BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 15 2010, pages 1834–1840
doi:10.1093/bioinformatics/btq305

Genome analysis

Advance Access publication June 6, 2010

Efficient learning of microbial genotype–phenotype association rules

Norman J. MacDonald and Robert G. Beiko*

Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Finding biologically causative genotype–phenotype associations from whole-genome data is difficult due to the large gene feature space to mine, the potential for interactions among genes and phylogenetic correlations between genomes. Associations within phylogentically distinct organisms with unusual molecular mechanisms underlying their phenotype may be particularly difficult to assess.

Results: We have developed a new genotype–phenotype association approach that uses Classification based on Predictive Association Rules (CPAR), and compare it with NETCAR, a recently published association algorithm. Our implementation of CPAR gave on average slightly higher classification accuracy, with approximately 100 time faster running times. Given the influence of phylogenetic correlations in the extraction of genotype–phenotype association rules, we furthermore propose a novel measure for downweighting the dependence among samples by modeling shared ancestry using conditional mutual information, and demonstrate its complementary nature to traditional mining approaches.

Availability: Software implemented for this study is available under the Creative Commons Attribution 3.0 license from the author at http://kiwi.cs.dal.ca/Software/PICA

Contact: beiko@cs.dal.ca

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

Received on April 13, 2010; revised on June 1, 2010; accepted on June 3, 2010

1 INTRODUCTION

Variation in an organism's observed physical and environmental attributes, known as its phenotype, is determined in part by its gene composition, known as its genotype. Organismal traits are often influenced not by a single gene, but by combinations of genes. Comparing profiles of orthologous genes or proteins (Tatusov *et al.*, 1997, 2003) with profiles of phenotypes is a common method for identifying gene correlations (Gaasterland and Ragan, 1998; Pellegrini *et al.*, 1999). Genes that confer a particular phenotype can be inherited from ancestors, created through duplication of existing genes and mutation, or acquired from other organisms through lateral gene transfer (LGT) events. Once thought to be restricted to a few exceptional events, LGT is now known to have played a central role in bacterial evolution, including in the emergence of traits such as thermophily (Beiko *et al.*, 2005), photosynthesis (Raymond *et al.*, 2002) and antibiotic resistance (Enright *et al.*, 2002).

The genomes of prokaryotic organisms typically contain 1000–6000 genes. Many have studied methods of associating single genes or clusters of genes to phenotypes, including with direct distribution comparisons (Raymond *et al.*, 2002), analysis of neural network parameters (Martin *et al.*, 2003), statistical significance (Goh *et al.*, 2006; Jim *et al.*, 2004; Liu *et al.*, 2006) and mutual information (MI; Slonim *et al.*, 2006). These approaches did not consider combinations of genes that do not share a similar pattern of distribution across genomes; if there is a gene that is only important when other genes have already been taken into consideration, it could be missed. Levesque *et al.* (2003) looked at several set-theoretic methods of combining cluster profiles. Several studies used more advanced machine learning techniques to build up multi-gene associations, such as recursive feature elimination (Guyon *et al.*, 2002) and filter (Wang *et al.*, 2005) methods on cancer microarray data. Recently, Kastenmüller *et al.* (2009) used metabolic profiles with wrapper, filter and embedded methods of feature selection. Depending on the classifier used, wrapper methods can be computationally expensive on feature-rich datasets as a classifier must be trained multiple times on various subsets of features.

Association rule mining (ARM; Agrawal *et al.*, 1993) is an alternative method to mine multi-feature associations. ARM is a data mining technique that finds associations among features, constructing rules of the general form [A,B] $\mapsto$ [C]. Classification using ARM is the basis of NETCAR (Tamura and D'haeseleer, 2008), a MI-based method of pruning the huge search space of gene combinations to identify genotype–phenotype relationships that would not be otherwise detected.

Feature selection is often used to find these predictive combinations of genes; however, good predictors do not necessarily have a causal influence on the property, but can simply be correlated with it (Guyon and Elisseeff, 2003). Common correlated factors, such as dependence among samples due to shared ancestry, can confound predictive analyses.

One of the strongest confounding patterns in phenotype prediction is the evolutionary ancestry of the organism (Harvey and Pagel, 1991). The default mode of genetic inheritance is vertical, with organisms typically acquiring genes from their direct ancestors. If a particular phenotypic trait is common or ubiquitous within a taxonomic group, then genes that correlate significantly with the presence of the trait may in fact be characteristic of the taxonomic group rather than causal of the trait. Figure 1 shows the uneven distribution of temperature environment preference over 387 bacteria and 40 archaea, subdivided by phylum. In this

*To whom correspondence should be addressed.

