Prediction of Human Protein Kinase Substrate Specificities

Javad Safaei*[1], Ján Maňuch [1,2] , Arvind Gupta[1] , Ladislav Stacho[2] and Steven Pelech[3,4]

[1]Department of Computer Science, University of British Columbia, Vancouver, Canada
[2]Department of Mathematics, Simon Fraser University, Burnaby, Canada
[3]Department of Medicine, University of British Columbia, Vancouver, Canada
[4]Kinexus Bioinformatics Corporation, Vancouver, Canada

Email: Javad Safaei*- jsafaei@cs.ubc.ca; Ján Maňuch - jmanuch@cs.ubc.ca; Arvind Gupta - arvindg@cs.ubc.ca; Ladislav Stacho - lstacho@sfu.ca; Steven Pelech - spelech@kinexus.ca;

*Corresponding author

Abstract

Background: Cell signaling pathways govern basic cellular activities and coordinates cell actions and it is believed that defects in cell signaling pathways are related to diseases such as cancer, autoimmunity, diabetes and at least 400 other human diseases. Therefore, it is critical to define and study cell signaling networks precisely in humans. The major components of cell signaling networks are protein kinases, substrate proteins, phosphatases, SH2 and PTB domain proteins. To simplify the modeling, in this work we limit ourselves on modeling kinase-substrate interactions within the cell, and for each kinase we predict the phosphorylation site specificity based on its catalytic domains.

Results: The proposed method predicts 500 different kinase-domains substrate specificity matrices in 488 kinases (among estimated 518 kinases in human) which is a dramatic progress comparing to the state of the art methods such as NetPhorest which covers 179 kinases in 76 groups and NetworKIN which covers 123 kinases. In terms of accuracy, we compared the predicted matrices with the experimental matrices for the kinases for which experimentally confirmed phospho-peptides are available and obtained very similar results, for instance all the serine, serine-threonine, and tyrosine kinases were correctly identified. Also we observed that the predicted profile matrices are in many cases more accurate than those profile matrices which are computed by aligning the phospho-peptides of each kinase. We also show in the result section that our method has higher accuracy than NetPhorest for predicting the best kinases of each input phospho-peptide.

Conclusions: In this work we presented a new predictor for human kinase specificities that covers all the typical kinases and four atypical kinases in human. We also showed that our predictor is more accurate and has higher coverage than NetPhorest which is one of the state of the art methods. We conclude that our predictor can be used for modeling cell signaling pathway for the other components such as phosphatases, SH2 and PTB domain proteins.

---

1 Introduction

Integrated cell signaling pathways contribute to a complex system of communication that governs basic cellular activities and coordinates cell actions [1]. The ability of cells to perceive and correctly respond to their micro environment is the basis of growth, development, tissue repair, and immunity as well as normal tissue homeostasis. Defects in cell signaling network are related to diseases such as cancer, autoimmunity, diabetes and some 400 other human diseases [2]. Therefore, it is critical to define and study cell signaling networks precisely in humans. The major components of cell signaling networks are protein kinases, substrate proteins, phosphatases, SH2 and PTB domain proteins which create cell phosphorylation network themselves.

Protein kinases transfer the gamma phosphate (PO4) of ATP to hydroxyl (-OH) groups found on amino acids in substrate proteins. Serine (S), Threonine (T) and Tyrosine (Y) represent the three amino acid residues most commonly targeted by these protein kinases [3–5]. Of the 23,000 proteins encoded by the human genome, two-thirds have already been demonstrated to be phosphorylated at over 93,000 phospho-sites[1]. Many of the targets of protein kinases include other protein kinases, and these enzymes can sequentially regulate each other in complex signaling networks. Our knowledge of the architecture of these kinase communications networks, which span from the cell plasma membrane to deep within the nucleus of cells, is very rudimentary. Most of the protein kinases are expressed in each cell in tens of thousands of copies, but a few are very restricted in their cellular expression patterns and have specialized functions. Under 10,000 kinase–substrate phospho-site interacting pairs have been identified empirically, but probably over 10 million exist.

Domains are substrings of protein sequences which can evolve, function, and exist independently of the rest

[1] www.phosphonet.com

of the protein chain. The most common domain in protein kinases is the catalytic domain which carries out the actual phosphorylation of protein substrates. Most of the kinases have only one catalytic domain, while others can have two catalytic domains or no known domain. Throughout the catalytic domain of the kinases *Specificity-determining residues* (SDRs) [2] are distributed which interact with the side chains of amino acid sequences surrounding phospho-sites in substrates [6], which we call a *phospho-site region* or phospho-peptides. Kinase–substrate binding resembles a lock and key model, where a semi-linear phospho-site sequence (surrounding the phospho-site) fits into a kinase active site in the vicinity of the SDRs.

