Gene 518 (2013) 114–123

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

GI-POP: A combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects

Chi-Ching Lee [a,b], Yi-Ping Phoebe Chen [c], Tzu-Jung Yao [b], Cheng-Yu Ma [b], Wei-Cheng Lo [d], Ping-Chiang Lyu [a,e,g,*], Chuan Yi Tang [b,f,g,**]

[a] Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan
[b] Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
[c] Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Australia
[d] Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan
[e] Department of Medical Science, National Tsing Hua University, Hsinchu, Taiwan
[f] Department of Computer Science and Information Engineering, Providence University, Taichung, Taiwan
[g] Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung, Taiwan

ARTICLE INFO

Article history:
Accepted 27 November 2012
Available online 12 January 2013

Keywords:
Genome annotation
Genomic island
Web server
Ongoing genome project
Microorganism

ABSTRACT

Sequencing of microbial genomes is important because of microbial-carrying antibiotic and pathogenetic activities. However, even with the help of new assembling software, finishing a whole genome is a time-consuming task. In most bacteria, pathogenetic or antibiotic genes are carried in genomic islands. Therefore, a quick genomic island (GI) prediction method is useful for ongoing sequencing genomes. In this work, we built a Web server called GI-POP (http://gipop.life.nthu.edu.tw) which integrates a sequence assembling tool, a functional annotation pipeline, and a high-performance GI predicting module, in a support vector machine (SVM)-based method called genomic island genomic profile scanning (GI-GPS). The draft genomes of the ongoing genome projects in contigs or scaffolds can be submitted to our Web server, and it provides the functional annotation and highly probable GI-predicting results. GI-POP is a comprehensive annotation Web server designed for ongoing genome project analysis. Researchers can perform annotation and obtain pre-analytic information include possible GIs, coding/non-coding sequences and functional analysis from their draft genomes. This pre-analytic system can provide useful information for finishing a genome sequencing project.

1. Introduction

In the research of pathogenesis and drug resistance, it has been found that genes or associated elements were clustered in chromosomal regions (Hacker et al., 1990). These DNA segments often have the ability to jump and incorporate into other bacterial genomes by an event termed horizontal gene transferring (HGT), which commonly occurs among microorganisms (Binnewies et al., 2006; Frost et al., 2005; Hacker and Kaper, 2000; Koonin et al., 2001; Mantri and Williams, 2004; Ou et al., 2006). The incorporated foreign DNA segments often have tRNA genes or repeated sequences at their boundaries (Hsiao et al., 2003; Lobry, 1996a; Yoon et al., 2007). Collectively referred to as genomic islands (GIs), these foreign DNA segments typically possess medical and environmental adaptability, and range from 5 to 500 kb in length (Schmidt and Hensel, 2004). For instance, secretion islands encode supplementary secretion systems, metabolic islands carry genes for secondary metabolism, and resistance islands bring antibiotic resistance to the host bacteria (Hacker and Kaper, 2000). The sequence composition of GIs (*e.g.*, guanine and cytosine contents) (*%G + C*), dinucleotide bias, and codon usage preferences differ from that of the host genome (Frost et al., 2005; Hacker and Kaper, 2000; Koonin et al., 2001). Of the several computational approaches developed for detecting GIs, some require pre-annotated information. For example, homologous genes appearing in detected GIs can act as the monitor of related GIs (Mantri and Williams, 2004; Ou et al., 2006; Yoon et al.,

Abbreviations: GI, genomic island; GI-POP, genomic island prediction by Genome Profile Scanning; GI-GPS, genomic island genomic profile scanning; HGT, horizontal gene transferring; DNA, deoxyribonucleic acid; %G + C, guanine and cytosine contents; PCR, polymerase chain reaction; tRNA, transfer RNA; COG, clusters of Orthology Groups; DIYA, do-it-yourself annotator; CDS, coding sequences; CDD, conserved domain databases; SWS, sliding window scanning; SVM, support vector machine; MGE, mobile genetic elements; CAI, codon adaption indices; *E. coli*, *Escherichia coli*; ROC, receiver-operating characteristic; AUC, area under curve; PE, probabilistic estimate; *S. enterica*, *Salmonella enterica*; NCBI, National Center of Biotechnology Information; PHP, PHP: hypertext preprocessor; RBF, radial basis function; ACC, accuracy; USD, unique signature-discovering; PAI, pathogenicity island.

* Correspondence to: P.-C. Lyu, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C. Tel.: +886 3 5742762; fax: +886 3 5715934.
** Correspondence to: C.Y. Tang, No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C. Tel.: +886 3 5715131 1077; fax: +886 3 572 3694.
E-mail addresses: d9662841@oz.nthu.edu.tw (C.-C. Lee), Phoebe.Chen@latrobe.edu.au (Y.-P.P. Chen), s9962598@m99.nthu.edu.tw (T.-J. Yao), s9962826@m99.nthu.edu.tw (C.-Y. Ma), WadeLo@nctu.edu.tw (W.-C. Lo), pclyu@mx.nthu.edu.tw (P.-C. Lyu), cytang@cs.nthu.edu.tw (C.Y. Tang).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.11.063

2007). Novel GIs are difficult to identify in this manner because of the lack of available genetic annotations or sequence homologies to known GIs. Most other methodologies use sequence compositional differences (*e.g.*, *%G+C*), dinucleotide frequencies, and amino acid and codon usage preferences between foreign and native DNA to identify GIs by assuming that different organisms vary in compositional chromosomal patterns (Hsiao et al., 2003; Lobry, 1996b; Mantri and Williams, 2004; Merkl, 2004; Nag et al., 2006; Rajan et al., 2007; Tu and Ding, 2003; van Passel et al., 2005). However, these methodologies are limited because even in the same organism, different chromosomal regions might vary in sequence compositions and gene expression levels (Schmidt and Hensel, 2004). Closely related organisms are assumed to share higher sequence similarities and expressional properties (Schmidt and Hensel, 2004). By adopting comparative genomic approaches, other strategies are independent of sequence compositional analysis. Comparing multiple chromosomes of closely related organisms leads to the assumption that unexpected phyletic sequences are horizontal transfer regions (Ragan, 2001). Comparison based method can identify GIs that sequence composition-based methods cannot detect because the sequence compositions of these GIs resemble the core chromosomes. GI-predicting methods can be applied to allocate possible GI regions in given microbial genomes. Genomes of pathogenesis, antibiotic resistance, or other researchable phenotypes can be sequenced by the high-throughput sequencing technique. Over the past decade, the high-throughput sequencing technique (or next generation sequencing) has made significant progress toward reducing the sequencing cost and handling time. The genome of the target organism is first separated into billions of short DNA elements called short reads. These short reads are then sequenced by sequencers and assembled into longer sequences called contigs. Finally, genes and proteins are annotated using annotating software, such as Glimmer (Salzberg et al., 1998).

However, finishing an entire genome using only computational methods continues to be difficult. Experimental procedures, including optical mapping and polymerase chain reaction (PCR), were performed to guarantee sequencing quality. These increased the cost and time. Furthermore, the present GI-predicting methods need whole genome sequences and require completed gene or protein annotations. Finding GIs in microorganisms can only be achieved by obtaining completed genomes. This observation suggests the need for a GI prediction and analysis method that can be used in ongoing genome projects.

