BIOINFORMATICS PREPRINT Vol. 00 no. 00 Pages 1–7

Predicting Homologous Signaling Pathways Using Machine Learning

Babak Bostan, Russell Greiner, Duane Szafron and Paul Lu

Department of Computing Science, University of Alberta, Edmonton, AB, Canada T6G 2E8
{bostan, greiner, duane, paullu}@cs.ualberta.ca

ABSTRACT

Motivation: In general, each cell signaling pathway involves many proteins, each with one or more specific roles. As they are essential components of cell activity, it is important to understand how these proteins work – and in particular, to determine which of a species' proteins participate in each role. Experimentally determining this mapping of proteins to roles is difficult and time consuming. Fortunately, many pathways are similar across species, so we may be able to use known pathway information of one species to understand the corresponding pathway of another.

Results: We present an automatic approach, Predict Signaling Pathway (PSP), that uses the signaling pathways in well-studied species to predict the roles of proteins in less-studied species. We use a machine learning approach to create a predictor that achieves a generalization F-measure of 78.2% when applied to 11 different pathways across 14 different species. We also show our approach is very effective in predicting the pathways that have not yet been experimentally studied completely.

Contact: bioinfo@cs.ualberta.ca

Supplementary information: The list of predicted proteins for all pathways over all considered species is available at www.cs.ualberta.ca/∼bioinfo/signaling.

1 INTRODUCTION

Intracellular and intercellular communications play a crucial role in cell life. We represent these communications by a directed graph of biochemical reactions. This network of reactions, called a *signal transduction pathway* or simply *signaling pathway*, is activated by receptors on the surface of the cell and includes secondary messenger molecules, proteins and compounds (small molecules). Understanding these pathways can help us discover previously unknown aspects of cellular life and may provide useful information for improving health. For instance, many known diseases, including diabetes and several cancers, are caused by cellular abnormalities linked to *signaling pathway* malfunctions (Seifter *et al.*, 2005). A better understanding of *signaling pathways* could lead to better treatments for these afflictions by aiding in drug design and development of other pathway interventions. Unfortunately, experimental approaches for investigating malfunctions in these networks are extremely difficult due to the large number of proteins in each cell and a lack of information about which proteins are involved in each pathway and their specific roles.

Computational approaches based on machine learning have become very popular for addressing complex biological challenges (Furey *et al.*, 2000; Ding and Dubchak, 2001; Guyon *et al.*, 2002; Park and Kanehisa, 2003; Lu *et al.*, 2004). Unfortunately, there is a dearth of research about applying computational techniques to help understand *signaling pathways*. Some published results use computational techniques to understand *metabolic pathways* – see Schilling *et al.* (1999). There have only been a few recent contributions; *e.g.*, Ma and Zeng (2003) find shortest paths between metabolites, Pireddu *et al.* (2006) use machine learning to predict the role of proteins in *metabolic* pathways and the MetaCyc group (Caspi *et al.*, 2008) provides two databases of organism-specific metabolic pathways: some experimentally elucidated and some predicted (BioCyc). However, none of these projects focus on *signaling pathways*. Previous work on signaling has been restricted to predicting *individual* signaling peptides and sorting signals (Nielsen *et al.*, 1999) or the effects of *single* genes on the overall functioning of signaling networks (Craven, 2002) or predicting *protein-protein interactions* (Yaffe *et al.*, 2001) that can be used to predict signaling pathways. While there are also many results that deal with a particular pathway or species (*e.g.*, Kim *et al.*, 2004), their narrow focus is not expandable to the other pathways or other parts of pathways. The Panther group (Thomas *et al.*, 2003) also provides a database of pathways, both metabolic and signaling, whose associated proteins are annotated by human experts. This is used to train HMMs that can then be used to classify novel proteins. Their system is a collection of proteins gathered by human experts, which uses a Hidden Markov Model (HMM) to classify functionality of novel protein sequences. While their computational approach has high coverage over mammalian protein-coding genes, it is not clear how to measure its accuracy, which makes it difficult to compare to other approaches.

The Predict Signaling Pathway (PSP) system presented in our paper uses an approach similar to Pireddu *et al.* (2006), but in the more complex domain of predicting *signaling* pathways. Both systems use homologous pathways and predict individual nodes in the graph structure. However, while Pireddu *et al.* use BLAST and HMM to predict enzymes in *metabolic pathways*, our PSP uses a very different technique, machine-learned classifiers, to predict proteins in the *signaling pathways*. We did not use HMMs as they did not perform as well for signalling pathways as they do for metabolic pathways: HMMs can accommodate the narrow variation in protein sequence occurring in enzymes but not the wider variation in protein sequences that serve in signalling pathways. Moreover, we employ a "retrospective" analysis that suggests that PSP's predictions are highly effective in predicting proteins that have not yet been experimentally verified. Fröhlich *et al.* (2008) provide another prediction

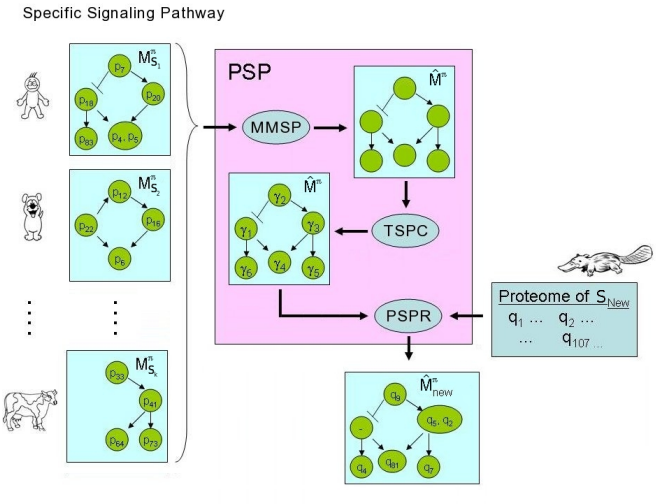

Fig. 1. Overview of Predict Signaling Pathway (PSP): Given the M_S^π pathways of various species S, *MMSP* builds the union model $\hat{M}^\pi$. *TSPC* uses the proteomes P_S of these species (not shown) to train one classifier γ_ρ for each role ρ of $\hat{M}^\pi$ (corresponding to each node). Finally, *PSPR* uses these classifiers along with the proteome of a new species S_{new} to predict which protein(s) will qualify for each role in this π signaling pathway in this new species.

