

An SVM-based algorithm for identification of photosynthesis-specific genome features

Gong-Xin Yu^a, Al Geist^a, George Ostrouchov^a, Nagiza F. Samatova^{a,b}

ABSTRACT

This paper presents a novel algorithm for identification and functional characterization of “key” genome features responsible for a particular biochemical process of interest. The central idea behind our algorithm is that individual genome features (or their combinations) are identified as significant “key” features if the discrimination accuracy between two classes of genomes with respect to a given biochemical process is sufficiently affected by the inclusion or exclusion of these features. In this paper, genome features are defined by high-resolution gene functions. The discrimination procedure utilizes the Support Vector Machine (SVM) classification technique. Changes in classification accuracy in response to addition or deletion of genome features measure the significance of these features. The application of this SVM-based feature identification algorithm to the oxygenic photosynthetic process resulted in a total of 126 highly confident candidate genome features. They cover not only dominant genome features (that always occur in oxygenic photosynthetic genomes but not in the other genomes) but also weak yet complementary genome features (their combinations make unique dominant genome features). While many of these features are well-known components in the oxygenic photosynthetic process, others are completely unknown, even including some hypothetical proteins. It is obvious that our SVM-based feature identification algorithm has the capability to discover novel genome features related to a targeted biochemical process.

Keywords: key genome features, oxygenic photosynthetic process, genome comparative analysis, support vector machines

INTRODUCTION

Genomes contain hundreds to thousands of genes; many of them encode multi-component, multi-faceted protein machines that function in a highly coordinated way for accomplishing various biochemical processes (Riley 1998; Zhou and Miller 2002). These biochemical processes cover all aspects of highly specialized yet coordinated (in structures, metabolic and regulatory networks) life supporting systems developed over millions, even, billions of years of evolution and speciation (Kasting and Siefert 2002). As a consequence, host organisms can survive and prosper in a diversity of hostile environments. Identification of “key” genes that are responsible for these processes is an essential step toward understanding the genetic and biochemical basis of entire life processes in general, and biologically important biochemical processes related to sustainable sources of energy, environmental management, carbon cycle and sequestration, and human health protection in particular (Heffelfinger *et al.* 2002). However, such a task presents a tremendous challenge for both experimental and computational scientists because of the complexity of gene networks involved in these processes. For experimental scientists, it is prohibitively expensive and labor-intensive to knock-out all possible genes to

^a Computational Biology Group, Computer Science and Mathematics Division, Oak Ridge National Laboratory*, P.O. Box 2008, Oak Ridge, TN 37830

^b To whom correspondence should be addressed: yug@ornl.gov and samatovan@ornl.gov

* Oak Ridge National Laboratory is managed by UT-Battelle for the LLC U.S. D.O.E. under Contract No. DE-AC05-00OR22725.

determine their effects on particular cellular processes (Hutchison *et al.* 1999). Function redundancies due to duplicated genes and alternative pathways may cause further complications in this analysis. For computational scientists, the evaluation of all possible candidate genes and their combinations presents a huge computationally infeasible challenge.

Current approaches to the problem of identifying “key” genes are based on genome comparative analysis, one of several sprouting and highly efficient methodologies (Tatusov *et al.* 1997; Bansal 1999; Bork *et al.* 1998). Advances in high-throughput genome sequencing have made it possible to use such approaches. At the time of this writing, approximately 126 genomes (16 Archaea, 93 Eubacteria, 17 Eukarya) have already been sequenced and made public^c. These genomes correspond to organisms of a tremendous biodiversity. They can be grouped into various genome classes such as Eubacteria vs. Archaea, pathogens vs. non-pathogens, thermophiles vs. mesophiles, aerobic vs. anaerobic organisms, oxygenic photosynthetic vs. non-oxygenic photosynthetic organisms. The comparative analysis of these genome classes opens an opportunity to identify genes that are highly specific to these processes, thus, improving our understanding of their genetic and biochemical basis.

The simplest and quickest strategy for genome comparative analysis is to directly compare genome sequences. This method has been utilized for understanding the genetic basis of pathogenesis (Strauss and Falkow 1997, Bansal *et al.* 1998, Glaser *et al.* 2001; Deng *et al.* 2002). The underlying principle is that if some genes occur in one genome (e.g. a pathogenic organism) but are completely absent in the other (e.g. a closely related, but non-pathogenic organism), it can be assumed that these genes may be responsible for targeted biochemical processes. However, this strategy is somewhat simplistic because the analysis does not have the capacity to differentiate between genes that make a decisive contribution to the cellular processes and those genes that just occur by chance. Furthermore, when applied only to two closely related genomes, the approach cannot be extended to multiple genomes with heterogeneous (environmental or evolutionary) backgrounds.