In this paper, we developed an annotating pipeline for ongoing genome projects. This pipeline integrates functional annotating and GI-predicting capabilities and can be used for analyzing incomplete genomes. It provides an annotating service where researchers can submit their own draft genomes. The pipeline assembles and annotates the submitted sequences, including contigs/scaffold assembly, gene prediction, tRNA or other non-coding RNA prediction, Clusters of Orthology Groups (COG) searching, and GI prediction. We believe that sequence compositional approaches continue to provide a strong foundation for developing GI detection methods. With an adequate length, any fragment of a GI should have sequence compositions resembling the remaining portion of the GI and differ from the core chromosome. Based on this assumption, we developed GI prediction by Genome Profile Scanning (GI-GPS), for example, a GI detection system that operates by scanning, filtering, and refining. Performing cross-validation on a published data set demonstrated the feasibility of using the genome profile and the prediction engine to distinguish between GI and non-GI sequences. Moreover, the GI-GPS requires only one organism's genome, which is advantageous to the identification of foreign DNA for newly sequenced organisms, especially the novel ones with few known related species. Moreover, GI-POP is the first combinational annotation and GI detecting Web server which provides pre-analytic information of ongoing genome project.

2. Results

2.1. GI-POP: the annotation platform with GI detecting modules

GI-POP is a Web server designed for online functional annotations, gene predictions, non-coding RNA predictions, and GI predictions of ongoing genome projects. As shown in Fig. 1, genomic sequences, including contigs, scaffolds, and chromosomal sequences, are first assembled by a do-it-yourself annotator (DIYA) assembler which is an annotating package. The coding sequences (CDS), non-coding region predictions, and analysis are then operated. Several subroutines participate in this stage; for example, Glimmer (Salzberg et al., 1998) is used to predict coding sequences from the chromosomal sequences. When the CDSs have been identified, sequence investigation is conducted to find the homologous genes (clusters of sequences from the UniProt Knowledge base, UniRef (Mulder et al., 2008)), conserved domains (conserved domain databases, CDD (Marchler-Bauer et al., 2011)), COGs (Tatusov et al., 2000, 2001), predicting genes of tRNA, and other non-coding RNAs (Lagesen et al., 2007). The GI prediction module GI-GPS is used to obtain the GI candidate sequences. Finally, the GBrowse, which is a widely used genome browser and annotation visualization toolkit (Donlin, 2007), is used to provide user-friendly experiences.

2.2. GI-GPS, the combinational GI-predicting method

We developed a hybrid method to detect GIs in an ongoing genome project. Most of the chromosomal sequences are a product assembled by the DIYA assembler from pieces of contigs or scaffolds. It is possible that several DNA regions can be missed. Therefore, detecting GIs by *S*liding *W*indow *S*canning (SWS) through the assembled chromosome is a feasible approach to tolerant error-assembled regions. The compositional approach is a logical method because the entire genome composition is not altered easily by truncating or misassembling some chromosomal regions. Based on these assumptions, we developed a three-stage GI-detecting pipeline, which includes (1) window scanning and SVM classifying, (2) refining by genome composition and the mobile genetic elements (MGE), and (3) boundary identification. The flowchart of all three stages is shown in Fig. 2. A diverse set of attributes, collectively referred to as a genome profile, has been defined to describe compositional differences thoroughly. A support vector machine (SVM) classifier, which functions as the scanning engine, has been trained by the genome profiles of a set of known GIs/non-GIs. In the first stage of GI-GPS, the entire chromosome is divided into numerous fragments of fixed lengths. The genome profile of each fragment is then extracted and subjected to SVM classification. This process can be regarded as if a sliding window were used to scan through the entire chromosome, subsequently extracting compositional propensities piecewise. As the sliding window progresses, compositional differences between the DNA fragment within the window and the entire genome are used by the SVM classifier to make predictions. Next, neighboring fragments determined to be GIs are merged with full-length GI candidates. In the second stage, the amounts of false positive identifications are reduced by re-examining and predicting the full-length GI candidates based on SVM classifications and homologous searching of mobile genetic elements. Finally, the boundaries of predicted GIs are refined by allocating the positions of probable tRNA genes and repeating elements, which are frequently found at the GI boundaries. The criteria of the SVM classifier used for refining stage are based on length of the segments and are identified by trial-and-error method.

2.3. Composition of genome profile

This study has defined a set of compositional indices, which consist of four classes: (1) codon usage preferences that represent the

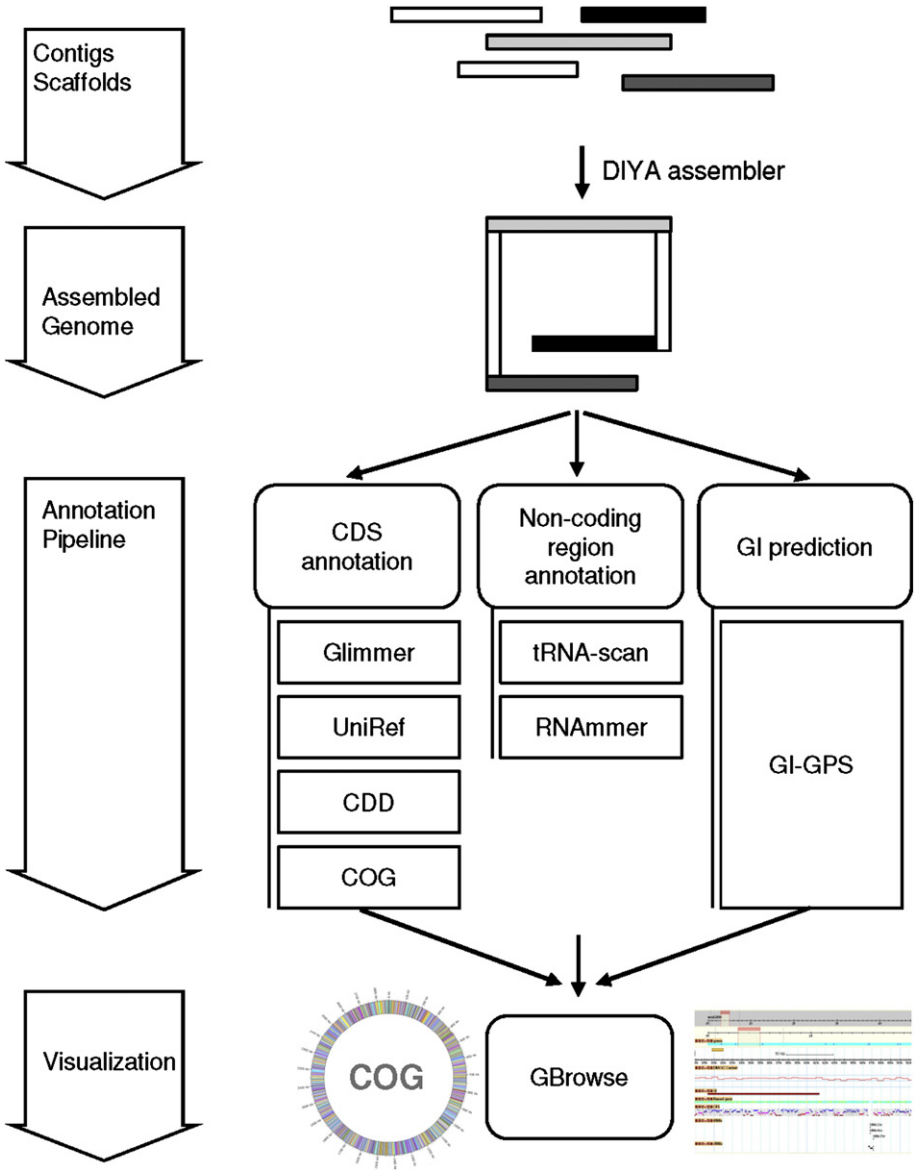

Fig. 1. The flowchart of GI-POP. The submitted draft sequences are assembled into assembled genome, and annotation can be selected in the annotation pipeline including CDS annotation, Non-coding region annotation, and GI prediction. After the annotation, the visualization modules including COG and genome browser are used for information representation.

