Distinguishing Species Using GC Contents in Mixed DNA or RNA Sequences

Evolutionary Bioinformatics
Volume 14: 1–4

DOI: 10.1177/1176934318788866
SAGE

Kamran Karimi[1,2], Daniel M Wuitchik[1], Matthew J Oldach[1] and Peter D Vize[1,2]

[1]Department of Biological Sciences, University of Calgary, Calgary, AB, Canada. [2]Department of Computer Science, University of Calgary, Calgary, AB, Canada.

ABSTRACT: With the advent of whole transcriptome and genome analysis methods, classifying samples containing multiple origins has become a significant task. Nucleotide sequences can be allocated to a genome or transcriptome by aligning sequences to multiple target sequence sets, but this approach requires extensive computational resources and also depends on target sequence sets lacking contaminants, which is often not the case. Here, we demonstrate that raw sequences can be rapidly sorted into groups, in practice corresponding to genera, by exploiting differences in nucleotide GC content. To do so, we introduce GCSpeciesSorter, which uses classification, specifically Support Vector Machines (SVM) and the C4.5 decision tree generator, to differentiate sequences. It also implements a secondary BLAST feature to identify known outliers. In the test case presented, a hermatypic coral holobiont, the cnidarian host includes various endosymbionts. The best characterized and most common of these symbionts are zooxanthellae of the genus Symbiodinium. GCSpeciesSorter separates cnidarian from Symbiodinium sequences with a high degree of accuracy. We show that if the GC contents of the species differ enough, this method can be used to accurately distinguish the sequences of different species when using high-throughput sequencing technologies.

KEYWORDS: classifying species, DNA, RNA, GC contents, SVM, C4.5 decision tree

RECEIVED: March 27, 2018. **ACCEPTED:** June 22, 2018.

TYPE: Software or Database Review

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science and Engineering Research Council (NSERC) of Canada and National Institutes of Health (NIH) grant P41 HD064556.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Kamran Karimi, Department of Biological Sciences, University of Calgary, 2500 University Dr. NW Calgary, AB T2N 1N4, Canada.
Email: kkarimi@ucalgary.ca

Background

In high throughput sequencing (HTS) experiments, it is difficult to determine the species of origin for sequenced reads because RNA or DNA isolated from biological samples rarely belongs to a single species. This can be due to various microorganisms living together. Often species are tightly linked through symbiosis and share tissue space. A popular method to determine species composition of a sample is to sequence specific regions containing internally transcribed spacers where single nucleotide polymorphism (SNP) differences in this region elucidate which species are present.[1] While this method is effective for describing community composition, it is unable to evaluate the species of origin from sequences generated when exploring other regions in the transcriptome or genome.

BLAST is a tried and true method for aligning mixed sequences to different genes and to various organisms.[2] Unfortunately, this method is very slow due to its computational requirements and is a hindrance when exploring millions of reads associated in HTS experiments. To overcome this volume of data, modern approaches using "mappers," such as Bowtie2,[3] can quickly align reads to a reference genome or transcriptome. These methods, however, are limited by the quality and availability of reference files and excludes sequences from different species.[4] Since there are limited numbers of quality reference genomes and transcriptomes available, classifying species presents a challenge in HTS experiments of non-model organisms and the communities within a single biological sample.

A prime example of wanting to classify a diverse range of sequenced reads to species of origin is with experiments studying symbiosis. For these experiments, it is relevant to characterize community compositions within a sample and to differentiate between reads that belong to host vs symbiont/parasite. This type of work has been addressed by Lehnert et al,[5] who developed a system called TopSort to predict anemone (host) versus *Symbiodinium* (symbiont) nucleotide sequences based on differing GC content and codon usage. Their system used Support Vector Machines (SVM) which classified sequences into their respective host vs symbiont groups. As TopSort is not publicly available, our goal was to generate a simple, fast, and accessible implementation of a classifier that operates similarly to TopSort that would be freely available to all.