Ortholog-based genome comparative analysis is another strategy for the whole genome analysis that is applicable to multiple genomes (Bansal 1999; Makarova *et al.* 1999; Raymont *et al.* 2002). Orthologs are defined as the exact functional counterparts in different genomes that have arisen from speciation (Fitch, 1970). This approach has been used to identify evolutionarily stable core gene sets for Archaea genomes, for genes with conserved functionalities among microbial genomes, and for genes specific to photosynthetic prokaryotes (Makarova *et al.* 1999; Bansal 1999; Raymont *et al.* 2002). In particular, the occurrences of orthologous genes among experimental genomes are evaluated to determine if these genes are evolutionarily conserved or specific to given genomes. This approach is also over-simplified. However, it is not necessarily true that all important gene components for certain biochemical processes are the ones that always occur only in the genomes with these processes. As a result, this approach has limitations for identification of genes essential for specific biochemical processes.

This paper presents a novel method for identification of key genes responsible for a particular biochemical process of interest. The central idea behind our method is that individual genome features (or their combinations) are identified as significant if the discrimination between two

^c <http://wit.integratedgenomics.com/GOLD/>

classes of genomes with respect to a given biochemical process is sufficiently affected by their addition or removal. The genes corresponding to significant features are reported as the key genes for this process. In this paper, genome features are defined by high-resolution gene functions organized into a hierarchical set of knowledge-driven protein function groups and families (Yu 2003). The discrimination procedure utilizes the Support Vector Machine (SVM) classification technique (Vapnik, 1998). Changes in classification accuracy in response to the addition or deletion of genome features are significance measures for these features.

We apply the proposed method to identification of key genes responsible for the oxygenic photosynthetic process. Understanding the genetic basis of this process is of considerable interest due to its extreme importance for transformation of solar energy into other forms of energy available to all living organisms. We show that our SVM-based feature identification algorithm has the capability not only to identify the well-known components in the oxygenic photosynthetic process, but also to discover new and novel candidate genome features that are completely unknown, even hypothetical proteins.

MATERIALS AND METHODS

This section describes the algorithm used to identify genes that are critical for a given biochemical process of interest.

Key genome features identification problem

Let P be a target biochemical process. Assume that a set $X = \{x_1, x_2, \dots, x_n\}$ of n genomes is partitioned into two classes: a positive class X_P^+ of genomes that have the biochemical process P and a negative class X_P^- of genomes in which P is absent. For example, if P is an oxygenic photosynthetic process then X_P^+ might include genomes such as *Prochlorococcus marinus* MED4, *Prochlorococcus marinus* MIT9313, *Synechococcus sp.* WH8102, *Nostoc sp.* PCC7120, *Anabaena sp.* PCC7120, *Thermosynechococcus_elongatus* BP1, and *Synechocystis* PCC6803 and X_P^- might include genomes such as *Saccharomyces cerevisiae*, *Thermotoga maritime*, *Shewanella*, *Streptococcus pyogenes* SF370. Let $F = \{f_1, f_2, \dots, f_m\}$ be a set of m genome features and 2^F be its power set, which includes all possible combinations of genome features from F . For instance, F might consist of all possible protein functions or protein domains. The *key genome features identification problem for process P* is the problem of identifying the set $S_P \subseteq 2^F$ of all genome features or their combinations that contribute to or are responsible for the process P in any of the genomes from X_P^+ . A *key genome feature* is any feature from S_P . There are three major steps in addressing the key genome features extraction problem:

1. Given a process P and a set of genomes X , define a set F of genome features.
2. Identify the set $S_P \subseteq 2^F$ of key genome features.
3. Validate that elements from S_P are key genome features or their combinations.

For the remainder of this paper, we let P denote specifically the oxygenic photosynthetic process and let X include 52 genomes partitioned into two classes: 7 genomes from X_P^+ and 45 genomes from X_P^- .