Atypical kinases have completely different structures when compared to the typical protein kinases. They do not possess a catalytic domain similar to those found in typical kinases and appear to have evolved separately. No equivalent catalytic domain has been computed for them using alignment techniques. As a result, SDRs of the atypical kinases have to be searched through the whole surface of the protein, while for typical kinases they are contained within their catalytic domains. In this work, we have predicted the locations of SDRs in 496 human kinase catalytic domains and generated position-specific scoring matrices (PSSM) for each kinase.

The organization of the paper is as follows. In Section 2 we explain previous works related to the prediction of kinase phosphorylation specificities. In Section 3 kinase phosphorylation specificity is mathematically formalized. In Section 4 we propose our prediction algorithm for kinase phospho-site specificities, based on consensus sequences of the phospho-site regions, and in Section 5 we improve the consensus idea by using profile matrices of each kinase, and finally in Section 6 we present our results.

2 Related Works

There are many works which predict kinase specificities for protein substrate recognition and identify potential phospho-sites. These methods are usually based on computing consensus kinase recognition sequences, PSSM matrices or machine learning methods. Scansite [7], artificial neural networks [8] and support vector machines [9], Conditional Random Fields [10], and voting based methods [11] are some of the examples of these works. A survey and comparison of the mentioned prediction methods are represented in [12]. In addition to that, NetworKIN [13] and NetPhorest [14] are two significant efforts for modeling cell phosphorylation networks. NetworKIN uses artificial neural networks and PSSMs to predict kinase domain specificities and uses protein-protein interaction databases such as STRING [15] to increase

[2] SDRs are also called hot-spots in many different works.

the accuracy of the prediction. Those kinases and substrates which are connected directly or indirectly (linked by a short path) in the STRING protein interaction graph are better candidates to be selected in the phosphorylation network. NetworKIN covers only 123 kinases of the 518 known human kinases, since it does not compute kinase phosphorylation specificities for those kinases where there are no experimentally confirmed phospho-sites. NetPhorest has wider coverage compared to NetworKIN and it covers 179 kinases. Similar to NetworKIN, NetPhorest uses a combination of ANN and PSSM matrices for predication, but it puts related kinases in the same group (76 groups in total) and assumes that all kinases in the same group have identical kinase phosphorylation specificities.

All the mentioned methods have two major problems: 1) they can only compute specificity of those kinases which are in the kinase–phospho-site pair databases; and 2) they are highly dependent on the number of confirmed phospho-sites available for each kinase. The training data for all these works is usually borrowed from PhosphoSitePlus[3] and Phospho.ELM [16] which store information on kinase–phospho-site pairs. At this juncture, PhosphoSitePlus has gathered 80,967 phosphorylation sites in 11,134 distinct proteins, while Phospho.ELM has about 42,000 sites in 8,718 proteins. It should be noted that for less than 9,012 of them the phosphorylating kinase is known.

3 Kinase phospho-site specificity

Generally, there is a pattern in the phospho-site regions that a specific kinase phosphorylates. We shall refer to this pattern as its *kinase phospho-site specificity*. This pattern is usually represented by profile (frequency) matrices which show the frequency of each amino acid at each position of the phospho-site region. Optimal consensus sequences are another way of representation of this pattern which rely on the most important amino acids at each position. Other methods such as PSSM matrices and machine learning methods (eg. ANN, HMM) generate a score for a given kinase and a phospho-site region. Higher scores show that the kinase is more likely to phosphorylate that phospho-site. In other words, the score is a measure of kinase phospho-site specificity. To represent the pattern properly at least 9 amino acids (centered at phospho-site with four amino acids to right and left of the site) should be considered [12]. We decided to work with regions of length 15 because by considering six more amino acids we may obtain further information about the specificities for some kinases. Indeed, after computing the profile matrices of several hundred kinases we found that some additional information can be obtained from the added positions -7, 6, -5, 5, 6, and 7 (where 0 is the phospho-site, $-$ means left and + means right of the

[3] http://www.phosphosite.org/

phospho-site). However, increasing the length of the phospho-site regions to more than 15 may lead to the higher noise in the training data, which would make the prediction task harder.

We introduce a new PSSM matrix to predict kinase phospho-site specificities, which is computed in 3 steps described below.

Profile matrix. We first compute the probability matrix, called the *profile matrix* for each kinase. Assume that it is experimentally known that kinase k phosphorylates n different phospho-site regions $\{p_1, p_2, ...p_n\}$ of length 15. The profile matrix P_k of kinase k is 21×15 matrix, where rows represent amino acids (including unknown amino acid 'x') and columns represent positions in the phospho-site regions. The reason of using symbol 'x' is that because in some positions of the primary structure of the proteins the exact amino acid is not known, and in addition to that some phospho-sites are located close to the N-terminus (C-terminus) of the proteins and as a result no amino acid can be considered for the left (right) of the phospho-site region. In both of these cases we use symbol 'x' to create a consistent training set.