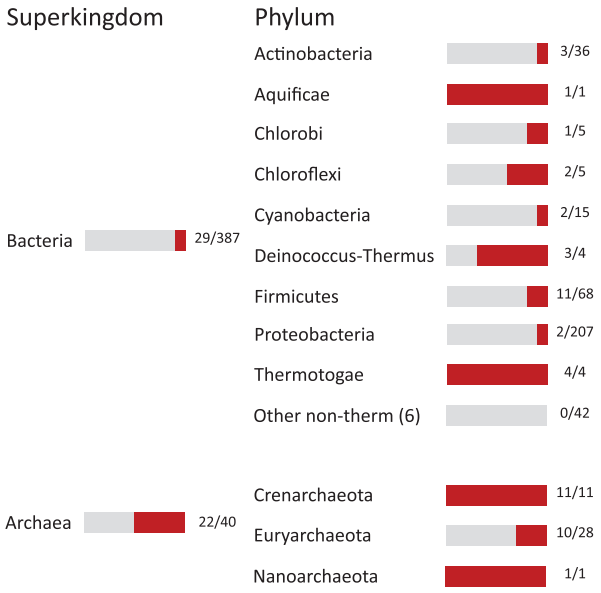

Fig. 1. Distribution of temperature preference of 387 bacteria, grouped by phylum. The bars represent the proportion of thermophiles (dark) and non-thermophiles (light). The number next to each bar is the number of microbes in that group. The uneven distribution of the phenotype class labels over the taxonomic groupings demonstrates the dependence among samples.

dataset, 62% of thermophiles are within three phyla: Crenarchaeota (21%), Eurarchaeota (20%) and Firmicutes (20%). The other 38% of thermophiles are unevenly distributed (from 0% to 8%) over 15 other phyla. As identifying predictors is a *needle-in-a-haystack* problem, with tens of thousands of possible genes over only hundreds of examples, ignoring confounding shared ancestry can lead to predictors of overrepresented groups (such as, in this case, the three main phyla). *Clostridium thermocellum* is a thermophilic member of an otherwise mesophilic genus. In this case, we would like to pay particularly close attention to those differences between *C.thermocellum* and the other clostridia. By subtracting genes common among all close relatives, we amplify the importance of genes that are different among closely related thermophiles and non-thermophiles. If these genes are present in other thermophiles, they will be further emphasized. By downweighting genes that are expected to be present due to common descent, we emphasize the genes with unusual distributional patterns due to selective gene loss or LGT and may therefore play a particularly important determining role in the phenotype. Phylogenetic dependency networks (Carlson *et al.*, 2008) have been used to mitigate shared ancestry of HIV viral proteins by constructing Bayesian networks conditioned on a phylogenetic tree. With our approach we do not depend on fine-grained and uncertain relationships within strict evolutionary hierarchies of microorganisms, as may be necessary with closely related viral samples such as HIV, and instead allow the exploration of shared ancestry within large clusters at various taxonomic depths.

We apply a predictive rule mining approach, which has not previously been tested on genotype–phenotype mapping problems, that achieves better predictive accuracy than NETCAR and has lower running time. To mitigate the effect of shared ancestry, we also propose a novel metric based on conditional MI (CMI) (Section 2.2) for incorporating information regarding shared ancestral descent into genotype–phenotype analyses. While this approach does not improve the overall classification accuracy, it highlights sets of genes with relevant biological functions, and highlights the strong influence of phylogenetic correlation.

2 APPROACH

2.1 Associative classification rule mining

Class ARM is useful for determining associations between sets of genes and phenotypes. Genes that do not independently correlate with a given trait may not be found in a gene-by-gene feature selection approach, but enumerating all possible subsets of genes is impossible for large feature spaces. Here we examine two algorithms designed to efficiently mine sets of genes associated with a phenotype with class ARM.

NETCAR (Tamura and D'haeseleer, 2008) first finds all single gene–phenotype associations with an MI score above a given threshold, known as the parent genes. It then builds a gene connectivity graph from these parents, where a connection implies the MI between a parent and another gene is above a second MI threshold. Child genes are defined as being within x steps of a parent on the connectivity graph, where x is user-defined. All such parent–child relationships are used to construct candidate associations between gene set and phenotype. It then uses MI between each gene set and the target phenotype to evaluate the most likely causative relationships. Consequently, NETCAR aims to discover sets of genes that show correlated patterns of occurrence. The authors reported various patterns of predicted gene interaction for six phenotypes.

Classification based on Predictive Association Rules (CPAR; Yin and Han, 2003) builds predictive rules that cover examples for each class label in a dataset. Initially, all samples (here, genomes) in a dataset are uniformly weighted. Next, counts of positive and negative examples of the class label are constructed, as well as counts for each gene in the database. The genes that increase the Foil gain (Quinlan and Cameron-Jones, 1993) on the positive set, thereby correctly classifying more of its members, are added to new rules, which are then extended in a similar manner until the information gain of adding a new gene is below a given threshold. At this point, the association between gene set and phenotype is stored, and each positive example that the gene set covers is downweighted for the next iteration of the algorithm. This process continues until a minimum weight threshold has been reached (coverage of the dataset) or no more single genes are above the Foil gain threshold. The weights are reset to a uniform level for each sample and the process is repeated for each class label (e.g. for both the 'yes' and 'no' class labels). In this manner, CPAR is constructed to find rules that make generalizations over as many samples as possible for a class, and then iteratively focus in on those samples that were not covered by the generalizations.

