BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 4 2015, pages 453–461
doi:10.1093/bioinformatics/btu407

Systems biology

Advance Access publication July 3, 2014

Inter-species prediction of protein phosphorylation in the sbv IMPROVER species translation challenge

Michael Biehl[1,*,†], Peter Sadowski[2,†], Gyan Bhanot[3], Erhan Bilal[4], Adel Dayarian[5], Pablo Meyer[4], Raquel Norel[4], Kahn Rhrissorrakrai[4], Michael D. Zeller[2] and Sahand Hormoz[5]

[1]Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, 9700 AK Groningen, The Netherlands, [2]University of California, Irvine, CA 92617, [3]Department of Physics and Department of Molecular Biology and Biochemistry, Busch Campus, Rutgers University, Piscataway, NJ 08854, [4]IBM T.J. Watson Research Center, Computational Biology, Yorktown Heights, NY 10598, [5]Kavli Institute for Theoretical Physics, University of California, Santa Barbara, CA 93106, USA

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: Animal models are widely used in biomedical research for reasons ranging from practical to ethical. An important issue is whether rodent models are predictive of human biology. This has been addressed recently in the framework of a series of challenges designed by the systems biology verification for Industrial Methodology for Process Verification in Research (sbv IMPROVER) initiative. In particular, one of the sub-challenges was devoted to the prediction of protein phosphorylation responses in human bronchial epithelial cells, exposed to a number of different chemical stimuli, given the responses in rat bronchial epithelial cells. Participating teams were asked to make inter-species predictions on the basis of available training examples, comprising transcriptomics and phosphoproteomics data.

Results: Here, the two best performing teams present their data-driven approaches and computational methods. In addition, *post hoc* analyses of the datasets and challenge results were performed by the participants and challenge organizers. The challenge outcome indicates that successful prediction of protein phosphorylation status in human based on rat phosphorylation levels is feasible. However, within the limitations of the computational tools used, the inclusion of gene expression data does not improve the prediction quality. The post hoc analysis of time-specific measurements sheds light on the signaling pathways in both species.

Availability and implementation: A detailed description of the dataset, challenge design and outcome is available at www.sbvimprover.com. The code used by team IGB is provided under http://github.com/uci-igb/improver2013. Implementations of the algorithms applied by team AMG are available at http://bhanot.biomaps.rutgers.edu/wiki/AMG-sc2-code.zip.

Contact: meikelbiehl@gmail.com

Received on April 3, 2014; revised on June 1, 2014; accepted on June 20, 2014

1 INTRODUCTION

Despite their limitations, animal models play an essential role in biomedical research, ranging from basic science to translational medicine, as human testing is severely limited by practical and ethical constraints. In the context of drug development, the usefulness of animal models obviously hinges on the extent to which results can be translated to human biology.

As an example, we consider here the response of bronchial epithelial cells to external chemical stimuli in rat and human. On one hand, organisms of common origin should arguably share many of the basic physiological mechanisms. On the other hand, different species may exhibit significant differences in the details of their cellular mechanisms such as signaling pathways. It is, therefore, essential to study this relationship systematically with the aim to develop reliable tools for the translation of results from animal models to human biology.

Recently, the *systems biology verification Industrial Methodology for Process Verification in Research* (sbvIMPROVER) initiative designed and organized the second sbv IMPROVER challenge, which was devoted to the question of species translation. In particular, sub-challenge 2 discussed in the following, concerned the protein phosphorylation status of normal human bronchial epithelial cells (NHBE) and normal rat bronchial epithelial cells (NRBE) exposed to the same set of chemical stimuli.

Herein, we present and discuss our studies of computational approaches for the prediction of stimulus-specific human protein phosphorylation levels based on gene expression and phosphorylation observed in the rat model. In section 2, we first give a brief description of the data and challenge design. Next, the computational approaches to the prediction task are described. After specifying the evaluation criteria applied by the challenge organizers, we describe methods used for the post hoc analysis of datasets and challenge results. Section 3 presents the results in terms of the predictions and their evaluation. In addition, results of the post hoc analysis concerning modifications of the prediction models and findings related to the phosphorylation kinetics are presented. We conclude with a discussion of the main results and an outlook on potential extensions and future studies.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

2 METHODS

Where applicable, the following is structured according to contributions from the challenge organizers and the two best performing teams: team AMG (Adel, Michael, Gyan) with members Bhanot, Biehl, Dayarian and Hormoz and team IGB (Institute for Genomics and Bioinformatics) with members Sadowski and Zeller.

2.1 Data acquisition and challenge design

The acquisition and preparation of the datasets are presented in greater detail in Rhrissorrakrai *et al.* (2015), which also provides further references. A detailed description is also available at www.sbvimprover.com. Here we give only a brief summary.

For subset (A) of chemical stimuli, gene expression and protein phosphorylation data in both species were made available to the challenge participants. Predictions of the protein phosphorylation status in human cells were to be made during the challenge for dataset (B), which corresponded to a different set of stimuli and comprised only the gene expression and phosphorylation data for rat.

Figure 1 illustrates the challenge setup and structure of the datasets. Each panel of data labeled as 'P' corresponds to the phosphorylation levels of 16 different proteins under 26 chemical stimuli in dataset (A) and 26 different stimuli in dataset (B). Phosphorylation measurements, using the Luminex xMap (TM) platform, were performed at 5 and 25 min after exposure to the stimuli. Repeated measurements provided two or three replicates per stimulus and protein. In addition, five or six DME (Dulbecco's modified Eagle's Medium) control measurements in absence of any stimulus were provided.

In the following, the 16 proteins are referred to by numbers: AKT1 (1), MP2K1 (2), CREB1 (3), MK03 (4), MK09 (5), MP2K6 (6), KS6B1 (7), MK14K11 (8), PTN11 (9), WNK1 (10), FAK1 (11), HSPB1 (12), KS6A1 (13), GSK3B (14), IKBA (15) and TF65 (16). For the full list of 26 stimuli in dataset (A) and 26 stimuli in dataset (B), see Rhrissorrakrai *et al.* (2015) or consult www.sbvimprover.com.

Gene expression was measured 6 h after exposure to the stimuli, yielding GCRMA normalized Affymetrix (TM) microarray data in two or three replicates. In addition, four or five replicates of DME controls were available. All measurements corresponded to 13 841 genes for rat and 20 110 genes for human, respectively.

