Title: Structure-sequence features based prediction of phosphosites of Serine/Threonine Protein Kinases of *Mycobacterium tuberculosis*
Short running title: Structure based phosphosites prediction

Full names of authors: Vipul V. Nilkanth; Shekhar C. Mande[1]

Institutional Affiliation: National Center for Cell Science, S.P. Pune University Campus, Ganeshkhind, Pune - 411 007, India

CORRESPONDING AUTHOR: Dr. Shekhar C. Mande

Correspondence to: Dr. Shekhar C. Mande: Secretary DSIR and Director General, Council of Scientific and Industrial Research, 2 Rafi Marg, New Delhi- 110 001, India

Email: dg@csir.res.in or shekhar@nccs.res.in

ACKNOWLEDGEMENTS: This work is supported by UGC-Fellowship to V.N. We would like to thank Dr. Vinay K. Nandicoori of National Institute of Immunology, New Delhi, for sharing their unpublished data of experimentally identified phosphorylation sites of *Mycobacterium tuberculosis*.

Abstract

Elucidation of signalling events in a pathogen is potentially important to tackle the infection caused by it. Such events mediated by protein phosphorylation play important roles in infection and therefore to predict the phosphosites and substrates of the serine/threonine protein kinases, we have developed a Machine learning based approach and predicted the phosphosites for *Mycobacterium tuberculosis* serine/threonine protein kinases using kinase-peptide structure-sequence data. This approach utilizes features derived from kinase 3D-structure environment and known phosphosite sequences to generate Support Vector Machine based kinase specific predictions of phosphosites making it suitable for prediction of phosphosites of STPKs with no or scarce data of their phosphosites. Support vector machine outperformed the four machine learning algorithms we tried (random forest, logistic regression, support vector machine and k-nearest neighbours) with aucROC value of 0.88 on the independent testing dataset and a ten-fold cross validation accuracy of ~81.6% for the final model. Our predicted phosphosites of *M. tuberculosis* STPKs form an useful resource for experimental biologists enabling elucidation of STPK mediated post-translational regulation of important cellular processes. The training features file and model files, together with usage instructions file, are available at: https://github.com/vipulbiocoder/Mtb-KSPP

Keywords: support vector machine, serine/threonine protein kinases, phosphorylation, *Mycobacterium tuberculosis*, computational prediction

1 Current affiliation: Council of Scientific and Industrial Research, 2 Rafi Marg, New Delhi- 110 001, India

1 INTRODUCTION

Mycobacterium tuberculosis (Mtb) genome encodes Serine/Threonine protein kinases (STPKs), which phosphorylate specific Serine/Threonine residues of substrate proteins, to post-translationally regulate the function of cellular proteins. In eukaryotes, STPKs play essential roles in majority of biological pathways, regulating cellular processes for efficient growth and survival of the cells. Similar to their roles in eukaryotes, these STPKs may facilitate the growth and survival of *M. tuberculosis*, especially in the host, by reprogramming its intracellular signalling network as well as that of the host cell. Understanding the roles of Mtb STPKs might therefore reveal important functions of these in the physiology of Mtb and moreover, in its ability to establish successful infection.

Ser/Thr protein phosphorylation involves recognition of distinct short peptide motifs on substrate proteins by the protein kinases leading to a phosphate moiety being transferred from Adenosine Triphosphate (ATP) to either Serine (Ser) or Threonine (Thr) residues. Recent studies of the STPKs of Mtb have revealed a high degree of structural homology of kinase domains between bacterial and eukaryotic STPKs. Moreover, they share similar mechanisms of substrate recognition and regulation. Mtb genome encodes 11 genes for STPKs, named PknA, PknB, PknD, PknE, PknF, PknG, PknH, PknI, PknJ, PknK, PknL[1]. However, all the in-vivo substrates of these STPKs and thus their role in regulation of cellular processes are not fully characterized experimentally. Some of the available information includes PknA and PknB being involved in regulation of cell shape and possibly cell division[2]. The activity of Mtb STPKs may be regulated in different stages of infection, for example, an in-vitro study has shown that the activity of PknB is reduced during latency and elevated upon resumption of replication[3].

In absence of comprehensive experimental data, computational prediction of substrates of STPKs of *M. tuberculosis* might reveal their functional roles. Several phosphosite prediction tools have been developed to predict phosphorylation sites from amino acid sequences of kinases, or those of shared substrates. These comprise identification of simple motifs using pattern search methods, or the use of machine learning methods such as Artificial Neural Networks (ANN)[4] and Support Vector Machines (SVM)[5]. Examples of such predictive algorithms include Scansite[6], NetPhos and NetPhosK[7,8], KinasePhos[9], DISPHOS[10], GPS[11], NetPhosBac[12] and PredPhospho[13]. Scansite converts the data from oriented peptide library and phage display experiments into a position-specific scoring matrix (PSSM) and generates protein sequence motifs recognized by the respective kinases. The NetPhos 3.1[7,8] server predicts generic and kinase specific serine, threonine or tyrosine

phosphorylation sites in eukaryotic proteins using ensembles of neural networks. KinasePhos[9] builds Profile Hidden Markov_Models (HMM) from each group of known phosphorylation site sequences corresponding to protein kinase classes and predicts phosphorylation sites within given protein sequences. DISPHOS[10] applies position-specific amino acid frequencies and intrinsic disorder score to build models for phosphosites prediction. Musite[14] incorporates three types of features viz, amino acid frequencies, k nearest neighbor (KNN) scores and disorder scores to build SVM models to predict phosphosites in six organisms and 13 kinase-specific phosphosites. All these tools assist in predicting substrates of STPKs.