2.2 Shared ancestry

MI is the measure of information shared between two random variables (Cover and Thomas, 2006). Formally, MI is defined as

$$I(X;Y)=\sum_{x,y} p_{x,y}(x,y)\log\frac{p_{x,y}(x,y)}{p_x(x)p_y(y)} \quad (1)$$

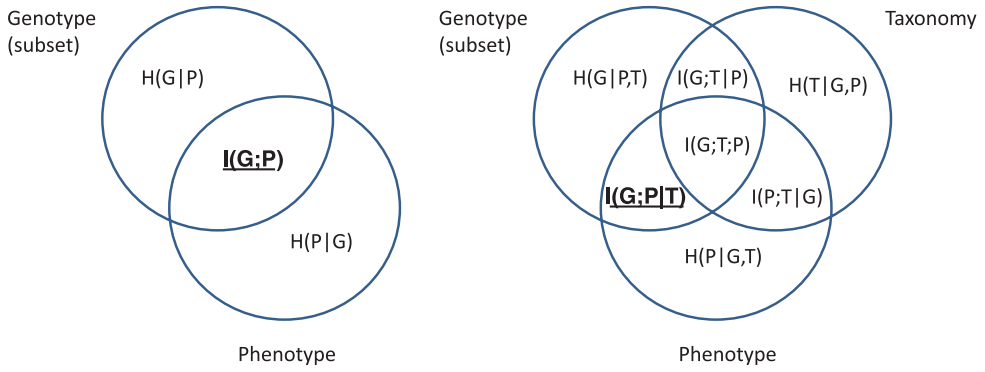

Fig. 2. Considering taxonomy as modeling a confounding factor. On the left is a typical information theoretic model for finding predictive attributes, where the goal is to choose G to minimize the entropy in our target, $H(P|G)$. On the right, we have an extra confounding attribute, taxonomy, which models genes that are correlated (and potentially predictive) with our target phenotype, but not necessarily biologically meaningful. As we are interested in understanding the system that produced our target class rather than only prediction, we would like to tell which predictive features have high $I(G;P|T)$, i.e. interaction between genes and phenotype that cannot be explained by the confounding taxonomic signal.

where subscripted p are the probability mass functions of the subscripts.

MI has been used as a method of finding features that follow a similar distribution as the target class, and are thus potentially predictive of that class. In genotype–phenotype association problems, we can find a predictive relationship between a gene and phenotype, but we would also like to downweight those genes that can be explained by common ancestry (Fig. 2). By modeling shared descent as a discrete random variable corresponding to a given taxonomic level, we can find the MI present between feature and class that is remaining after subtracting the information both share with the taxonomic level. This measure is known as CMI.

CMI has been examined recently for selecting non-redundant features (Fleuret, 2004; Wang and Lochovsky, 2004). CMI has been used to select features that are non-redundant with previously chosen features. The CMI evaluates the MI shared between the feature in question and the class in the context of the already known features. In this study, we are interested in finding features that have a biological influence on the class, and wish to evaluate the effect of taxonomy as a confounding factor.

Formally, CMI is defined as

$$I(X;Y|Z)=\sum_{x,y,z} p(x,y,z)\log\frac{p(x,y,z)p_z(z)}{p_{x,z}(x,z)p_{y,z}(y,z)} \tag{2}$$

where $p(x,y,z)$ is the joint probability mass function of random variables X,Y and Z, and subscripted p are the probability mass functions of the subscripts. With a log base of 2, the units of CMI are bits.

The information shared between random variables X and Y in the context of Z is given by CMI. For the phenotype problem, let X be the presence or absence of a gene set, Y be the phenotype, and Z be the taxonomic group, our confounding variable. The CMI gives us the amount of MI between the genes and the phenotype that cannot by explained by common ancestral signal. We can also think of the CMI as the amount of new information a feature brings for the prediction of a phenotype, if we have used the taxonomic label as the first step in a decision tree classifier.

2.3 Weighting with CMI

When MI is high, we know that there is a strong correlation between the variables. When accounting for the confounding effects of common descent, we are looking for variables that appear to influence the phenotype in ways that are not easily explained with common descent. Once we identify these variables, we want to report the strength of the dependence between the phenotype and genotype, regardless of taxonomy, i.e. the MI. Here we define a score that balances these two considerations, Conditionally Weighted MI (CWMI).

