BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction Model

Volume 7(6)

EGID: an ensemble algorithm for improved genomic island detection in genomic sequences

Dongsheng Che[1]*, Mohammad Shabbir Hasan[1], Han Wang[1], John Fazekas[1], Jinling Huang[2], Qi Liu[3]

[1]Department of Computer Science, East Stroudsburg University, East Stroudsburg, PA 18301; [2]Department of Biology, East Carolina University, Greenville, NC 27858; [3]College of Life Science and Biotechnology, Tongji University, Shanghai, 200092, China; Dongsheng Che - Email: dche@po-box.esu.edu; Phone: 1-570-422-2731; *Corresponding author

Received November 16, 2011; Accepted November 17, 2011; Published November 20, 2011

Abstract:
Genomic islands (GIs) are genomic regions that are originally transferred from other organisms. The detection of genomic islands in genomes can lead to many applications in industrial, medical and environmental contexts. Existing computational tools for GI detection suffer either low recall or low precision, thus leaving the room for improvement. In this paper, we report the development of our Ensemble algorithm for Genomic Island Detection (EGID). EGID utilizes the prediction results of existing computational tools, filters and generates consensus prediction results. Performance comparisons between our ensemble algorithm and existing programs have shown that our ensemble algorithm is better than any other program. EGID was implemented in Java, and was compiled and executed on Linux operating systems. EGID is freely available at http://www5.esu.edu/cpsc/bioinfo/software/EGID.

Keywords: Bacterial genomes; Ensemble algorithm; Genomic islands.

Background:
Genomic islands are chromosomal regions that have the evidence of horizontal gene transfer. The studies of genomic islands are extremely important to biomedical research, due to the fact that such knowledge can be used to explain why some strains of bacteria within the same species are pathogenic while others are not, or the phenomena that some strains of bacteria can adapt to extreme environments while others cannot.

Current approaches of detecting genomic islands include comparative genomic analyses and sequence composition analyses. The comparative genome analysis consists of collecting the genome sequences of phylogenetically closely related species, aligning these genome sequences, and then considering those genome segments present in a query genome but not in others to be GIs **[1].** Since this type of approach does not apply to the genomes that do not have enough number of phylogenetically closely related genomes for reference, it cannot be applied to all genomes. The second kind of approach, sequence composition-based approach, does not require reference genomes and can be applied to any genome. It is generally believed that each genome has a unique genomic sequence signature, and thus genomic islands can be detected by analyzing sequence composition. Existing sequence composition based tools include AlienHunter **[2],** Centroid **[3],** COLOMBO SIGI-HMM**[4],** IslandPath **[5],** INDeGenIUS **[6],** and PAI-IDA **[7].** The assessment of these computational tools in recent studies has shown that none of these tools can predict genomic islands accurately in all genomes **[1]**. Langille **[8]** further suggested that a computational framework that combines multiple prediction results of existing programs should be developed for more accurate genomic island prediction.

In this paper, we present our ensemble program for improved genomic island prediction, based on predicted results of five

existing GI programs. The framework of our approach includes: **[1]** collecting prediction results from existing programs; (b) analyzing and filtering on predicted results; and c) generating final consensus GI results **(Figure 1).** Experimental tests on benchmark datasets have shown that our ensemble program could improve prediction accuracy, and thus it may be used for the future GI prediction.

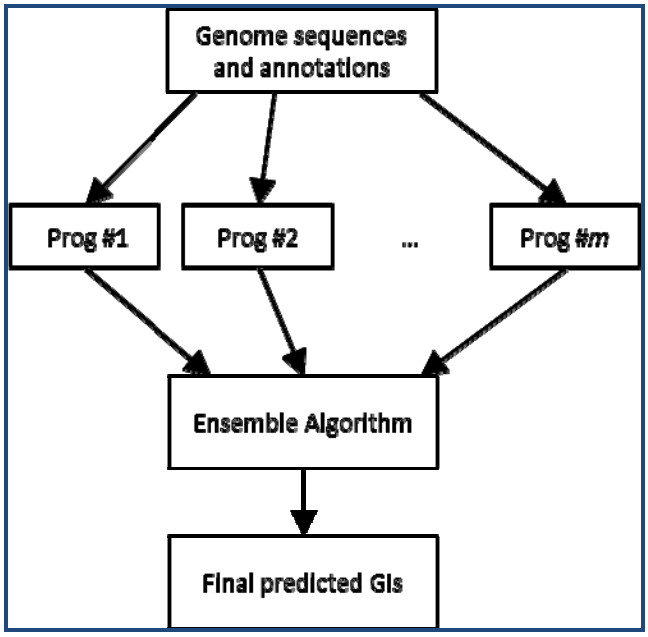

Figure 1: The flowchart of our computational framework for GI prediction.