Genome features

We refer to a set F of genome features as a set of high-resolution *protein function groups* (*PFGs*), as we did in an earlier study (Yu 2003). They were automatically generated from the analysis of 111,046 protein sequences in the Swiss-Prot (Bairoch and Apweiler 2000) database, version of 6/25/2002, and this was followed by manual curation. This resulted in $m=21,656$ PFGs, or genome features. These groups often represent the smallest biochemical or evolutionary units encoded by single genes. They also represent universal symbols for given functions that occur in multiple genomes. Some of these groups are broadly shared by species across their phylogenetic tree or by species that live in different environmental niches, while others are species-, environment- or biochemical process-specific. As a result of function annotation of 52 genomes, the PFGs can be utilized as a common genome feature system for the genome comparative analysis. PFGs are further organized into 2157 protein families.

Every genome $x_i \in X = \{x_1, \dots, x_{52}\}$ either has a genome feature $f_j \in F = \{f_1, \dots, f_{21,656}\}$ with some confidence value v_{ij} ($v_{ij} \neq 0$) or does not ($v_{ij} = 0$). If there are protein sequences in genome x_i that belong to the PFG f_j , then v_{ij} equals the highest value among the confidence values of protein sequence assignments to the PFG f_j . Otherwise, $v_{ij} = 0$. The confidence values v_{ij} are arranged into a genome matrix. A *genome matrix* $V_{n,m}$ is a matrix of n genome vectors, each with m vector coordinates corresponding to PFGs.

An algorithm for identification of key genome features

We begin with a biochemical process P , a set of genomes X , some with and some without P , and a set of genome features F . A set $S_P \subseteq 2^F$ of key genome features can be identified by a brute-force approach, namely by evaluating all possible combinations of genome features whether they sufficiently affect discrimination between the two sets of genomes. However, this approach is feasible only for a small number of genome features. Even for $m=10$ features, there are 2^{10} combinations to try. Since the number of possible genome feature combinations grows exponentially with the number of features, a heuristic computational algorithm for key genome features identification is proposed.

The underlying assumption behind the algorithm is that individual genome features can be identified as key features if the discrimination between two classes of genomes with respect to P is sufficiently affected by the addition or deletion of these features. We use Support Vector Machines, a supervised classification technique (Vapnik 1998; Jaakkola *et al.* 1999), as the discrimination procedure. Changes in classification accuracy in response to the deletion/addition of genome features measure the contribution of these features to the process P .

In brief, the SVM-based feature identification algorithm pursues feature addition and deletion in a hierarchical manner. First, the set $X = X_P^+ \cup X_P^-$ of 52 genomes, partitioned into a positive class (7 genomes) and a negative class (45 genomes) with respect to the oxygenic photosynthetic process P , is used as a training set by the SVM classifier. The rows of $V_{n,m}$ begin with the full m -dimensional genome feature space. The leave-one-out error of classification performance is assessed (see the “Performance measures” section). If the error is below some cutoff value (10% is used as a default), then it can be assumed that the set F of genome features is sufficient for

classification. In this case, the SVMs are able to capture an internal structure in the training data. Otherwise, either the set F of genome features needs to be redefined or a forward-feature-propagation procedure should be explored (see below).

Identifying target protein families. Next, the algorithm selects target protein families (each with several PFGs) for more detailed analysis. Since the size of the search space is enormous, namely $2^{21,656}$, only the target protein families are used to make identification of key genome features in that space computationally feasible. Each individual protein family k defined by a subset $F_k \subset F$ of PFGs is evaluated for providing a small (below some threshold value; 4% is used as default) leave-one-out classification error by using the SVM classifier on a set of genomes X feature space F_k . This increases the opportunity of identifying weak candidates for key genome features in the subsequent feature extraction process by reducing the chance of multiple candidate features confounding each other.

Identifying key protein function groups within the target protein families. Finally, the algorithm identifies key PFGs (or their combinations) for each target protein family selected in the previous step using the following exhaustive search. For each PFG or combination of PFGs in a protein family, its contribution to the overall performance of the SVM-based classifier is measured. Again, the underlying assumption is that if these genome features can make a significant difference in the performance of the learning system, then it is highly possible that they represent features that can be considered key for the biochemical process of interest.

Two approaches are used to measure the contribution of individual genome features or combinations of features: *backward-feature-propagation* and *forward-feature-propagation*. With backward-feature-propagation, an individual genome feature $f \in F_k$ is removed with replacement from a genome matrix defined for a target protein family k selected in the “Identifying target protein families” section. The leave-one-out classification error of the SVM classifier applied to a set of genomes X in k -dimensional feature space is calculated. The larger the degradation of performance is, the greater the contribution of the removed feature to the process P can be assumed. Once the highly contributing individual genome features are selected (based on a threshold value; 5% degradation is a default), the combined contribution of selected features to the process P is evaluated after all the selected features are removed.