For any given target phenotype, there is a maximum possible CMI, as it is impossible to share more information than is present in either factor. Formally, the maximum CMI for this problem is the entropy remaining in phenotype having subtracted the MI of phenotype and taxonomic groups, $H(Y|Z)$, defined as

$$H(Y|Z)=\sum_{y,z} p_{y,z}(y,z)\log p_{y|z}(y|z). \tag{3}$$

In order to factor in the amount of MI in the context of taxonomy to the amount of MI as a whole, we use a normalized weight formula, defined when $H(Y|Z)$ is positive,

$$\alpha(X,Y,Z)=\frac{I(X;Y|Z)}{H(Y|Z)} \tag{4}$$

$$\mathrm{CWMI}(X;Y|Z)=\alpha(X,Y,Z)I(X;Y). \tag{5}$$

Here, we use CMI normalized between 0 and 1 to provide a measure of our confidence that the correlation between X and Y is not confounded by Z. Several benefits of this normalization include the ability to scale other predictive scores with α such as the Laplace error estimate used by Yin and Han (2003), its unit-less nature and its interpretation as the confidence that a given relationship is not confounded by common descent. Note that for a given problem, scaling $I(X;Y)$ by α is proportional to scaling $I(X;Y)$ by $I(X;Y|Z)$, as $H(Y|Z)$ is constant over a phenotype distribution.

When $\alpha=0$, then there is no evidence that X and Y share any more information than X and Z or Y and Z alone. When $\alpha=1$, then the information shared between X and Y is the maximum amount possible given the taxonomic grouping. Thus, CWMI

simultaneously covers the cases in which either the candidate gene or the target phenotype are conserved in the taxonomic group.

3 MATERIALS AND METHODS

3.1 Data

Clusters of orthologous groups (COG; Tatusov *et al.*, 2003) are manually curated groups of orthologous genes constructed by merging triangles of reciprocal best BLAST hits. Non-supervised orthologous groups (NOG: Jensen *et al.*, 2008) are constructed in a similar manner but are extended to more genomes. The phylogenetic COG/NOG (herein referred to collectively as COG) profiles were downloaded from the STRING v8.2 database (Jensen *et al.*, 2009) and COG descriptions were downloaded from the NCBI COG database (Tatusov *et al.*, 2003). Of the 630 available COG profiles from STRING, 55 eukaryotes were removed as only one had associated phenotype data and it was not a microbe. Of the 46 archaea and 529 bacteria remaining, 427 had data for at least 1 of the 10 phenotypes considered (see below). Before executing on each phenotype, organisms with a null label for that phenotype were removed. Redundant COG profiles were detected and removed, for example, the 47 615 COGs in the 427-microbe thermophily dataset were automatically reduced to 26 290 distinct COG patterns within 5 s.

Ten phenotypes, consisting of aerobic, anaerobic, facultative, gram-negative, halophilic, thermophilic, motile, photosynthetic, psychrophilic and endospore-forming traits (identified in this work as AEROBE, ANAEROBE, FACULT, GRAMNEG, HALO, THERM, MOTILE, PHOTO, PSYCHRO, and SPORE, respectively), were obtained from the Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) (Markowitz *et al.*, 2008). In Supplementary Table S1 the distribution of each phenotype is given. In the case of halophily, only positive examples are provided by JGI IMG. In this case, the negative set was downloaded from the list of sequenced prokaryotic organisms at NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

Taxonomic assignments of all genomes used in this study were downloaded from the NCBI taxonomy database (http://www.ncbi.nlm.nih.gov/Taxonomy). For taxonomic ranks not defined for a given organism, the ranks were propagated downward, from superkingdom to genus. For example, the undefined family and order of *Dehalococcoides ethenogenes* 195 were each set to its class, Dehalococcoidetes.

3.2 Classifier training

The Java implementation of NETCAR from Tamura and D'haeseleer (2008) was used for this study. Phylogenetic profiles were converted into NETCAR's sparse binary matrix format. Only the top 30 MI-ranked parent COGs [as explained by Tamura and D'haeseleer (2008), online Supplementary Material] were used as seeds to build rules composed of two and three COGs (size-2 and size-3 rules), as choosing an appropriate MI cutoff is dependent on the phenotype and problem size, and not intuitive to choose. Only positive class label rules are mined by NETCAR by default. In order to evaluate the predictive capability of rules mined by NETCAR, negative predictors were mined by switching the positive (1) and negative (0) class labels and re-executing NETCAR. For each phenotype and class label, NETCAR was executed for size-1, size-2 and size-3 rules, and the resulting lists are combined and sorted by F-score on the training set, the default sorting by NETCAR. Since the size-1 and size-2 rules are necessarily mined before the size-3 rules, and the pre-specification of rule size is just a limitation of the interface, the runtime performance of only the size-3 rule sets are compared with CPAR. Due to the pre- and post-processing required to compare with the NETCAR algorithm, a single, 2-fold cross-validation was performed between CPAR and NETCAR.