Methodology:

Data sets

Genomic sequences used for GI prediction were collected from the National Center for Biotechnology Information (NCBI) FTP server (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). The genomic islands used for performance evaluation of GI tools were obtained by IslandPick **[1].**

Prediction of GIs with existing tools

In our framework, we used the predicted GI results from five GI tools, AlienHunter, COLOMBO SIGI-HMM, INDeGenIUS, IslandPath, and PAI-IDA. All five programs use genome sequences as program inputs, with some individual programs requiring additional inputs such as gene annotations. The prediction results from these programs were used in our ensemble method.

Ensemble method

Since GIs could range in size from several kilo base pairs (kb) to several hundred kb, it is very unlikely that two different GI prediction tools predict exactly same genomic islands. Thus, the predicted GIs by different tools often overlap, making it difficult to vote predicted results simply based on their predicted GIs. To handle this problem, we considered the genes within the predicted GI regions to be *GI genes*, and *non-GI genes* otherwise. We collected GI and non-GI gene information based on the prediction results by multiple GI tools.

A simple voting scheme could be applied by selecting a threshold value, and considering the region, where all contained genes meeting the threshold requirement, to be a GI region, as shown in **Figure 2.** This approach may work fine for those candidate GI regions that are far away **(Figure 2B),** but not for those which are close each other **(Figure 2A),** which are supposed to be a big GI region **[8].**

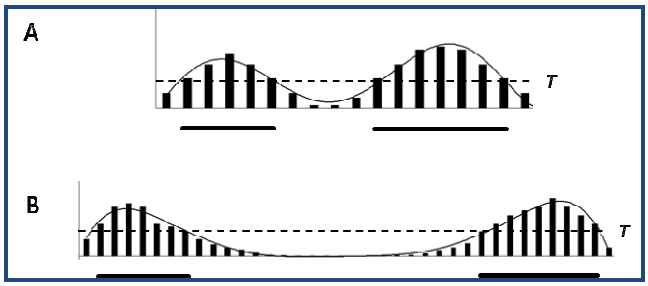

Figure 2: Illustrative examples of candidate GI regions, where two candidate GI regions are close in (A) and distant in (B). Each vertical bar represents the vote of a GI gene by multiple GI tools. The candidate GI regions meeting the threshold value are underlined.

To resolve this problem, we proposed a measure for GI or non-GI based on the overall score of all genes in the region, rather than individual gene scores. To do so, we first form candidate GI regions, G_1, G_2, G_3, ..., G_m, where two neighboring GI regions, G_i and G_{i+1}, are separated by a non-GI region (*i.e.*, none of programs predicted the region to be a GI). We then merge any two neighboring GI regions, G_i and G_{i+1}, if the average score of all genes (including the genes in G_i and G_{i+1}, and between the two regions) meets a predefined threshold value T_1. By applying this measurement, we should merge two close GI regions (as shown in **Figure 2A**), but not for distant GI regions **(Figure 2B).** If two GI regions are merged into a newly formed GI region $G_{i,\ i+1}$, then $G_{i,\ i+1}$ and G_{i+2} will be picked for the next merging test. Otherwise, G_{i+1} and G_{i+2} will be selected for merging test. The merging process will be repeated until it reaches to the last GI region, and we can obtain a set of GI regions, G'_1, G'_2, G'_3, ..., and G'_n.

We further filter out GIs from the previous step if (a) the GI is short (*i.e.*, containing < eight genes in the GI); and (b) the percentage of high GI gene scores (*i.e.*, >1) does not meet a threshold value T_2, so that we can guarantee that predicted GIs are supported by multiple programs. The determination of threshold values, T_1, and T_2 was described in Supplementary Material.