occurrence of synonymous codons, (2) oligonucleotide bias indices describing how the foreign and native DNA differ in sequence, (3) codon adaption indices (CAI), which are theoretical estimations of gene expression levels [18], and (4) GC contents of the individual positions of codons. This study encodes DNA regions by genome profiles. Several GIs were selected from *E. coli* strain 536 as testing cases. Figs. 3(a–d) indicate that the genome profiles of GIs and whole genomes have different patterns. Some individual and classes of indices may offer little information for distinguishing GIs and non-GIs; in this case, some indices still contribute to the discrimination. For instance, the dinucleotide frequencies of the GI shown in Figs. 3(a–d) are not different from those of the host genome, but the positional $\%G+C$ and the preference of codon usage are different between the two. The testing results confirmed our assumptions that the genome profile can be used to represent DNA and highlight the foreign DNA.

Fig. 3(e) shows that all 93 features of the given microbial genomes were normalized by the amount of genes, and were integrated to generate a genome profile. We used an IslandPick data set (Langille et al., 2008) and selected sequences adapting the GI criteria from PAIDB (Yoon et al., 2007) to perform the feature selection. To build the SVM classifier, 85 features were selected. Comparing to all 93 features, the average accuracy was increased from 0.86 to 0.89 in our 32 times experiments using IslandPick data set. The *F*-score used in the feature selection is described in the Methods section. Four dinucleotide features (AG, CA, GA and GT), three codon adaption indices (CAI) (0–1, 1–2, and 3–4), and one codon usage (TAA) were removed. The SVM model was generated by 5-fold cross-validation using an IslandPick data set.

2.4. Basic assessment of the SVM classifier

Because the genome profile is composed of many indices with divergent tendencies in various GIs, this study attempted to integrate their effects efficiently by using an SVM classifier. The SVM classifier was evaluated by using the data set provided by the IslandPick Web site, in which 118 bacterial strains constituted 12 orders. Using

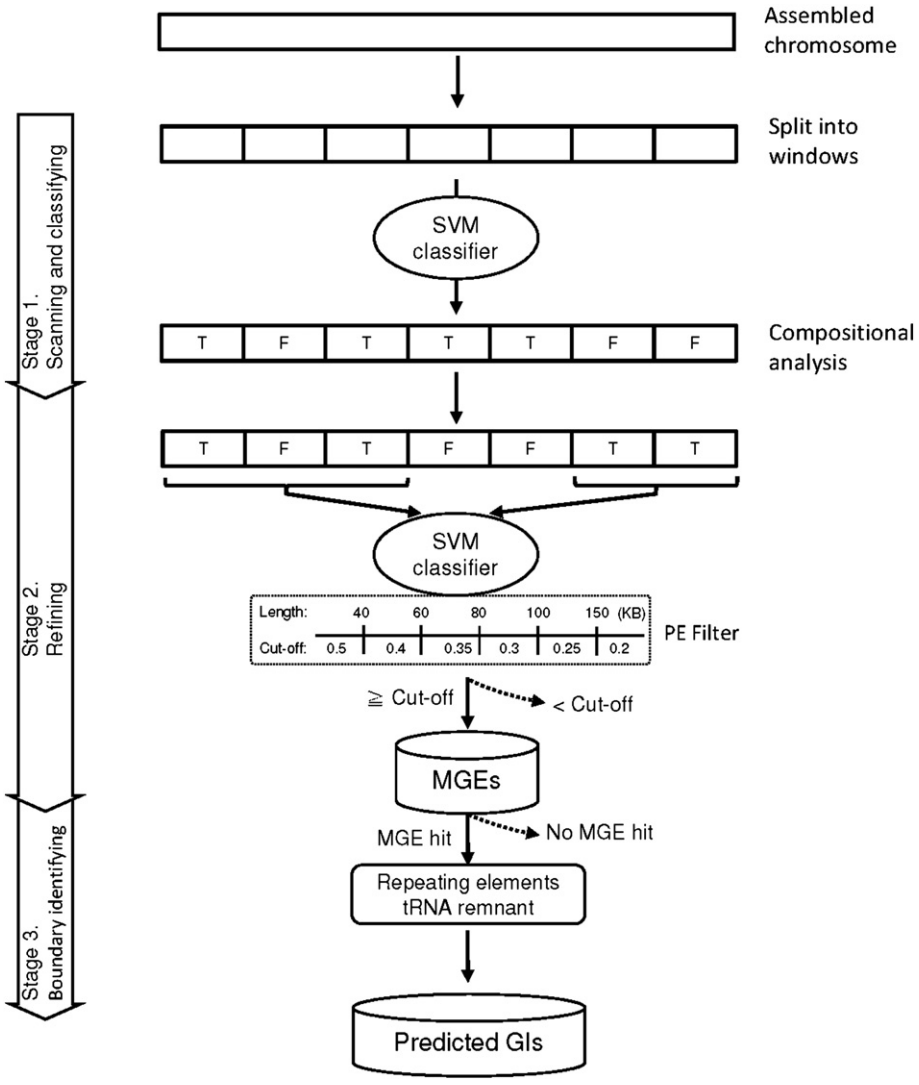

Fig. 2. The flowchart of the GI-GPS; a combinational GI-predicting method. Stage 1: Scanning and classification; the assembled chromosome sequences are divided into segments. These segments are then classified by a well-trained SVM classifier into two groups. The positive group (T) represents the high-sequence compositional variants containing a whole chromosome, and the negative group (F) shows their similarities to the chromosomes. In the refining stage, the neighboring segments, which are classified as positive, are merged into one large segment; in some cases, few negative predicted segments between a set of continuous positive predicted windows is acceptable. The merged segments are reclassified by the SVM, and the results are filtered by the PE score (probability estimation) filter which the cut-offs are based on the length of the segment. After filtering, the segments which higher than the cut-off values are filtered by the mobile genetic element (MGE) filter. Finally, hotspots of the GI boundaries, including tRNA and repeating elements, are used to assist boundary identification.

