BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 16 2010, pages 2051–2052
doi:10.1093/bioinformatics/btq299

Genome analysis

Advance Access publication June 10, 2010

cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data

Fengfeng Zhou[1] and Ying Xu[1,2,*]

[1]Computational Systems Biology Laboratory, Department of Biochemistry and Molecular Biology, and Institute of Bioinformatics, and BioEnergy Science Center (BESC), University of Georgia, Athens, GA 30602, USA and [2]College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China

Associate Editor: Alex Bateman

ABSTRACT

Summary: Huge amount of metagenomic sequence data have been produced as a result of the rapidly increasing efforts worldwide in studying microbial communities as a whole. Most, if not all, sequenced metagenomes are complex mixtures of chromosomal and plasmid sequence fragments from multiple organisms, possibly from different kingdoms. Computational methods for prediction of genomic elements such as genes are significantly different for chromosomes and plasmids, hence raising the need for separation of chromosomal from plasmid sequences in a metagenome. We present a program for classification of a metagenome set into chromosomal and plasmid sequences, based on their distinguishing pentamer frequencies. On a large training set consisting of all the sequenced prokaryotic chromosomes and plasmids, the program achieves ∼92% in classification accuracy. On a large set of simulated metagenomes with sequence lengths ranging from 300 bp to 100 kbp, the program has classification accuracy from 64.45% to 88.75%. On a large independent test set, the program achieves 88.29% classification accuracy.

Availability: The program has been implemented as a standalone prediction program, cBar, which is available at http://csbl.bmb.uga.edu/~ffzhou/cBar

Contact: xyn@bmb.uga.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 16, 2010; revised on May 10, 2010; accepted on June 3, 2010

1 INTRODUCTION

One of the major challenges in analyzing a metagenome lies in the reality that metagenomic data are generally very complex in terms of their composition, often consisting of sequence fragments from numerous genomes possibly from different kingdoms (McHardy and Rigoutsos, 2007). While analyses of a metagenome as a whole could definitely reveal useful information about a microbial community, more detailed analyses generally requires the binning of the metagenomic sequences into multiple taxonomical groups according to some criteria (McHardy and Rigoutsos, 2007). There have been a number of published computer programs designed to bin metagenomic sequences into multiple groups, each of which consists of sequences from the same taxonomical group (Chan *et al*., 2008; Diaz *et al*., 2009; McHardy and Rigoutsos, 2007), e.g. from the same genera.

Another important metagenome binning problem that has not been well addressed by metagenome sequence analysts is to separate chromosomal from plasmid sequences. This problem is important because sequence features associated with various classes of genomic elements in chromosomal and plasmid genomes are generally different (Zhou *et al*., 2008); hence separation of these two classes of sequences is the prerequisite to accurate computational identification of such genomic elements. For example, identification of genes in chromosomal and plasmid sequences represents two different problems even when they are from the same organism, since their di-codon frequency biases in coding versus non-coding regions, the basis for computational gene finding, could be substantially different (Davis and Olsen, 2010). For genome sequencing projects, this has not been a general problem since plasmids have been typically extracted using alkaline lysis (Li *et al*., 2008) or removed using nuclease (Kock *et al*., 1998) so they can be sequenced separately from chromosomal sequences. However, microbial community sequencing projects typically do not separate chromosomes from plasmids in advance due to various technical reasons. So the sequenced metagenome is generally a mixture of chromosomal and plasmid sequences, hence raising the needs to computationally separate them.

Here, we present the first computer program for classification of a given metagenome into chromosomal and plasmid sequences, based on observed differences in their pentamer frequencies. Using the pentamer frequencies collected from 881 completely sequenced prokaryotic genomes, including both chromosomal and plasmid sequences, we have trained an sequential minimal optimization (SMO)-based model (Cessie and Houwelingen, 1992) to separate chromosomal from plasmid sequences, which achieved consistent accuracies at ∼90% in our 10-fold cross-validations (10FCVs) on our training data, on an independent testing dataset and on the simulated metagenomes.

2 MATERIALS AND METHODS

We downloaded the genome sequences of all the 881 completely sequenced prokaryotic genomes from the NCBI Microbial Genome Projects at http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, as of May 4, 2009. This dataset was partitioned into a training dataset (denoted as *Training*) consisting of all 808 prokaryotic genomes released before January 1, 2009 and a testing dataset (denoted as *Testing*) consisting of the remaining 73 genomes, all released after January 1, 2009.

*To whom correspondence should be addressed.

We calculated the frequencies of each k-mer and its reverse complement (considering each such pair as one entity) on the whole sequences from both sets. The collection of 512 frequencies for each genome sequence is called its *k-mer profile*, which is used as the input of the classification models. We have used the classification tool provided in the Weka package (Frank *et al*., 2004) to train our classifier. We chose $k=5$, due to its significant improvement in the overall accuracy, as described in details in the Supplementary Material.

We have used the following measurements to evaluate the classification accuracy of a model: sensitivity (Sn), specificity (Sp), the overall accuracy (Ac) and the Matthews correlation coefficient (MCC) (Matthews, 1975). In addition, the area under the receiver operating characteristic curve (AUC) is also used to measure the overall performance of a prediction model. The detailed definitions of the aforementioned measurements can be found in the Supplementary Material.

