. (2002) Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. Nucleic Acids Res., 30, 3732-3738.

[23] Vapnik,V. (1998) Statistical Learning Theory. Wiley Interscience, New York.

[24] Visco,D.P.,Jr, Pophale,R.S., Rintoul,M.D. and Faulon,J.L. (2002) Developing a methodology for an inverse quantitative structureÍCactivity relationship using the signature molecular descriptor. J. Mol. Graph. Model, 20, 429-438.

[25] Vogel,J., Bartels,V., Tang,T.H., Churakov,G., Slagter-Jager,J.G., Huttenhofer,A. andWagner,E.G.H. (2003) RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. Nucleic Acids Res., 31, 6435-6443.

[26] Wassarman,K.M., Zhang,A. and Storz,G. (1999) Small RNAs in Escherichia coli. Trends Microbiol., 7, 37-45.

[27] Wassarman,K.M., Repoila,F., Rosenow,C., Storz,G. and Gottesman,S. (2001) Identification of novel small RNAs using comparative genomics and microarrays. Genes Dev., 15, 1637-1651.

[28] Wassarman,K.M. (2002) Small RNAs in bacteria: diverse regulators of gene expression in response to environmental changes. Cell, 109, 141-144.

[29] Zhang,Y., Zhang,Z., Ling,L., Shi,B. and Chen,R. (2004) Conservation analysis of small RNA genes in Escherichia coli. Bioinformatics, 20, 599-603.