these organisms, receiver-operating characteristic (ROC) curve analysis with 5-fold cross-validations was performed (refer to the Methods section). The average AUC (area under curve) after the cross-validations was 0.93 (see Supplementary Fig. 1), demonstrating that our classifier can accurately distinguish between GIs and non-GIs. Because the genome profile is used to describe the compositional differences between GI sequences and the core chromosomes, these results indicate that sequence compositional propensities are still an appropriate basis for developing GI-detecting methods. In our GI-GPS system, the boundaries of a predicted GI are refined by allocating probable short tRNA genes and repeating elements. When several candidate termini are detected many combinations can be formed. For instance, 6 repeating elements at the 5′ end and 7 tRNA genes at the 3′ end result in *42 (6×7)* different versions of the same GI; they differ from one another at the terminal region. Fig. 4 shows all the possible structures of the island boundaries. To determine which version is the most natural form for a GI, each combination must be scored. This study takes advantage of the probabilistic estimate (PE) provided by LIBSVM (Chang and Lin, 2011), which describes, on the basis of the trained model, how a subject sequence is a piece of GI in percentages. Fig. 5(a), which shows a summary form of the PEs of 600 GIs and non-GIs randomly selected from the IslandPick data set (Langille et al., 2008), indicates that the distributions of the PEs of the GI and non-GI are different. In Fig. 5(b), we also evaluated this cut-off on published GIs which were retrieved from PAIDB. In the early GI detecting stage, the PE cut-off is set to 0.5 which can incorporate more possible island candidates than higher cut-off.

2.5. The performance of GI detection on two synthetic draft genomes

Because there are no other GI prediction tools designed for draft genomes, we used the synthetic draft genomes as the evaluating sequences. Two microbial genomes (*E. coli* and *S. enterica*) that have been well-studied and have published GIs were selected for this experiment. The synthetic genomes were generated by randomly

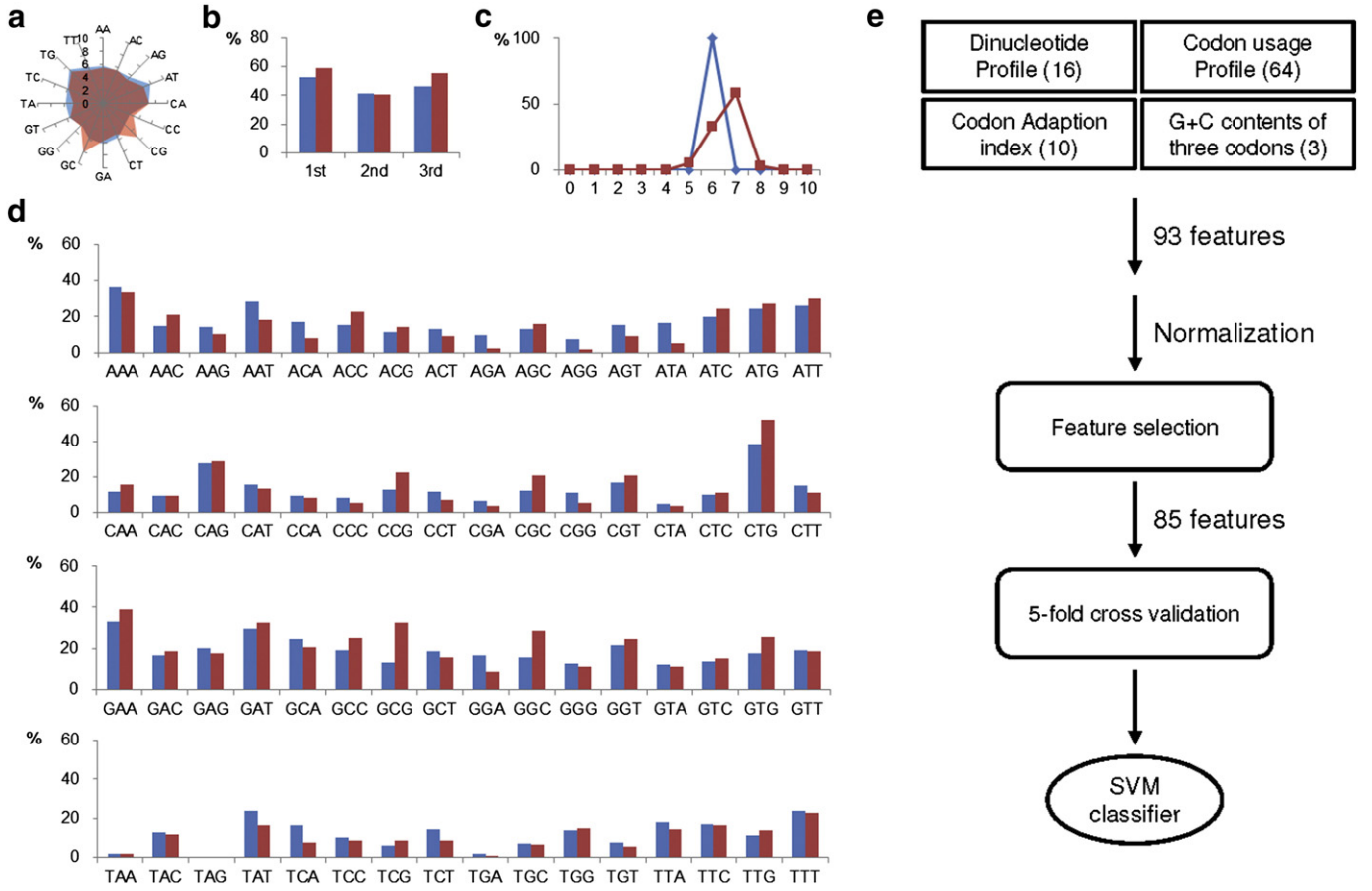

Fig. 3. An example of profile differences of GI and host genomes and the flowchart used to build the SVM classifier. The profile of the whole genome and GI of *E. coli* strain 536 includes (a) codon usage frequencies, (b) dinucleotide frequencies, (c) codon adaption indices, and (d) GC contents of individual positions of particular codons. The blue bar represents the genome profile of PAI I, and the red bar represents the genome profile of the whole genome. (e) The SVM-classifier-generated flow is integrated, normalized, and performs featured selections to remove the less discriminative features, and the generated flow is used for GI predictions.

removing a few segments from the genome sequences that mimic the draft genomes from the in-complete genome projects. For example, an 80% remaining genome represents that 20% of the total length of the genome was removed. As shown in Table 1, the accuracy of the GI prediction for the results for synthetic *S. enterica* draft genome was higher than 85%. For the 80%, 60%, and 10% remaining genomes, the accuracy of the prediction results was approximately 80%, as was the prediction result for *E. coli*. According to these synthetic prediction results, the proposed method implies the potential for GI detecting of ongoing genome projects.

2.6. User interface

GI-POP is an online Web server designed for easily annotating, analyzing, and visualizing. Fig. 6 shows the interface and functional units of the Web site. Users create accounts on the registration page. Users may submit multiple annotation jobs. Each job has uniquely identified sessions. The genomic sequences or contigs/scaffolds are uploaded on the sequence uploading page. The contigs/scaffolds are assembled by the DIYA assembler into one chromosomal sequence. The analytic results, including coding/non-coding regions and GIs, are visualized by a genome browser that is modified from the GBrowse. The COG results are visualized in a circular form, where each color represents a different type of COG hit. All the analytic and annotated sequences can be downloaded into a text format file. Multiple annotating and GI detecting jobs can be submitted simultaneously. The progress of each job can be monitored in the home page of each user.