system that predicts pathways using protein domains, which are subsequences of a peptide string that are intended to be functional units that can act independently from the rest of the protein chain. They predict whether a protein is involved in a particular pathway or not, but do not provide a way to predict the specific *role* of each protein. While their system can determine whether a protein belongs to a subset of roles in some *signaling pathways*, they limit their predictions to annotated *human* genes that have domain signatures. Testing their system on 10 out of the 11 *human* signaling pathways available in the KEGG database (Kanehisa *et al.*, 2008), they obtained an F-measure of approximately 80%. Our PSP system, on the other hand, can predict pathways using the whole proteome of *any* species and can predict the exact role of each protein within *arbitrary* signaling pathways, with an overall F-measure of 90.4% over all 11 human signaling pathways available in KEGG.

2 SYSTEM AND METHODS

Given a species' proteome (*i.e.*, the set of its proteins) and a specified set of signaling pathways, PSP predicts which of these proteins play which role in each of these signaling pathways for that species. As shown in Figure 1, our PSP system has three sub-systems, Make Model Signaling Pathway (*MMSP*), Train Signaling Pathway Classifiers (*TSPC*) and Predict Signaling Pathway Roles (*PSPR*).

2.1 Pathway Representation

In general, a pathway structure is a directed graph that describes the relations between proteins[1] in a signaling pathway. Each node of the graph represents a *role* of the pathway and involves a set of proteins, and each arc represents a

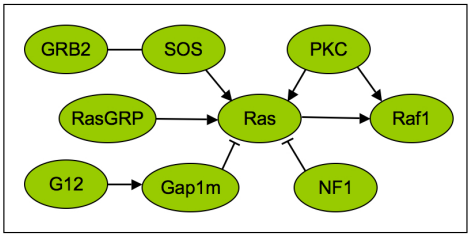

Fig. 2. A small part of MAPK signaling pathway structure in human.

relation (activation, inhibition, binding, etc.) between its source node and its target node. Figure 2 depicts a small part of the MAPK pathway structure in human, showing nine relation arcs, of three different types. An activation arc "$\alpha \rightarrow \beta$" indicates that proteins in the source node α can activate proteins in the target node β; an inhibition arc "$\dashv$" indicates that proteins in the source node inhibit proteins in the target node; and a bind-to arc "–" means that any protein in the source node can bind to any protein in the target node. For example, Figure 2 shows that proteins in the SOS, RasGRP and PKC nodes can activate proteins in the Ras node, while proteins in the Gap1m and NF1 nodes can inhibit proteins in the Ras node.[2] We take our pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2008), but other sources could be used as long as they have an appropriate graph structure.

We represent signaling pathways using the following notation. Each pathway $M = \langle N, A \rangle$ is a graph where each node $n \in N$ has an associated "role", and a set of associated proteins P^n, and the arcs $A \subset N \times N$ are a subset of pairs of nodes, each labeled with a type $a(\langle n', n'' \rangle) \in \{$ activation, inhibition, phosphorylation, dephosphorylation, binding $\}$. M_S^π denotes the instance of the π pathway for the species S. For example, the MAPK pathway for *H. sapiens* is denoted $M^{MAPK}_{H.sapiens}$; it may be different from homologous pathways in other species — *e.g.*, $M^{MAPK}_{H.sapiens} \neq M^{MAPK}_{M.mulatta}$. The symbol n^ρ_S denotes the node with role ρ associated with species S; *e.g.*, $n^{Ras}_{H.sapiens}$ is the human node with the Ras role. A specific node n^ρ_S is identified with a single species S and it can appear in several pathway graphs of that species and in the pathways of many species. For example, the $n^{Raf1}_{H.sapiens}$ node appears in $M^{MAPK}_{H.sapiens}$, $M^{VEGF}_{H.sapiens}$, $M^{erbB}_{H.sapiens}$ and there are many nodes with role Raf1 in different organisms: $n^{Raf1}_{H.sapiens}$, $n^{Raf1}_{R.norvegicus}$ and $n^{Raf1}_{M.mulatta}$. However, a role can appear only once in a single pathway of a single species. For example, $n^{RAS}_{H.Sapiens}$ can appear only once in the MAPK pathway of H.Sapiens.

We let $R^\pi_S = \{\rho \mid n^\rho_S \in N^\rho_S\}$ denote the set of roles that appear in pathway instance $M^\pi_S = \langle N^\pi_S, A^\pi_S \rangle$. For example, Ras is a member of $R^{MAPK}_{H.sapiens}$. P_S denotes the proteome of species S and $P^\rho = \cup_S P^\rho_S$ denote the union of all the proteins associated with the role ρ across all available species. Figure 3 presents a summary of these, and other, terms.