SVMs are supervised classifiers, which work by establishing a linear hyperplane to separate data into classes based on a predetermined training set. SVMs are widely used for text classification as they are flexible and are very accurate.[6] While there are numerous machine learning methods for classifying data, a comparative study using microarray gene expression data found that SVM classifiers were consistently more accurate than radial basic function neural nets, multi-layer perception neural nets, Bayesian and decision trees classification methods.[7]

Since we are using a single criterion, theoretically other classifiers may also provide good results. To test this theory, we added

support for another classifier, C4.5[8] which is an entropy-based decision tree and decision rule generator, to GCSpeciesSorter. C4.5 is a multi-objective classifier. It examines the data and creates a decision tree, where each decision node in the tree reduces the entropy (randomness) of the data. This decision tree is then pruned and converted to a set of decision rules, which can be applied to new and unseen data. In terms of configurability and supported command line options, GCSpeciesSorter's support for C4.5 is less extensive than that of SVM classification.

The general principles and application of the GCSpeciesSorter presented here are similar to that of TopSort in that both use differences in species GC content in order to train a classifier and sort sequences into species' origins. GCSpeciesSorter's use of C4.5 emphasizes the inherent differences between two species' GC contents, since C4.5 and SVM operate using very different methods, but as seen below, the results are similar.

Implementation

GCSpeciesSorter is a binary classification package for distinguishing between two or more species based on the GC contents of their DNA or RNA sequences. It includes source code in Python and is released under the GNU General Public License (GPL). Beyond unpacking, there is no special installation step necessary. Python, LIBSVM[9] and/or C4.5, and optionally, BLAST are needed to run the scripts. A README file in the package provides more details about running the scripts. The package includes all the input files mentioned in this paper to use as a tutorial, including test sequence files and BLAST database files.

For both SVM and C4.5, there are two phases to using the software, implemented as two Python tools. The first phase accepts a 'Target' species and an 'Other' species as two nucleotide sequence files. They are used to train an SVM classifier or a C4.5 decision tree to distinguish between the Target and Other species. The second phase accepts a file containing unknown sequences, which are fed to the classifier generated in the first phase to determine the species. For SVM, the output of phase 2 consists of Target and Other sequence files. With C4.5, the output is a set of decision rules that can be used to distinguish between samples.

We computed the GC content as a normalized value relative to the size of the sequence. To do so, the number of G and C bases were added together and then divided by the total length of the sequence. Shorter sequences may not have a GC content representative of the hole genome, so GCSpeciesSorter allows the user to specify a minimum sequence length. Any sequence shorter than this minimum is ignored in GC computations. Unknown bases, often represented by the letter N, are also left out of the computations to make sure the values are not biased toward lower GC content.

GCSpeciesSorter supports FASTA or FASTQ inputs, and optionally uses BLAST as an additional step to remove sequences that have variant GC content, such as those found in organelles like mitochondria and chloroplasts. In our experiments, we found FASTA files with longer sequences to be a more reliable source of GC statistics that short-read sequences in a FASTQ file. In this article, we apply GCSpeciesSorter to transcript sequences containing mixed *Acropora millepora* coral and *Symbiodinium* samples.

Results

All the reported results in this article were obtained without BLAST filtering. Known coral and *Symbiodinium* nucleotide sequence sets were assembled based on our target coral species, *Acropora millepora*,[10] and predicted *Symbiodinium* clades (clade A from Bayer et al,[11] clades B and C from Ladner et al[12]). The coral training set contained 100 entries, and the *Symbiodinium* set contained 299 entries. We used 90 known coral and 270 known *Symbiodinium* samples to train the SVM and the decision tree, with the remaining sequences were used to evaluate the accuracy of the trained SVM and tree. These numbers are chosen such that the classifiers are trained with roughly the same proportion of Target and Other species as are found in biological samples. Both the resulting SVM and the decision tree had 100% accuracy on the 39 samples left out of training.