3. Discussion

GIs or segments of foreign DNA cause morphological changes in its host microorganisms. Such changes are an impetus of microbial evolution. Elucidating how GIs and host organisms are related sheds light into the adaptations of microbes and their living environments. Studies have analyzed and compared sequence compositions to identify foreign DNA segments (Binnewies et al., 2006; Frost et al., 2005; Koonin et al., 2001). A training set is required to apply these compositional methods to detect GIs. However, neither standard GI data sets nor clear universal definitions are available for selecting the training GI sequences. Moreover, foreign DNA may still have similar sequence compositions with the core chromosomes, which could lower accuracy and increase the rate of false negatives. False positive cases may increase when segments with variant sequence compositional patterns appear in core chromosomes. The effectiveness of GI-detecting methods is evaluated using a well-defined standard benchmark. Langille et al. designed comparative genomic approaches based on whole genome sequence alignment (Yoon et al., 2007). Their basic premise is that closely related organisms should share conserved sequences. Unique segments among closely related organisms are, thus, possible genomic islands. By using this method, GIs with similar compositions to core chromosomes can be identified. However, this method is limited because clearly defining "close relatedness" of organisms is complex, and multiple closely related chromosomes are required for comparative analysis. The same GIs may appear in closely related organisms; for example, the genomic island SPI-2 can be found in *Salmonella enterica* subsp. enterica serovar Choleraesuis str. SC-B67

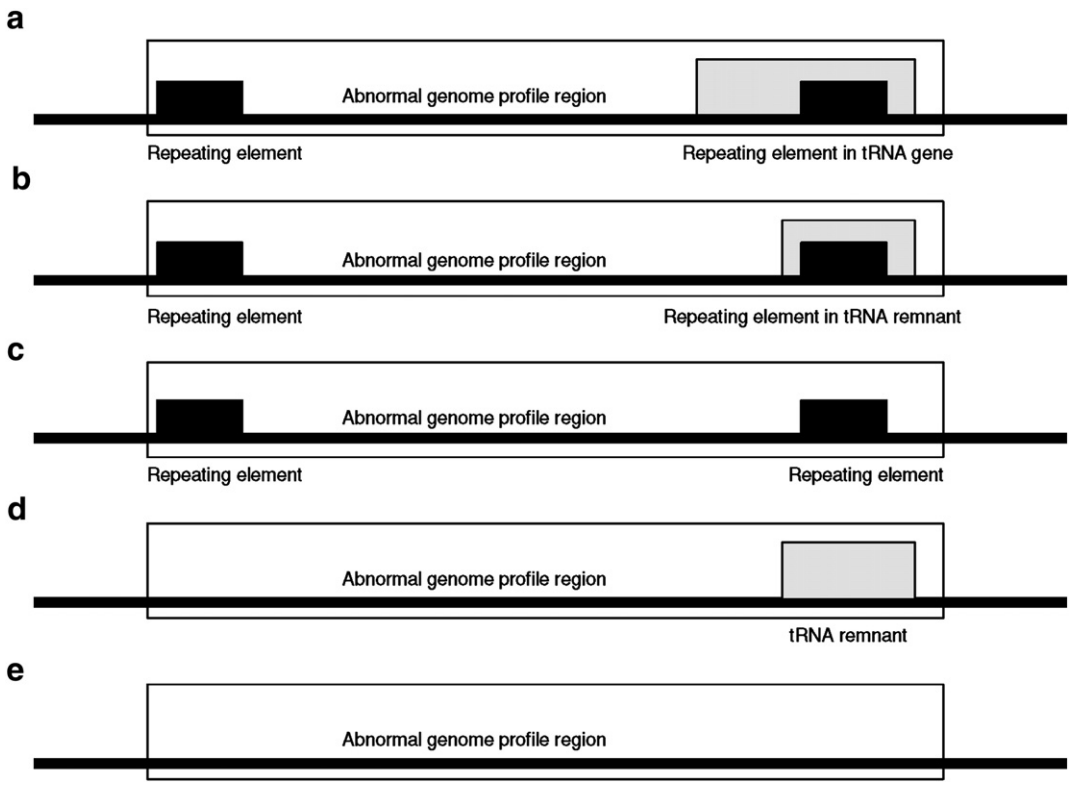

Fig. 4. All the possible boundaries of given GIs. (a) The repeat elements appear at both the boundaries of the GIs and the full-length tRNA genes. (b) Repeating elements appear, but only tRNA remnants appear in the boundaries. (c) Repeating elements appear on both sides of GI without tRNA-associated sequences. (d) Repeating elements are lost but the tRNA-associated sequences remain. (e) No hotspot sequences appear.

and *S. enterica* subsp. enterica serovar Typhi CT18. Therefore, the comparative method is less sensitive. The compositional and comparative approaches are alternative means of detecting foreign DNA, each with limitations and disadvantages. This study attempts to integrate both approaches to overcome these limitations. GIs and non-GIs defined using the comparative method are translated into our genome profile. The differences in profiles between GIs and core chromosomes in which these GIs reside are calculated.

3.1. The combinational indices

This study describes a combinational approach to detect GIs in a microbial genome, including compositional, comparative, and annotation-based methods. The proposed SVM model is checked by cross-validation and an independent approach. A comparison is also made of the accuracy of each feature set generated by reported data sets. The proposed method is compatible with both IslandPick's sets, which makes the accuracy of GI-GPS higher than when testing one data set. The genome profiles indicate that some differences arise between core chromosomes and genomic islands; in addition, these four types of compositional features (codon usage preferences, oligonucleotide bias, codon adaption indices, and GC contents of the individual positions of codons) compensate for individual weaknesses. For example, when the DNA dinucleotide pattern is similar to the core chromosome, other feature sets may not exhibit the same similarities, as in the testing island profile shown in Figs. 3(a–d).

3.2. The evolutionary insight

What is the original and what is the key factor of environmental adaptation of microbes? The mosaic structure of GIs, where many homolog segments in different taxa are found, may suggest a possible evolutionary pathway. Without traditional sequence alignment methods, GI-GPS are a highly effective alternative to searching for island homologous segments in all sequenced microbial genomes. GI-GPS can be used to build the network of these island sequences, and the original donor organisms of these islands can be identified.

3.3. The boundary problem

The hybrid hypothesis (Hou, 1999) espouses that the sequences of 3′-end of tRNAs can be hybridized to DNA to stabilize the double helix during the recombination process. The tRNA sequences show a high degree of conservation among organisms, which can facilitate transfer across different species of microorganisms. Once recombination occurs, the short repeating sequences should be identified on both sides of the island. Therefore, the repeating sequences and tRNA gene sequences could be markers for island identification. GI prediction tools such as Islander and IslandPath identify GIs based on methods that use tRNA genes as hotspots (Hsiao et al., 2003; Mantri and Williams, 2004; Ou et al., 2006). These markers are used to search for the possible footprints of GI insertion to help refine the determined boundaries of GI candidates more precisely. Our results further demonstrate that the short direct repeating sequences appear in chromosomes frequently, which may pose a challenge in isolating the boundaries. Research indicates that the length of the direct repeating elements can range from 4 to more than 20 bases. How crossing events can choose such short direct repeating elements and avoid longer repeating elements warrants further research.