The two best performing teams (AMG and IGB) submitted purely data-driven predictions based on the available phosphorylation data only. The specific approaches used by the two teams and the evaluation of submitted predictions are outlined in the following subsections.

It was disclosed to the challenge participants that a protein should be considered *activated* if its phosphorylation level, compared with DME controls, was greater than a threshold value of 3 in absolute value. The available training data displayed a strong prevalence of inactive proteins: considering the median over replicates, the resulting panel of $416 = 16 \times 26$ rat phosphorylation levels in dataset A contained 61 (14.7%) values above threshold and 48 (11.5%) in dataset B. The available human data in set A contained 35 (8.4%) positive samples.

2.2 Team AMG: naive and Learning Vector Quantization-based prediction

2.2.1 Processing of phosphorylation data The noise level observed over replicates in the phosphorylation data appeared to be roughly constant and of order O(1), independent of the mean phosphorylation level. To address the issue of outliers, we decided to use the median of the three replicate values. This effectively removed outliers because at least two of the three replicates were close together in value for all measurements. After exposure to the stimuli, activation might have occurred before 5 min, between 5 and 25 min or later. As there is no objective method to decide which of these options is correct, we chose to combine the data from both time points, using only the larger phosphorylation level (in absolute value and compared with controls) for each protein–stimulus pair. Finally, the mean level of phosphorylation across all available control data was subtracted. In the following, ratP and humP denote phosphorylation levels after these simple preprocessing steps.

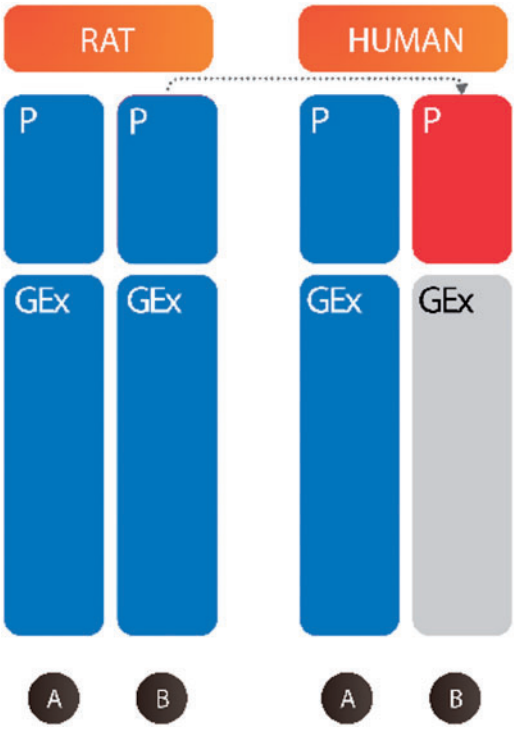

Fig. 1. Schematic illustration of the sub-challenge structure and datasets. The objective was to predict the phosphorylation status (P) of human phosphoproteins to stimuli subset B, shown in red, given the gene expression (GEx) and phosphorylation data for rat under the same stimuli. Available data (blue) also comprised the measurements of phosphorylation and gene expression in rat and human under a different set of stimuli A, which served as the training data. Human GEx data under the set of stimuli B was unavailable (shown in gray)

2.2.2 Prediction methods We applied two different methods to predict human protein activation in dataset B: a baseline prediction was obtained by simply assuming equal activation in both species, i.e. by setting *'human = rat'* after appropriate thresholding. In a second approach, we used a linear classifier trained on rat and human phosphorylation data in set A. Eventually, both results were combined, taking into account their prediction performance as assessed within dataset A.

2.2.3 Naive prediction A first prediction was based on the simple hypothesis that human phosphorylation levels (humP) and rat phosphorylation (ratP) should be similar under the same stimuli. Taking into account a threshold of 3.0 in absolute value for protein activation, we obtained heuristic certainty values by means of a non-linear monotonic transformation of the form

$$c_{naive} = \frac{1}{2}\left[1 + \tanh\left(\frac{|ratP| - 3}{5}\right)\right] \in [0, 1] \qquad (1)$$

The corresponding crisp classification can be achieved by thresholding at $c_{naive} = 0.5$, i.e. $|ratP| = 3$.

Performance measures like Receiver Operating Characteristics (ROC) or Precision Recall (PR) (Davis and Goadrich, 2006; Fawcett 2006)

depend only on the ranking of certainties and are insensitive to the precise choice of the monotonic non-linear function. Here, the arbitrary scaling factor 5 in the argument of tanh was chosen to yield a reasonable spread of values c_{naive} in the interval [0,1]. Measures that consider the accuracy of crisp classification depend on the choice of a threshold. Tuning the threshold could have, for instance, the aim to achieve a number of positive predictions that matches the occurrence of positive human samples in the training data. We did not tune or adapt the certainty threshold for crisp classification explicitly.

2.2.4 Learning Vector Quantization As an alternative to the naive prediction, we applied a simple version of Learning Vector Quantization (LVQ), specifically the so-called LVQ1 scheme (Biehl *et al.* 2007; Kohonen 1990, 1997). To this end, the training dataset (A) was interpreted as to provide 26 feature vectors $\vec{x}^{\mu} \in I\!R^{16}$, which comprise the known rat phosphorylation levels *ratP* under stimuli $\mu = 1, 2, \ldots, 26$. We considered 16 classification problems separately, corresponding to the protein-specific phosphorylation levels in human. Binary target labels 0 or 1 were defined according to the comparison of $|humP|$ with the threshold value of 3.0.

We used the simplest possible LVQ system, using only one prototype per class, i.e. $\vec{w}_0$ and $\vec{w}_1$, which were adapted iteratively under random sequential presentation of example data. For a given labeled feature vector $\vec{x}$, the prototype $\vec{w}_i$ with the smallest Euclidean distance from $\vec{x}$ was updated according to the standard LVQ1 prescription [Biehl *et al.* (2007); Kohonen (1990)]

$$\vec{w}_i \leftarrow \vec{w}_i + \eta\Psi(\vec{w}_i, \vec{x})(\vec{x} - \vec{w}_i). \quad (2)$$

Here $\Psi(\vec{w}_i, \vec{x}) = +1$ if $\vec{x}$ belonged to class i and $\Psi(\vec{w}_i, \vec{x}) = -1$ otherwise. Prototypes were initialized in the class conditional means of the actual training set. Updates (2) were performed at constant learning rate ($\eta = 0.005$) over 1000 single example presentations. The procedure yielded two prototypes $\vec{w}_o, \vec{w}_1 \in I\!R^{16}$, which represent inactive or activated proteins, respectively. A crisp Nearest Prototype LVQ classifier assigns a feature vector $\vec{x}$ to class 1 (activation) if

$$d(\vec{w}_1, \vec{x}) \leq d(\vec{w}_0, \vec{x})$$

and to class 0 (inactive) otherwise. Here, the squared Euclidean measure $d(\vec{w}, \vec{x}) = (\vec{w} - \vec{x})^2$ was used to quantify the distance of feature vectors and prototypes; hence, the simple system with one prototype per class parameterizes a linear class boundary (Biehl *et al.*, 2007). It is important to note that a separate LVQ classifier was obtained for each of the 16 target proteins.