2.2 *MMSP* Constructs the Model Pathway

There are many known cellular signaling pathways, including the eleven KEGG pathways, each with its own function. However, these pathways are vary over different species; *e.g.*, the VEGF pathway in *P. troglodytes* involves $|R^{VEGF}_{P.\ troglodytes}| = 30$ roles and 32 arcs, while the same pathway in *X. laevis* has only $|R^{VEGF}_{X.\ laevis}| = 27$ roles and 31 arcs. In fact, there are roles and arcs in *P. troglodytes* that are not in *X. laevis*, and vice versa. Our goal is to use the pathways of a set of studied species to predict the pathways of less known species. For example, we might use the VEGF pathways of both *P. troglodytes* and *X. laevis* to find the proteins participating in various roles in the VEGF pathway of a third species. This species might have some roles and arcs that correspond to only *P. troglodytes*, and other roles and arcs that correspond

[1] In general, these graphs can also include compounds – *i.e.*, small molecules. However, we ignore them for this paper, focusing on only the proteins.

[2] Here we identify each node with its associated role. Hence, the "Ras node in human" refers to the node whose role is Ras, which we write $n^{Ras}_{H.sapiens}$

S: species; n: node; π: (signalling) pathway; ρ: role	
$M_S^\pi = \langle N_S^\pi, A_S^\pi \rangle$	instance of π pathway for the species S; graph structure involving nodes N_S^π and arcs A_S^π
$\mathcal{M}^\pi$	all models associated with pathway π, across species
$\hat{M}^\pi$	the union pathway for the π pathway
R_S^π	set of roles in pathway instance M_S^π
n_S^ρ $n_{\hat{M}}^\rho$	node with role ρ associated in species S; in model pathway $\hat{M}$
P P_S P^ρ P_S^ρ $P_{\hat{M}}^\rho$	all proteins ... across species; ... in species S; ... associated with role ρ; ... role ρ in species S; ... role ρ in model pathway $\hat{M}$
$\hat{P}_S^\rho$	for the set of proteins predicted to serve role ρ in species S
$\gamma_\rho(p)$	classifier associated with role ρ
PSP($\mathcal{M}^\pi$, P)	"Predict Signaling Pathway"
MMSP($\mathcal{M}^\pi$)	"Make Model Signaling Pathway"
TSPC($\hat{M}^\pi$, P)	"Train Signaling Pathway Classifiers"
PSPR($\hat{M}^\pi$, P_S)	"Predict Signaling Pathway Roles"

Fig. 3. Glossary of Terms used

only to *X. laevis*. (In fact, the VEGF pathway in *H. Sapiens* indeed has some roles and arcs corresponding to only *P. troglodytes* and some other roles and arcs corresponding to only *X. laevis*.)

MMSP (a sub-system of PSP) starts by building a general model pathway by combining the pathway versions for a set of model species. This requires creating a "graph union" of the graphs for each species pathway. This approach not only creates a diverse set of roles and arcs in the pathway structure, but also increases the number of proteins associated with each role.

MMSP constructs the model pathway $\hat{M}$ by taking the union of all pathway instances. The model pathway has a node $n_{\hat{M}}^\rho$ for each role ρ occurring in any of the species models, whose associated proteins $P_{\hat{M}}^\rho$ are the union of all of the proteins associated with the same role in any species. $\hat{M}$ also includes an arc of type a between two nodes $n_{\hat{M}}^{\rho_1}$ and $n_{\hat{M}}^{\rho_2}$ if nodes with these two roles are connected by the same type of arc in any of the individual species. [3]

Figure 4 shows an example of the union of two trivial pathways, where each node n^ρ is labeled by its role ρ (on upper left bump) and shows the associated set of proteins P^ρ. The set of proteins in role b of the model pathway is the union of the set of proteins of role b in species A and B: $P_{\hat{M}}^b = P_A^b \cup P_B^b$. The set of arcs of the union pathway is the union of the sets of arcs from the two pathways. Note that there is only *one* arc from $\langle n_{\hat{M}}^b, n_{\hat{M}}^a \rangle$, of type $\rightarrow$ corresponding to both $\langle n_A^b, n_A^a \rangle$ and $\langle n_B^b, n_B^a \rangle$.

2.3 *TSPC* Learns a Set of Classifiers

After producing this model pathway $\hat{M}$, *TSPC* learns a set of classifiers, one for each of $\hat{M}$'s roles. Each of these role-specific classifiers $\gamma_\rho(p)$ predicts whether each protein p in a species plays the ρ role in the pathway for that species. We let $\gamma_\rho : P_S \rightarrow \{Y, N\}$ denote the classifier associated with role ρ, where $\gamma_\rho(p) = Y$ if the protein p plays role ρ in S and $\gamma_\rho(p) = N$ otherwise. For example, the $\gamma_a(\cdot)$ classifier for the role a (not shown in Figure 4) returns Y if it predicts that the protein plays role a in the pathway, and N otherwise.

There are many supervised learning methods that can learn a classifier from a data sample whose instances are each labeled either positive or negative. Like most standard classifiers, our γ_ρ's take as input a fixed-size vector of

[3] The same pair of nodes could be connected many times in the union pathway $\hat{M}^\pi$ if they appeared with different labels in different species.

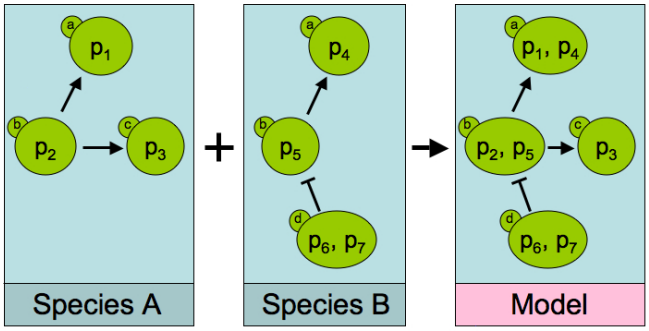

Fig. 4. Building the model pathway. Two instances of a pathway in two different species A (left) and B (middle). The model pathway (right) is created by combining pathway instances in species A and B. The labels a, b, c, and d (appearing in the warts) indicate the roles of each node; hence upper node in Species A is $n_{\hat{M}}^a$, with proteins $P_{\hat{M}}^a = \{p_1, p_4\}$. Similarly we see $P_A^b = \{p_2\}$ and $P_B^d = \{p_6, p_7\}$.