3.4. Compositional and homologous approaches working together

In the GI-predicting methods, the compositional approaches are the simplest strategies. However, these methods require complete genomic information as the genomic compositional standard. Microbial genomes do not always represent the same sequence compositions as entire chromosomes. Some regions show higher sequence variations. The compositional approach might show a higher false positive rate

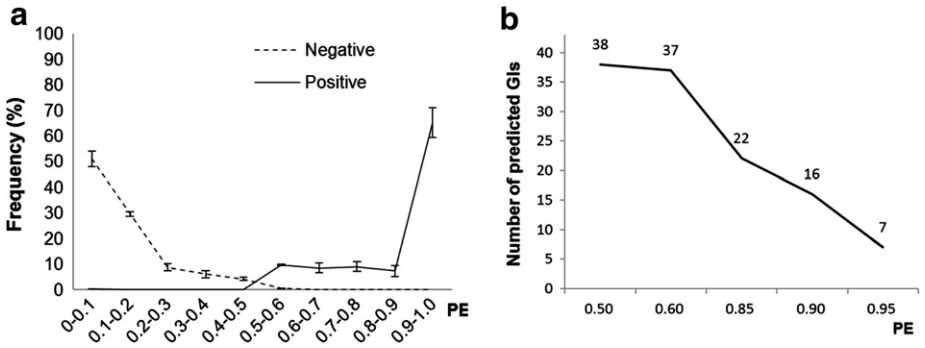

Fig. 5. The PE score of the SVM classifier. The PE (probability estimation) scores generated by SVM classifier of two data sets are tested. (a) We tested the positive and negative GI sequences of IslandPick data sets. The *x*-axis represents the PE score, and the *y*-axis represents the amount of GIs divided by the number of all GI sequences. The error bar represents the standard deviations of three times experiments. (b) The SVM classifier used in real GI data sets is downloaded from PAIDB (see Supplementary Table 1). The *x*-axis represents the cut-off PE score, and the *y*-axis represents the accuracy under the PE cut-off score.

than other GI predicting methods. Therefore, we integrated the mobile genetic elements downloaded from the ACLAME database (Leplae et al., 2010). We designed a homologous sequence filter of mobile genetic elements. All positive sequences predicted using our SVM classifier are filtered by an MGE filter. Only the sequences with MGE evidence are possible genomic island sequences. The profiles of host chromosomes and GI sequences used to build the SVM models were extracted from the island sequences predicted using comparative GI-detecting methods (Langille et al., 2008). The GI-GPS uses the sequence compositional characteristics from the comparative approach, and also used the mobile elements as a refining data set, which makes the GI-GPS a combinational GI-detecting method.

3.5. The completed or in-complete genome: the perspective on genomic island detecting

According to the synthetic draft genome experiments (Table 1), the GI-GPS showed high GI prediction accuracies for the different percentages of the remaining genomes. This well performance may have been caused by the compositional differences of the foreign and native chromosomal regions, which are consistent regardless of how long the sequences are. This characteristic is useful for pre-analyses for ongoing genome projects in which the genome sequences are in-complete and some regions have not been sequenced. The GP-GPS module uses small segments as detecting units in which the lengths are smaller than the general genomic islands. In addition, the easy-to-use modules for assembling and gene predicting are incorporated in the GI-POP system. Furthermore, researchers can use the GI-POP to determine the possible GIs of the draft genomes and to perform pre-annotation. This is useful for ongoing genome projects because the pre-analytic data provides knowledge for additional wet-lab experiments. For example, deciding which gaps of the in-complete genome cannot be assembled by high throughput sequencing should be filled by PCR (polymerase chain reaction) first.

Table 1
The performance evaluation of synthetic draft genomes*.

Salmonella enterica serovar Typhimurium LT2 uid57799[a]			*Escherichia coli* O157 H7 Sakai uid57781[a]		
Remaining[b] Genome (%)	Remaining[c] GI (%)	Accuracy (%)	Remaining[b] Genome (%)	Remaining[c] GI (%)	Accuracy (%)
100	100.0	81.9	100	100.0	82.8
80	81.8	84.4	80	80.0	79.3
60	61.0	83.4	60	59.8	82.5
40	50.3	80.3	40	45.2	77.7
20	23.2	83.1	20	20.9	82.3
10	14.9	78.9	10	14.9	69.8

*All the scores are the averages of three-times repeating experiments.
[a] The length of the GI of *E. coli* is 882 kb, and 833 kB of *S. enterica*.
[b] The remaining genome represents the length of synthetic draft genome divided by the length of the whole genome sequence.
[c] The remaining GI is the length of GIs of synthetic genome divided by total length of GIs.

4. Methods

4.1. Data preparation

The sequence files of prokaryotic complete genomes were obtained from the National Center of Biotechnology Information (NCBI) FTP server [ftp://ftp.ncbi.nih.gov/refseq/release/complete/]. For SVM training, the 771 positive and 3700 negative GIs belonging to 118 bacterial chromosomes were downloaded from the IslandPick (Langille et al., 2008) Web site [http://www.pathogenomics.sfu.ca/islandpick_GI_datasets/].

4.2. Databases and software packages used for annotation pipeline

The DIYA is a Perl-based package designed for microbial genome annotation and visualization (Stewart et al., 2009). The DIYA can perform the gene annotation efficiently. The input genomic sequences can be either complete genomes or unfinished genomic sequences such as contigs or scaffolds. We modified the scripts of the DIYA to require it to perform the GI prediction, and developed an in-house PHP script to process the Meta data, which was generated by each subroutine and functional module in our pipeline. Fig. 1 shows the flowchart of annotation and GI prediction in an ongoing genome project. The programs used in the annotation pipeline are as follows:

Glimmer (Salzberg et al., 1998): For predicting genes in microbial DNA.
TRNAscan-SE (Lowe and Eddy, 1997): To search for tRNA genes.
RNAmmer (Lagesen et al., 2007): For predicting ribosomal RNA genes.
BLAST (Brodmann et al., 2001): To find regions of local similarities between sequences.
CDD (Marchler-Bauer et al., 2011): Conserved Domain Database provides conserved domains of proteins.
UniRef (Mulder et al., 2008): Clusters of sequence from UniProt Knowledge base.
COG (Tatusov et al., 2000, 2001): Clusters of Orthologous Groups.
MGE (Leplae et al., 2010): Mobile genetic elements.
Circus(Krzywinski et al., 2009): Visualization of COG analysis.
GBrowse(Donlin, 2007): Visualization of annotated sequences.