While probabilistic and fuzzy variants of LVQ have been suggested in the literature, see e.g. [Schneider *et al.* (2010); Seo *et al.* (2003)], we resorted here to the heuristic computation of a certainty similar to (1):

$$c_{LVQ} = \frac{1}{2}\left[1 + \tanh\left(\frac{d(\vec{w}_o, \vec{x}) - d(\vec{w}_1, \vec{x})}{200}\right)\right] \in [0, 1]. \quad (3)$$

Crisp classification is obtained by thresholding at $c_{LVQ} = 0.5$. Again, the scaling factor was set manually to achieve a variation of certainties similar to the naive prediction. Whenever a particular training set contained negative examples only, an LVQ system could not be determined and the naive prediction was used, instead.

Estimates of the expected classification performance were obtained using a standard Leave-One-Out (L-O-O) validation procedure (Duda *et al.*, 2001; Hastie *et al.*, 2009). The 26 slightly different LVQ classifiers per protein were applied to dataset (B), inserting the corresponding *ratP* values as feature vectors and obtaining predictions analogous to Equation (3). The final LVQ predictions $c_{LVQ} \in [0, 1]$ were computed as averages over the 26 L-O-O results.

2.2.5 Validation and combination of predictions We evaluated the naive prediction by directly comparing the certainties cnaive obtained from ratP in dataset A with the corresponding phosphorylation levels humP. To this end, the latter were binarized by thresholding at $|humP| = 3$. As one of the many possible criteria, we considered the ROC (Fawcett, 2006) over the full panel of $16 \cdot 26 = 416$ values, which yielded an area under the ROC curve (AUROC) of $AUROC_{naive} \approx 0.83$. This value reflects a relatively high predictive power of ratP for the observation of protein activation, $|humP| > 3$, within dataset A.

For the LVQ classifier, we computed the ROC within dataset A as obtained from the 26 L-O-O runs (Fawcett, 2006). The corresponding area under curve was found to be $AUROC_{LVQ} \approx 0.88$, suggesting a slightly better performance of LVQ as compared with the naive prediction.

Eventually, the final prediction in terms of certainties for $|humP| > 3.0$ in the test dataset B was obtained as the weighted combination

$$c_{AMG} = \frac{c_{naive}(AUROC_{naive} - 0.5) + c_{LVQ}(AUROC_{LVQ} - 0.5)}{AUROC_{naive} + AUROC_{LVQ} - 1}. \quad (4)$$

The specific form reflects the fact that $AUROC = 0.5$ corresponds to the baseline for random guesses. For the prediction submitted to the challenge, we had computed protein-specific combinations. However, the corresponding prediction and its test set performance were virtually identical with the results reported here. An unweighted mean of the two predictions would have given similar results, because of the relatively small difference of the $AUROC$.

As post hoc analysis, we also tested our results using the criteria that were ultimately used by the challenge organizers (see Section 2.4). For the training set (A) performance, we found the values summarized in Table 1. These alternative criteria also suggest a comparable or slightly superior quality of the LVQ system compared with the naive prediction.

Table 1. Performance of the naive prediction within dataset (A) and L-O-O estimate of the corresponding performance of LVQ, respectively, as evaluated according to the four measures defined in Section 2.4

prediction	*AUROC*	*AUPR*	*PCC*	*BAC*
c_{naive}	**0.83**	**0.34**	**0.72**	**0.75**
c_{LVQ}	**0.88**	**0.36**	**0.74**	**0.73**

2.3 Team IGB: neural network-based prediction

Team IGB's pipeline consisted of two parts: an artificial neural network (NN) trained to predict human phosphorylation status from rat data, and a statistical analysis that aggregated evidence from the replicated measurements.

Both phosphorylation and gene expression status was provided with the rat data, so it was possible to use both in predicting the human phosphoprotein status. However, validation experiments indicated that the rat gene expression data only increased overfitting during training, so these features were removed. Thus, the submitted predictions used a NN with 32 inputs, corresponding to the 16 protein phosphorylation levels measured at both 5 and 25 min, a single hidden layer of 1000 logistic units and 32 logistic outputs with a cross entropy loss function. The training data comprised every possible input–target pair for each stimulus; a stimulus with three rat measurements and three human measurements contributed nine samples to the training set. As a preprocessing step, the log phosphorylation measurements were clipped to be between the values of 3 and 7, then translated and scaled to be in the range [0,1], allowing us to interpret these values as the probability that a particular protein is phosphorylated. The same transformation was applied to the DME controls.

To avoid overfitting, the NN was trained with stochastic neurons. We tested both the dropout learning algorithm (Baldi *et al.*, 2013; Hinton *et al.*, 2012) and a variant in which Gaussian noise was added to each

neuron (Baldi *et al.* (2014)); i.e. at each forward propagation through the network during training, a different random value $\epsilon \sim N(0, 0.2)$ is added to the output of each neuron independently. The additive Gaussian noise algorithm was used to train the final network because it performed slightly better than the dropout algorithm on a validation set. The final network was trained for 6000 epochs with a learning rate that started at 0.1, and decayed exponentially by 1.000004 after each batch. The momentum term increased linearly from 0.5 to 0.99 over the first 500 epochs, then remained constant. The weights were initialized randomly from $U(-0.01, 0.01)$ for the first layer and $U(-0.001, 0.001)$ for the second layer. All NN training was performed using the Pylearn2 and Theano software libraries (Bergstra *et al.*, 2010; Goodfellow *et al.*, 2013).