Figure 1 shows the GC contents for all 360 training samples used above. *Symbiodinium* GC contents are plotted as class −1 on the left, while coral GC values are plotted as class + 1 on the right. There is some overlap between the values, but the majority of the samples are clearly skewed in opposing directions.

To test the program on unverified data, we obtained 26,275 presumed *A. millepora* coral nucleotide sequences,[13] and 47,014 presumed *Symbiodinium* sequences.[14] Both the coral and the *Symbiodinium* files were fed to the SVM trained with verified sequences as described above. With coral samples, 95.40% were classified as coral. With Symbiodinium samples, 97.35% were classified as *Symbiodinium*. To test the symmetry of the classifier, we used the *Symbiodinium* data as the Target, trained a new SVM using the verified samples, and reran the test. The same results were observed. On a virtual machine running on an Intel Xeon E5-2650 CPU at 2 GHz and 16 GB RAM, training the SVM and testing the above unverified samples took less than 15 seconds per dataset.

With the C4.5 tree created from verified data, the same tests resulted in 95.5% correct classification of corals, and 93.0% correct classification of *Symbiodinium* samples. We did a similar test of switching Target and Other categories, and the decision tree's results were also symmetric. Training and testing on the same virtual machine took less than 10 seconds.

The results above show that both the SVM and C4.5 can learn the GC differences using a relatively small number of verified samples. In the rest of the article, we show how the application performs with bigger and unverified data sets.

In the above experiments, we had access to positively identified coral and *Symbiodinium* samples to train the classifiers. It is, however, possible to use unverified samples to train a classifier. To illustrate this case, we used 10,000 presumed coral and

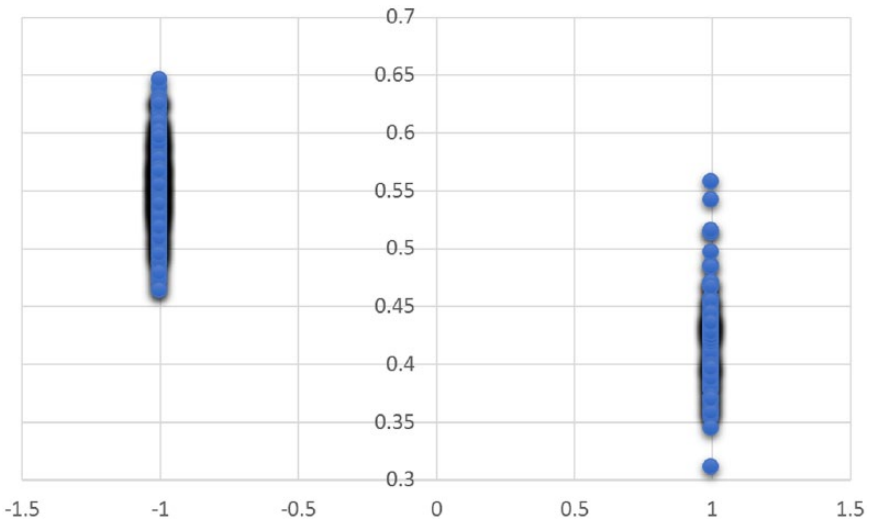

Figure 1. GC contents for *Symbiodinium* (left) and coral (right) samples.

Table 1. SVM accuracy test results.

INPUT	SVM FROM VERIFIED SAMPLES	SVM FROM UNVERIFIED SAMPLES
Unverified Coral	95.40%	93.78%
Unverified *Symbiodinium*	97.35%	98.53%

Table 2. C4.5 accuracy test results.

INPUT	RULES FROM VERIFIED SAMPLES	RULES FROM UNVERIFIED SAMPLES
Unverified Coral	97.3%	94.0%
Unverified *Symbiodinium*	93.0%	98.4%

18,000 presumed *Symbiodinium* samples to train an SVM, and tested it on the remaining presumed coral and *Symbiodinium* sequences. The accuracy for the total 45,289 samples left out of training was 97.19%, showing that this method can be quite effective even without verified samples.