Performance evaluation

To evaluate the performance of our model, we compared the predicted GIs with the benchmark dataset **[1]**. The benchmark dataset contains picked GIs from 118 genomes, and we predicted GIs using our EGID algorithm on these 118 genomes. True positives (TP) are the nucleotides in the positive benchmark dataset predicted to be genomic islands. True negatives (TN) are the nucleotides in the negative benchmark dataset predicted to be non-genomic islands. False positives (FP) are the nucleotides in the negative benchmark dataset predicted to be genomic islands. False negatives (FN) are the nucleotides within the positive benchmark dataset not

predicted to be genomic islands. We focus on four validation measures, recall = TP/(TP+FN), precision = TP/(TP+FP), performance coefficient (PC) = TP/(TP+FP+FN) and F-Measure = 2*recall*precision/(recall + precision).

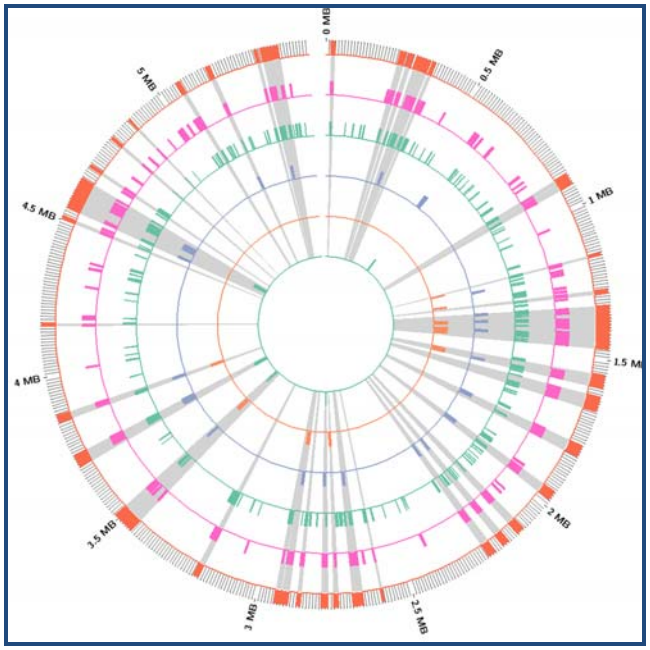

Figure 3: Circular representations of the *Escherichia coli* O157:H7 str. Sakai (NC_002695) showing predicted GIs, with each circle predicted by each program. The predicted GIs from the outer to the inner circle are EGID, AlienHunter, COLOMBO SIGI-HMM, INDeGenIUS, Island-Path, and PAI-IDA. The shaded parts show the predicted GIs by EGID, and evidenced GIs by other programs.

Discussion:

We collected 118 prokaryotic genomes from the National Center for Biotechnology Information (NCBI) FTP server, ran our EGID program, and generated GI locations (http://www5.esu.edu/cpsc/bioinfo/software/EGID) for each genome. We used genomic islands obtained by IslandPick **[1]** as benchmark, to evaluate the predicted GIs by EGID. We also collected predicted GI results of five component programs in EGID, and summarized all performance results in (**Table 1, see supplementary).** As we can see from **Table 1 (see supplementary)**, both COLOMBO SIGI-HMM and IslandPath have relative high precision rate, but with low recall rate. On the other hand, AlienHunter has relative high recall rate, but with low precision rate. EGID makes the balance between recall and precision, and it reaches relative high recall (0.630) and precision rate (0.630). Since PC and F-measure capture both recall and precision in a single accuracy measurement, their values reflect overall performance more accurately. EGID improves 12.14% over the best existing program AlienHunter in PC, and 7.88% in F-measure, suggesting the performance improvement of our ensemble method.

In order to view the predicted GIs, we displayed the GI locations through one of the popular visualization tools, circus **[9].** As we can see from **Figure 3,** EGID always picks GIs predicted by multiple programs, thus guaranteeing the reliability of GIs selected. The circular representations of other 117 genomes can also be found in our website.

Conclusion:

In this paper, we have reported the development of an ensemble algorithm EGID for more accurate GI detection. We hope our improved GI prediction program could aid in molecular evolution and horizontal gene transfer studies.