features to describe each instance (protein). We therefore compute a fixed set of features based on the primary amino-acid sequence of the protein. The first feature for each protein (with regard to the classifier $\gamma_\rho(\cdot)$) is a measure of the similarity between that protein and the most similar protein in the model pathway that is associated with role ρ. We use the BLAST algorithm (Altschul *et al.*, 1997) to compute this similarity measure. In general, BLAST takes as input a specific protein p and a database of proteins D, and returns a mapping, $\text{BLAST}_{p,D}$, from each protein $p' \in D$ to $\mathbb{R}$, where $\text{BLAST}_{p,D}(p')$ is a measure of how similar p' is to p. The similarity result is dependent on the database, including its size. For our computations, we set D to be the set of proteins described in the KEGG website. For each protein, BLAST actually returns a vector of values, including a similarity score, percent identity and an *e-value*. We use only the *e-value*, where smaller values indicate a higher similarity: In particular, the first feature value for protein p, wrt classifier γ_ρ, is $e_\rho(p) = \min_{q \in P^\rho} e_{p,q}$ where $e_{p,q}$ is the *e-value* of protein q wrt protein p:

$$e_{p,q} = \text{BLAST}_{p,D_{\text{KEGG}}}(q) \quad (1)$$

Although the first feature depends on the union model for the role as well as the protein, the other features depend only on the protein. The next nine features of protein p correspond to its subcellular locations. *TSPC* uses the Proteome Analyst system (Lu *et al.*, 2004) to predict in which of the nine cellular locations this protein does its main work: nucleus, cytoplasm, peroxisome, mitochondrion, plasma membrane, lysosome, golgi, endoplasmic reticulum and/or extracellular. Note that a protein can be in more than one location; hence *TSPC* uses 9 subcellular features (each a single bit) to encode this information. *TSPC* also uses the Phobius system (Käll *et al.*, 2007) to predict two more features for protein p: the number of membrane spanning regions (a non-negative integer) and whether it is a signal peptide or not (a bit). These features are relevant characteristics of roles in signaling pathways. For example, each signaling pathway should have one or more roles whose proteins each have a positive number of membrane spanning regions since the signal must pass through some cell membrane. All together, *TSPC* computes twelve features for each protein: the real-valued $e_\rho(p)$, nine binary subcellular values, the number of membrane regions and one binary feature that indicates if the protein p is a signal peptide or not.

For training examples, *TSPC* uses the model pathway $\hat{M}^\pi$ to obtain labeled positive instances — *e.g.*, for the role a in Figure 4, p_1 and p_4 serve as positive examples. For negative examples, we use all other proteins from the set of species that were used to produce the model pathway. In this implementation, we consider $k = 14$ species $\{S_1, \cdots, S_k\}$ (Table 2), with proteome sizes varying between 7,126 proteins and 29,445 proteins, with a total of 278,201 proteins, $P = \cup_{i=1}^k P_{S_i}$ across all species. The number of proteins $|P_S^\rho|$ in any single node n_S^ρ varied from 1 to 45 across the 5,608 nodes in the 11 different pathways (of the 14 species) that we considered.

To train a $\gamma_\rho(\cdot)$ classifier for each role ρ, we must identify many training instances, both positive and negative. The most straightforward way to train a classifier for role ρ is to use a set of proteins, P^ρ, as positive examples and the complimentary set, $P - P^\rho$, as negative examples. However, this leads to a very imbalanced training set. Therefore *TSPC* used a quick cut-off on the *e-value* to define our negative training set, including only those proteins p where $e_\rho(p) < 10$ as negative training examples. This reduced the number of negative examples to approximately 1,000 instances, which creates more balanced training sets. The remaining negative training instances are the proteins that are most similar to the positive training instances, but do not play the appropriate role.

For each of the 5,608 roles in the 11 union pathways, *TSPC* trained a Support Vector Machine (SVM) classifier (Bishop, 2006) using these labeled training instances. This system constructs a hyperplane that approximately separates the classes, by maximizing the margin between the two data sets. We used the `libsvm 2.86` implementation of SVM (Chang and Lin, 2001) with default settings and chose either linear or radial basis function kernels for each role, selecting the one with the larger "in-fold" training accuracy obtained by cross validation; see Section 3.2. We also tried polynomial basis functions with a range of degrees and a sigmoid basis function. However, these basis functions did not perform as well as the radial basis function where a non-linear function performed better.

2.4 *PSPR* uses the Model Pathway to Make Predictions about Novel Proteins

PSPR is the component of PSP that uses the classifiers built by *TSPC*, within the model pathway $\hat{M}^\pi$, to predict which proteins of a given proteome P_S, from a new species S, play which roles in this π pathway. For each role ρ in the model pathway $\hat{M}^\pi$, *PSPR* applies ρ's classifier γ_ρ, to each protein in P_S to produce the set $\hat{P}_S^\rho = \{\gamma_\rho(p) = Y \mid p \in P_S\}$, which is the set of S's proteins predicted to play role ρ. For some roles, this set is empty. The "predicted pathway roles" $R_S^\pi = \{\rho \in \hat{M}^\pi \mid \hat{P}_S^\rho \neq \{\}\}$ include all roles ρ in $\hat{M}^\pi$ for which $\hat{P}_S^\rho$ is non-empty. We then let N_S^π be the associated nodes in this predicted pathway, with those roles. In addition, M_S^π inherits all arcs from $\hat{M}^\pi$ that connect nodes in M_S^π. That is, if $n_S^\alpha, n_S^\beta \in N_S^\pi$ and $\langle n_{\hat{M}}^\alpha, n_{\hat{M}}^\beta \rangle \in A_{\hat{M}}^\pi$ then $\langle n_S^\alpha, n_S^\beta \rangle$ is in A_S^π; moreover, it will have the same label: $a(\langle n_S^\alpha, n_S^\beta \rangle) = a(\langle n_{\hat{M}}^\alpha, n_{\hat{M}}^\beta \rangle)$.