We developed an efficient, user-friendly, plug-in extensible framework implemented in Python for comparing genotype–phenotype algorithms (PICA). PICA handles input of data from human-readable files, efficiently compresses identical profiles and handles randomized, paired, cross-validated comparisons of user-defined training algorithms, while providing complete classification logs for identifying difficult-to-classify genomes. A CPAR plugin was built into our classification framework and executed with the settings from Yin and Han (2003).

3.3 CMI

Ten replicates of paired 5-fold cross-validation comparing features selected with MI, CMI, CWMI and an even combination of features from MI and CWMI were generated and tested using a Python interface to LIBSVM (Chang and Lin, 2001).

For each training set, the MI between each COG and phenotype was calculated. The CWMI was calculated among each COG, phenotype and taxonomic rank, where the ranks included genus, family, order, class, phylum and superkingdom. A binary flag of whether a COG was predictive of the positive or negative set was calculated based on observed frequencies, or assigned to the positive set in the case of equal observed frequencies.

The top x unique positive and the top x unique negative features from MI, CMI, CWMI and a combination of MI and CWMI were chosen, where x is in (2, 4, 10, 20, 50, 500).

3.4 Testing

Normally, association rule classifiers such as the classification based on association rules (CBA) algorithm (Liu *et al.*, 1998) or CPAR (Yin and Han, 2003) would be used to test classification accuracy of association rule sets. Since these decision-tree type classifiers are generally suboptimal compared with support vector machines (SVMs), we have instead used the rules as a feature-selection step, allowing the SVM to develop its own model of interaction among genes for positive and negative examples. Rules were broken down into component features, and the top unique x were taken from each set, balanced for positive and negative predictors, and used to train LIBSVM on that set using a linear kernel and cost parameter $C=5$, which controls the cost of misclassifying points when determining the margin between class labels. Each prediction made was recorded and summarized over all replicates. In all cases, the Q_α score (herein referred to as balanced accuracy) is used as the accuracy measure, which is an average of the positive set accuracy with the negative set accuracy (Baldi *et al.*, 2000).

4 RESULTS

4.1 Comparison of CPAR with NETCAR

CPAR and NETCAR were compared with 1 replicate of 2-fold cross-validation over 10 phenotype datasets. Each algorithm was trained and the resulting rules were broken down into single items, with equal numbers of positive and negative predictive genes for both approaches. The number of each positive (negative) set of genes was dependent on the minimum of the positive (negative) CPAR or NETCAR sets. The feature sets were then evaluated using SVM training and testing.

Table 1 contains balanced accuracies of CPAR and NETCAR trials, as well as the number of positive and negative genes used in each case. Since we performed cross-validation on the same training and test sets, we perform a paired t-test to evaluate differences in balanced accuracy. Considering all datasets, the difference in accuracy is not significant at the 0.05 level ($N=20$, paired $t=1.27$, $P=0.220$). However, there is a strong outlier effect due to the small psychrophile dataset: neither algorithm performs well on this dataset, and there are very few positive examples (17, less than half of the next largest, halophily, with 35). The observed difference in accuracy is due to NETCAR classifying 5/17 correctly versus

Table 1. Balanced accuracy comparison of CPAR and NETCAR

Phenotype	CPAR	NETCAR	P	N	S P	S N
AEROBE	**0.875**	0.861	103.5	89	15	16
ANAEROBE	**0.893**	0.860	129	70.5	15.5	5.5
FACULT	**0.833**	0.826	26.5	19.5	12	9.5
GRAMNEG	**0.994**	0.975	26.5	19.5	7.5	7
HALO	0.641	**0.655**	30	38.5	4	3
THERM	**0.843**	0.793	157	45	13	5
MOTILE	0.827	**0.833**	32.5	113	10	8.5
PHOTO	**0.909**	0.892	92.5	41	13.5	6.5
PSYCHRO	0.586	**0.645**	132.5	15	8	1
SPORE	**0.877**	0.830	83.5	42.5	13.5	1.5

P is the average number of distinct positive genes used (e.g. 'aerobe'), *N* is the average number of distinct negative genes used (e.g. 'not an aerobe') and *S* P and S N are the average number of distinct genes shared among the CPAR and NETCAR feature sets over the positive and negative sets, respectively. Best accuracy scores for each phenotype are shown in boldface.

Table 2. CPAR and NETCAR runtimes in seconds

Phenotype	CPAR	NETCAR
AEROBE	**15.6**	1317.2
ANAEROBE	**17.3**	1299.2
FACULT	**19.5**	2831.6
GRAMNEG	**5.0**	1186.6
HALO	**6.3**	1901.0
THERM	**10.6**	1250.9
MOTILE	**15.3**	1186.6
PHOTO	**8.0**	1516.8
PSYCHRO	**10.3**	8525.3
SPORE	**12.9**	1438.3

Best runtimes for each phenotype are shown in boldface.