With the forward-feature-propagation procedure, the genome matrices used for the SVM-based classification are built bottom-up from the individual PFGs within each target protein family. Initially, every genome matrix corresponding to a target protein family k consists of exclusively one PFG. The contribution of this PFG is assessed similarly based on the performance of the SVM classifier. Once the highly contributing individual protein function groups are selected (based on a threshold value; 80% performance accuracy), an individual pair of the selected protein function groups is included into a genome matrix. The contribution of the pair to the biochemical process P is evaluated.

Validation of key genome features

To provide preliminary evidence for the relationships between key genome features and given biochemical processes, we investigate the co-occurrence (or clustering) of gene corresponding to

key genome features on the genomes. The underlying assumption is that if these genes are clustered together on the genome as well as with some well-known gene components of the biochemical process, it is very likely that these key genome features are related to this process (by function coupling: Overbeek *et al.* 1999; Zheng *et al.* 2002).

RESULTS

Discrimination of genome classes using protein function groups as genome features

A comprehensive assessment of the leave-one-out error has shown that our SVM-based classification algorithm can achieve maximum classification performances on the genome data (Table 1). All five measurements reach the highest value of 1 using a polynomial kernel function of any power up to 4 as well as using a radial kernel function. A reasonably good performance but with much smaller number of support vectors (48% of the total size of the training set) is attained for the polynomial kernel of power 6. Obviously, our decision to use the protein function groups as our genome features is the right choice for discrimination of genomes into two classes based on the oxygenic photosynthetic process.

Table 1. The performances of SVM on genome data

Kernel	SV (%)	ROC	Fp	Fn	Tp	Tn	N	SP	SN	CC	CS
I. Genome data											
Plynm1	94.20%	1	0	0	7	45	52	1	1	1	1
Plynm 2	82.70%	1	0	0	7	45	52	1	1	1	1
Plynm 3	80.80%	1	0	0	7	45	52	1	1	1	1
Plynm 4	80.80%	1	0	0	7	45	52	1	1	1	1
Plynm 5	78.80%	0.99683	1	0	7	44	52	0.88	1	0.92	0.98
Plynm 6	48.10%	1	0	1	6	45	52	1	0.86	0.92	0.98
Plynm 7	38.50%	0.71746	4	6	1	41	52	0.2	0.14	0.06	0.81
Plynm 8	34.60%	0.74603	15	4	3	30	52	0.17	0.43	0.07	0.63
Radial	100%	1	0	0	7	45	52	1	1	1	1

Plynm: kernel function of polynomial, SV: Support vectors, Fp: False positives, Fn: False negatives, Tp: True positives, Tn: True negatives, N: Total number of genomes, SP: Specificity, SN: Sensitivity, CC: Correlation coefficient and CS: Cost savings. ROC: A score of the normalized area under a curve that plots true positives as a function of false positives for varying decision thresholds.

Identification of target protein families

The assessment of SVM performance on the family-based genome data demonstrates that only a small proportion of the protein families have a differentiation capability for the genome classes. A total of 2172 protein families are examined. Less than a quarter of the protein families (510 out of 2170) provide classification performance under a training ROC cutoff of 0.5. Similarly, less than 3.5% (72 out of 2172) have classification performance under a training ROC cutoff of 0.96. This result indicates that the majority of key genome features will be covered within 510 protein families at a confidence of 0.5 (training ROC) and within 72 protein families at a confidence of 0.96 (training ROC).

Identification of key genome features within the target protein families

For the purpose of demonstrating the efficacy of our SVM-based feature extraction algorithm, protein families with confidence of 0.96 or greater (training ROC) are chosen as our target protein families. Two approaches are used to identify the key genome features (i.e., protein

function groups) in these families: backward-feature-propagation and forward-feature-propagation (see *materials and methods* for more details). They are applied in a coordinated way so that our algorithm can achieve a maximum capability to identify candidate genome features responsible for the biochemical process of interest.

The backward-feature-propagation approach can recognize both single key genome features as well as combinations of genome features that make a key genome feature. A total of seventy-six key genome features are identified. The classification accuracy drops by at least 5% on the removal of these genome features from the genome matrix. In some cases, however, the backward-feature-propagation approach fails to detect some potential key genome features. It either misses some candidate genome features or cannot identify any key features at all even though the classification accuracy for the target protein family is very high. In this case, even accumulated ROCs are still much smaller than that of their family ROCs. The accumulated ROCs are obtained when all potential key features are removed from the genome matrix.