Background frequencies of amino acids. Next, we compute the probabilities of each amino acid to appear on the surface of proteins. We call these probabilities *background frequencies of amino acids* and denote them by $B(i)$, where $1 \leq i \leq 21$. To compute background frequency we use all the 93,000 confirmed phospho-site regions in human proteome. The reason is that all of these confirmed regions are on the surface of proteins, and hence, they can be a good sample of the protein surface. By examining the profile matrices of the kinases we have determined that positions -3, -2, 0, and 1 are particularly biased for kinase recognition, since all of them had a very low entropy. Therefore, we excluded these positions for the computation of the background frequency of each amino acid.

PSSM matrix. Now having profile matrix of each kinase and the background frequency of amino acids, the PSSM matrix for kinase k is typically computed using log odds ratio measure:

$$M_k(i,j) = \log \frac{P_k(i,j)}{B(i)}, \tag{1}$$

where $1 \leq i \leq 21$ and $1 \leq j \leq 15$. The problem with this method is that since the profile matrix P_k computed using experimental data contains many zeros, the resulting PSSM matrix M_k has many $-\infty$ values, and consequently, M_k is not smooth enough for the prediction. Various smoothing techniques [17] are applied here to avoid zeros and $-\infty$ values, but we use a different approach which produces better PSSM matrices for prediction:

$$M_k(i,j) = \text{sgn}(P_k(i,j) - B(i)) \cdot |P_k(i,j) - B(i)|^{1.2} \tag{2}$$

The exponent 1.2 was determined experimentally to achieve the best results.

The logic behind this method is similar to log odds ratio. If the probability of amino acid i at position j of profile matrix is bigger than the background frequency of i then that amino acid is a positive determinant, while if it is less than the background frequency it is a negative determinant for the phospho-site region containing i at position j to be recognized by that specific kinase. For a given candidate phospho-site region we are interested to see more positive and less negative determinant amino acids to predict it as a phospho-site.

Score of phospho-site region. Having PSSM matrix M_k for kinase k, we can compute how likely a given candidate phospho-site region $r = r_1 r_2 \dots r_{15}$ is going to be phosphorylated by kinase k. This value is called *kinase specificity score* S and is computed as follows.

$$S(k, r) = \sum_{j=1}^{15} M_k(r_j, j). \tag{3}$$

4 Prediction of PSSM for kinases without substrate data.

In this section, we present our algorithm for prediction of PSSM matrices based on their catalytic domains. The idea is that those catalytic domains in different kinases which have similar SDRs tend to have similar patterns in the phospho-site regions. To quantify the similarity of catalytic domains of kinases we perform multiple sequence alignment (MSA) of catalytic domains using ClustalW algorithm [18]. The result of the MSA is not quite accurate as it has many gaps, therefore, the alignments were manually modified. We perform this alignment on 496 catalytic domains of the typical protein kinases. The length of each kinase catalytic domain after MSA is 247. For 224 domains in the alignment we compute *consensus sequences* using 6,515 confirmed kinase–phospho-site pairs. Figure 2 represents portions of the catalytic domain after MSA of some of the best characterized kinases for which the most phospho-sites have been identified.
To generate the consensus sequence of each kinase, profile matrix of each kinase is computed using the confirmed phospho-site regions of each kinase. For each position in the consensus sequence the amino acids with the maximum probability in that position is selected. If the probability is bigger than 15% then a capital letter is used to represent that amino acid, if it is less than 15% and bigger than 8%, a small letter is used, and if it is less than 8%, symbol 'x' is used in that position of the consensus sequence. 'x' here is a "don't care" letter and it means that any amino acid can appear in that position of the phospho-site region of a kinase. Therefore, those kinases that have more 'x' in their consensus sequence are more general and can phosphorylate more sites than the others. In Figure 1 consensus sequences are shown with the

combination of small and capital letters.

In what follows we use the example in Figure 2 to explain how mutual information and charge information are used to find SDRs on the catalytic domains of the kinases.

Mutual Information. Each position in catalytic domains or consensus sequences can be considered as a random variable which can take 21 different values. Both random variables can take any of the 20 amino acids. In addition, the random variables in domains can also take the gap value $\sim$, while the random variables in consensus sequence can take the unknown value 'x'. In information theory the mutual information of two random variables is a quantity that measures the mutual dependence of the two variables [19]. We can use this measure here to find out which two positions in consensus and catalytic domain are highly correlated. Formally, the mutual information of two discrete random variables X and Y is defined as:

$$I(X,Y) = \sum_{x \in A} \sum_{y \in B} p(x,y) \log \frac{p(x,y)}{p_1(x)p_2(y)},$$

where $p(x,y) = \mathbf{P}(X = x, Y = y)$, $p_1(x) = \mathbf{P}(X = x)$, and $p_2(y) = \mathbf{P}(Y = y)$. The higher mutual information, the more the random variables are correlated. In our context, X is a position in the kinase catalytic domain, Y is a position in the consensus sequence, A is a set of amino acids plus $\sim$ and B is a set of amino acids plus 'x'.