CPAR classifying 3/17 correctly on the positive set, and 409/410 versus 410/410 on the negative set. After psychrophily is removed, the differences in accuracy are significant ($N = 18$ paired $t = 2.53$, $P = 0.021$); however, the effect size is small (0.855 versus 0.836). The mean runtimes, however, differ substantially: CPAR with 12.2 s and NETCAR with 1410.0 s (Table 2).

The difference in accuracy between CPAR and NETCAR seems to be attributable to CPAR covering taxonomically distinct groups. For example, in the thermophily group, CPAR was able to correctly classify five more thermophiles than NETCAR, covering both superkingdoms. NETCAR found more DNA replication, recombination and repair COGs than CPAR (fold 1: 36 versus 14, fold 2: 20 versus 13). Since CPAR tends to reduce redundancy when producing rule sets, it could be the case that some of the DNA repair COGs are found by NETCAR but not CPAR, but redundant with the original COGs.

We conclude that the rules mined with CPAR are marginally better predictors than those mined with NETCAR, and that CPAR has a significantly smaller runtime (100 times faster on average). CPAR also has the advantage of not requiring pre-specification of target rule sizes to mine. Thus, if a single gene is predictive in and of itself, the algorithm does not necessarily build a larger rule with other genes which would obfuscate the predictive gene's central role.

We further note that increased predictive accuracy on previously unseen testing data does not necessarily mean increased biological relevance, but it is an important filtering step of determining genes for further investigation. Following an investigation with CPAR, it is recommended that other COGs that are linked with the discovered rules are investigated, as these could be left out due to the built-in redundancy compensation of CPAR, but may play important biochemical or structural roles in contributing to the phenotype of interest. These linkages can be discovered with a simple similarity metric such as MI.

In order to gain a better understanding of the classification results produced by CPAR, 10 replicates of 5-fold cross-validation were performed. We note that some organisms were particularly difficult to classify (Supplementary Table S2). Some of these misclassifications were potentially not annotated correctly or were near the borderline between classes in the original dataset. For example, *Steptococcus thermophilus* LMD-9, a thermotolerant mesophile, is misclassified as a thermophile in 7 of 10 replicates; *Sulfurovum* sp., classified by JGI as a thermophile, was consistently classified as a mesophile in all 10 trials. In this case the CPAR predictions, rather than the phenotype label, appear to be correct: although *Sulfurovum* sp. was isolated from a hydrothermal vent environment, its optimum growth temperature appears to be 30–35°C (Nakagawa *et al.*, 2007).

Each of the 10 misclassified thermophiles was derived from taxonomic groups which contained predominantly mesophilic organisms. Of the 40 correctly classified thermophiles, 29 were part of homogeneous taxonomic ranks (to order, class or phylum). The strong correlation of classifier performance with the heterogeneity of taxonomic ranks implies that our classification technique may be finding genes spuriously correlated with phenotype, and are rather predictive of closely related organisms who happen to share the phenotypic trait thermophily.

4.2 Ranking with CWMI

We now demonstrate our novel method of correcting for shared ancestry, focusing on thermophily and the point at which homogeneity and heterogeneity seem to impact classifier performance overall, the taxonomic rank of order. Table 3 (Panel a) and Table 3 (Panel b) display results of the two measures on the same dataset, thermophily (THERM). The top 10 ranked COGs by MI for the thermophile class are given in Table 3 (Panel a), while the top 10 ranked COGs by CWMI are given in Table 3 (Panel b) (complete list available in Supplementary Table S3). Since it is known that DNA repair proteins are often specialized in organisms that live at high temperatures, we expect to see a high proportion of these genes labeled as having an influence on that phenotype. In the top 10 MI-ranked scores, we do in fact see two COGs that have been confirmed as being involved in DNA replication, recombination and repair. COG1110 is also found in the top 10 list for CWMI, and COG1857 was discovered by CWMI five times. Overall, 55 of the top 100 COGs identified as having high agreement according to MI have a CWMI of 0. This means that the COG and phenotype share no more information than COG and taxonomic rank order. Given two hypotheses, the first (H_0) that either the COG and the phenotype have simply been conserved in the ancestral signal and share no other biological dependency, or the other (H_A) that the COG biologically influences the phenotype, the CWMI score indicates that there is no

Table 3. The top 10 COGs for thermophile at the taxonomic rank of order

COG	F^n	MI	CWMI	Description
(a) Ranked by MI.				
1756[a]	S	50	42	*Uncharacterized*
1110[a]	**L**	50	37	Reverse gyrase
2250[a]	S	50	30	*Sacsin-like chaperone*
1318	K	49	7	*Transcriptional regulator*
1353[a]	R	48	50	*HD superfamily hydrolase*
1980[a]	G	48	32	Fructose 1,6-bisphosphatase
1618	F	45	6	*Nucleotide kinase*
2248	R	37	19	*Metallo-β-lactamase*
1818	R	26	0	*RNA-binding protein*
1857	**L**	17	5	*DNA repair*
(b) Ranked by CWMI.				
1353[a]	R	48	50	*HD superfamily hydrolase*
1337	**L**	5	48	*DNA repair (RAMP)*
1336	**L**	15	47	*DNA repair (RAMP)*
1604	**L**	0	43	*DNA repair (RAMP)*
1756[a]	S	50	42	*Uncharacterized*
1583	**L**	4	42	*DNA repair (RAMP)*
1110[a]	**L**	50	37	Reverse gyrase
1980[a]	G	48	32	Fructose 1,6-bisphosphatase
2250[a]	S	50	30	*Sacsin-like chaperone*
2248[a]	R	37	19	*Metallo-β-lactamase*

