BIOINFORMATICS

Vol. 19 no. 1 2003
Pages 125–135

Whole-proteome interaction mining

*Joel R. Bock and David A. Gough**

Department of Bioengineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA

Received on Augutst 23, 2001; revised on March 11, 2002; accepted on June 27, 2002

ABSTRACT

Motivation: A major post-genomic scientific and technological pursuit is to describe the functions performed by the proteins encoded by the genome. One strategy is to first identify the protein–protein interactions in a proteome, then determine pathways and overall structure relating these interactions, and finally to statistically infer functional roles of individual proteins. Although huge amounts of genomic data are at hand, current experimental protein interaction assays must overcome technical problems to scale-up for high-throughput analysis. In the meantime, bioinformatics approaches may help bridge the information gap required for inference of protein function. In this paper, a previously described data mining approach to prediction of protein–protein interactions (Bock and Gough, 2001, *Bioinformatics*, **17**, 455–460) is extended to interaction mining on a proteome-wide scale. An algorithm (the *phylogenetic bootstrap*) is introduced, which suggests traversal of a phenogram, interleaving rounds of computation and experiment, to develop a knowledge base of protein interactions in genetically-similar organisms.

Results: The interaction mining approach was demonstrated by building a learning system based on $1,039$ experimentally validated protein–protein interactions in the human gastric bacterium *Helicobacter pylori*. An estimate of the generalization performance of the classifier was derived from 10-fold cross-validation, which indicated expected upper bounds on precision of 80% and sensitivity of 69% when applied to related organisms. One such organism is the enteric pathogen *Campylobacter jejuni*, in which comprehensive machine learning prediction of all possible pairwise protein–protein interactions was performed. The resulting network of interactions shares an average protein connectivity characteristic in common with previous investigations reported in the literature, offering strong evidence supporting the biological feasibility of the hypothesized map. For inferences about complete proteomes in which the number of pairwise non-interactions is expected to be much larger than the number of actual interactions, we anticipate that the sensitivity will remain the same but precision may decrease. We present specific biological examples of two subnetworks of protein–protein interactions in *C. jejuni* resulting from the application of this approach, including elements of a two-component signal transduction systems for thermoregulation, and a ferritin uptake network.

Contact: dgough@bioeng.ucsd.edu

1 INTRODUCTION

The recent publication of the Human Genome Working Draft Sequence (Lander *et al.*, 2001; Venter *et al.*, 2001) is an unequivocal landmark in the advancement of biological knowledge. However, even a completely-sequenced genome presents only a coarse specification for an organism's proteomic complement, and cannot provide understanding of biological function. A major post-genomic scientific and technological pursuit is to describe the exceedingly diverse functions performed by the proteins encoded by the genome. Within the cell, proteins assemble into complex and dynamic macromolecular structures, recognize and degrade foreign molecules, regulate metabolic pathways, control DNA replication and progression through the cell cycle, synthesize other chemical species (Alberts *et al.*, 1989), facilitate molecular recognition, localize and 'scaffold' other proteins within signal transduction cascades (Pawson and Scott, 1997), and participate in other important functions.

To appreciate the role of protein function, a description of protein–protein interactions is a necessary first step. After identifying the proteomic constituents, a rational research strategy should then proceed in the direction of information flow represented by Kanehisa (2000)

$$\text{Interaction} \rightarrow \text{Network} \rightarrow \text{Function}$$

The combinatorial expansion of information advancing along this pathway is enormous. Given the volume of proteomic data generated by high-throughput technologies (Uetz and Hughes, 2000), description of protein function must rely on the integration of empirical data with bioinformatic comparative and predictive analyses.

The workhorse of experimental proteomics has been the two-hybrid screen (Fields and Song, 1989). Although criticized based on the accuracy of results and its labor-

*To whom correspondence should be addressed.

intensive nature (Enright *et al.*, 1999; Ito *et al.*, 2001), it presently stands as the most viable technique for large-scale characterization of protein interactions in complete genomes (LeGrain and Selig, 2000). Protein chips may eventually provide large-scale simultaneous protein–protein interaction data (MacBeath and Schreiber, 2000), but technical problems (denaturing, substrate biocompatibility) must be overcome to scale-up for high-throughput analysis. Other approaches will undoubtedly become prominent as proteomics technology continues to evolve. A review of technological advances on this front can be found in Mann *et al.* (2001).

In the meantime, bioinformatics approaches may help bridge the information gap required for inference of protein function.

1.1 Bioinformatic approaches to protein–protein interactions

A number of different strategies have been proposed, including network inference based on a reference map of interacting domain profile pairs (Wojcik and Schächter, 2001), conserved gene-pairs and correlated prokaryotic interacting gene products (Dandekar *et al.*, 1998), clusters of orthologous proteins (Tatusov *et al.*, 1997), phylogenetic profile (Pellegrini, 2001) or tree similarity (Pazos and Valencia, 2001), gene fusion events (Marcotte *et al.*, 1999), location within a functional cluster map (Schwikowski *et al.*, 2000), and others. Because investigators concentrate on different organisms, or reporting is confined to partial hypothesized interaction results, it is difficult to compare the predictive power of these various computational methods on an objective basis.

We previously reported a data mining technique (Bock and Gough, 2001) wherein a Support Vector Machine (SVM) learning system was trained on a limited, heterogeneous data set to recognize and predict protein interactions based solely on primary structure and associated physicochemical properties. Testing against previously unseen test samples, the system predictive accuracy exceeded 80% over the ensemble of statistical experiments. It was argued that such a system might be used as a screening method to focus experimental assessment of protein interactions. The remarkable success of the methodology reported in Bock and Gough (2001) has provided motivation for the present work, which is more ambitious in scope. Our present objective is to expand the range of prediction to whole-proteome 'interaction mining' using computational statistical learning theory.