Charge Information. Negatively charged amino acids interact with positively charged, and hydrophobic amino acids with hydrophobic ones. Therefore, if a position in the catalytic domains (see Figure 2) tends to have many negatively charged amino acids and a position in the consensus sequences tends to have more positively charged amino acids, it is likely that these two positions are interacting with each other. Therefore, we define *charge dependency* $C(X,Y)$ of two positions (random variables), one in kinase catalytic domains (X) and the other in consensus sequences (Y), as follows.

$$C(X,Y) = \sum_{i=1}^{n} R(x_i, y_i), \tag{4}$$

where n is the number of kinases with consensus pairs (in our case 224). R is also residue interaction score of two different amino acids, cf. Figure 3, x_i is the amino acid of the i^{th} kinase at position X of the catalytic domain and y_i is the amino acid of the corresponding consensus sequence at position Y. Residue interaction matrix shown in Figure 3 estimates the strength of a bond created between amino acids in the average case independent of their distance. Negatively (positively) charged amino acids repel themselves (score -2 in the interaction matrix R) and they attract positively (negatively) charged amino acids (score +2). Histidine (H) has a smaller positive charge than Lysine (K) and Arginine (R). Therefore,

scores for it are +1 for interacting with negatively charged amino acids and -1 for interacting with positively charged amino acids. Hydrophobic amino acids attract each other (score +2) while they repel both positively and negatively charged amino acids. S, T and Y residues have a weak tendency to bind to each other (score +0.5), while they are completely neutral with the other amino acids (score 0). For all the amino acids discussed so far, it is not relevant whether they are in the kinase catalytic domain or phospho-site region. In both situations the score is the same, which makes the interaction matrix symmetric. However, Glycine (G) is favored to be in the phospho-site region, because it is a small amino acid that creates a pocket on the surface of the region that permits the catalytic domain of the kinase come closer to the region. The reason that we do not consider effect of G in the catalytic domain is that we are unclear about the 3D structure of the most kinase catalytic domains, while phospho-site regions are linear or semi-linear.

If we look at Figure 2 we observe that columns 69, 135, and 161 are quite conserved with negatively charged amino acids. Since at (-3) position of the consensus sequences of the substrates mostly positively charged amino acids (e.g. Arginine (R)) appear, these positions have a high charge dependency score C and are strong candidate positions for interaction with (-3) position of the phospho-site regions. On the other hand, these positions are very conserved and they seem to be uncorrelated with the (-3) position of the phospho-site regions (e.g. when the (-3) position is positively charged or neutral position 69, 135, and 161 are still negatively charged). Therefore, we need a criterion to combine the correlation and charge dependency measures. The following equation combines these two measures.

$$C_c(X,Y) = \sum_{x\in A}\sum_{y\in B} R(x,y)\cdot p(x,y)\log\frac{p(x,y)}{p_1(x)p_2(y)}, \tag{5}$$

where $C_c(X,Y)$ is called *correlation–charge dependency* of two positions X in catalytic domains and Y in consensus sequences.

Using this hybrid criterion $C_c(X,Y)$ in our example, column 120 gets the maximum correlation charge dependency in Figure 2. It is usually preferred that for a particular position in consensus sequences, SDRs in catalytic domain stay near each other, because they can easily interact with that position in consensus sequences. For example, positions 120 and 121 should be preferred to positions 120 and 220. However, in the 3D structure of the protein kinase domain, amino acids that are well separated in the sequence could be situated next to each other. In view of such exceptions, we did not include this preference in our model. Algorithm 1 computes the best SDRs (positions X in the catalytic domain) for each kinase consensus sequence position Y and their interaction probabilities $\mathbf{P}(Y|X) = \frac{\mathbf{P}(X,Y)}{\mathbf{P}(X)}$ using correlation–charge

dependencies.

Algorithm 1 Computing SDRs

Input: 224 human kinase catalytic domains and their consensus sequences. Parameter $m \leq 247$.
Output: SDRs and their interaction probabilities for each position in the phospho-site region.

1: **for** $j \leftarrow 1, 15$ **do**
2: Let Y_j be the j^{th} position in consensus sequences
3: **for** $i \leftarrow 1, 247$ **do**
4: Let X_i be the i^{th} position in catalytic domains
5: Compute $C_c(X_i, Y_j)$
6: **end for**
7: Order positions X_i based on $C_c(X_i, Y_j)$ (decreasingly). Let $Z_{j,k}$ be the k^{th} position in this order.
8: **Output** $Z_{j,1}, \ldots, Z_{j,m}$ as SDRs for position Y_j and interaction probabilities $\mathbf{P}(Y_j|Z_{j,1}), \ldots, \mathbf{P}(Y_j|Z_{j,m})$.
9: **end for**

By examination of the x-ray crystallographic 3D structures of 11 protein kinases co-crystallized with peptide substrates, we determined that usually at most seven SDRs may interact with an amino acid position on the substrate phospho-site region, therefore we set the value m in Algorithm 1 to 7.