3 EMPIRICAL RESULTS AND DISCUSSION

Our experiments were based on the KEGG Pathway database, using the eleven pathways shown in Table 1, on the fourteen species shown in Table 2. As each pathway varies in size for different species, the second and third columns of Table 1 give the minimum and maximum number of roles appearing in each pathway across the different species. For evaluation purposes, we used two different versions of KEGG, one from 2006 and one from 2008. Table 1 contains summary data from both versions, with some new roles being discovered after 2006 there are only included in the 2008 data and one role (in MAPK) being removed between 2006 and 2008. The information about ErbB pathway is not included in the KEGG-06 section because this pathway was added to KEGG after 2006. In all 11 pathways, the number of roles in each model pathway matches the maximum number of roles for that pathway, as in each case there happened to be at least one species that had all of the roles.

3.1 Evaluation

As shown in Figure 1, PSP takes as input a proteome P_S from a novel species S and a set of known pathways $\mathcal{M}^\pi = \{M_{S_1}^\pi, \ldots, M_{S_k}^\pi\}$, corresponding to the same π signaling pathway across multiple different species. Let $\hat{M}^\pi$ be the model pathway produced by *MMSP*; M_S^π be the correct[4] signaling pathway for this species S; and $R^\pi = R_{\hat{M}}^\pi \cup R_S^\pi$ be the union of the roles that appear in either $\hat{M}^\pi$ or M_S^π. Then PSP computes a set of predictions. For each role $\rho \in R$, *PSPR* predicts a set of proteins $\hat{P}_S^\rho \subset P_S$ that (appear to) qualify for this role. If this ρ is not in $R_{\hat{M}}^\pi$, then PSP sets $\hat{P}_S^\rho := \{\}$. Let $\hat{\mathcal{P}}_S^\pi = \{\hat{P}_S^\rho\}_\rho$ be the entire collections of these protein-sets, one for each role. Similarly let $P_S^\rho \subset P_S$ is the true set of proteins associated with this role ρ, and $\mathcal{P}_S^\pi := \{P_S^\rho\}_\rho$. We again set $P_S^\rho := \{\}$ if this ρ is not in $\hat{\mathcal{P}}_S^\pi$.

Table 1. For each signaling pathway π used: the minimum and maximum number of roles $|R_{S_i}^\pi|$ across the 14 species S_i, in both the 2006 and 2008 versions of the KEGG database. Note that ErbB did not appear in KEGG 2006.

	KEGG-08		KEGG-06	
Pathway	Min	Max	Min	Max
MAPK	26	124	21	125
Wnt	12	67	11	67
ErbB	12	60	–	–
TGF-beta	26	54	18	54
Calcium	18	43	10	41
Phosphatidylinositol	10	31	7	30
mTOR	4	29	4	29
VEGF	6	28	6	28
Jak-STAT	12	26	6	26
Notch	2	22	3	22
Hedgehog	2	18	4	18

Ideally, if PSP worked perfectly, then $R_{\hat{M}}^\pi$ would match R_S^π, and for each role ρ, the predicted set $\hat{P}_S^\rho$ would exactly match the true set of proteins P_S^ρ. To compare $R_{\hat{M}}^\pi$ with R_S^π, we therefore compute their similarities over all of their roles, based on

$$\hat{q} = \bigcup_{\rho \in R_{\hat{M}}^\pi, p \in \hat{P}_S^\rho} \langle \rho, p \rangle \qquad q = \bigcup_{\rho \in R_S^\pi, p \in P_S^\rho} \langle \rho, p \rangle$$

which are each a set of pairs whose first component is the role and whose second is one of the proteins of that role. We then define the similarly between $R_{\hat{M}}^\pi$ and R_S^π based on the F-measure of the associated $\hat{q}$ and q:

$$F(\hat{q}, q) = \frac{2 \cdot |\hat{q} \cap q|}{|\hat{q}| + |q|} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

which uses

$$\text{Precision} = \frac{|\hat{q} \cap q|}{|\hat{q}|} \qquad \text{Recall} = \frac{|\hat{q} \cap q|}{|q|} \tag{2}$$

Note this F-measure ranges from 0 to 1, and is only 1 if $\hat{P}_S^\rho = P_S^\rho$ for all $\rho \in R^\pi$.

We use "leave out one species" cross-validation to estimate the accuracy of PSP. We start with $k = 14$ species $\{S_1, \ldots, S_k\}$, with known proteomes $P_i = P_{S_i}$ and let pathway M_i^j be the j^{th} pathway of the i^{th} species. Here, for each pathway M^j, for each species

[4] That is, "currently accepted"; see Section 3.3.

Table 2. From left to right: Species; Number of proteins in the species; Number of roles in each species over all considered pathways; Precision/Recall/F-measure of pathway predictions (averaged over all considered pathways).