The forward-feature-propagation procedure is a more exhaustive method undertaken for feature selection. This approach is able to identify key genome features within the protein families when the backward-feature-propagation fails either partially or completely. About 50 additional key genome features have been identified, demonstrating an enhanced sensitivity.

Validation of key genome features

The application of the SVM-based feature selection algorithm to the oxygenic photosynthetic process results in the identification of a total of 126 highly confident key genome features. They cover not only dominant key genome features (that always occur in oxygenic photosynthetic genomes but not in the others) but also weak yet complementary key genome features (their combinations make unique dominant key genome features). While many of these features are easily recognized gene components in the oxygenic photosynthetic process (features in italic in **Table 2**), the others are less clear (features in bold in **Table 2**). Some of them are even genes that encode proteins annotated as “hypothetical”. To provide preliminary evidence for the relationships between these genome features and the oxygenic photosynthetic process, we investigate gene clustering among those (in **Table 2**) and other (in supplementary tables) key genome features on the genome of *Synechococcus sp.* Strain: WH 8102. Two out of five established genome feature clusters are shown (**Table 3**). Each of the clusters is built around one particular genome feature as indicated by the name prefix of the cluster. For example, “Erythromycin-biosynthesis-sensory-transduction-protein-cluster” is constructed around “Erythromycin biosynthesis sensory transduction protein” under protein family of PR00344.

A case-by-case investigation of these genome feature clusters demonstrates that the majority of the genes in these clusters correspond to the key genome features identified by our SVM-based feature selection algorithm. “Erythromycin-biosynthesis-sensory-transduction-protein-cluster” is one of the biggest key genome feature clusters. A total of 16 genes are covered and five are considered as dominant features. Among these five, four are the genes that are known to be involved in the oxygenic photosynthetic biochemical process. This result indicates that these key genome features are potential components of gene networks for the oxygenic photosynthetic biochemical process. It is worthwhile to mention that both the gene at the genome position 1591 in this cluster and the Ycf52 protein at the position 1615 in Cluster II are identified as key genes,

although they are annotated as hypothetical genes. Apparently, our algorithm has an ability to provide some functional clues about these hypothetical proteins.

Table 2. Functions of candidate genome features and their class categorizations

Pfg index	Fp	Fn	Tp	Tn	N	Pfg	
20361	0	0	7	45	52	Ycf52 protein	IPB000182
2611	0	0	7	45	52	<i>Carbon dioxide concentrating mechanism protein</i>	IPB000249
4874	0	0	7	45	52	<i>Light-independent protochlorophyllide reductase iron-sulfur ATP-binding protein (EC 1.18.-.-) (LI-POR subunit L) (DPOR subunit L).</i>	IPB000392
15885	0	0	7	45	52	<i>Photosystem D2 protein</i>	IPB000484
16343	0	0	7	45	52	<i>Proteasome</i>	IPB000484
5372	0	0	7	45	52	Sulfite reductase (Ferredoxin) (EC 1.8.7.1)	IPB000660
15877	0	0	7	45	52	<i>Photosystem 44 kDa reaction center protein</i>	IPB000932
15886	0	0	7	45	52	<i>Photosystem P680 chlorophyll A apoprotein</i>	IPB000932
3647	0	0	7	45	52	<i>Cytochrome</i>	IPB001417
6527	0	0	7	45	52	Mannose-1-phosphate guanylyltransferase (GDP) (EC 2.7.7.22) (GDP-mannose pyrophosphorylase) (GMP)	IPB001538
15859	0	0	7	45	52	<i>Photosystem</i>	IPB002628
15894	0	0	7	45	52	<i>Photosystem reaction center J protein</i>	IPB002682
5523	0	0	7	45	52	Precorrin-3B C17-methyltransferase (EC 2.1.1.131) (Precorrin-3 methyltransferase) (Precorrin-3 methylase)	IPB003043
15888	0	0	7	45	52	<i>Photosystem assembly protein</i>	IPB003359
15895	0	0	7	45	52	<i>Photosystem reaction center L protein</i>	IPB003372
15864	0	0	7	45	52	<i>Photosystem</i>	IPB003375
15866	0	0	7	45	52	<i>Photosystem</i>	IPB003666
13188	0	0	7	45	52	<i>Magnesium-chelatase</i>	IPB003672
15867	0	0	7	45	52	<i>Photosystem</i>	IPB003685
15915	0	0	7	45	52	<i>Photosystem reaction center protein</i>	IPB003687
8895	0	0	7	45	52	Precorrin-8X methylmutase (EC 5.4.1.2) (Precorrin isomerase)	IPB003722
5039	0	0	7	45	52	Precorrin-6X reductase (EC 1.3.1.54)	IPB003723
15868	0	0	7	45	52	<i>Photosystem</i>	IPB003757
8431	0	0	7	45	52	<i>Ribulose biphosphate carboxylase small chain (EC 4.1.1.39) (RuBisCO small subunit).</i>	PR00152
15871	0	0	7	45	52	<i>Photosystem</i>	PR00353
4904	0	0	7	45	52	Ferredoxin--NADP reductase (EC 1.18.1.2) (FNR)	PR00371
1574	0	0	7	45	52	<i>Apocytochrome</i>	PR00610
3687	1	0	7	44	52	<i>Cytochrome B6-F complex</i>	IPB000179
9800	1	0	7	44	52	Erythromycin biosynthesis sensory transduction protein	IPB000653
4642	1	0	7	44	52	<i>Cytochrome B6-F complex iron-sulfur subunit (EC 1.10.99.1) (Rieske iron-sulfur protein) (RISP)</i>	IPB001281
19852	1	0	7	44	52	Urease accessory protein	IPB002639
19853	1	0	7	44	52	Urease accessory protein	IPB002669
19854	1	0	7	44	52	Urease accessory protein	IPB002894
2683	1	0	7	44	52	<i>Cation-transporting</i>	IPB003581
6183	1	0	7	44	52	Adaptive-response sensory-kinase sasA (EC 2.7.-.-)	PR00344