Interaction mining uses analogy between the proteomes of two closely related organisms to predict protein–protein interactions. A 'template' or design organism provides a network of experimentally derived interactions, and this pattern is used to infer the structure of an interaction network in a related organism.[†] Given a list of experimental interactions, all that is required to infer the proteome-wide interaction map are the amino acid sequences of the target organism. We refer to this approach as 'interaction mining', in association with the concept of data mining, which concentrates on the application of specific algorithms for extracting structure from data (Bradley *et al.*, 1998).

To demonstrate the approach, we trained a learning system to recognize correlated patterns of primary structure within protein interaction pairs taken from the human gastric bacterium *Helicobacter pylori*, associated with peptic ulcers. A compendium of over 1, 200 *H. pylori* interactions were recently reported (Rain *et al.*, 2001). *Helicobacter pylori* interaction data are used to train the system, and to estimate the standard error of its generalization capability. Primary structure data from a close phylogenetic neighbor within the Bacteria Kingdom, *Campylobacter jejuni*, comprise the prediction data set. *C. jejuni* is an enteric pathogen causing common symptoms of food poisoning. Its infection is a precursor to a form of neuromuscular paralysis known as Guillain–Barre syndrome (Parkhill *et al.*, 2000). Both *H. pylori* and *C. jejuni* are microaerophilic, gram-negative, flagellate, spiral bacteria. These orthologous bacteria represent model systems for demonstration of the proteome-wide interaction mining approach.

2 SYSTEM AND METHODS

The Support Vector Machine (Vapnik, 1995; Burges, 1998) can be trained to classify labeled empirical data points by constructing an optimal high-dimensional decision surface that simultaneously maximizes the separation between data classes, and minimizes the 'structural risk'

$$R(\alpha) = \int_Z Q(z, \alpha) \mathrm{d}F(z), \ \alpha \in \Lambda \qquad (1)$$

with respect to parameters α using an independent, identically distributed (i.i.d.) sample $Z = \{z_1, z_2, \ldots, z_l\}$ generated by an (unknown) underlying probability distribution F, where Q is an indicator function, and Λ is a set of parameters.

The sample points $z_i = (x_i, y_i)$ comprise protein features $x_i \in \mathbb{R}^n$ and their classifications $y_i \in \{-1, +1\}$. In practice, the learning task converges rapidly as a constrained quadratic programming is solved. The resultant decision function h represents an hypothesis generator for inference on novel data points, mapping them onto the

[†] After the original submission of this manuscript, the authors were made aware of conceptually similar work reported in Wojcik and Schächter (2001). In that investigation, a reference map of interacting protein domains was combined with sequence similarity and clustering analysis to predict a new interaction map in another organism.

discrete set y, or $h : x \rightarrow y$. This is a binary decision ($+1 \Rightarrow$ interaction, $-1 \Rightarrow$ nointeraction).

2.1 Phylogenetic bootstrap

Building on previous work (Bock and Gough, 2001), we propose that the support vector machine-learning approach may be used to extrapolate from a protein interaction map in one organism to a complete map in a related organism. Let us establish a framework for prediction of whole-proteome interaction maps. The assumption in Equation (1) of a fixed generative probability distribution $F(Z)$ is a key issue in the design of this data mining application. A direct consequence of this assumption is that a decision function h, developed from a training sample Z_a taken from species S_a, may be used to predict protein–protein interactions on a sample Z_b *from another species* S_b, provided that features of their respective protoeomes are not too dissimilar in some sense, or

$$\rho\,(F(Z_a), F(Z_b)) \leq \delta \tag{2}$$

where ρ is a measure of distance between its arguments, and δ is a constant. The statistic ρ is general, and may be taken to signify cross-species similarity based on genome-level 'edit distance' (Sankoff *et al.*, 1992), whole-proteomic content (Tekaia *et al.*, 1999), or proximity within phylogenies constructed from multi-domain orthologous protein sequences (Brown *et al.*, 2001), to cite only three of many possibilities. For this discussion, it is assumed that δ varies as $0 \leq \delta < \infty$, where $\delta = 0$ is a proteome's self-distance, and extreme mutual divergence between two organisms is expressed in the limit as $\delta \rightarrow \infty$.

We introduce here the *phylogenetic bootstrap* algorithm. Bootstrap methods in applied statistical inference are numerical techniques for estimating the standard error of arbitrary test statistics (Efron and Gong, 1983). The phylogenetic bootstrap for protein–protein interaction mining does not compute a statistic *per se*, but suggests a method for incrementally 'walking' laterally across a phenogram, interleaving rounds of computation and experiment, to develop a knowledge base of protein–protein interactions in genetically related organisms. Using the hypothesis $h : x \rightarrow y$ (based on an assumed common probability distribution $F(Z)$), we infer the interactions within a sample taken from a distinct, evolutionarily similar proteome. These predictions are a function of the generalization confidence level derived from 10-fold cross-validation error estimation (Stone, 1974). The probability of correctness of a novel prediction may be estimated by

$$\Pr\{\hat{y} = y \mid h\} = g(\delta)(1 - \epsilon_{cv}) \tag{3}$$

where $\hat{y}$ is the predicted interaction for a putative interacting protein pair, y is the true state of nature, ϵ_{cv} is the cross-validation error rate, and $g(\delta)$ is a decreasing function of the interproteomic distance (Equation (2)). A simple plausible (and conservative) form for the function g is an exponential

$$g(\delta) = \mathrm{e}^{-\lambda\delta} \tag{4}$$

where λ is the rate of decay. Substituting this function in Equation (3), the prediction confidence becomes

$$\Pr\{\hat{y} = y \mid h\} = \mathrm{e}^{-\lambda\delta}(1-\epsilon_{cv}), \quad \lambda > 0,\ \delta \in [0, \infty) \tag{5}$$

Note that this representation is schematic. The value of the decay parameter λ and calibration of the distance in Equation (2) can only be determined after experimental validation of the numerical predictions.