Species	Proteins	Roles	Precision	Recall	F-measure
H. sapiens	24200	500	0.877	0.938	0.904
C. familiaris	19807	474	0.847	0.914	0.877
M. musculus	29445	501	0.825	0.923	0.868
P. troglodytes	25185	449	0.862	0.879	0.864
M. mulatta	23964	451	0.831	0.908	0.864
G. sallus	18115	433	0.857	0.870	0.861
M. domestica	19114	441	0.798	0.920	0.852
R. norvegicus	26160	476	0.793	0.920	0.850
O. anatinus	16387	406	0.804	0.722	0.751
X. laevis	10623	322	0.606	0.940	0.732
B. taurus	22327	399	0.592	0.921	0.716
X. tropicalis	8228	242	0.588	0.916	0.712
**D. rerio*	27520	383	0.533	0.824	0.642
**S. scrofa*	7126	131	0.315	0.902	0.457
Total	278201	5608	0.724	0.893	0.782

Table 3. Precision/Recall/F-measure of pathway predictions (averaged over all considered species).

Pathway	Precision	Recall	F-measure
TGF-beta	0.816	0.955	0.871
mTor	0.781	0.896	0.821
ErbB	0.761	0.894	0.809
Jak-STAT	0.826	0.796	0.797
Phosphatidylinositol	0.716	0.920	0.785
Wnt	0.677	0.934	0.768
Notch	0.792	0.906	0.768
MAPK	0.695	0.897	0.762
VEGF	0.702	0.867	0.757
Hedgehog	0.648	0.936	0.754
Calcium	0.645	0.819	0.713
Total	0.724	0.893	0.782

$i = 1..k$, we compute

$$\hat{M}_i^j = \hat{M}^j(P_i) = PSP(\{M_1^j, \ldots, M_{i-1}^j, M_{i+1}^j, \ldots, M_k^j\}, P_i)$$

then compute the overall score – the (precision, recall, F-measure) triple – for each of the roles in $\hat{M}_i^j$ and M_i^j:

$$s_i^j = \text{Score}(M_i^j, PSP(\{M_1^j, \ldots, M_{i-1}^j, M_{i+1}^j, \ldots, M_k^j\}, P_i))$$

Finally, for each species S_i with m known signaling pathways $\{M_i^1, \ldots, M_i^m\}$, we then compute the average triple $\text{ES}(S_i) = \frac{1}{m}\sum_{j=1}^{m} s_i^j$.

3.2 Empirical Results

Table 2 shows the results of our predictions, listing the average precision, recall and F-measure scores for each of the fourteen species $\text{ES}(S_i)$. The "Total" row is based on the average of the 14 values. This table also includes the number of proteins in P_S and the total number of roles, over the pathways considered. We see that, in 12 of the 14 species, the average recall is over 0.85, meaning that PSP is able to find essentially all of the relevant proteins for the roles (average recall = 0.913); the average precision of 0.724 shows that it occasionally included a few too many proteins. Moreover, the F-measures of only the two "*" ed species — *i.e.*, *S. scrofa* (pig) and *D. rerio* (zebra fish) — are below 0.70; in both cases due to low precision (*i.e.*, many false positives). We discuss this result in Section 3.3.

Table 3 presents our results from another point of view. Here we categorized the results based on *pathways* instead of species — *i.e.*, this is the average over all the species for each of the pathways. This shows our prediction is accurate for almost all of the pathways and the overall high F-measure (seen in Table 2 for species) is not just due to some specific pathways.

Table 4 shows the effect of each of the features used by our $\gamma_\rho(\cdot)$ classifiers. The first row, *e-value*, shows our predictive accuracy using only the single feature, *e-value*, which measures the highest similarity between the target protein and the proteins in this role. The values in the table are the average precision, recall and F-measures scores for all the fourteen species. The second row shows the effects of adding the 9 sub-cellular localization features; we see this slightly improves all three of the measures. The third row shows the effect of also adding the number of membrane spanning regions to the features — which makes essentially no difference. The values of the fourth row are obtained after adding signal peptide as the last feature of our classifier. While we can see that F-measure has dropped by a small value, this is not statistically significant (1-sided paired t-test,

Table 4. Precision/Recall/F-measure of pathway predictions (averaged over all considered species and pathways). Each row provides the accuracy of the prediction after using the feature mentioned along with the features mentioned in upper rows.

Feature/approach	Precision	Recall	F-measure
e-value	0.696	0.854	0.752
+ *sub-cellular localization*	0.698	0.858	0.755
+ *membrane regions*	0.692	0.865	0.756
+ *signal peptide*	0.679	0.862	0.746
+ *kernel selection*	0.724	0.893	0.782

Table 5. (left) Number of arcs in each species over all considered pathways; (right) Precision/Recall/F-measure of pathway predictions calculated for *arcs* (averaged over all considered pathways)

Species	#Arcs	Precision	Recall	F-measure
C. familiaris	492	1.000	0.995	0.997
M. musculus	443	1.000	0.960	0.980
H. sapiens	536	1.000	0.960	0.980
M. mulatta	536	0.967	0.977	0.972
M. domestica	460	0.965	0.967	0.966
R. norvegicus	500	0.939	0.957	0.948
G. sallus	437	0.958	0.937	0.945
P. troglodytes	458	0.951	0.888	0.913
D. rerio	346	0.840	0.999	0.911
B. taurus	369	0.759	1.000	0.858
X. laevis	233	0.735	1.000	0.847
O. anatinus	384	1.000	0.657	0.785
X. tropicalis	131	0.553	0.995	0.710
S. scrofa	49	0.329	1.000	0.494
Total	5374	0.870	0.948	0.888

$p \approx 0.90$). However, when combined with our final change (kernel selection), this feature turns out to give a higher F-measure than if it is not used. The final row represents our most accurate classifier. It shows the advantages of allowing *PSPR* to decide which kernel to use in the support vector machine: linear versus radial basis function (rbf). We see that this made an improvement to the F-measure, from 0.746 to 0.782 which is statistically significant — p < 3E-05. Even though adding the "signal peptide" feature had not improved the F-measure (third "addition" of Table 4), we found that the average F-measure of the classifiers that exclude this feature (but include "kernel selection") is only 0.768, which is inferior to the classifiers that include it, at the 0.782 shown in Table 4. This is true in general: kernel selection helps increase our accuracy due to the features we are using.