We conducted the 10FCV on the dataset *Training* and *All*. We studied whether the model has consistent prediction performance on an independent dataset *Testing*, by training the model on the dataset *Training*. We also tested our model on simulated metagenome sequences, as discussed in the Supplementary Material.

3 RESULTS

We have tested five types of classification approaches, namely C4.5 decision tree, Bayes Network, Support Vector Machine, SMO and Nearest neighbor, for solving our classification problem. The results of our three evaluations, as described above, on the five approaches are listed in Table 1. On the 10FCVs on the *Training* dataset, four out of five algorithms, except for Bayes network, worked well with the overall accuracy better than 83% and the SMO approach has the highest performance in all the performance measurements with overall accuracy 91.82%. On the *Testing* dataset, the performance of the trained classifier dropped slightly as shown in Table 1, and the SMO approach remains the best performer with Ac = 88.29%. Using the 10FCVs on the combined *Training* and *Testing* set, the SMO-based classifier has approximately the same level of performance accuracy, still the best among the five. The ∼90% AUC-value strongly suggests that the SMO-based classifier is accurate and robust.

Some bacteria harbor chromids, whose lengths are between those of chromosomes and plasmids (Harrison *et al*., 2010). cBar predicted 84.85% of them as chromosomes, due to their similarities in the nucleotide compositions to the host chromosomes (Harrison *et al*., 2010).

A few other validations, including the comparison with a BLAST-based strategy, can be found in the Supplementary Material.

Table 1. The prediction performance by five classification approaches, C4.5 decision tree, Bayes network, SVM with the RBF kernel, SMO and nearest neighbor

Strategy	Algorithm	Sn	Sp	Ac	MCC	AUC
10FCV *Training*	C4.5	0.8766	0.7832	0.8371	0.6646	0.832
	Bayes net	0.6594	0.8034	0.7203	0.4585	0.807
	RBF	0.9463	0.7613	0.8681	0.7315	0.854
	SMO	0.9474	0.8783	0.9182	0.8321	0.913
	NN	0.9177	0.8128	0.8734	0.7395	0.865
Testing	C4.5	0.7667	0.8627	0.8108	0.6280	0.811
	Bayes net	0.6333	0.8039	0.7117	0.4398	0.764
	RBF	0.8500	0.8235	0.8378	0.6735	0.837
	SMO	0.8667	0.9020	0.8829	0.7664	0.884
	NN	0.8667	0.8039	0.8378	0.6730	0.835
10FCV *All*	C4.5	0.8759	0.8020	0.8445	0.6809	0.841
	Bayes net	0.6877	0.7905	0.7314	0.4730	0.815
	RBF	0.9358	0.7746	0.8672	0.7290	0.855
	SMO	0.9422	0.8887	0.9195	0.8349	0.915
	NN	0.9112	0.8165	0.8709	0.7349	0.864

These algorithms were evaluated using the 10FCV on the *Training* dataset, the *Testing* dataset and the *All* dataset.

ACKNOWLEDGEMENTS

We would like to thank the CSBL members for helpful discussions and three anonymous reviewers for their constructive comments.

Funding: National Science Foundation (DBI-0354771, ITR-IIS-0407204, DBI-0542119 and CCF0621700, in part); National Institutes of Health (1R01GM075331 and 1R01GM081682, in part); Distinguished Scholar grant from the Georgia Cancer Coalition (in part); grant for the BioEnergy Science Center, which is a U.S. Department of Energy BioEnergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science (in part).

Conflict of Interest: none declared.

REFERENCES

Cessie,S.L. and Houwelingen,J.C.V. (1992) Ridge estimators in logistic regression. *Appl. Stat.*, **41**, 191–201.

Chan,C.K. *et al*. (2008) Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*, **9**, 215.

Davis,J.J. and Olsen,G.J. (2010) Modal codon usage: assessing the typical codon usage of a genome. *Mol. Biol. Evol.*, **27**, 800–810.

Diaz,N.N. *et al*. (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, **10**, 56.

Frank,E. *et al*. (2004) Data mining in bioinformatics using Weka. *Bioinformatics*, **20**, 2479–2481.

Harrison,P.W. *et al*. (2010) Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol*., **18**, 141–148.

Kock,J. *et al*. (1998) Duck hepatitis B virus nucleocapsids formed by N-terminally extended or C-terminally truncated core proteins disintegrate during viral DNA maturation. *J. Virol.*, **72**, 9116–9120.

Li,X. *et al*. (2008) A continuous process to extract plasmid DNA based on alkaline lysis. *Nat. Protoc.*, **3**, 176–180.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys. Acta*, **405**, 442–451.

McHardy,A.C. and Rigoutsos,I. (2007) What's in the mix: phylogenetic classification of metagenome sequence samples. *Curr. Opin. Microbiol.*, **10**, 499–503.

Zhou,F. *et al*. (2008) Barcodes for genomes and applications. *BMC Bioinformatics*, **9**, 546.