Fp: False positive, Fn: false native, Tp: True positive, Tn: True negative. N: total number of genomes.

Table 3. Gene clustering of selected key genome features on *Synechococcus* sp. genome

Gene Cluster	GFSA	ROC	PC	Geneid	Gene Position	Function information	
I	<i>BWFP</i>	0.68	-100%	or2283	1583	Deoxyribodipyrimidine photolyase (DNA photolyase) (Photoreactivating enzyme). E.C.4.1.99.3 IPB002081	
	-	-	-	or2284	1584	S-adenosylmethionine synthetase IPB002133	
	-	-	-	or2286	1585	Hypothetic protein	
	<i>FWFS</i>	0.97		or2287	1586	Erythromycin biosynthesis sensory transduction protein IPB000653	
	<i>FWFP</i>	0.65		or2289	1587	Multidrug resistance protein IPB001140	
	-	-	-	or2290	1588	Helix-turn-helix protein, CopG family	
	-	-	-	or2291	1589	PIN (Pfam domain) IPB002716	
	<i>BWFP</i>	0.84	-40%	or2292	1590	Magnesium-chelatase IPB002078	
	<i>BWFP</i>	0.86	-100%	or2293	1591	Hypothetic protein IPB002882	
	<i>BWFP</i>	0.57	-100%	or2295	1592	Rubredoxin (Rd). Rubredoxin IPB001052	
	-	-	-	or2296	1593	Hypothetic protein	
	<i>BWFP</i>	1	-100%	or2297	1594	Cytochrome B559 alpha chain IPB001417	
	<i>BWFP</i>	1	-100%	or2298	1595	Cytochrome B559 beta chain. Cytochrome IPB001417	
	<i>BWFP</i>	1	-100%	or2299	1596	Photosystem reaction center L protein IPB003372	
	<i>BWFP</i>	1	-100%	or2300	1597	Photosystem reaction center J protein IPB002682	
	<i>FWFP</i>	0.8		or2305	1598	UDP-glucose 4-epimerase (Galactowaldenase) EC 5.1.3.2 IPB001509	
	II	<i>FWFP</i>	0.74		or3145	2225	Iron(III)-transport ATP-binding protein IPB001140
		<i>BWFP</i>	0.44	-100%	or3146	2226	ATP phosphoribosyltransferase EC 2.4.2.17 IPB001348
<i>FWFP</i>		0.86		or3147	2227	Hemolysin IPB001140	
<i>BWFP</i>		1	-70%	or3148	2228	Ycf52 protein IPB000182	

GFSA: Genome feature selection approach, PCC: Performance percentage change, -:no testing data available, BWFS: Backward feature propagation, FWFS: Backward feature propagation. If *FWFP*, the feature selection confidences are determined only by its performance (ROC) (the single feature SVM performance), otherwise, the confidences are measured by both family-based ROC and the percentage change in leave-one-out experiments