Upon completion of this process, predicted protein–protein interactions in the novel organism may be used to design successive genetic or biochemical experiments. The results of these selected experiments are fed-back to refine the current model, and flesh out empirical protein interactions within the new proteome. This iterative process may continue as long as certain criteria on acceptable estimated prediction error rate and proteome similarity remain satisfied. The steps comprising the phylogenetic bootstrap as proposed in this investigation may be distilled into an algorithm, described in Section 3.

2.2 Generalization potential

We estimate the expected value of the error rate of the classifier $h(\alpha, x)$ using k-fold cross-validation on the training sample Z_a. Here, we take $k = 10$, producing a 10-fold cross-validation prediction error estimate. The expected generalization error is taken as the average of the classification error observed on each of the k data folds. Averaging reduces the variance of this estimate (Perrone, 1993). The prediction error derived from 10-fold cross-validation is known to have low bias, and precision approximating that of leave-one-out error estimation, at lower computational cost (Martin and Hirschberg, 1996).

In this procedure, an SVM decision rule $h(\alpha, x)$ is constructed k times, each time training on a different set of example data points $\{Z_m \mid Z_m \subset Z_a, m \in 1, \ldots, (k-1)\}$, and testing prediction accuracy on the omitted set $\{Z_n \mid Z_n \subset Z_a, n \neq m\}$, where $Z_m \bigcup Z_n = Z_a$. The number of prediction errors for each model is accumulated, and the k-averaged expected value of the individual data sets' inferred classifiers is taken as the system error rate estimate ϵ_{cv}. Note that the statistic ϵ_{cv} is an estimate of the expected prediction error rate, and is itself a random function of population, the sample taken from that population, and the inference method. (Martin and Hirschberg, 1996).

'Prediction accuracy' as used here means that a correct declaration is made by the decision rule, or $\hat{y} = y \mid h$. This can represent either a positive or a negative predicted

protein interaction. If the cross-validation error rate is expressed as a fraction assuming values $0 \leq \epsilon_{cv} \leq 1.0$, the confidence level expected for predictions of putative protein–protein interactions is given by the probability expression of Equations (3)–(5).

3 ALGORITHM

The phylogenetic bootstrap algorithm is summarized in this section.

(1) *Input*. First, it is necessary to specify the species S_a, S_b subject to investigation. In general, some existing protein interaction data may be at hand for each proteome, although their relative cardinality may be quite skewed. Our line of thought assumes that no interaction data are available for S_b; we have only a set of labels $\{Y_a\}$ corresponding to experimentally verified interactions sampled from the proteome of species S_a. These labels, along with the amino acid sequence sets $\{s_a\}$ and $\{s_b\}$ comprising the species respective proteomes, are inputs to the algorithm. Other inputs required are the inter-proteome distance δ (Equation (2)), and the maximum acceptable rate of generalization error, $\epsilon_{cv}^{\max}$, where $0 < \epsilon_{cv}^{\max} < 0.5$.

(2) *Construct features from training sample*, based on attributes of the primary structure sequences s_a from the training data set. Encoded attributes X_a for entire proteomes may be derived from tabulated residue properties including charge, hydrophobicity, and surface tension as described previously (Bock and Gough, 2001). At this stage, data preprocessing including normalization and filtering should be performed to produce a useful sampled attribute set $\{x|x \in \mathbb{R}^n, x \subset X\}$. A total of l data points z are constructed by adding labels y to the accepted feature vectors $\{x\}$, or $z_i = (x_i, y_i), i = 1, \ldots l$. The union of positively- and negatively-labeled examples constitutes the training sample $\{Z_a\}$.

(3) *Compute decision rule*. Design an optimal support vector machine to classify data points in the sample $\{Z_a\}$. After learning, the system builds a decision rule h that maps input data vectors x_i onto the classification space $y_i \in [+1, -1]$. The numerical sign of y_i is interpreted as the likelihood that the two proteins represented by x_i will interact.

(4) *Estimate CV error*. Perform k-fold cross-validation experiments on the training set. Segregate the observations $\{z^k\}$ within each data fold k, and train a different SVM using data $\{z^m\}$ from each of the $k-1$ disjoint data folds $\{z^m | z^m \in Z_a, m \neq k\}$. Predict the class membership of the omitted points $\{z^k\}$. Accumulate the total number of misclassifications observed in this process. Take the final k-fold average cross-validation error as the estimated expectation of generalization error rate ϵ_{cv} of the learner h. The magnitude of this error estimate in practice will be extended by some function of interproteomic distance, say $g(\delta)$.

(5) *Construct features from novel sample*. Construct features $\{X_b\}$ from sequences $\{s_b\}$ for the unlabeled proteome S_b. All-vs-all pairwise interactions may be represented in the prediction set. The same data preparation process should be applied as carried out in Step 1.

(6) *Predict novel interaction network*. Predict a new network of protein–protein interactions $\{\hat{Y}_b\}$ via the trained system $h(\alpha) : x_b \rightarrow Y_b$, where α are parameters of the model. To the extent that the assumption of proteomic similarity $\rho\,(F(Z_a), F(Z_b)) < \delta$ is satisfied, each point estimate is expected to be accurate with a probability $g(\delta)(1 - \epsilon_{cv})$, or $\Pr\{\hat{y} = y \mid h\} = g(\delta)(1 - \epsilon_{cv})$.

(7) *Validate sample experimentally*. Take a random sample from the protein interaction prediction set $Z_b = \{(x, \hat{y})|x \subset X_b, \hat{y} \subset Yb\}$ and verify the predicted protein interactions (both positive and negative) using experimental proteomics techniques. Compare the experimentally observed and calculated estimated prediction error rates. Assert that the following statement holds true: $\epsilon_{cv}^{\mathrm{v}} \leq \epsilon_{cv} < \epsilon_{cv}^{\max}$, where the superscript v denotes validation by biological experiment.