The number for MI and CWMI represent the number of times, out of a possible 50 (5 folds x 10 replicates), that the COG was selected. [a]Common to lists (a) and (b); underlined COGs: predicted thermophile-specific by Makarova *et al.* (2002); italicized descriptions: uncertain function; NCBI COG function (F^n) codes: L: replication, recombination and repair; K: transcription; F/G: metabolism; S/R: poorly characterized.

reason to prefer H_A over H_0. It is certainly possible that COGs with a CWMI = 0 influence the phenotype, but we cannot deconvolute the effects of taxonomy and phenotype in such cases without further taxonomic sampling or experimentation.

Five of the CWMI-ranked COGs for thermophily are involved in DNA replication, recombination and repair. Four of the remaining five either have only general functional prediction, or the function has not yet been identified. Overall, 40% of the unique COG patterns in this dataset have this poorly characterized functional classification. Some of the genes found have previously been predicted as being part of a thermophile-specific DNA repair system through neighborhood, sequence and structural analysis (Makarova *et al.*, 2002; Table 3).

Interestingly, both thermophilic strains of *S.thermophilus* were classified correctly the majority of the time by CPAR, but almost never correctly by CWMI or MI. The three rules extracted by CPAR that pertain to these two organisms (and 14 other thermophilic organisms) are in Table 4. The individual COGs do not stand out as differentiating between thermophile and non-thermophiles, e.g. COG2189 has near zero MI/predictive value, but when taken with COG1337 and COG2723, they cover the thermophilic *Streptococci* and 8 other thermophiles, and no non-thermophiles. This illustrates the need to examine multi-element associative rules.

On the other hand, *C.thermocellum* ATCC 27405 and *Synechococcus* sp. JA-3Ab were classified incorrectly the majority of the time with CPAR and with MI, but with CWMI applied at the taxonomic rank of order, both are classified correctly the majority of the time. This could be due to the downweighting of COGs that are confounded with the taxonomic rank order, and thus the highlighting of COGs that span multiple taxonomic groups. For instance, these organisms both have COG1337 (noted by; Makarova *et al.*, 2002), which is highly ranked with CWMI (Table 3, Panel b) but not with MI.

Table 4. CPAR rules for the two thermophilic *Streptococci* and the total number of thermophiles covered for each

Rule				#
1337 [L]	**2723** [G]	**2189** [L]		10
0428 [P]	1847 [R]	0153 [G]	0425 [O]	9
3547 [L]	3270 [S]	1774 [S]		8

In braces are the NCBI function codes; in bold are those not found in the non-thermophilic (thermotolerant) *S.thermophilus*. NCBI COG function codes: L: replication, recombination and repair; K: transcription; G/P: metabolism; O: post-translational modification, protein turnover, chaperones; S/R: poorly characterized.

Even though the discovered COGs and the organisms that each classifies correctly are different (Supplementary Tables S2 and S3), overall, there is no difference in classification accuracy of the two measures ($N = 10$, paired $t = 0.46$, $P = 0.659$; Supplementary Tables S4 and S5). This means that after highly predictive (though taxonomically confounded) COGs have been downweighted, we have discovered a subset of COGs that are just as predictive of the given phenotype, and are relevant across taxonomic groups.

5 CONCLUSIONS

Predictive machine learning methods are being used to reverse engineer the genetic underpinnings of phenotypes, but mining the large feature space for interactions leading to traits is computationally demanding. We have demonstrated that our implementation of the predictive associative-rule mining algorithm, CPAR, outperforms a recent genotype–phenotype classifier in terms of accuracy and is on average 100 times faster.

It is widely known that attributes with high predictive power are not necessarily causal, and the uneven taxonomic distribution and sampling of genomes with various phenotypes can potentially generate spurious correlations. Our analysis using CWMI produced a remarkably different ranking of genes than did more-traditional ranking criteria such as MI and the known functional roles of many genes favored by CWMI indicate that phenotypically relevant genes are indeed being highlighted.