Algorithm 2 computes the profile and PSSM matrices for 496 catalytic domains in 484 different human kinases[4] using the SDRs determined by Algorithm 1. The formula in Line 5 of the Algorithm 2 is based on the observation that those interactions which have higher correlation–charge dependency are more important in estimation of profile matrices.

Algorithm 2 Prediction of PSSM matrices of all kinases.

Input: SDRs and interaction probabilities from Algorithm 1 and 496 catalytic domains.
Output: Profile and PSSM matrices of all kinase catalytic domains.

1: ▷ Estimation of the profile matrix of each kinase.
2: **for** $k \leftarrow 1, 496$ **do**
3: **for** $j \leftarrow 1, 15$ **do**
4: ▷ Estimation of interaction probabilities
5: Compute $\mathbf{P}(Y_j|Z_{j,1}, Z_{j,2}, ..., Z_{j,m})$ as $\frac{\sum_{\ell=1}^{m} C_c(Z_{j,\ell}, Y_j)\mathbf{P}(Y_j|Z_{j,\ell})}{\sum_{\ell=1}^{m} C_c(Z_{j,\ell}, Y_j)}$
6: **end for**
7: Store 21×15 computed values in profile matrix P_k
8: **end for**
9: ▷ Computing PSSM matrices.
10: Compute the background frequencies B using the idea mentioned in Section 3.
11: Compute the PSSM matrix of each kinase using Equation (2).

[4]From 518 known human protein kinases, 484 kinases are typical kinases with 496 known catalytic domains and the remaining 34 kinases are all atypical kinases and we have phospho-site specificity data only for four of them.

5 Using profile matrices instead of consensus sequences.

In Section 4 we used phospho-peptide consensus sequences of each kinase to compute correlation charge dependency and SDRs, because it was easier to describe. Another idea is to use profile matrices of each kinase in Algorithm 1 without building consensus sequence. In this strategy we use more information and it can help us for better prediction, while on the other hand it may lead us to overfitted results. In Section 6 we will test both of these algorithms (1. consensus based and 2. profile based), and compare the results. In profile matrix based method the main difficulty is that for the random variable Y_j (column j in the aligned consensus sequences) we do not have the correlated values of the random variable X_i (column i in the aligned catalytic domains). Instead, for each value in X_i we have 21 different amino acid probabilities of Y_j. Assume $a_{k,i}$ is the amino acid in the aligned catalytic domain of the kinase k, also let $p^k_{l,j}$ be the probability of the l^{th} amino acid ($1 \leq l \leq 21$) at position j ($1 \leq j \leq 15$) of the profile matrix of kinase k. Figure 4 represents these notations in a visual manner. Before computing the charge dependency correlation of two columns (or random variables) X_i and Y_j we compute the probability of amino acids in each random variable. $\mathbf{P}(X_i = x)$ is computed by maximum likelihood estimation using $a_{k,i}$ amino acids as follows:

$$\mathbf{P}(X_i = x) = \frac{\sum_{k=1}^{w} \langle x = a_{k,i} \rangle}{w}, \tag{6}$$

where $\langle x = a_{k,i} \rangle$ is the indicator variable taking values ones (when $a_{k,i}$ is equal to x) and zeros otherwise, w is the number of all kinase catalytic domains. $\mathbf{P}(Y_j = l)$ or $\mathbf{P}(Y_j = y)$ is also computed by the following equation:

$$\mathbf{P}(Y_j = y) = \sum_{k=1}^{w} p^k_{l,j}, \tag{7}$$

Similar to the previous section we replace $\mathbf{P}(X_i = x)$ and $\mathbf{P}(Y_j = y)$ with $p_1(x)$ and $p_2(y)$ respectively. $p(x, y)$ is also computed using maximum likelihood estimation as follows:

$$p(x, y) = \frac{\sum_{k=1}^{w} \langle x = a_{k,i} \rangle \cdot p^k_{y,j}}{w}, \tag{8}$$

having $p_1(x)$, $p_2(y)$, and $p(x, y)$ we can now compute the charge correlation dependency using Equation (5) and pick top values for SDRs.

Another modification which should be applied on the consensus method is to change conditional probability of phospho-peptide positions given SDRs (which is shown by $\mathbf{P}(Y_j|Z_{j,l})$) in Line 5 of

Algorithm 2. This probability according to Bayes' theorem equals to $\frac{P(Y_j,Z_{j,l})}{Zj,l}$ and both numerator and denominator can be computed similar to Equations (8) and (6) respectively.

6 Results

In this study we perform four different experiments, the first experiment is to evaluate the accuracy of the predicted profile matrices by consensus and profile based modules of the predictor. The second experiment is to build PSSM based on the predicted profile matrices, use it as a classifier for each kinase and then compute the confusion matrices and accuracy of the predictor. The third experiment is to compare NetPhorest with our predictor based on our kinase phospho-site pairs. Finally the last experiment is to compare NetPhorest and our method with NetPhorest data sets. Each of these experiments are explained thoroughly in the following subsections.