(8) *Input*. Select sequences $\{s_c\}$ from a new, related organism $\{S_c\}$. The similarity assumption $\rho\,(F(Z_a), F(Z_b)) < \delta$ must still be maintained.

(9) *Update training sample*. Add sequences from the validated prediction set to the training set, and consider this expanded set as the training set for the next iteration: $\{s_a\} = \{s_a\} + \{s_b\}$. Update the class labels by adding the prediction label set $\{Y_a\} = \{Y_a\} + \{\hat{Y}_b\}$. Protein interactions for organism $\{S_c\}$ will now be computed.

(10) *Iterate*. Return to Step 1 and repeat the process. The stopping condition for this iteration is violation at any time of the assertions regarding the generalization error rate, i.e. when the error rate from cross-validation, ϵ_{cv}, exceeds the specified limit $\epsilon_{cv}^{\max}$, or when the experimental observations contain more frequent errors than the calculated rate, or $\epsilon_{cv}^{v} \geq \epsilon_{cv}$.

4 IMPLEMENTATION

4.1 Primary structure features

Our objective is to gain insight into protein interactions, if possible using strictly amino acid sequence information.

To teach a learning machine, it is necessary to portray salient aspects of the data (the 'features') that intuition or hypotheses suggest will contribute to effective learning of the concept. The problem of feature selection is to define descriptors which discriminate between two classes of data, while inhibiting the irrelevant and redundant features (Mangasarian, 1996).

Here, we sought to find the interacting protein pairs within a complete proteome, for which experimental data representing a negligible percentage of the total possible pairwise interactions are available. We built feature vectors for SVM training as described previously (Bock and Gough, 2001), using native proteins directly sampled from the proteome of *Helicobacter pylori*. The protein interaction data were obtained from the online resource as described in Section 2. Construction of the negative examples was carried out following Assumption 2 (see Section 4.3), which maintains that any pair of proteins not labeled as mutually interacting in the design sample Z are assumed to not interact. This represents another strong assumption: we assume that the *H. pylori* design sample reported in Rain *et al.* (2001) is complete in the sense that all possible protein–protein interactions comprising the proteome were discovered. Non-interacting protein pairs are designated as negative interactions. In the absence of further information, we must make this assumption, cognizant that by labeling the sample in this manner we may inadvertently commit a logical fallacy of *argumentum ad ignorantiam* (argument from ignorance).

4.2 Proteome data quality control

Protein interaction examples are filtered to ensure high-quality representation in the learning machine. In Step 1 of the phylogenetic bootstrap algorithm (cf. Section 3), data preprocessing is performed. This preprocessing typically includes (1) scaling the feature vectors to equalize relative numerical magnitudes of the disparate features, and may be followed by (2) curation based on predefined criteria or prior knowledge impacting confidence in the data set. Scaling techniques are well-documented in the machine learning literature, and will not be further discussed here (a succint summary for applications can be found in Swingler (1996)).

With regard to the second cited aspect of preprocessing, we selected only positive samples for *H. pylori* interactions where the estimated probability that the observed interaction was found purely by chance (as a two-hybrid artifact) was at most $1.0\mathrm{E}-6$. In this case the originators of the data set assigned degrees of confidence to the various interactions comprising the sample, according to a model of competition for bait-binding between prey fragments (Rain *et al.*, 2001).

Commonly, a large percentage of the open reading frames (ORFs) in a given genome remain experimentally unobserved, and if sequential homology to a protein of known function is not discovered, these proteins are labeled as 'hypothetical'. The machine learning investigator might be tempted to consider excluding such sequences from the design sample. An overriding argument against such action is the recognition of the fundamental objective of assigning functional roles to the so-called 'hypothetical' protein sequences. Consequently, a concession must be made to incorporate possible numerical artifacts, learned from experimental data which may be fraught with false positive and false negative interaction data. As structural proteomics continues to fill in the gaps in our knowledge in the future, these hypothetical proteins will eventually be confirmed or invalidated experimentally.

4.3 Assumptions

Interaction mining analysis makes certain assumptions about the distributions of proteomic data in the design sample Z (recall discussions in the context of Equation (2)). Other assumptions inherent in this approach include Bock and Gough (2002):

4.3.1 Static intracellular state. If proteins A and B interact in the design species S_d, they will also interact if co-occurring in a novel species S_n. This assumption may not be generally valid for where physiological conditions present in S_n differ relative to S_d.

4.3.2 Coverage of design sample. Any pair of proteins (A, B) not labeled as interactors in the design sample Z are assumed to not interact. This is a subtle but significant point that must be held in mind when interpreting prediction results.

4.3.3 Physical proximity. The all-vs.-all interacting mining technique selects interaction pairs based on correlated patterns of primary structure, and does not discriminate protein subcellular location. In particular cases, additional information regarding subcellular location might offer insight regarding prediction practicability. Such analysis could be done in a separate post-mining filtering step.

4.3.4 Simple interactions. Only binary interactions are represented; complexes of proteins with more than two components are only inferred indirectly in post-mining analysis. Dynamic multiprotein complexes (Gavin *et al.*, 2002) are not directly resolved (but, may be inferred after the fact, with details of each component protein's interaction surface characteristics (Finley and Brent, 1994)). Also, pairwise interactions predicated upon modifications to protein A (e.g. phosphorylation, glycosylation, proteolytic cleavage) prerequisite to its recognition by B are excluded from the prediction space.

5 DISCUSSION

For the design organism *Helicobacter pylori* strain 26695, a total of 1039 protein interactions were selected for analysis. Interactions were identified from the database provided online at http://pim.hybrigenics.com. From the nominal *H. pylori* proteomic complement of $N = 1555$ sequences, a sample of 1039 non-interacting sequences was selected according to the various data filtering procedures described in Section 4, and following the assumption of comprehensive coverage in the positive design sample (Section 4.3). This created a balanced representation of each data class to train the learning system, the total sample length being $l = 2078$ observations. Each sample point $z_i = (x_i, y_i), i = 1, \ldots, l$ was constructed from primary structure features $x_i \in \mathbb{R}^n$ and their interaction class labels $y_i \in \{-1, +1\}$ (see Section 2).