Note that this selection (linear or rfb) is done completely automatically, without any human adjustment. Here, for each species S_i in Table 2, we remove S_i from the training set and run cross validation on the rest of the data (for the 13 other species). For each such fold, *PSPR* excludes species S_j from the training set (in addition to S_i) and for each role $\rho \in R^{\pi}_{\hat{M}}$ in each pathway π, *PSPR* trains two classifiers (SVM-linear and SVM-rbf) on the remaining $14 - 2 = 12$ species, and compares the accuracy of these classifiers on S_j. After repeating the process for all the species in the training set, *PSPR* calculate the average prediction accuracy for each of SVM-linear vs SVM-rbf for this role ρ in this pathway π, then selects the kernel function with the highest average performance value. The final classifier for that role ρ uses this kernel function. Across all classifiers, 1792 linear kernels and 5262 radial-basis functions kernels were selected.

The first row of Table 4 shows that running the SVM learner on the e-value, alone, gives a fairly high F-measure. This suggests two other, simpler approaches: First, we could just use this e-value directly to identify the proteins. Here, for each role ρ with associated proteins P^{ρ}, for each $q \in P^{\rho}$ we compute $e_{p,q}$ (Equation 1) with respect to each $p \in P_S$ for the proteome P_S of the novel species S, and simply set $\hat{P}^{\rho}_S = \{p \in P_S \mid \exists q \in P^{\rho},\ e_{p,q} < \text{1E-100}\}$ to be those proteins in the novel species with an e-value less than 1E-100 to some protein the model pathway. (We used 1E-100 as the threshold as was empirically determined this was the best cut-off value in {1E-200, 1E-100, 1E-50, 1E-25, 1E-10, 1E-5}.) This produced an average leave-out-one-species F-measure (over the 11 pathways and 14 species) of only 0.650, which is 10.2% less than using SVM on the e-value alone (Table 4), and 13.2% worse than our best system. Second, we can view e-value as the basis for a nearest-neighbor classifier. For each protein $q \in P^{\rho}$, we find the $p \in P_S$ with the smallest e-value: $\text{nn}(q\,;\,P_S) = \arg\min_{p \in P_S}\{e_{p,q}\}$, then let $\hat{P}^{\rho}_S = \{\text{nn}(q\,;\,P_S) \mid q \in P^{\rho}\}$. The average F-measure here was 0.730, which was significantly lower than our best result (1-sided paired t-test, $p < \text{3E-3}$).

We also considered the multi-class classifier approach, of learning a single classifier for each pathway, that maps each protein to a role. The classifier returns one of $|R^{\pi}_{\hat{M}}| + 1$ values for each protein (the extra "1" accounts for "none of the above"). However, this would force us to predict *(at most) one* role for each protein, rather than a set of roles. This is problematic as many proteins (1359 in our training set) belong to more than one role, which is why we could not use this approach.

Table 5 shows the average precision, recall and F-measure scores for predicting the *arcs* in the various pathways. Here, PSP includes an A^{π}_S arc in a predicted model if it predicts at least one protein for each end. For example, if it was seeking the Model pathway (from Figure 4) within the proteome of species C, P_C, we would include the "b → c" arc if at least protein from P_C qualified for the "b" role, and at least one P_C protein qualified for the "c" role – $\hat{P}^{a}_C$ and $\hat{P}^{b}_C$ are non-empty. (Note that we do not require that these qualifying proteins are correct.) This would be a false positive if the Model pathway of species C did not include this "b → c" arc.

While the focus of this system is predicting which proteins fill which roles of the pathways, Table 5 shows our system does accurately identify most of the arcs in the examined species as well as the proteins in the associated roles.

We see that PSP can effectively predict the roles, and arcs, of essentially all available signaling pathways in all species, except possibly the two marked with *'s in Table 2. However, the result for these two species may actually be better than they appear; see the next section.

3.3 Alternative Historical Evaluation

The predictive accuracies in Table 2 show the precision is relatively low for two species (the "*"ed ones) — which are species that have not been extensively studied. We therefore wondered if PSP's accuracy for these species might actually be higher than these reported rates. That is, our F-measure scores are based on the (allegedly) "true" set of proteins associated with each role. However, many signaling pathways, especially those in understudied species, are not yet complete; researchers are still updating these pathways, typically by adding new proteins to roles. This means our evaluation may be wrong when it declares a predicted protein to be a false positive: *i.e.*, when PSP predicts a protein qualifies for a role ρ, but this protein is not in the current P^{ρ}_S. It is possible that PSP is actually correct as P^{ρ}_S is incomplete, in that this protein *should be* a member of P^{ρ}_S. Counting this protein as a false positive will (incorrectly) reduce PSP's precision score for this pathway of this species.

To test the possibility, we ran our PSP system on *historical* data: *i.e.*, we trained classifiers based on the 2006 version of KEGG (KEGG-06), and used these classifiers to make predictions on the (2006) proteomes of various species. The "KEGG-06/KEGG-06" columns in Table 6 provide these scores, when using the "2006 versions of the truth". (This involves only 7 of the 14 species, as the 2006 KEGG database did not include the data required for the other seven species.) Note especially the abysmal precision values for *C. familiaris*, *P. troglodytes* and *B. taurus* (the "*"ed ones). Our argument suggests this may be because, in 2006, these species had not been well annotated. If so, then we anticipate our predictions would better match the "2008 versions of the truth" — *i.e.*, the P^{ρ}_S for each role of these species based on KEGG-08.