6.1 Comparison of profile matrices

In this section we compare the accuracy of the predicted matrices with the original matrices computed based on experimental kinase-phospho site pairs. For 308 kinases we could gather 9,012 confirmed phospho-sites from PhosphoSitePlus, Phospho.ELM and the scientific literature. The confirmed kinase-phospho site pairs are partitioned into two training and test sets. The test set contains top five kinases that have the most phospho-site data. The reason is that by choosing those kinases in the test set we will be almost confident that the experimentally computed profile matrices are correct and reasonable to compare with the predicted matrices. The training set contains 302 kinases with 6,515 experimentally confirmed phospho-peptides. To start running our predictor on the training data we needed to generate reliable consensus sequences (for the consensus based module of the predictor) of phospho-peptide for each kinase first, therefore we eliminated those kinases having less than 10 phospho-peptides. 224 kinases among all the 302 kinases in the training set had more than 10 phospho-peptides and we could compute 224 consensus sequences for each using the process explained in Section 4. On the other hand, from about 518 kinases in human we gathered 496 catalytic domains in 484 human typical protein kinases and aligned the catalytic domains and used Algorithms 1 and 2 to compute SDRs and profile matrices. To evaluate these predicted matrices, we also computed the profile matrices of 302 kinases in the training set computed by the method described in Section 3 (empirical matrices), and the results were compared using sum of squared differences. Figure 5 illustrates how we set up this comparison, and Figure 6 shows the distribution of these errors. This figure presents the results for the profile matrix based module of the

predictor as well. It is evident that the majority of the predicted matrices are extremely similar to those generated by known substrate alignments. Interestingly the results on the test set are more accurate (with sum of squared error less than 1) than the predicted results on the training set (which can be up to 10 to 15) for both modules. The reason is that for each kinase in the test set there are more experimental peptides, and as a result empirically computed matrices are closer to correct specificity of each kinase, however in the training set there are many kinases with less than 20-30 confirmed phospho-peptides and we may expect the empirically computed matrices are not close the correct profile of each kinase. It is also worth mentioning that although the profile matrix based module is using more information than the consensus module, not only it does not overfit but also it has more accuracy on both test (with total 1.99 sum of squared error (SSE) for all kinases) and training set (with total 494.22 SSE), while as we see in Figure 6 consensus based method has higher errors with total 584.22 SSE for all kinases in training set and total 2.66 SSE for all kinases in the test set.

6.2 Computation of confusion matrices and accuracy

In this experiment we decide to measure the accuracy of each predicted kinase phospho-site specificity for those kinases in the test set. Therefore, we need to build the classifier for each kinase and prepare positive and negative instances for each classifier. We propose to use PSSM matrices of each kinase as a classifier and we use the confirmed phospho-peptides of each kinase in the test set as positive instances. For negative instances unlike [9] and [10], we randomly generate negative instances for each kinase in the test set equal to the number of its positive instances using the uniform distribution. The reason is that if we choose those phospho-peptides which are not experimentally confirmed as phospho-sites, it is probable that in future in lab experiment (e.g. mass spectrometry) they are proved to be positive instance. Afterward, for any given phospho-peptide we compute the score of the PSSM matrix as in (3) and if the score is less than zero we declare that phospho-peptide as negative, otherwise we accept the given phospho-peptide as a candidate phosphorylatable peptide by the kinase where PSSM matrix belongs to. Figure 5 is also helpful to show the flow of the data for preparing positive and negative phospho-peptides for the top five test kinases. For all the kinases in the test set similar results are computed, and the classifiers were successful to identify most of the negative instances (low false positive) a while they were somehow inefficient to identify all the positive instances (high false negative). Approximately 77% accuracy, 60% sensitivity and 95% specificity was computed for all the classifiers in the test set. Figure 7 represents the exact confusion matrix, accuracy, sensitivity and specificity values for each kinase in the test set for both consensus and profile

matrix based sub-modules of our predictor. As we can see the sensitivity for all classifiers in the consensus based method is low, while this disadvatage is elimianated in the profile matrix based method with 10% higher accuracy comparing to the consensus method.

6.3 NetPhorest vs. our predictor

In this part we compare the accuracy of NetPhorest and our method on the same kinase– phospho-site pairs. To fulfill this task we extracted 1978 distinct phospho-peptide–kinase pairs from the total 9,012 pairs discussed at the beginning of this section. This set was used afterward, for measuring the accuracy of each predictor. For each phospho-peptide in this set we stored the best kinase (highest score) suggested by NetPhorest and our method. Afterward, we measured how many of the predicted kinases are matching with the original kinases in the 1978 pairs. Because NetPhorest works based on kinase groups and predicts the best kinase group for the input phospho-peptide, anytime that the original kinase falls into the predicted kinase group we consider it as matching. For instance, if the input pair is <TRKLMEFpSEHCAIIL, TGFbR2> and NetPhorest predicts kinase group ACTR2_ACTR2B_TGFbR2_group for the input phospho-peptide 'TRKLMEFpSEHCAIIL' we accept it as a hit. After running this experiment, we observed that NetPhorest was successful in 72 of the pairs and our proposed method in 82 cases. By this experiment we show that our method outperforms NetPhorest for three reasons: firstly, it has higher matches, secondly it, covers 500 different kinase domains, while NetPhorest matchings are based on 67 kinase groups, thirdly and finally 2497 training phospho-peptide pairs for five kinases PKACa, PKCa, CK2a1, ERK2 and CDK1 are not used in training the classifiers and they are solely kept for testing the algorithm, while NetPhorest uses most of these data to improve its accuracy.