5.1 Cross-validation results from *H. pylori*

The learning machine generates an interaction hypothesis $\hat{y}$ for each data point x via the computed decision surface $h : x \rightarrow y$. Define the null hypothesis H_0 to mean that no interaction is present between a pair of proteins, or $H_0 : y \mid x = -1$. The alternative hypothesis is $H_A : y \mid x = +1$. There are two types of statistical errors that may occur on each decision $\hat{y}$. (1) If H_0 is true and is rejected ($\hat{y} = +1, y = -1$), the machine commits a Type I error, or 'false positive' decision. (2) If H_0 is false (interaction present) and is not rejected ($\hat{y} = -1, y = +1$), a Type II, or 'false negative' error, is made.

The 10-fold cross-validation prediction error estimates obtained on the design sample are presented in Table 1. Results are shown for three conventional statistical instruments used to evaluate the performance of classifiers in machine learning applications. These include the *sensitivity*, *precision* and *accuracy* (Kohavi and Provost, 1998). Sensitivity is calculated as $S = TP/(TP + FN)$, where $TP =$ number of true positive interaction decisions, and $FN =$ number of Type II errors. Precision is computed as $P = TP/(TP + FP)$, where FP is the number of Type I errors made by the system. Accuracy expresses the overall correctness rate of the system, and is computed as $A = (TP + TN)/(TP + TN + FP + FN)$. Here, TN represents the number of true negative classifications.

The cross-validation measurements summarized in Table 1 are comparable to previously published predictive results (Bock and Gough, 2001). On average, three of four SVM predictions were correct when applied to the unseen data partition. The precision was 80%, suggesting a strong level of confidence in positive interactions detected by the system. Precision expresses the rate of Type I error suppression. Sensitivity observed in cross-validation was 69%, which indicates the true positive rate of the decision function.

Table 1. 10-fold cross-validation performance estimate derived from classifiers trained on examples from the design organism *H. pylori*

Precision	Sensitivity	Accuracy
80.2	68.6	75.8

High precision indicates the suppression of Type I (false positive) errors. High sensitivity means that Type II errors are suppressed by the decision function (i.e. low false negative rate). Numbers are expressed as percentages. Data sample size $N = 1880$.

Recalling Equations (3)–(5), the expected precision of the classifier's performance in the novel organism will be less than 80%. The actual performance decrement cannot be evaluated until biological experiments validate or invalidate the testable hypotheses comprising the network of interactions. At present we can only estimate the upper bound on the precision of this set of generated hypotheses.

For inferences about complete protemes in which the number pairwise non-interactions is expected to be much large than the number of actual interactions, we anticipate that the sensitivity will remain the same but the precision may decrease.

5.2 *C. jejuni* interaction hypotheses

The level of estimated generalization obtained from leave-one-out analysis of the *H. pylori* proteome supports confidence in the prediction of protein–protein interactions in *Campylobacter jejuni*. *C. jejuni* and *H. pylori* are close phylogenetic relatives (see, e.g. Figure 1 in Eisen (2000)), displaying highly-similar constitutent protein domains [‡] and genomic content (Tekaia *et al.*, 1999, Figure 2). The *C. jejuni* proteome contains 1613 proteins, of which all possible unique pairwise protein–protein interactions (1 300 078 pairs) were encoded as features and added to the sample X_b for interaction mining. Using one of the 10 classifiers $h(\alpha, x)$ developed during cross-validation analysis on the design organism, an interaction hypothesis was generated for each data point in this sample. A total of 5367 distinct protein–protein interactions were declared by the decision function. Each protein comprising the *C. jejuni* interaction map was predicted to have, on average, biological connections with 3.33 other proteins.

By way of discussion of the predicted *C. jejuni* protein interaction network, we first discuss general scaling properties of the map, comparing these to investigations appearing in the literature. Secondly, some specific biological examples produced by the interaction mining procedure will be examined in greater detail.

[‡] Source: EBI Proteome Analysis Database http://www.ebi.ac.uk/proteome/comparisons.html.

5.3 Scaling properties of map

Objects in nature which are invariant with respect to certain transformations are said to *scale* (Mandelbrot, 1977). We observed here that the inferred *C. jejuni* protein–protein interaction map shares a key topological scaling property in common with previous proteome-wide investigations: the average connectivity of the interaction network. The agreement between the present results and the cited works, which represent a variety of investigations on different organisms, offers strong evidence supporting the biologically feasibility of the hypothesized map. Another scaling property, namely the distribution of sizes of 'clusters' of binary protein–protein interactions, varied significantly between the present investigation and a previous study (Jeong *et al.*, 2001).

5.3.1 Network connectivity. A basic, large-scale architectural statistic describing a protein interaction map is the average number of connections between a given protein and other proteins in the map. Let us call this the 'average connectivity' of the map. Table 2 lists data collected from several different proteome-scale investigations on different organisms. It can be seen that on average, 3.33 proteins are linked to each protein in the *C. jejuni* interaction map. This level of connectivity compares favorably to the other investigations cited in the table, especially to the experimental data from (Rain *et al.*, 2001), which provided the design sample for training the learning system in the present investigation.

Table 2 contains a column entitled 'Proteome coverage', defined here as the estimated number of distinct proteins involved in interactions as a fraction of either the total proteomic complement or assay depth for a given organism. Note that the inferred network of interactions in this investigation has full coverage, that is, each protein is expected to participate in at least one biological interaction. Although this level of coverage is higher when compared to estimates made from other investigations in the table, a recent investigation focused on elucidating multiprotein complexes in *S. cerevisiae* indicates higher connectivity densities (0.78) than previously observed (Gavin *et al.*, 2002).