The "KEGG-06 / KEGG-08" columns of Table 6 show the results of using KEGG-08 as ground truth to evaluate the predictions made by the "KEGG-06 classifier". We find a statistically significant improvement in the average precision (compared to KEGG-06/KEGG-06): from 0.615 to 0.773 — due largely to huge improvements in precision for those three species, coupled with minimal reductions in recall. This shows that many of the predictions based on KEGG-06 were correct, even though they involved claims that were not included in KEGG-06. In total, KEGG-08 included 1,620 proteins that we considered false positives when we evaluated the predictions based on KEGG-06 data, that turned out to be true positives. This is why we suspect that many of the false-positives found using KEGG-08 may actually be correct — *i.e.*, that PSP's actual precision may be higher than the 0.782 found when training

Table 6. (left) Number of proteins in each species, P_S, and number of roles over all considered pathways, for each of the 2006 and 2008 versions of KEGG; Precision/Recall/F-measure of pathway predictions (average over all considered pathways): (middle) trained and tested on KEGG-06; (right) trained on KEGG-06 and tested on KEGG-08.

	KEGG-06		KEGG-08		KEGG-06 / KEGG-06			KEGG-06 / KEGG-08		
Species	Proteins	Roles	Proteins	Roles	Precision	Recall	F-measure	Precision	Recall	F-measure
S. scrofa [Pig]	1,062	104	7,126	131	0.845	0.878	0.852	0.686	0.667	0.670
H. sapiens [Man]	25,719	440	24,200	500	0.885	0.814	0.842	0.879	0.805	0.835
M. musculus [Mouse]	30,172	436	29,445	501	0.839	0.832	0.827	0.839	0.816	0.820
R. norvegicus [Rat]	26,259	379	26,160	476	0.691	0.893	0.776	0.776	0.881	0.822
**B. taurus* [Bull]	22,854	218	22,327	399	0.345	0. 869	0.479	0.637	0.861	0.721
**C. familiaris* [Dog]	19,808	155	19,807	474	0.205	0.857	0.318	0.812	0.838	0.821
**P. troglodytes* [Chimpanzee]	21,825	139	25,185	449	0.186	0.653	0.262	0.840	0.634	0.715
Total (7 species)	147,699	1,871	154,250	2,930	0.563	0.827	0.615	0.783	0.788	0.773

and testing on KEGG-08, as shown in Table 2. Note also that the annotations present in 2008 but not 2006 are very likely to be experimentally determined; if they were purely analytic, we suspect they would have been present in 2006. Hence, this "train on 2006, test on 2008" measure is tested on annotations that are probably more accurate than the alternative of just removing a random subset of the 2008 data.

4 CONCLUSION

This article provides a new technique for learning to predict signaling pathways in novel species, based on known signaling pathways in familiar species. This technique is completely automated – *i.e.*, it does not need human adjustments at any level and is based on various automatically-computed properties of proteins.

We have shown that our approach produces accurate predictions over all of the species and pathways we considered — *i.e.*, total precision, recall and F-measure of 0.724, 0.893 and 0.782 respectively. We have also used historical data to indicate why we think that the actual accuracy of our prediction might be even higher than reported here, due to incompleteness of the test sets. The webpage (http://cs.ualberta.ca/∼bioinfo/signaling) provides the complete set of these PSP's predictions; it will be interesting to see which of these predicted roles turn out to be correct. Moreover, our overall PSP system is expandable, as other species and other pathways can easily be added to the system. In addition, new features (perhaps, based on protein-protein interaction, or protein domains) may be used along with the described features to potentially increase the accuracy of its predictions.

ACKNOWLEDGEMENT

We gratefully acknowledge the insights from our colleagues in the Proteome Analyst team, including Alex Ebhardt, Tadaaki Hiruki, Yifeng Liu and Chris Chen, as well as the suggestions of the anonymous referees.

Funding: This work was supported by Alberta Ingenuity Centre for Machine Learning and Natural Sciences and Engineering Research Council of Canada.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, R. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., Walk, T. C., Zhang, P., and Karp, P. D. (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway Genome Databases. *Nucleic Acids Research*, **36**(suppl_1), 623–631.

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Craven, M. (2002). The genomics of a signaling pathway: a kdd cup challenge task. *SIGKDD Explorations*, **4**, 97–98.

Ding, C. H. Q. and Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, **17**, 349–358.

Fröhlich, H., Fellmann, M., Sültmann, H., Poustka, A., and Beibbarth, T. (2008). Predicting pathway membership via domain signatures. *Bioinformatics*, **24**(19), 2137–2142.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**(10), 906–914.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**(1-3), 389–422.

Käll, L., Krogh, A., and Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction–the phobius web server. *Nucleic acids research*, **35**(Web Server issue).

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, **36**(suppl_1).

Kim, J. H., Lee, J., Oh, B., Kimm, K., and Koh, I. (2004). Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**(17), 3179–3184.

Lu, Z., Szafron, D., Greiner, R., Lu, P., Wishart, D., Poulin, B., Anvik, J., Macdonell, C., and Eisner, R. (2004). Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics*, **20**(4), 547–556.

Ma, H. and Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, **19**(2), 270–277.

Nielsen, H., Brunak, S., and von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**(1), 3–9.

Park, K.-J. and Kanehisa, M. (2003). Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, **19**(13), 1656–1663.

Pireddu, L., Szafron, D., Lu, P., and Greiner, R. (2006). The path-a metabolic pathway prediction web server. *Nucleic Acids Research*, **34**, 714–719.

Schilling, C. H., Schuster, S., Palsson, B. O., and Heinrich, R. (1999). Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, **15**(3), 296–303.

Seifter, J., Sloane, D., and Ratner, A. (2005). *Concepts in Medical Physiology*. Lippincott Williams and Wilkins.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, **13**(9), 2129–2141.

Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., and Cantley, L. C. (2001). A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nature Biotechnology*, **19**(4), 348–353.