6.4 NetPhorest and our predictor based on NetPhorest confirmed sites

In this part we compare consensus module of the predictor with NetPhorest based on confirmed phospho-peptides exist in NetPhorest database. At this juncture, NetPhorest contains 10,261 confirmed phospho-sites and has 76 specified groups for a total of 179 kinases linked to phosphorylation of 8,746 of those sites. In this dataset, some phospho-sites had more than one kinase phosphorylating them. To compare our predictor with NetPhorest easier we retained only the best kinase for each phospho-site. As a result, the number of kinase–phospho-site pairs was reduced to 6,299. To examine how many of these kinase-phospho-site pairs were consistent with our predictor, we subjected these 6,299 phospho-sites to our predictor algorithm to determine which individual kinases were more likely to phosphorylate these sites.

We ranked 492 protein kinases (488 kinase domains, and 4 atypical kinases for which we had PSSM matrices using phospho-site regions) based on their calculated PSSM scores for each NetPhorest confirmed phospho-site region. It was desirable that the experimentally confirmed kinases for each phospho-site region had high PSSM scores in our predictor. However, we cannot expect these confirmed kinases always have maximum PSSM scores, because although these kinases were experimentally demonstrated to phosphorylate those phospho-sites, it is unclear that they are always the best possible matches. Figure 8 shows that 1058 NetPhorest kinase groups were similarly predicted by our algorithm as the best kinase groups for the specific phospho-peptides, and 651 kinase groups were predicted as the second best kinase groups, etc. On average each NetPhorest kinase family has 3.3 kinases and because our algorithm works based on individual kinases and not a group, we adjusted the ranks and intervals for the results from our algorithm accordingly to provide direct comparison. It is evident that 35 percent of the NetPhorest predicted kinases groups corresponded to the top 10 candidate kinases proposed by our algorithm. Therefore, our predictor had similar prediction accuracy to NetPhorest, but we achieved coverage with three times as many different protein kinases and with individual assignments rather than groups of kinases. This result is also shown in our previous work in BIBM 2010 [20].

Acknowledgment

This work was supported in part by CRD grant from the Natural Sciences and Engineering Research Council of Canada and the MITACS Accelerate Internship Program.

References

1. Kostich M, English J, Madison V, Gheyas F, Wang L, Qiu P, Greene J, Laz TM: **Human members of the eukaryotic protein kinase family**. *Genome Biology* 2002, **3**:research00.
2. Via M: **Kinases: From targets to therapeutics**. *Cambridge Health Institutes Insight Reports* 2003, **1**:1–124.
3. Pelech S: **Kinase profiling: The mysteries unraveled**. *Future Pharmaceuticals* 2006.
4. Pelech S: **Dimerization in protein kinase signaling**. *Journal of Biology* 2006, **5**:12+.
5. Linding R, Jensen LJJ, Ostheimer GJ, van Vugt MA, Jørgensen C, Miron IM, Diella F, Colwill K, Taylor L, Elder K, Metalnikov P, Nguyen V, Pasculescu A, Jin J, Park JGG, Samson LD, Woodgett JR, Russell RB, Bork P, Yaffe MB, Pawson T: **Systematic discovery of in vivo phosphorylation networks.** *Cell* 2007, **129**(7):1415–1426.
6. Saunders NFW, Brinkworth RI, Huber T, Kemp BE, Kobe B: **Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites**. *BMC Bioinformatics* 2008, **9**:245+.
7. Obenauer JC, Cantley LC, Yaffe MB: **Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs.** *Nucleic acids research* 2003, **31**(13):3635–3641.
8. Blom N, Gammeltoft S, Brunak S: **Sequence and structure-based prediction of eukaryotic protein phosphorylation sites**. *Journal of Molecular Biology* 1999, **294**(5):1351–1362.