5.3.2 Cluster size distribution. In Jeong *et al.* (2001), it is argued that the most highly-connected proteins within a cell are also the most critical for its survival. In studies involving the protein interaction network of *Saccharomyces cerevisiae*, they derived scaling laws describing the distribution of numbers of connections between proteins in the network. Power-law scaling characteristics were found common to both *S. cerevisiae* and *H. pylori*, indicating the possibility of a universal large-scale structure in biological networks.

Table 2. Comparision of proteome-wide interaction map connectivities for different organisms found in the literature

References	Organism	Method	Proteomic coverage	Average connectivity
1	*S. cerevisiae*	Experiment	0.55	1.388
2	*S. cerevisiae*	Experiment	0.26	1.523
3	*E. coli*	Prediction	0.10	2.14
4	*C. jejuni*	Prediction	1.00	**3.33**
5	*H. pylori*	Experiment	0.47	3.36
6,7	*S. cerevisiae*	Experiment	0.17	$3.2, 4.5 - 5.8$
8	*C. elegans*	Experiment	??	5.4

'Proteome coverage' is the estimated number of distinct proteins involved in interactions as a fraction of either the total proteomic complement or assay depth for a given organism. 'Average connectivity' refers to the average number of interaction partners per protein comprising the map. References: 1. (Ito *et al.*, 2001); 2. (Schwikowski *et al.*, 2000); 3. (Wojcik and Schächter, 2001); 4. Present investigation; 5. (Rain *et al.*, 2001); 6. (Uetz *et al.*, 2000); 7. (Tucker *et al.*, 2001); 8. (Walhout *et al.*, 2000a). Note: in Tucker *et al.* (2001), a retrospective reanalysis of data originally reported in Uetz *et al.* (2000) resulted in an updated estimated average connectivity of $4.5 - 5.8$ for *S. cerevisiae*

In that investigation, network architectural details for *S. cerevisiae* showed that the largest and smallest clusters of connected proteins constituted 0.7 and 93% of the total number of proteins comprising the map, respectively. A large interaction cluster was defined as one with > 15 links, while small clusters had ≤ 5 binary connections to other proteins. In the present investigation, we found similar connectivity distribution properties in the predictions for *C. jejuni* only for the largest clusters, i.e. those where $n > 15$ partners per protein mode were predicted. The inferred map has a much larger distribution of small- to medium-sized clusters by comparison, as summarized in Table 3. One explanation for this variance might be represented in arguments put forth in Hasty and Collins (2001), where it is noted that the power-law cluster size distribution is characteristic of networks in a state of transitory expansion. It follows that protein interaction network connectivity is a dynamic feature; different connection properties would be expected at different states in an organisms' evolution.

5.4 Selected biological examples

In this section, we present specific biological examples of protein–protein interactions predicted for *C. jejuni*, exemplifying the type of information that may be extracted from the application of this approach. This represents only a sampling of the subnetworks automatically generated by the interaction mining procedure.

5.4.1 Thermoregulation. Two-component signal transduction systems are essential in the regulation of many bacterial functions, including chemotaxis,

Table 3. Distribution of protein interaction cluster sizes compared to Jeong *et al.* (2001)

Ref.	Large clusters %	Medium clusters %	Small clusters %
1	0.7	6.3	93
2	1.054	38.0	60.9

A cluster size represents the average number of interactions (edges) each protein (node) shares with other proteins. 'Large' clusters refer to instances of proteins with a large number of partners ($n > 15$); 'medium' cluster nodes have $5 < n \leq 15$, and in 'small' clusters each protein has, on average, $n \leq 5$ connections to other proteins. Numbers are expressed as percentage of total number of proteins comprising the map. References: 1. Jeong *et al.* (2001); 2. Present investigation

metabolism, and the response to environmental stress. The two-component mechanism constitutes a membrane environmental sensor and a cytoplasmic regulator. This mechanism typically involves autophosphorylation of histidine residues on the sensor protein, which then acts as a kinase for the regulator, the phosphorylation of which induces transcriptional activation appropriate to the chemical or thermal stimulus (Klumpp and Krieglstein, 2002).

Elements of a hypothesized a two-component thermoregulation signalling pathway in *C. jejuni* are presented in Figure 1 and Table 4. The figure displays only a subnetwork of interactions comprising the primary interaction partners of the sensor and regulator proteins. Each protein node is labeled by its corresponding ORF designation. The two-component sensor (Q9PN36) is functionally linked to the putative heat-shock regulator (Q9PN67) via an intermediary protein Q9PMG7. Heat-shock proteins are known to solubilize misfolded or denatured proteins in case of extreme thermal insult to the cell (Alberts *et al.*, 1989).

The intermediate protein Q9PMG7 is designated as 'hypothetical', meaning it has sequential similarity to other proteins of unknown function. This 180-residue protein contains two possible sites for phosphorylation (casein kinase II, tyrosine) as detected by PROSITE search (Bairoch *et al.*, 1997). It is a feasible hypothesis that this previously uncharacterized protein may play a role in transferral of the message from sensor to regulator in the *C. jejuni* thermoregulation signalling pathway.

If elements of this inferred pathway are validated in wet-biological studies, we suggest the possibility of its manipulation or obstruction using antibiotic agents. As recently noted, targeted inhibition of histidine kinase signal transduction pathways in bacteria may have beneficial effects for host mammals, in which cellular signal transduction proceeds according to a different mechanism (Matsushita and Janda, 2002).