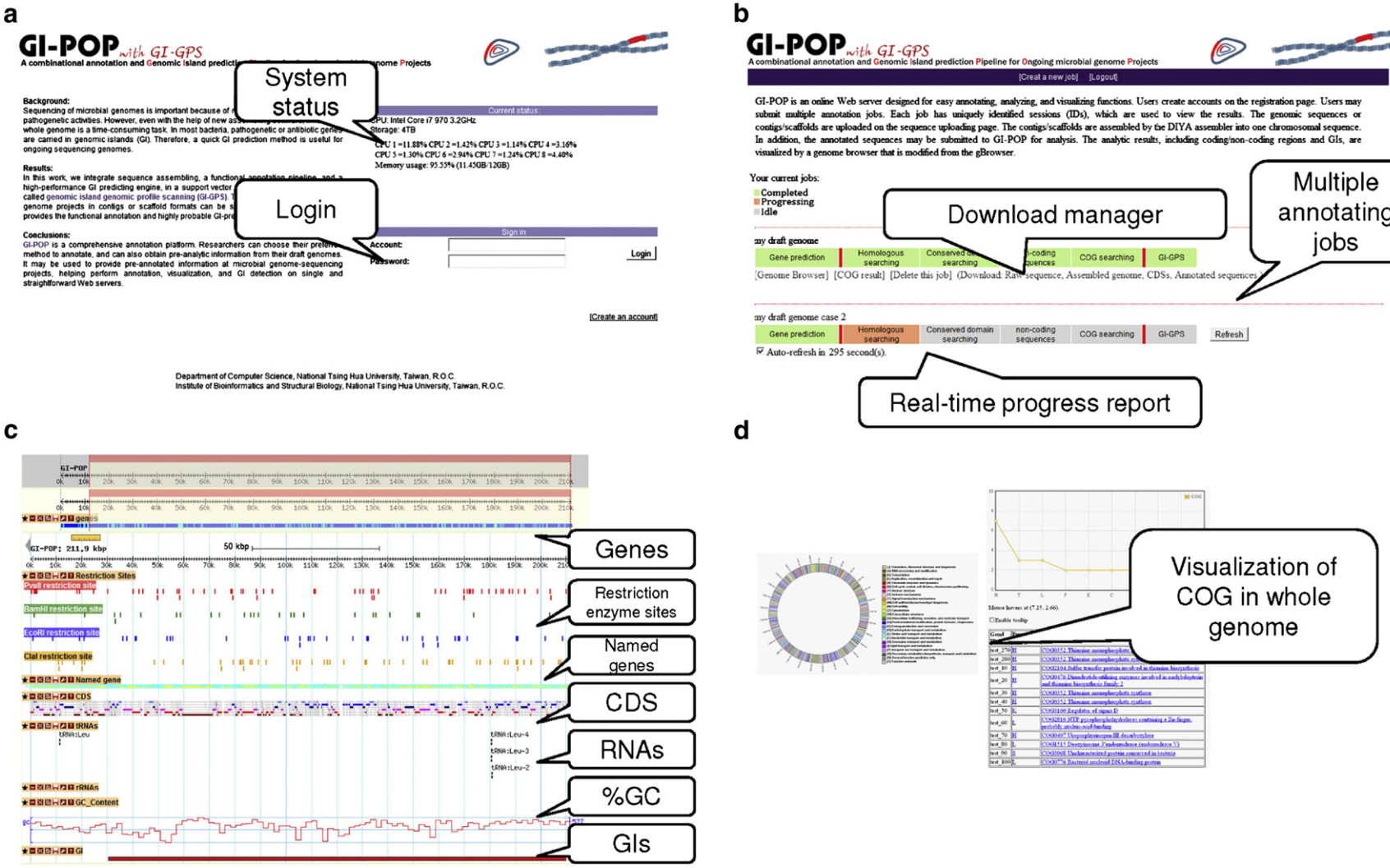

Fig. 6. The Web interface of GI-POP. (a) GI-POP has the abilities for multiple users to perform online genome annotation and GI prediction. (b) The home page for each user provides a dashboard to monitor all the submitted jobs; the progress of each stage is real-time updated. The annotated sequences can be downloaded from the download manager page when the process is finished. (c) A user friendly genome browser provided in our Web server including predicted genes, coding/non-coding regions and possible GIs. The functional predictions by CDD and UniRef database are also represented in this page. Users can click on any predicted gene, the detailed information of functional prediction is shown. (d) COG mapper offers the distributions of the Cluster of Orthologous Groups in the chromosome.

4.3. Components of genome profiles

Creating a genome profile involves calculating various indices belonging to four types. Class 1 includes 16 dinucleotide bias indices; Class 2 consists of the usage preferences of 64 codons; Class 3 contains codon adaption indices (CAI) that measure theoretical gene expression levels (Sharp and Li, 1987); and Class 4 includes positional GC contents of codons. The oligonucleotides, amino acid usages, and GC contents of different positions of codons are determined from sequenced genomes. The codon adaptation indices are then calculated using program CAI and cusp of the Emboss package. Finally, all results are normalized by the number of genes of a given genome to avoid the effect of sequence length variations, allowing us to compare the core chromosomes with targeted foreign DNA segments. The *F*-score provided by the LIBSVM package was used as the scoring for feature selection. The *F*-score of the *i*th feature is defined by the following formula:

$$F_i = \frac{\left(\overline{x}_i^p - \overline{x}_i\right)^2 + \left(\overline{x}_i^q - \overline{x}_i\right)^2}{\frac{\sum_{k=1}^{P}\left(x_{k,i}^p - \overline{x}_i^p\right)}{P-1} + \frac{\sum_{k=1}^{Q}\left(x_{k,i}^q - \overline{x}_i^q\right)}{Q-1}}$$

The amounts of positive and negative instances are *P* and *Q*, where $\overline{x}_i^p$, $\overline{x}_i^p$ and $\overline{x}_i$ are the average values of positive, negative, and complete data sets. $x_{k,i}^p$ is the *i*th feature of the *k*th positive instance, and $x_{k,i}^q$ represents the negative instance. The *F*-score represents the discrimination of a feature in a given feature set; for example, a larger score is more discriminative. The features with smaller *F*-scores are removed in the feature selection process.

4.4. SVM training and testing

Initially, 600 positive and negative GIs were randomly selected from 118 bacterial chromosomes, adapted from the IslandPick data set (Langille et al., 2008). The compositional indices of these islands and of their core chromosomes were translated into a genome profile. The performance of the SVM model was then evaluated by selecting all islands larger than 5 kb and calculating the frequencies of probable estimates retrieved from the SVM outcome. All SVM calculations were performed using the LIBSVM package (*i.e.*, a general library for support vector classification and regression (Chang and Lin, 2011)). During training, the sequences were translated into genome profiles by PHP programs and the EMBOSS package (Rice et al., 2000). The results were obtained with five-fold cross-validation. The radial basis function (RBF) kernel was used for all experiments. The leave-one-out approach, which measures the prediction accuracy (ACC), was systematically implemented by isolating a chromosome from the data set during training, and then testing the classifiers against the GIs in this single chromosome.

4.5. Evaluation of the predictive performance of the SVM classifier

Various SVM models were developed using different classes of feature sets, including nucleotide composition and/or amino acid composition indices. Each SVM model was tested and trained by a five-fold cross-validation approach. The data set was then evaluated using these SVM models. Finally, the accuracy of different lengths and coverage rates of the GIs was calculated.