9. Kim JH, Lee J, Oh B, Kimm K, Koh I: **Prediction of phosphorylation sites using SVMs.** *Bioinformatics* 2004, **20**(17):3179–3184.
10. Dang THH, Van Leemput K, Verschoren A, Laukens K: **Prediction of kinase-specific phosphorylation sites using conditional random fields.** *Bioinformatics (Oxford, England)* 2008, **24**(24):2857–2864, [http://dx.doi.org/10.1093/bioinformatics/btn546].
11. Wan J, Kang S, Tang C, Yan J, Ren Y, Liu J, Gao X, Banerjee A, Ellis LB, Li T: **Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection.** *Nucleic acids research* 2008, **36**(4).
12. Miller ML, Blom N: **Kinase-specific prediction of protein phosphorylation sites.** *Methods in molecular biology (Clifton, N.J.)* 2009, **527**.
13. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P, Yaffe MB, Pawson T: **NetworKIN: a resource for exploring cellular phosphorylation networks.** *Nucleic Acids Res* 2008, **36**(Database issue).
14. Miller MLL, Jensen LJJ, Diella F, Jørgensen C, Tinti M, Li L, Hsiung M, Parker SA, Bordeaux J, Sicheritz-Ponten T, Olhovsky M, Pasculescu A, Alexander J, Knapp S, Blom N, Bork P, Li S, Cesareni G, Pawson T, Turk BE, Yaffe MB, Brunak S, Linding R: **Linear motif atlas for phosphorylation-dependent signaling.** *Science signaling* 2008, **1**(35):ra2+.
15. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C: **STRING 8–a global view on proteins and their functional interactions in 630 organisms.** *Nucleic acids research* 2009, **37**(Database issue):D412–416.
16. Diella F, Gould CM, Chica C, Via A, Gibson TJ: **Phospho.ELM: a database of phosphorylation sites update 2008**. *Nucl. Acids Res.* 2007, :gkm772+.
17. Chen SF, Goodman J: **An empirical study of smoothing techniques for language modeling**. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, Morristown, NJ, USA: Association for Computational Linguistics 1996:310–318.
18. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23**(21):2947–2948.
19. Cover TM, Thomas JA: *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2 edition 2006.
20. Safaei J, Manuch J, Gupta A, Stacho L, Pelech S: **Prediction of human protein kinase substrate specificities**. In *BIBM* 2010:259–264.

Figure 1 - Kinase Consensus Sequences.

Consensus sequences of some of the kinases where we have the most number of confirmed phosphorylation sites (except PDK1 which we show it because it is threonine specific kinase). Phosphorylation sites are marked by bold font at the center of consensus sequence. Number of phosphorylation sites for each kinase which are confirmed experimentally are also showed in the last column.

Figure 2 - Kinase Catalytic Domain Alignment.

Some of the well characterized protein kinases with critical amino acids in their catalytic domains. In the right most column, (-3) position of the consensus sequence of each kinase is shown. Strongly positively charged amino acids (R, K) are represented as blue, weakly positively charged histidine as light blue, strongly negatively charged amino acids (E, D) as red, hydrophobic amino acids (L, V, I, F) as green, and Proline (P) as brown.

Figure 3 - Residue interaction matrix.

Residue interaction matrix R. Rows show the amino acids in the phospho-site regions and columns are amino acids in the catalytic domain of the kinases. Negatively charged amino acids (D, E) are red, positively charged amino acids (K, R, H) are blue, hydrophobic amino acids (F, I, L, V) green, proline as orange, and phosphorylation site residues (S, T, Y) are represented as gray. 'x' also corresponds to the absence of an amino acid, which occurs for phospho-sites located at the N-and C-termini of proteins. This figure was derived from knowledge of the structure and charge of the amino acid side chains.

Figure 4 - Computing correlation charge dependency in profile matrix based module of the predictor.

In the left part of the figure the aligned catalytic domain of the 302 training kinases is shown, and on the right hand side for each kinase the profile matrix is drawn. It is clear that the same columns in all the kinase profile matrices create only one random variable, where its correlation to the aligned catalytic domain should be studied.

Figure 5 - Data and process flow of the experiments.

The figure shows the order of creating the datasets for the experiments and comparison of our predictor results with experimental and current state of the art methods such as NetPhorest and NetworKIN.

Figure 6 - Comparison of predicted vs. experimentally computed profile matrices.

The figure contains four different histograms, where each diagram represent the sum of squared error of the predicted profile matrices and experimentally computed profile matrices based on confirmed phospho-peptide pairs for each kinase. x-axis is the sum of squared error, and y-axis is the frequency or number of matrices which have the specified error. Left histograms are based on consensus based module of the predictor and right histograms related to the profile matrix based module of the predictor. Top histograms show the training set, and the bottom histograms are the results on the five test kinases. Total sum of squared error (SSE) for the consensus based module on training data is 584.89, and on the test set is 2.66, while total SSE for the profile matrix based on training data is 494.22, and on the test set is 1.99.

Figure 7 - Confusion matrices for two modules .

The figure includes two tables and each table represents the classification power or consensus module (on the left) and profile matrix module of the predictor (on the right). In each table confusion matrix is

represented by true positive (TP), false positive (FP), false negative (FN), and true negative (TN). Also accuracy (AC), sensitivity (SE), and specificity (SP) metrics are computed based on the confusion matrices.

Figure 8 - Comparison with NetPhorest predictions.

This table shows how many times NetPhorest kinases groups fall to the rank 1, 2, to rank 10 in our predictor.