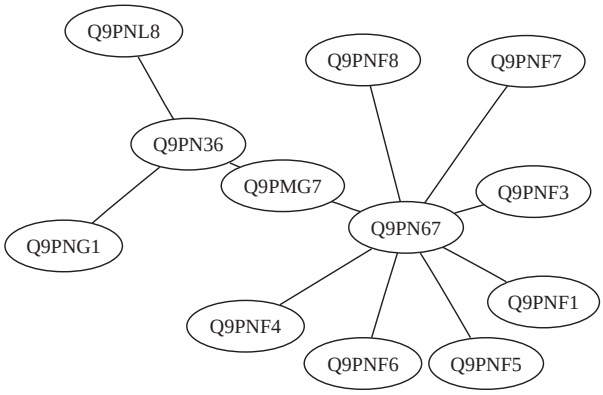

Fig. 1. Principal components of an hypothesized two-component thermoregulation signalling pathway in *C. jejuni*. Shown is a subnetwork of interactions comprising the primary interaction partners of the sensor (Q9PN36) and regulator (Q9PN67) proteins. Each protein node is labeled by its corresponding ORF designation. The previously uncharacterized protein Q9PMG7 may play a role in transferral of the message from sensor to regulator in the thermoregulation signalling pathway.

Table 4. Principal components of an hypothesized a two-component thermoregulation signalling pathway in *C. jejuni*

ORF	Status	Annotation	Partners
Q9PN36	A	Two-component sensor	Q9PNL8,Q9PNG1,Q9PMG7
Q9PN67	P	Heat shock regulator	Q9PMG7,Q9PNF8,Q9PNF7, Q9PNF3,Q9PNF1,Q9PNF5, Q9PNF6,Q9PNF4
Q9PMG7	H	Protein Cj1495c	Q9PN36,Q9PN67

'Status' refers to the functional annotation status of the ORF, with H = hypothetical, P = putative, A = annotated

5.4.2 Ferric uptake and regulation. The storage and regulation of iron levels is a fundamental aspect of cellular survival for Gram-negative bacteria. Iron is a non-abundant essential nutrient that is toxic in excessive concentrations, necessitating its regulation within the cell. In *C. jejuni*, ferritins (iron-storage proteins) are also involved in oxidative stress resistance (Andrews, 1998).

A subnetwork of putative protein interactions integral to ferric uptake and regulation processes is shown in Figure 2. This interaction group comprises proteins linking the extracellular signal (Q9PJA5, putative integral membrane protein) to the regulatory (P48796, ferric uptake regulation) and transcriptional machinery (Q9PNK3, leucyl-tRNA transferase; Q9PN44, polyribonucleotide nucleotidyltransferase) within the cell. Such a connection is required to respond to dynamically changing requirements for iron storage or removal. Q9PNK3 is predicted to interact with Q9PMS3, a putative ferredoxin that may play a role in the intracellular redox system.

Table 5. Principal components of an hypothesized ferric uptake regulation pathway in *C. jejuni*

ORF	Status	Annotation	Partners
P48796	A	Ferric uptake regulation protein	Q9PNK3,Q9PNK2,Q9PNK1,Q9PNG1,Q9PMG7
Q9PNK3	A	Leucyl-tRNA synthetase	Q9PMS3,Q9PN43,Q9PMS4,Q9PN44,Q9PJA5
Q9PMD5	A	Possible bacterioferritin	Q9PI17,Q9PHR6,Q0ZI13,Q9PI37,Q9PMG7

'Status' refers to the functional annotation status of the ORF, with H = hypothetical, P = putative, A = annotated

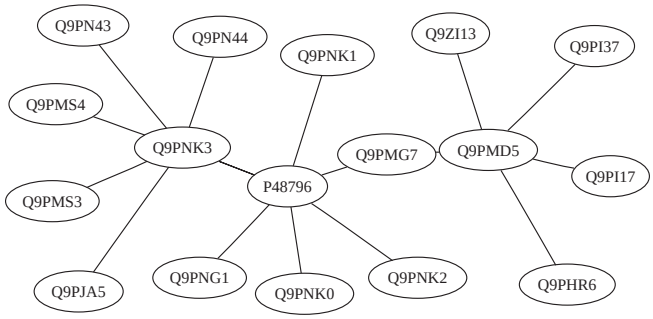

Fig. 2. Principal components of an hypothesized ferric uptake regulation pathway in *C. jejuni*. Each protein node is labeled by its corresponding ORF designation. The figure shows a subnetwork of predicted protein interactions linking the extracellular signal (Q9PJA5, putative integral membrane protein) to the regulatory (P48796, ferric uptake regulation) and transcriptional machinery (Q9PNK3, leucyl-tRNA transferase; Q9PN44, polyribonucleotide nucleotidyltransferase). Such connection is required to respond to changing requirements for iron storage or removal. Protein Q9PMD5 (possible bacterioferritin) may participate in redox stress resistance, by storing iron in a soluble, non-toxic form. Q9PMD5 is linked to a 30S ribosomal protein (Q9PI17) suggesting that this system may be involved in protection of the ribosomal machinery from iron toxicity.

Another key protein in this figure is Q9PMD5 (possible bacterioferritin) that may be instrumental in redox stress resistance, by storing iron in a soluble and non-toxic form. Q9PMD5 is linked to a 30S ribosomal protein (Q9PI17) which may suggest that this system is also involved in protection of the ribosomal machinery from iron toxicity. It is of interest to note that the hypthetical protein Q9PMG7 appears again in this inferred scenario of iron regulation. While a functional role has not been assigned for this protein, is it possible that it participates in many pathways within the cell. Recall Jeong *et al.* (2001), where it was argued that the most highly-connected proteins in protein interaction networks are most crucial to a cell's viability. Perhaps this protein carries such signficance within *C. jejuni*. This question awaits further proteomic study and validation.

The protein components central to the hypothesized ferric uptake interaction cluster are summarized in Table 5.

ACKNOWLEDGEMENTS

We thank Charles Elkan for his timely suggestions.

REFERENCES

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J.D. (1989) *Molecular Biology of the Cell*, 2nd edition, Garland Publishing, New York.