At prediction time, the NN was used to make predictions from each individual replicate. To combine these predictions, and to use the statistical properties of the background distribution, we performed an additional statistical analysis step to test for a significant difference in the predicted phosphorylation of the proteins compared with the DME controls. For each of the 32 predictions, we perform a one-tailed two-sample *t*-test for equal means against the DME controls. To make a final prediction in the range [0,1], we compute $\log_{10}(-\log_{10}(P\text{−value}) + 1)$.

2.4 Evaluation of predictions

The choice of evaluation criteria had to take the pronounced prevalence of inactive prosphoproteins into account, to avoid artifacts. For instance, all-negative predictions would appear competitive with non-trivial schemes when taking only overall accuracies into account.

The precise assessment criteria were not disclosed beforehand. This prevented participants from fine-tuning their results according to the expected evaluation process. For a more complete discussion of the evaluation criteria, see also (Rhrissorrakrai *et al.*, 2015). Participants submitted their predictions in terms of certainty scores ranging from 0 (certainly not activated) to 1 (certainly activated). These predictions were evaluated by comparison with a binarized *gold standard,* where 1 or 0 indicated that the phosphorylation level (median over replicates) was above or below a threshold of 3 in absolute value, respectively. This choice corresponds to ±3 SDs calculated across all phosphorylation values including DME controls. Details of the preprocessing and definition of the gold standard are also given in Rhrissorrakrai *et al.* (2015).

The organizers decided to consider a combination of three different quality measures:

- **AUPR:** The area under the Precision Recall curve (AUPR) is accepted as an appropriate measure for biased classification problems (Davis and Goadrich, 2006). It was obtained for the panel of 416 predictions and yielded a quantity between 0 and 1.
- **PCC:** The Pearson correlation coefficient (PCC) was based on the correlation *corr* of certainties with the binarized *gold standard* and was evaluated over the full panel of prediction and then scaled to obtain values in [0,1]:

$$PCC = (1 + corr)/2.$$

- **BAC:** The Balanced Accuracy (BAC) takes into account the number of true-positive (TP) and true-negative (TN) predictions separately (Brodersen *et al.*, 2010):

$$BAC = (TP/P + TN/N)/2$$

with $0 \leq BAC \leq 1$, where T and N are the total number of positive and negative samples, respectively. Team predictions were thresholded at $c = 0.5$ for this measure.

Team submissions were assessed and ranked according to the sum $AUPR + PCC + BAC$. In addition, tests for the statistical significance of differences between team performances were performed. The $AUPR$ depends only on the order of the certainties $c \in [0, 1]$, whereas for the BAC, only the comparison with the threshold 0.5 matters. However, the Pearson correlation may depend significantly on the precise values of the certainties, as for instance controlled by the non-linearities in the predictions (1) and (3).

Table 2. Rough representation of the activation kinetics, comparing human and rat phosphorylation measured at 5 min (early) and/or 25 min (late), respectively

rat ↓\ hum. →	Early	Both	Late	Inactive
Early	**0**	**0**	**1**	**2**
Both	**0**	**0**	**1**	**2**
Late	**–1**	**–1**	**0**	**2**
inactive	**–2**	**–2**	**–2**	**0**

The values were chosen to enhance the differences in timecourse phosphorylation between rat and human.

2.5 Post hoc analyses

The release of the *gold standard* after completion of the challenge made possible further investigations. The performance measures used in the challenge ranking were also applied to the individual predictions of team AMG, i.e. the naive and LVQ classifiers. Similarly, predictions based on single time point measures (at 5 and 25 min, respectively) were evaluated separately along the same lines. In addition to the challenge criteria, AUROC were determined analogously.

Several modifications of the classification scheme were considered by team IGB, including attempts to make use of the gene expression data as additional input features to the NN classifier.

The time point-specific data were also exploited in an analysis provided by the challenge organizers. All pairs formed of the 52 stimuli and 16 proteins can be labeled according to Table 2, which compares the timing of activation in human and rat according to the measurements at 5 min (early) and 25 min (late). After assigning labels as specified in Table 2, a hierarchically clustered heatmap of proteins and stimuli was generated using the routine `pheatmap` in the R package (distance: 'Euclidean', clustering method: 'average').

3 RESULTS

3.1 Test set predictions and evaluation

Figure 2 displays the naive, the LVQ based and the combined prediction of team AMG, as well as team IGB's prediction as color-coded certainties for dataset B.

The performance of the final prediction with respect to the test set data was evaluated by the challenge organizers in terms of the three different quality measures presented above. After disclosure of the sub-challenge outcome and the *gold standard*, we also assessed the corresponding performance of the individual predictions c_{naive} and c_{LVQ}, separately, in terms of the measures $AUROC$, $AUPR$, PCC and BAC. The outcome is summarized in Table 3 and shows that the naive prediction showed the best performance among all individual methods; it was only outperformed by the weighted combination c_{AMG}.

3.2 Further post-challenge analysis

After completion of the challenge and disclosure of the *gold standard,* additional aspects of the data could be investigated in a post hoc analysis as summarized in the following.

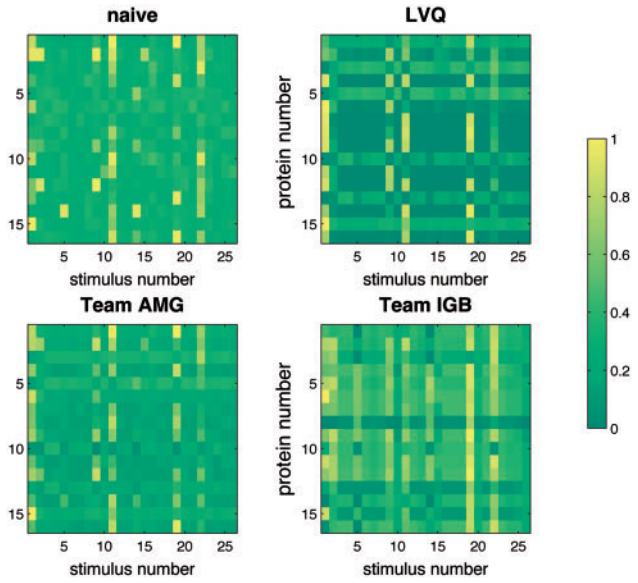

Fig. 2. Color-coded visualization of the predictions for $|humP| > 3$ in dataset B: c_{naive} (upper left panel), c_{LVQ} (upper right) and the combination c_{AMG} of team AMG (lower left). The lower right panel displays the prediction by team IGB. Proteins are numbered according to the list given in Section 2.1