On the same virtual machine, training the SVM using the unverified samples took about 3 minutes. Reading and processing the input files took about 2 minutes and 20 seconds, around 35 seconds were spent training the SVM model, and the rest was used in testing the model.

We then tried the complete unverified data sets against the SVM trained using the verified sequences and the SVM trained using the unverified data. The results appear in Table 1. As can be seen, sorting coral samples is more error-prone, likely due to more variability in transcript GC content.

We created a C4.5 decision tree using unverified samples in a manner similar to the SVM training above. We then tried the unverified samples on the decision rules derived from the tree, and also on the rules generated using verified samples. The results appear in Table 2.

C4.5's execution times were in general lower than LIBSVM, although the accuracy results are comparable.

In another round of experiments, we used the application to create classifiers for separating two related frog species, the *Xenopus laevis* and *Xenopus tropicalis*, using data available from Xenbase.[15] As expected, the GC contents in these frogs are very similar, and the classifiers could not distinguish the two species reliably.

Conclusion

We showed that the method presented here is useful in analyses of symbiosis. Our results indicate that species with differentiated GC contents can be accurately and quickly classified using GCSpeciesSorter with an SVM or a decision tree. This method also may be of value to multi-species sorting in metagenomic studies. While the working example presented here is binary in nature, it can be applied to sorting more than two species. In such cases, the user simply needs to set the Target to a different species each time and repeat the classification process. Both LIBSVM and C4.5 support multiple classes, so direct classification of multiple species is in principal possible, and an interesting topic for future work. GCSpeciesSorter is available for download from ftp://xenbaseturbofrog.org/GCSpeciesSorter for free.

Author Contributions

KK designed and developed the code and performance tests, DMW and MJO prepared and processed the data, PDV developed the ideas and methodologies. All authors contributed to the manuscript.

Ethical Approval

No human subjects or data were involved. No animals were involved.

REFERENCES

1. Hebert PDN, Cywinska A, Ball SL, DeWaard JR. Biological identifications through DNA barcodes. *Proc Biol Sci*. 2003;270:313–321.
2. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–410.
3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–359.
4. Shomron N, ed. *Deep Sequencing Data Analysis*. New York, NY: Springer; 2013.
5. Lehnert EM, Mouchka ME, Burriesci MS, Gallo ND, Schwarz JA, Pringle JR. Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. *G3 (Bethesda)*. 2014;4:277–295.
6. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C. Text classification using string kernels. *J Mach Learn Res*. 2002;2:419–444.
7. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*. 2008;9:S13. http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-9-S1-S13.
8. Quinlan JR. *C4.5: Programs for Machine Learning*. San Francisco, CA: Morgan Kaufmann Publishers Inc.; 1993.
9. Chang C, Lin C. LIBSVM : a library for support vector machines. *ACM Trans Intell Syst Technol*. 2011;2:1–27.
10. Moya A, Huisman L, Ball EE, et al. Whole transcriptome analysis of the coral Acropora millepora reveals complex responses to CO2-driven acidification during the initiation of calcification. *Mol Ecol*. 2012;21:2440–2454.
11. Bayer T, Aranda M, Sunagawa S, et al. Symbiodinium transcriptomes: genome insights into the dinoflagellate symbionts of reef-building corals. *PLoS ONE*. 2012;7:e35269.
12. Ladner JT, Barshis DJ, Palumbi SR. Protein evolution in two co-occurring types of Symbiodinium: an exploration into the genetic basis of thermal tolerance in Symbiodinium clade D. *BMC Evol Biol*. 2012;12:217.
13. http://marinegenomics.oist.jp/coral/viewer/download?project_id=3.
14. http://marinegenomics.oist.jp/symb/viewer/download?project_id=21.
15. Karimi K, Fortriede JD, Lotay VS, et al. Xenbase: a genomic, epigenomic and transcriptomic model organism database. *Nucleic Acids Res*. 2018;46:D861–D868.