DISCUSSION AND CONCLUSION

In this paper, we present an SVM-based genome feature selection algorithm for the identification of genes that contribute to the oxygenic photosynthetic process. The algorithm is applied in a hierarchical fashion so that the search space for key genome features is gradually narrowed from full-feature genome data matrix of 52 rows (annotated genomes) and 21,656 columns (genome features, or function groups) to protein family-based matrices with a few features. As a result, as many as 126 highly confident key genome features are identified. Twenty-seven of those features occur only among oxygenic photosynthetic genomes. The majority of these genome features (20 out of 27) have already been documented as functions specific to the biochemical process. This result provides strong evidence that the key genome features are indeed the potential components of oxygenic photosynthetic gene networks.

Additional evidence comes from the clustering of genes based on their co-occurrence on the genome. Each of the five gene clusters established around a single key genome feature includes

the genes that either correspond to the predicted key genome features or are known to be the gene components of the oxygenic photosynthetic process. Specifically, the Ycf52-protein-cluster was identified in 3 out of 7 annotated cyanobacterial genomes. The majority of genes from this cluster are the components of the biochemical process. They are highly conserved across multiple genomes. These results provide an additional support for the capability of our algorithm to identify key genome features directly related to the targeted biochemical process.

Dominant vs. weak key genome features

The predicted key genome features can be partitioned into two categories: the dominant features and the weak ones. The dominant genome features make a significant contribution to the SVM's classification performance. Using a single dominant feature is sufficient for building a highly confident separating hyperplane. They occur only in the positive class of genomes (in which the biochemical process is present) and are completely absent in the negative class. The identification of such dominant genome features is straightforward (Jaakkola *et al.* 1999). The weak candidate genome features do not make a significant contribution to the classification performance but they do contribute significantly once they are combined with other weak genome features. The prediction of such weak key genome features is especially important for building genetic networks. Most of the existing approaches reported in literature fail to identify such weak key genome features. For example, in Raymond *et al.* (2002), only 50 genes were identified as photosynthesis-related (occurred in all photosynthetic genomes and some or all of non-photosynthetic genomes) and 3 genes as photosynthesis-specific (occurred only in photosynthetic genomes). Thus, our SVM-based feature selection algorithm provides the capability far beyond the reach of any exiting genome comparative analysis approach.

Future development

Our SVM-based feature identification algorithm finds candidate key genome features at various confidence levels. The evaluation of the reliability of predictions without a biology wet lab experimental support presents a tremendous problem not only for our algorithm but also for any computational algorithm. To alleviate this problem and improve the prediction accuracy, we plan to text-mine the InterPro data (Mulder *et al.* 2003) for the purpose of extracting all possible functions that are related, either directly or indirectly, to our targeted biochemical cellular processes. Once such a knowledge base is built, we can estimate the specificity of our algorithm by compare the functions from the knowledge base with the predicted ones. Statistical analysis of these data will give us a rough estimation of the confidence level and provide a scientific basis for determining the threshold parameters in future applications of our algorithm to genome features selection.

ACKNOWLEDGEMENTS

This work was funded in part or in full by the US Department of Energy's Genomes to Life program (www.doegenomestolife.org) under project, "Carbon Sequestration in Synechococcus Sp.: From Molecular Machines to Hierarchical Modeling," (www.genomes-to-life.org). The work of G.O. was sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory. This research used resources of the Center for Computational Sciences at Oak Ridge National Laboratory.