Table 3. Test set performances for the prediction of activated human proteins, as evaluated according to four different measures

Prediction	*AUROC*	*AUPR*	*PCC*	*BAC*
c_{naive}	**0.85**	**0.45**	**0.74**	**0.79**
c_{LVQ}	**0.79**	**0.37**	**0.69**	**0.76**
c_{AMG}	**0.83**	**0.54**	**0.75**	**0.77**
c_{IGB}	**0.84**	**0.41**	**0.68**	**0.76**

3.2.1 Team IGB: influence of model details on performance Various choices existed at each stage of IGB's prediction pipeline, and ultimately only a single model was chosen for final submission. Table 4 shows various combinations of these choices and their influence on the final performance evaluation on the gold standard. Rough predictions for human phosphorylation were made using (i) a NN trained on phosphorylation data alone (P); (ii) a NN trained on phosphorylation and gene expression data (P + GE) or (iii) by using the rat phosphorylation as a prediction for that of human directly. Next, the NN outputs were transformed using the logistic function to spread the data away from a fixed value 0.5, the max predicted value for DME in the training data (max DME) or not at all. Last, we varied the method of aggregating the NN output on the replicates, using either a *t*-test (standard or Bayesian) or a simple mean over replicates.

Results suggest that performing *t*-tests increased performance on the BAC performance metric, but in fact lowered performance on the other metrics. The overall performance of IGB's final submission could have been improved by eliminating the standard *t*-test in favor of averaging the raw NN output, but not enough to rank above the top-scoring submission from Team AMG, which had a sum of 2.06 versus a sum of 1.93 using the original Team IGB submission without a *t*-test. Additionally, the original model used by Team IGB could have been improved if the aggregation of time points was performed using the mean rather than the max over both time points, as the sum of the three scoring methods increased for predictions using the human (P) and for rat (rat P), while BAC decreases, when a standard *t*-test was performed. If the standard *t*-test is eliminated, the original choice of taking the max over both time points performed best. Further, a marginal gain of 0.0002 could have been obtained in the final summation of scores by transforming the NN output using the logistic function centered on 0.5, but this gain is not consistent with the choice of training data used in the NN (P, P + GE or Rat P).

Table 4. A post-challenge comparison of Team IGB's modeling choices, with model combinations on the left (see text for details) and performance metrics on the right

Method	Logistic Position	*t*-test	Aggregate 5 and 25	AUPR	Pearson	BAC
P	0.5	None	Max	**0.4989**	**0.7525**	0.6806
P	None	None	Max	**0.4990**	**0.7522**	0.6806
P	Max DME	None	Max	**0.5017**	**0.7249**	0.6657
P	None	Standard	Mean	**0.4115**	**0.7039**	0.7436
P	0.5	None	Mean	**0.4607**	**0.7500**	0.6348
P	**None**	**Standard**	**Max**	**0.4075**	**0.6778**	**0.7595**
P	0.5	Standard	Max	**0.4078**	0.6775	0.7595
P	None	Standard	Min	0.3904	**0.6996**	0.7501
Rat P	Max DME	None	Max	**0.4090**	**0.7113**	0.7160
P	None	Bayes	Max	0.3131	**0.6821**	**0.7634**
Rat P	0.5	None	Max	0.2864	**0.7100**	0.7417
Rat P	None	None	Max	0.3006	**0.7123**	0.7160
Rat P	Max DME	Standard	Max	0.3251	**0.7072**	0.6902
Rat P	0.5	None	Mean	0.3262	**0.7025**	0.6889
P	Max DME	Standard	Max	0.3799	**0.6841**	0.6186
Rat P	None	Bayes	Max	0.2514	0.6592	0.7343
Rat P	None	Standard	Mean	0.3399	0.6574	0.6288
Rat P	0.5	Standard	Max	0.2804	0.6519	0.6778
Rat P	None	Standard	Max	0.2704	0.6537	0.6842
Rat P	None	Standard	Min	0.2995	0.6329	0.6121
P + GE	0.5	None	Max	0.1292	0.5418	0.5057
P + GE	None	None	Max	0.1292	0.5398	0.5070
P + GE	0.5	None	Mean	0.1346	0.5345	0.4909
P + GE	0.5	Standard	Max	0.1179	0.5384	0.5031
P + GE	None	Standard	Max	0.1162	0.5371	0.5031
P + GE	None	Standard	mean	0.1222	0.5325	0.4896
P + GE	None	Standard	Min	0.1273	0.5194	0.4948
P + GE	max DME	None	Max	0.1385	0.4924	0.4987
P + GE	max DME	Standard	Max	0.1041	0.5218	0.4987
P + GE	None	Bayes	Max	0.0601	0.4639	0.5044

The combination that Team IGB used for submission is highlighted in bold, along with all performance scores that exceed the performance of the submitted predictions. All combinations are ranked according to the sum of the three metrics, as was done for the subchallenge scoring.

3.2.1 Team AMG: time-specific naive predictions In addition to the naive prediction described in Section 2.2, we considered *ratP*-based certainties $c^{(5)}$ and $c^{(25)}$ as obtained following the

same naive scheme, cf. Equation (1), but from the measurements at 5 and 25 min separately. The total c_{naive} discussed above could be recovered as $c_{naive} = \max\{c^{(5)}, c^{(25)}\}$ from these time-resolved predictions. All results presented in this subsection correspond to the test set data (B).

The naive certainties yield 31 positive predictions with $c^{(5)} \geq 0.5$, 34 cases with $c^{(25)} \geq 0.5$ and 48 positive predictions in total ($c_{naive} \geq 0.5$). In comparison, the number of human protein activations in the target data is 21 at 5 min, 20 at 25 min and 31 in total. Hence, the naive predictions tend to overestimate the number of active proteins in the test set.

We compared the individual predictions with the thresholded $humP^{(5)}$ and $humP^{(25)}$ values at 5 and 25 min, respectively. In addition, we considered $c^{(5)}$ and $c^{(25)}$ as separate predictions for the binarized *gold standard*, which corresponds to thresholding $|humP| = \max\{|humP^{(5)}|, |humP^{(25)}|\}$. For the time point-specific predictions analogous to Equation (1), we obtained the test set performances summarized in Table 5. The results show that the agreement between *ratP* and *humP* appears to be slightly stronger for the measurements at 25 min. Moreover, the naive $c^{(25)}$ yields a test set performance similar to that of the total c_{naive} already, cf. Table 3.