4.6. Generate tRNA and direct repeating sequences

The tRNAscanSE program was used for tRNA gene prediction (Lowe and Eddy, 1997), and the unique signature-discovering (USD) tool (Lee et al., 2004) was used for efficient sequence searching. All possible short sequences were initially generated. All matched sequences in a given region were located using USD; the mismatching tolerance was then set to one base.

4.7. Sliding window mechanism

Short sliding windows increase the rate of false positives because the sequence composition does not always represent similar distributions in every chromosome region. Additionally, a sliding window shortened by more than the average gene length causes inaccuracies in the genome profile. We chose 5 kb as the sliding window size representing the highest degree of accuracy. The sliding window scans along the genome sequentially; whether the region covered by the window was a GI was determined using a well-trained SVM classifier. Neighboring fragments determined as GIs were then joined to full-length GI candidates. Additionally, a program was designed to detect repeating elements. Fifty nucleotides in the 3′ of the tRNA genes were then cut as footprints to detect the boundaries of candidate GIs; its complement sequences were also detected. One nucleotide mismatch was allowed in the searching process. The boundaries of candidate GIs were refined based on the location of the 3′ end of identified tRNA genes and repeating elements. Finally, to obtain the highest scored GI candidate, each candidate GI region was ranked by probabilistic estimation.

5. Conclusions

In this study, we developed a combinational annotation and genomic island prediction pipeline for ongoing microbial genome projects (GI-POP). To the best of our knowledge, the GI-POP is the first integrated genome annotating and GI predicting server that provides genome assembly, coding/non-coding region prediction, homologous gene searching, COG alignment, and genomic island detecting. In many genome projects, the genomes are unfinished. For example, draft genomes are often composed of contigs or scaffolds. There are no extant GI detecting methods that can process unfinished genomes. For this reason, we developed a compositional based method called GI-GPS to detect the GI regions of in-complete genome sequences. The GI-GPS integrates the sequences compositional GI detecting approach using an SVM classifier with the homologous search approach, which helps reduce the false positive cases generated by the SVM classifier. We believe that GI-POP can provide simple and easy annotating, as well as GI detecting assistance. Finally, it can help researchers obtain pre-analytic information for their ongoing genome projects.

Conflict of interest statement

The authors declare that they have no competing interests.

Contributors

CCL, TJY and CYM designed and carried out this study. TJY designed and performed the synthetic draft genome experiment. CCL and WCL drafted the manuscript and analyze the data. PCL, CYT, WCL and YPPC conceived the study, participated in its design and helped draft the manuscript.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.11.063.

Acknowledgments

This work is funded by the National Science Council, Taiwan, R.O.C. with grant numbers 100-2221-E-126-010-MY3, and 101-2319-B-400 -001.

References

Binnewies, T.T., et al., 2006. Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. Funct. Integr. Genomics 6, 165–185.

Brodmann, P.D., Nicholas, G., Schaltenbrand, P., Ilg, E.C., 2001. Identifying unknown game species: experience with nucleotide sequencing of the mitochondrial cytochrome b gene and a subsequent basic local alignment search tool search. Eur. Food Res. Technol. 212, 491–496.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. 2 (27), 1–27 (27).

Donlin, M.J., 2007. Using the Generic Genome Browser (GBrowse). Curr. Protoc. Bioinformatics (Chapter 9:Unit 9.9).

Frost, L.S., Leplae, R., Summers, A.O., Toussaint, A., 2005. Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. 3, 722–732.

Hacker, J., Kaper, J.B., 2000. Pathogenicity islands and the evolution of microbes. Annu. Rev. Microbiol. 54, 641–679.

Hacker, J., et al., 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various extraintestinal Escherichia-coli isolates. Microb. Pathog. 8, 213–225.

Hou, Y.M., 1999. Transfer RNAs and pathogenicity islands. Trends Biochem. Sci. 24, 295–298.

Hsiao, W., Wan, I., Jones, S.J., Brinkman, F.S., 2003. IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19, 418–420.

Koonin, E.V., Makarova, K.S., Aravind, L., 2001. Horizontal gene transfer in prokaryotes: quantification and classification. Annu. Rev. Microbiol. 55, 709–742.

Krzywinski, M., et al., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19, 1639–1645.

Lagesen, K., Hallin, P., Rodland, E.A., Staerfeldt, H.H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 35, 3100–3108.

Langille, M.G.I., Hsiao, W.W.L., Brinkman, F.S.L., 2008. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9.

Lee, H.P., Sheu, T.F., Tsai, Y.T., Shih, C.H., Tang, C.Y., 2004. An efficient algorithm for unique signature discovery on whole-genome EST Databases. 2004 IEEE Computational Systems Bioinformatics Conference, Proceedings, pp. 650–651.

Leplae, R., Lima-Mendez, G., Toussaint, A., 2010. ACLAME: a classification of mobile genetic elements, update 2010. Nucleic Acids Res. 38, D57–D61.

Lobry, J.R., 1996a. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Lobry, J.R., 1996b. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Mantri, Y., Williams, K.P., 2004. Islander: a database of integrative islands in prokaryotic genomes, the associated integrases and their DNA site specificities. Nucleic Acids Res. 32, D55–D58.

Marchler-Bauer, A., et al., 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 39, D225–D229.

Merkl, R., 2004. SIGI: score-based identification of genomic islands. BMC Bioinformatics 5.

Mulder, N.J., Kersey, P., Pruess, M., Apweiler, R., 2008. In silico characterization of proteins: UniProt, InterPro and Integr8. Mol. Biotechnol. 38, 165–177.

Nag, S., Chatterjee, R., Chaudhuri, K., Chaudhuri, P., 2006. Unsupervised statistical identification of genomic islands using oligonucleotide distributions with application to Vibrio genomes. Sadhana Acad. Proc. Eng. Sci. 31, 105–115.

Ou, H.Y., et al., 2006. A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. Nucleic Acids Res. 34.

Ragan, M.A., 2001. Detection of lateral gene transfer among microbial genomes. Curr. Opin. Genet. Dev. 11, 620–626.

Rajan, I., Aravamuthan, S., Mande, S.S., 2007. Identification of compositionally distinct regions in genomes using the centroid method. Bioinformatics 23, 2672–2677.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16, 276–277.

Salzberg, S.L., Delcher, A.L., Kasif, S., White, O., 1998. Microbial gene identification using interpolated Markov models. Nucleic Acids Res. 26, 544–548.

Schmidt, H., Hensel, M., 2004. Pathogenicity islands in bacterial pathogenesis. Clin. Microbiol. Rev. 17, 14.

Sharp, P.M., Li, W.H., 1987. The Codon Adaptation Index — a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Stewart, A.C., Osborne, B., Read, T.D., 2009. DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics 25, 962–963.

Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 28, 33–36.

Tatusov, R.L., et al., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29, 22–28.

Tu, Q., Ding, D.F., 2003. Detecting pathogenicity islands and anomalous gene clusters by iterative discriminant analysis. FEMS Microbiol. Lett. 221, 269–275.

van Passel, M.W.J., Bart, A., Thygesen, H.H., Luyf, A.C.M., van Kampen, A.H.C., van der Ende, A., 2005. An acquisition account of genomic islands based on genome signature comparisons. BMC Genomics 6.

Yoon, S.H., et al., 2007. Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic Acids Res. 35, D395–D400.