REFERENCES

- Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Bansal, A.K. (1999). An automated comparative analysis of 17 complete microbial genomes. *Bioinformatics*. **15**:900-908.
- Bansal, A.K., Bork.P. and Stuckey, P.J. (1998). Automated pair-wise comparison of microbial genomes. *Math. Model. Sci. Comput.*, **9**:1-23.
- Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *J Mol Biol.*, **283**:707-725.
- Deng, W., Burland, V., Plunkett III, G., Boutin, A., Mayhew, G. F., Liss, P., Perna, N. T., Rose, D. J., Mau, B., Zhou, S., Schwartz, D. C., Fetherston, J. D., Lindler, L. E., Brubaker, R. R., Plano, G. V., Straley, S. C., McDonough, K. A., Nilles, M. L., Matson, J. S., Blattner, F. R., and Perry, R. D. (2002). Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184**:4601-4611.
- Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. *System. Zool.*, **19**:99-113.
- Glaser, P., Frangeul, L., Buchrieser, C., Rusniok, C., Amend, A., Baquero, F., Berche, P., Bloecker, H., Brandt, P., Chakraborty, T., Charbit, A., Chetouani, F., Couvé, E., de Daruvar, A., Dehoux, P., Domann, E., Domínguez-Bernal, G., Duchaud, E., Durant, L., Dussurget, O., Entian, K.D., Fsihi, H., Garcia-Del Portillo, F., Garrido, P., Gautier, L., Goebel, W., Gómez-López, N., Hain, T., Hauf, J., Jackson, D., Jones, L.M., Kaerst, U., Kreft, J., Kuhn, M., Kunst, F., Kurapat, G., Madueño, E., Maitournam, A., Mata Vicente, J., Ng, E., Nedjari, H., Nordsiek, G., Novella, S., de Pablos, B., Pérez-Diaz, J.C., Purcell, R., Remmel, B., Rose, M., Schlueter, T., Simoes, N., Tierrez, A., Vázquez-Boland, J.A., Voss, H., Wehland, J., and Cossart, P. (2001). Comparative Genomics of *Listeria* Species, *Science* **294**: 849-852.
- Heffelfinger, G.S., Martino, A., Gorin, A., Xu, Y., Rintoul III, M.D., Geist, A., Al-Hashimi, H.M., Davidson, G.S., Faulon, J.L., Frink, L.J., Haarland, D.M., Hart, W.E., Jakobsson, E., Lane, T., Li, M., Locascio, P., Olken, F., Olman, V., Palenik, B., Plimpton, S.J., Roe, D.C., Samatova, N.F., Shan, M., Shoshoni, A., Strauss, C.E.M., Thomas, E.V., Timlin, J.A., and Xu, D. (2002). Carbon Sequestration in *Synechococcus* Sp.: from molecular machines to hierarchical modeling. *OMICS, A Journal of Integrative Biology*, **6**:305-330.
- Hutchison III, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., and Venter, J.C. (1999). Global Transposon Mutagenesis and a Minimal Mycoplasma Genome. *Science*, **286**:2165-2169.
- Jaakkola, T., Diekhans, M., and Haussler, D. (1999). Using the Fisher kernel method to detect remote protein homologies. *Proc Int Conf Intell Syst Mol Biol*: 149-58.
- Kasting, J.F., and Siefert, J.L. (2002). Life and the Evolution of Earth's Atmosphere. *Science*, **296**: 1066-1068.
- Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I. and Koonin, E.V. (1999). Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell, *Genome Res.* **9**: 608-628.
- Mulder N.J., Apweiler ., Attwood T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R.R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A.,

- Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S.E., Pagni M., Peyruc D., Ponting C.P., Selengut J.D., Servant F., Sigrist C.J.A., Vaughan R, and Zdobnov E.M. (2003). *The InterPro Database, 2003 brings increased coverage and new features. Nucl. Acids. Res.*, **31**: 315-318.
- Overbeek, R., Fonstein. M., D'Souza, M., Pusch, G.D., and Maltsev, M. (1999). The use of gene clusters to infer functional coupling. *Genetics*, **96**, 2896-2901.
- Raymond, J., Zhaxybayeva, O., Gogarten, J.P., Gerdes, S.Y., and Blankenship, R.E. (2002). Whole-Genome Analysis of Photosynthetic Prokaryotes. *Science* **298**: 1616-1620.
- Riley, M., (1998). Systems for categorizing functions of gene products. *Curr. Opin. Struct. Biol.*, **8**:388-92
- Strauss, E.J., and Falkow, S. (1997). Microbial Pathogenesis: Genomics and Beyond. *Science*, **276**:707-712.
- Tatusov, R.L., Koonin, E.V., Lipman, D.J. (1997). A genomic perspective on protein families. *Science*. **278**:631-637.
- Vapnik, V. (1998). *Statistical Learning Theory*. Wiley.
- Xiong, J., and Bauer, C.E. (2002). Complex evolution of photosynthesis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, **53**:503–521.
- Zheng, Y., Roberts, R.J., and Kasif, S. (2002). Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.* **3**, 1-9.
- Zhou, J.Z., and Miller, J.M. (2002). Microbial genomics – challenges and opportunities: the 9th international conference on microbial genomes. *J. of Bacteriology*, **184**: 4327-4333.
- Yu, G. X. (2003). A rule-based Knowledge system for Function Annotations in High-throughput Sequence Analysis. *Bioinformatics* (to be submitted).