3.2.2 Other challenge results and meta-analysis The challenge organizers analyzed and compared predictions provided by 13 different teams. A detailed discussion is presented in Rhrissorrakrai *et al*. (2015). Arguably the most remarkable findings were the following:

Among the 13 participating teams, 8 based their predictions on phosphorylation data only. This turned out advantageous, as also five of the six top-ranked submissions did not use the GEx data.

All submitted predictions were based on data-driven approaches, applying a variety of computational methods including NNs, linear discriminant analysis and support vector machines. A universally superior approach or family of algorithms could not be identified with respect to the achieved rankings.

Ten teams submitted predictions that were significantly better than random concerning at least two of the three applied performance measures. However, most predictions failed to outperform the naive approach of equating human with rat phosphorylation. This simple baseline strategy would have achieved rank 2 in the sub-challenge.

Averaging all team predictions, exploiting the potential *wisdom of the crowd,* did not outperform the top-ranked teams, but scored better than the second best performer with respect to AUPR and PCC.

3.2.3 Phosphorylation kinetics in rat and human To detect phosphorylation patterns that differ in timing and activity in rat and human, the challenge organizers looked for changes in phosphorylation at the two different time points, 5 min and 25 min, after the cells' exposure to a stimulus and computed the state of phosphorylation of the 16 measured proteins for each of the 52 stimuli (training and test sets) at both time points. Figure 3 shows that, overall, stimuli cause more activation in rat than in human cells except for two specific phosphoproteins KS6A1 and HSPB1. Differences between the kinetics of activation at time points 5 and 25 min are minimal. This difference could be owing to a more homogeneous biological sample in rats than in humans, or simply to higher sensitivity and faster signaling of the NRBE cells compared with NHBE cells.

Table 5. Test set performance of naive prediction schemes $c^{(5)}$ and $c^{(25)}$ obtained from the *ratP* measurements at 5 and 25 min separately, compared with the corresponding time-specific (top) and total (bottom) binarized human protein activation

Time-specific Naive prediction	*AUROC*	*AUPR*	*PCC*	*BAC*
$c^{(5)} \rightarrow humP^{(5)}$	0.76	0.23	0.65	0.66
$c^{(25)} \rightarrow humP^{(25)}$	0.88	0.46	0.76	0.82
$c^{(5)} \rightarrow humP$	0.75	0.29	0.66	0.63
$c^{(25)} \rightarrow humP$	0.82	0.45	0.75	0.77

To test whether signaling pathways are used similarly in rat and human, we describe in Figure 3B and C, two pathways taken from literature and whose components are present in our dataset. AKT1 is known to activate KS6B1/p70-S6 kinase, an mTOR activation marker shown to phosphorylate and activate CREB1 (Xing *et al*., 1996) but KS6B1/p70-S6 is destabilized by TNFα signaling through IKBA and TF65/RelA (Gao *et al*., 2009). Conversely, AKT1 has been shown to phosphorylate and negatively regulate GSK3 β (Cantley 2002) and positively regulate WNK1 (Jiang *et al*., 2005). EGF activates AKT1 (Cantley 2002) and PTN11, the protein tyrosine phosphatase non-receptor type 11 also known as PTP2C (Schulze *et al*., 2005).

In Figure 3B, we note that, besides the overall larger number of phosphorylated proteins in rat compared with human, from the four stimuli, i.e. EGF (Epidermal Growth Factor), TGFA (Transforming Growth Factor alpha), EPGC (Epigallocatechin), and PMA (Phorbol-12-Myristate-13-Acetate), that activate CREB1 through KS6B1 in rat, only EGF does so in human. Also all the stimuli that activate CREB1 independently of KS6B1 in rat do not do so in human, showing a large divergence in signaling. We observe that TNFα signaling through IKBA and TF65/RelA occurs as expected in both human and rat, as KS6B1 is not active in presence of TNFα probably being degraded through the phosphorylated IKBA. Interestingly, differences arise with IL1B and EPGC that also activate IKBA in human, but in rat, EPGC does not activate IKBA and KS6B1 is active. Conversely, PMA activates IKBA only in human, but in both species, KS6B1 is active.

In Figure 3C, AKT1 activation of WNK1 is conserved as from the four stimuli (EGF, TGFA, EPGC, INS) activating WNK1 in human and rat, only EGF signaling is not similar in rat and human cells. Once again more activity is shown in rat, as five stimuli activate AKT1 and inactivate GSK3β in rat but not in human. As shown in the wiring diagram of Figure 3c, EGF turns on AKT1 and PTN11 and AKT1 activation represses GSK3. However, contrary to this, for five stimuli in rat and four in human, GSK3 is active even though its repressor AKT1 is also active. PTN11 activation is concordant in human and rat only for 2 of 13 stimuli (NACL and EPGC). Overall, EGF seems to activate AKT1 in rat but PTN11 in human and GSK3 seems active in both organisms. The EGF activation diagram is respected for human (i.e. when AKT1 is inactive, GSK3 is