An advantage of our approach is that it is not tied to the COG methodology or phylogenetic profiles in general. Phenotypes such as temperature adaptation are conferred not only by the presence of certain genes, but also by adaptive mutations in existing genes that change their physical or chemical properties. For instance, the DNA polymerase from the thermophile *Thermus aquaticus* is an example of a protein that is stable at high temperatures (Chien *et al.*, 1976). However, this protein is homologous to polymerases from non-thermophiles and its presence would therefore not discriminate between thermophiles and non-thermophiles. By encoding variations in DNA or protein sequence in a matrix, we will be able to evaluate gene content and gene sequence variation in a combined framework, thereby gaining a much more comprehensive picture of the genetic adaptations that are necessary for different phenotypic traits.

Funding: Natural Sciences and Engineering Research Council of Canada (to N.J.M.); Genome Atlantic, the Canada Foundation for Innovation, and the Canada Research Chairs program (to R.G.B.).

Conflict of Interest: none declared.

REFERENCES

Agrawal,R. *et al*. (1993) Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on management of data*. ACM, New York, NY, pp. 207–216.

Baldi,P. *et al*. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.

Beiko,R.G. *et al*. (2005) Highways of gene sharing in prokaryotes. *Proc. Natl Acad. Sci. USA*, **102**, 14332–14337.

Carlson,J.M. *et al*. (2008) Phylogenetic dependency networks: inferring patterns of CTL escape and codon covariation in HIV-1 Gag. *PLoS Comput. Biol.*, **4**, e1000225.

Chang,C.-C. and Lin,C.-J. (2001) *LIBSVM: a Library for Support Vector Machines*. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm (last accessed date February 24, 2010).

Chien,A. *et al*. (1976) Deoxyribonucleic acid polymerase from the extreme thermophile thermus aquaticus. *J. Bacteriol.*, **127**, 1550–1557.

Cover,T.M. and Thomas,J.A. (2006) *Elements of Information Theory*, 2nd edn. Wiley-Interscience, New York, NY.

Enright,M.C. *et al*. (2002) The evolutionary history of methicillin-resistant staphylococcus aureus (MRSA). *Proc. Natl Acad. Sci. USA*, **99**, 7687–7692.

Fleuret,F. (2004) Fast binary feature selection with conditional mutual information. *J. Mach. Learn. Res.*, **5**, 1531–1555.

Gaasterland,T. and Ragan,M.A. (1998) Constructing multigenome views of whole microbial genomes. *Microb. Comp. Genomics*, **3**, 177–192.

Goh,C. *et al*. (2006) Integration of curated databases to identify genotype-phenotype associations. *BMC Genomics*, **7**, 257.

Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.

Guyon,I. *et al*. (2002) Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422.

Harvey,P.H. and Pagel,M.D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK.

Jensen,L.J. *et al*. (2008) eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.*, **36**, D250–D254.

Jensen,L.J. *et al*. (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Jim,K. *et al*. (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Res.*, **14**, 109–115.

Kastenmüller,G. *et al*. (2009) Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes. *Genome Biol.*, **10**, R28.

Levesque,M. *et al*. (2003) Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Curr. Biol.*, **13**, 129–133.

Liu,B. *et al*. (1998) Integrating classification and association rule mining. In *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, New York, NY, pp. 80–86.

Liu,Y. *et al*. (2006) An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. *PLoS Comput. Biol.*, **2**, e159.

Makarova,K.S. *et al*. (2002) A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.*, **30**, 482–496.

Markowitz,V.M. *et al*. (2008) The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.*, **36**, D528–D533.

Martin,M. *et al*. (2003) Comparing bacterial genomes through conservation profiles. *Genome Res.*, **13**, 991–998.

Nakagawa,S. *et al*. (2007) Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. *Proc. Natl Acad. Sci. USA*, **104**, 12146–12150.

Pellegrini,M. *et al*. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285.

Quinlan,J. and Cameron-Jones,R. (1993) FOIL: a midterm report. In *Proceedings of the 1993 European Conference on Machine Learning*, Vienna, Austria, pp. 3–20.

Raymond,J. *et al*. (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science*, **298**, 1616–1620.

Slonim,N. *et al*. (2006) Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Mol. Syst. Biol.*, **2**, 2006.0005.

Tamura,M. and D'haeseleer,P. (2008) Microbial genotype-phenotype mapping by class association rule mining. *Bioinformatics*, **24**, 1523–1529.

Tatusov,R. *et al*. (1997) A genomic perspective on protein families. *Science*, **278**, 631.

Tatusov,R. *et al*. (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

Wang,G. and Lochovsky,F.H. (2004) Feature selection with conditional mutual information maximin in text categorization. In *Proceedings of the thirteenth ACM international conference on information and knowledge management*. Association for Computing Machinery, New York, NY, pp. 342–349.

Wang,Y. *et al*. (2005) Gene selection from microarray data for cancer classification-a machine learning approach. *Comput. Biol. Chem.*, **29**, 37–46.

Yin,X. and Han,J. (2003) CPAR: Classification based on predictive association rules. In *Proceedings of the Third SIAM International Conference on Data Mining*. San Francisco, CA.