Acknowledgement:

This research was partially supported by President Research Fund, and FDR major grant at East Stroudsburg University, USA.

References:

[1] Langille MG. *BMC Bioinformatics*. **9** 329 [PMID: 18680607]
[2] Vernikos GS & Parkhill J. *Bioinformatics*. 2006 **22**: 2196 [PMID: 16837528]
[3] Rajan I *et al. Bioinformatics*. 2007 **23**: 2672 [PMID: 17724060]
[4] Waack S *et al. BMC Bioinformatics*, 2006 **7**: 142 [PMID: 16542435]
[5] Hsiao W et al. *Bioinformatics*. 2003 **19**: 418 [PMID: 12584130]
[6] Shrivastava S *et al. J Biosci*. 2010 **35**: 351 [PMID: 20826944]
[7] Tu Q & Ding D. *FEMS Microbiol Lett*. 2003 **221**: 269 [PMID: 12725938]
[8] Langille MG *et al. Nat Rev Microbiol*. 2010 **8:** 373 [PMID: 20395967]
[9] Krzywinski M *et al. Genome Res*. 2009 **19**: 1639 [PMID: 19541911]

Edited by P Kangueane
Citation: Che *et al.* Bioinformation 7(6): 311-314 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Performance summary of GI prediction tools.

Tools	Recall	Precision	nPc	F-Measure
AlienHunter	0.650	0.530	0.412	0.584
COLOMBO SIGI-HMM	0.270	0.870	0.256	0.412
INDeGenIUS	0.340	0.610	0.276	0.437
IslandPath	0.230	0.880	0.224	0.365
PAI-IDA	0.190	0.720	0.175	0.301
EGID	0.630	0.630	0.461	0.630

Parameter determination:

The genomic islands predicted by our EGID algorithm are determined by two threshold values, T_1 and T_2. To determine the optimal threshold values that can be used for GI prediction, we ran EGID on 118 genomes. We compared our predicted results with the benchmark dataset, and measured performance metrics of our predicted GIs. As shown in Figure 1, the prediction results are relatively stable when T_1 is between 1.2~2.2. T_2 seems to have little effect on performance when T_1 becomes large (e.g., 1.5), indicating that T_1 and T_2 are highly related. In general, a GI candidate region that has high average label values should have a high percentage of multiple predictions in that region.

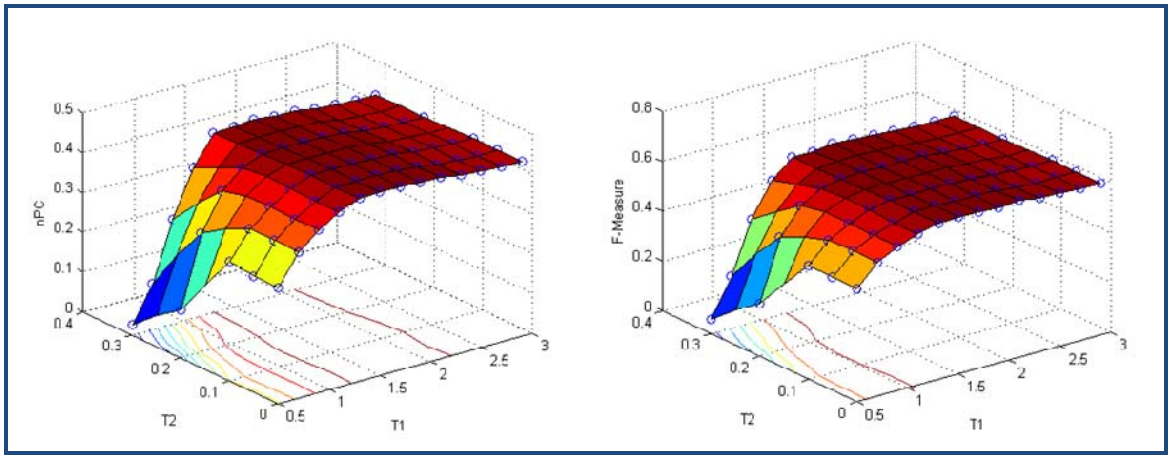

Figure 1: Parameter (T1, T2) determination used in EGID. The left figure shows the PC values, while the right figure shows the F-measure values.