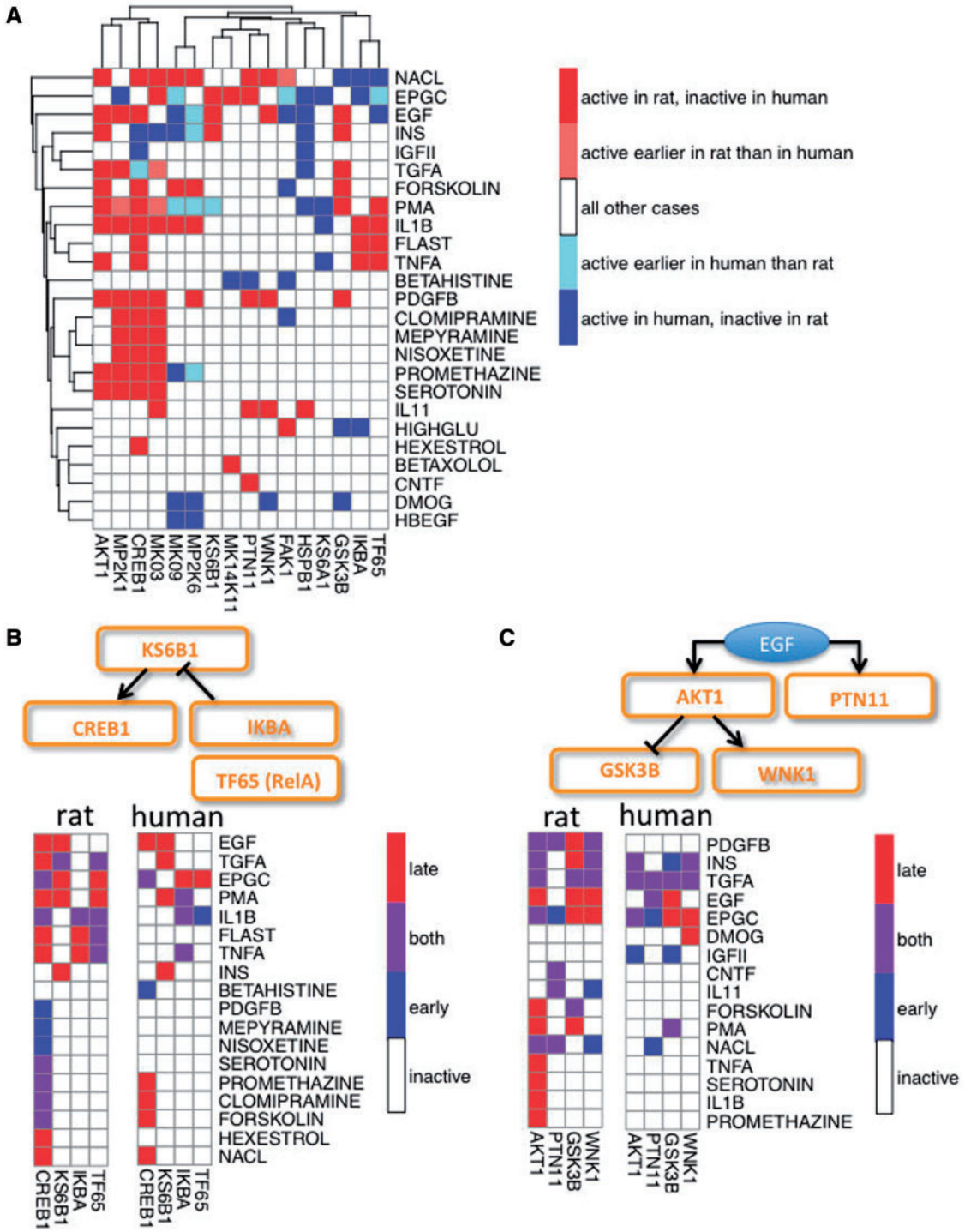

Fig. 3. Comparison of signaling pathways in rat and human. (**A**) Heatmap showing the clustering for directionality of phosphoprotein activation (columns) after a given stimulus (rows) for the two species, showing which phosphoproteins were activated early or late in each species by the different stimuli. Only stimuli with at least one non-zero entry according to Table 2 are shown. (**B** and **C**) Top: in orange are shown potential pathway activation diagrams for phosphoproteins activated by RPKB6S1 (B) and AKT1 (C). Bottom: left heatmaps show the clustering of the rat phosphorylation activation status of the phosphoproteins shown in the diagrams for all active stimuli. Right heatmaps display human phosphorylation activation status of the phosphoproteins shown in the diagrams for stimuli using the same clustering structure obtained from the rat data to ease comparison among species. Only stimuli where activation is present in at least one species are shown. Protein phosphorylation states are defined as inactive, active early (active only at 5 min), active at both time points (active at 5 and 25 min) and active late (active only at 25 min)

active) but not for rat. However, independent of the wiring diagram the activation profile of AKT1 and GSK3 in human and rat are similar.

4 DISCUSSION

With respect to the predictions required in the sub-challenge, results summarized in Table 3 indicate that the naive prediction was already competitive and yielded the best prediction performance of all individual methods. On the other hand, the findings demonstrate that the combination of different methods had the effect of improving the test set performance with respect to some of the criteria. In the comparison across the primary methods shown in Table 4, the ranking of the different choices also confirms that assuming $humP \approx ratP$ yields a decent baseline performance as compared with the NN trained on just phosphorylation data.

It is also interesting to note that team AMG's naive prediction outperformed the LVQ method with respect to the test set B, while LVQ appeared superior in the training set validation. Apparently, the relatively small sample sizes do not allow for more reliable performance estimates by means of the L-O-O method.

The organizers of the challenge compared all the submissions, and the details of this analysis are described in Rhrissorrakrai *et al.* (2015). Like most participants of the sub-challenge, the two top-ranked teams chose to predict protein activation in human exclusively from the rat phosphorylation data. The immediate reaction of the cell in terms of protein phosphorylation can be expected to be similar. However, species-specific details in the regulation processes may be significant along the complex pathways from phosphorylation to gene expression levels. The use of gene expression data is further complicated by high dimension compared with the relatively small number of samples. Both teams decided to avoid this complexity, which, combined with measurement noise and unknown thresholds for gene activation, may contribute to significant prediction errors.

One possible strategy for the inclusion of gene expression data in the analysis would require several steps: from rat phosphorylation levels to rat gene expression data to human gene expression to human phosphorylation levels, with the possibility of significant and unknown systematic and stochastic errors. Alternatively, as explored by team IGB, rat GEx data could be used as additional input to the prediction model directly. Results summarized in Table 4 show that, interestingly, a NN trained on P in combination with gene expression (GEx) performs worse than using rat P as a naive prediction for human P. This is not unexpected, primarily because of overfitting on the numerous GEx features using only a handful of training examples. Additionally, no cross-validation methods were used by Team IGB to specifically avoid overfitting, and therefore the trained model performs well on the training data but not on the test data, for which it performs about as poorly as random and ranks among the worst performing submissions.

The post hoc evaluation of time-specific rat phosphorylation values as naive predictors for human protein activation at the corresponding time points, cf. Table 5, reveals that the agreement is slightly better for the measurement at 25 min. This could simply reflect a more pronounced variability in the early stages of protein activation, or that most of the activation happens between 5 and 25 min.

It is difficult to come to a general conclusion concerning the comparison of signaling pathways and their kinetics in rat and human. Closer inspection of the phosphorylation kinetics revealed significant differences between rat and human with respect to the activation patterns, but pathways that are activated similarly in the two species could also be identified.

The differences in regulation of signaling pathways in these species seem to be stimulus and pathway-dependent, but can be identified from a well-structured training set as shown by the challenge results.