Andrews,S.C. (1998) Iron storage in bacteria. *Adv. Microbial Physiol.*, **40**, 281–351.

Bairoch,A., Bucher,P. and Hofmann,K. (1997) The PROSITE database, its status in 1997. *Nucleic Acids Res.*, **25**, 217–221.

Bock,J.R. and Gough,D.A. (2001) Predicting protein–protein interactions from primary structure. *Bioinformatics*, **17**, 455–460.

Bock,J.R. and Gough,D.A. (February 2002) Machine learning inference of protein–protein binding. In review.

Bradley,P.S., Fayyad,U.M. and Mangasarian,O.L. (1998) Mathematical programming for data mining: Formulations and challenges. *Technical Report MSR-98-01*. University of Wisconsin, Data Mining Institute, Madison, WI.

Brown,J.R., Douady,C.J., Italia,M.J., Marshall,W.E. and Stanhope,M.H. (2001) Universal trees based on large combined protein sequence data sets. *Nature Genet.*, **28**, 281–285.

Burges,C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.

Efron,B and Gong,G (1983) A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, **37**, 36–48.

Eisen,J.A. (2000) Assessing evolutionary relationships among microbes from whole-genome analysis. *Curr. Opin. Microbiol.*, **3**, 475–480.

Enright,A.J., Ililopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Fields,S. and Song,O-K (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.

Finley,R.L. and Brent,R. (1994) Interaction mating reveals binary and ternary connection between *Drosophila* cell cycle regulators. *Proc. Natl Acad. Sci. USA*, **91**, 12980–12984.

Gavin,A.C., Bosche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A.M., Cruciat,C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Hasty,J. and Collins,J.J. (2001) Protein interactions: Unspinning the web. *Nature*, **411**, 30–31.

Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hydrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.

Jeong,H., Mason,S.P., Barabási,A-L and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Kanehisa,M. (2000) *Post-genome Informatics*. Oxford University Press, Oxford, UK.

Klumpp,S. and Krieglstein,J. (2002) Phosphorylation and dephosphorylation of histidine residues in proteins. *Eur. J. Biochem.*, **269**, 1067–1071.

Kohavi,R. and Provost,F. (1998) Glossary of terms. *Machine Learning*, **30**, 271–274.

Lander,E.S., Linton,L.M., Birren,B. and Nusbaum,C. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 806–921.

LeGrain,L. and Selig,L. (2000) Genome-wide protein interaction maps using two-hybrid systems. *FEBS Lett.*, **480**, 32–36.

MacBeath,G. and Schreiber,S.L. (2000) Printing proteins as microarrays for high-throughput function determination. *Science*, **289**, 1760–1763.

Mandelbrot,B.B. (1977) *The Fractal Geometry of Nature*. W.H. Freeman, New York, NY.

Mangasarian,O.L. (1996) Mathematical programming in data mining. *Technical Report 96-05*. University of Wisconsin, Madison, WI.

Mann,M., Hendrickson,R.C. and Pandey,A. (2001) Analysis of proteins and proteomes by mass spectrometry. *Ann. Rev. Biochem.*, **70**, 437–73.

Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Martin,J.K. and Hirschberg,D.S. (1996) Small sample statistics for classification error rates I: Error rate measurements. *Technical Report ICS-TR-96-21*. Department of Information and Computer Science, University of California, Irvine.

Matsushita,M. and Janda,K.D. (2002) Histidine kinases as targets for new antimicrobial agents. *Bioorganic and Medicinal Chemistry*, **10**, 855–867.

Parkhill,J., Wren,B.W., Mungall,K., Ketley,J.M., Churcher,C., Basham,D., Chillingworth,T., Davies,R.M., Feltwell,T., Holroyd,S. *et al.* (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**, 665–668.

Pawson,T. and Scott,J.D. (1997) Signaling through scaffold, anchoring, and adaptor proteins. *Science*, **278**, 2075–2080.

Pazos,F. and Valencia,A. (2001) Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Engg*, **14**, 609–614.

Pellegrini,M. (2001) Computational methods for protein function analysis. *Curr. Opin. Chem. Biol.*, **5**, 46–50.

Perrone,M. (1993) *Improving Regression Esimation: Averaging methods for variance reduction with extensions to general convex measure optimization*, PhD thesis, Brown University.

Rain,J.C., Selig,L., De Reuse,H., Battaglia,V., Reverdy,C., Simon,S., Lenzen,G., Petel,F., Wojcik,J., Schächter,V. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.

Sankoff,D., Leduc,G., Paquin,B., Lang,B.F and Cedergren,R. (1992) Gene order comparisons of phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl Acad. Sci. USA*, **89**, 6575–6579.

Schwikowski,B., Uetz,P. and Fields,S. (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.

Stone,M. (1974) Cross-validatory choices and assessment of statistical predictions. *J. Roy. Statist. Soc.*, **36**, 111–147.

Swingler,K. (1996) *Applying Neural Networks: A Practical Guide*. Academic Press, London.

Tatusov,R.I., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.

Tekaia,F., Lazcano,A. and Dujon,B. (1999) The genomic tree as revealed from whole proteome comparisons. *Genome Res.*, **9**, 550–557.

Tucker,C.L., Gera,J.F. and Uetz,P. (2001) Towards an understanding of complex protein networks. *Trends Cell Biol.*, **11**, 102–106.

Uetz,P., Goit,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.

Uetz,P., Hughes,R.E. (2000) Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.*, **3**, 303–308.

Vapnik,V.N. (1995) *The Nature of Statistical Learning Theory*. Springer, Heidelberg, DE.

Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Walhout,A, Boulton,S, Vidal,M. (2000a) Yeast two-hybrid systems and protein interaction mapping projects for yeast and worm. *Yeast*, **17**, 88–94.

Wojcik,J., Schächter,V. (2001) Protein–protein interaction map inference using interacting domain profile pairs. *Bioinformatics*, **17**, S296–S305.