5 CONCLUSION AND OUTLOOK

A detailed comparison and analysis of predictions submitted by 13 different teams is provided in Rhrissorrakrai *et al.* (2015). In general, the challenge results indicate that the stimulus-dependent protein phosphorylation displays significant correlation between rat and human, which facilitates direct inter-species prediction. Indeed, the analyses provided by the best performing teams were purely data driven and based on the rat phosphorylation status only.

In forthcoming projects, more sophisticated classifiers and training schemes should be exploited for the prediction. As just one example, the application of more advanced variants of LVQ using adaptive distance measures appears promising (Bunte *et al.*, 2012; Schneider *et al.*, 2009).

Improved training schemes and careful control of overfitting effects may allow for the beneficial inclusion of gene expression. Similarly, low-dimensional representations of the GEx data or the consideration of ortholog gene sets could be used. The latter strategies have proven useful in the related sub-challenges concerning intra-species predictions of phosphorylation (Dayarian *et al.*, 2014) and inter-species prediction of gene set activation (Hormoz *et al.*, 2014), respectively.

In all approaches presented here, target proteins were considered independently. Correlations or anti-correlations between different phosphoproteins as observed in the training set could prove useful in more sophisticated prediction techniques.

Going beyond a purely data-driven analysis and prediction by taking into account available domain knowledge should provide further insights into the mechanisms that control the activation kinetics and help us better understand the differences and similarities between the two species.

ACKNOWLEDGEMENTS

P.M., R.N. and K.R. helped to edit the manuscript and generate figures. E.B., R.N., P.M. and K.R. helped to develop and organize the challenge. S.H. and A.D. thank the Kavli Institute of Theoretical Physics at UC Santa Barbara, USA, and in particular Boris Shraiman, for support. G.B. thanks the Kavli Institute of Theoretical Physics at UCSB for its support during the early stages of this project, the Jülich Supercomputing Centre at the Forschungszentrum Jülich, Germany, for their support when this research was completed, and the Tata Institute of Fundamental Research, Mumbai, India, for support and hospitality during the writing of the manuscript. M.B. thanks the Institute of Advanced

Studies, University of Birmingham, UK, for a visiting fellowship in the final phase of completing the manuscript. M.D.Z. and S.H. contributed equally to this work.

Funding: S.H., G.B. and A.D. were supported in part by the National Science Foundation under Grant No. NSF PHY11-25915. P.S. and M.Z. were supported by grants from the National Institutes of Health under NIH LM010235, NIH NLM T15 LM07443 and NSF IIS-0513376.

Conflict of Interest: The data and organization of challenge was performed under a joint research collaboration between IBM and Philip Morris International R&D (PMI), and was funded by PMI.

REFERENCES

Baldi,P. and Sadowski,P. (2013) Understanding dropout. *Adv. Neural. Inf. Process. Syst.*, **26**, 2814–2822.

Baldi,P. and Sadowski,P. (2014) The Dropout learning algorithm. *Artificial Intelligence*, **210**, 78–122.

Bergstra,J. *et al.* (2010) Theano: a CPU and GPU math expression compiler. In: *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Austin, TX.

Biehl,M. *et al.* (2007) Dynamics and generalization ability of LVQ algorithms. *J. Mach. Learn. Res.*, **8**, 323–360.

Brodersen,K.H. *et al.* (2010) The balanced accuracy and its posterior distribution. In: *Proceedings of the 2010 20th International Conference on Pattern Recognition (ICPR'10)*. IEEE Computer Society, Washington, DC, USA, pp. 3121–3124.

Bunte,K. *et al.* (2012) Limited rank matrix learning, discriminative dimension reduction, and visualization. *Neural Netw.*, **26**, 159–173.

Cantley,L.C. (2002) The phosphoinositide 3-kinase pathway. *Science*, **296**, 1655–1657.

Davis,J. and Goadrich,M. (2006) The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine learning (ICML'06)*. ACM, New York, pp. 233–240.

Dayarian,A. *et al.* (2014) Sbv Improver sub-challenge 1: learning and predicting phosphorylation levels of upstream effectors in rat lung epithelial cells. *Bioinformatics*, [Epub ahead of print].

Duda,R.O. *et al.* (2001) *Pattern Classification*. 2nd edn. Wiley, New York.

Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.

Gao,Z. *et al.* (2009) Inactivation of NF-κB p50 leads to insulin sensitization in liver through post-translational inhibition of p70-S6K. *J. Biol. Chem.*, **284**, 18368–18376.

Goodfellow,I.J. *et al.* (2013) Pylearn2: a machine learning research library. *arXiv*, e-print 1308.4214.

Hastie,T. *et al.* (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer, New York, NY.

Hinton,G.E. *et al.* (2012) Improving neural networks by preventing co-adaptation of feature detectors. *arXiv*, e-print 1207.0580.

Hormoz,S. *et al.* (2014) Inter-species inference of gene set enrichment from large data sets. *Bioinformatics*, [Epub ahead of print].

Jiang,G.Y. *et al.* (2005) Identification of WNK1 as a substrate of AKT/protein kinase B and a negative regulator of insulin-stimulated mitogenesis in 3T3-L1 cells. *J. Biol. Chem.*, **280**, 21622–21628.

Kohonen,T. (1990) Improved versions of learning vector quantization. In: *International Joint Conference on Neural Networks*. Vol. 1, pp. 545–550.

Kohonen,T. (1997) *Self-Organizing Maps*. Springer, New York, NY.

Rhrissorrakrai,K. *et al.* (2015) Understanding the limits of animal models as predictors of human biology: lessons learned from the sbv IMPROVER Species Translation Challenge. *Bioinformatics*, **31**, 471–483.

Schneider,P. *et al.* (2009) Adaptive relevance matrices in Learning Vector Quantization. *Neural Comput.*, **21**, 3532–3561.

Schneider,P. *et al.* (2010) Hyperparameter learning in probabilistic prototype-based models. *Neurocomputing*, **73**, 1117–1124.

Seo,S. *et al.* (2003) Soft nearest prototype classification. *IEEE Trans. Neural Netw.*, **14**, 390–398.

Schulze,W.X. *et al.* (2005) Phosphotyrosine interactome of the ErbB-receptor kinase family. *Mol. Syst. Biol.*, **1**, 2005.0008.

Xing,J. *et al.* (1996) Coupling of the RAS-MAPK pathway to gene activation by RSK2, a growth factor-regulated CREB kinase. *Science*, **273**, 959–963.