Most of the computational phosphosite predictors described above are not organism-specific and their ability to predict the phosphosites of bacterial STPKs is not well known since they are based only on known phosphosites of eukaryotic STPKs. Further, most of them do not take into account the phosphosites nor the structural information of the bacterial STPKs. To address this issue, we have developed a Support Vector Machine based tool which uses structural information of Mtb STPKs and sequences of eukaryotic phosphosites to predict phosphosites for Mtb STPKs. The generalized workflow of our method has been shown in Figure 1.

An interesting aspect of signalling mechanisms is that the eukaryotic STPKs exhibit a high degree of cross-regulation. For example, CK2 and AKT cross-regulate their respective functions, through phosphorylation and also by cross-talk among downstream signalling effectors to cause sustained AKT activation[15]. In another case, the kinase PKCε plays a role in lipid-induced insulin resistance through cross talk with p70S6K via shared substrates[16]. Similarly, there may also be cross-talks between Mtb STPKs and among their predicted substrates. Such cross-talks would enable co-ordinated regulation of the network components and the cellular processes driven by them. Consequently, there is a possibility that the STPKs may modulate the activity of hubs of the various cellular processes of Mtb.

In this paper, we present MTB-KSPP (*Mycobacterium tuberculosis* Kinase Specific Phosphosites Predictor), a method which incorporates structural information from peptide-binding cleft of the kinases which enables prediction of kinase-specific phosphosites with limiting amounts of data of experimental phosphosites. Moreover, our method predicts the phosphosites using only seven-residue long phosphopeptide sequences centered on the potential phosphosite. We use the window size of only seven-residues since our method is based on information from the peptide-binding cleft of each kinase and in most of the kinase-peptide complex structures in the Protein Data Bank (PDB)[17] have three-residues on the C-terminal side of the phospho-acceptor (Ser/Thr) residue.

Moreover, we observed that it is usually the three residues on the either side of the central Ser/Thr phospho-acceptor, which make maximum contacts with the kinase residues in the peptide binding pocket. Therefore, to maintain consistency we used hepta-peptide sequences (centered on the known or potential phospho-acceptor residue) for all our analyses.

2 MATERIALS AND METHODS

2.1 Homology modelling of Mtb STPKs

For building homology models, a template structure of Ser/Thr protein kinase was selected with the query STPK domain sequence determined by performing a Protein BLAST (blastp)[18] search using the PDB as the reference database. In case of multiple protein structure hits with the same percentage of identity, protein structure with better resolution was chosen as the template. Among all the selected template STPK structures, one was understandably that of a STPK from Mtb but lacked structural information for the activation segment region of the kinase. For example, Mtb PknB lacking structural information for the activation segment has been shown in Figure 2. We therefore applied loop modelling to generate a favourable conformation of the activation segment region. The input sequence alignment files were generated and used to build models using the command-line version of the software Modeller 9.14[19]. For each STPK, 100 models were built and DOPE score[19] was calculated for them. The details of homology models of Mtb STPKs are listed in Supplementary Table 1. The model to be used for further analysis was selected based on their DOPE score ranking and favourable conformation of the activation loop. The geometry of the best structure model of each Mtb STPK was evaluated using the MOLPROBITY[20] and QMEAN[21] tools for determining structure quality.

2.2 Modelling of substrate peptide into catalytic cleft of Mtb STPKs

In order to model a substrate peptide in the active site of each of the Mtb STPKs, we extracted the structures of eukaryotic STPK-peptide complexes, each with a unique sequence of a bound peptide. Among the many STPK structures available in the PDB, only some of these had a peptide bound at the enzyme's catalytic cleft and there were multiple (redundant) structures corresponding to the same STPK-peptide complex. After removing such redundancy and other structures such as those in complex with other proteins (e.g. Cyclins), 10 unique STPK-peptide complex structures were retained. The selected structures were meticulously analysed by visualizing them in PyMol molecular viewer software. These eukaryotic STPK-peptide complexes were used to calculate the average coordinates of C^{α} and C^{β} atoms of the peptides bound at the catalytic cleft of the eukaryotic STPKs.

We used an experimentally determined structure of phosphorylase kinase (PDB ID: 2PHK)[22] to superimpose each of the models of Mtb STPKs and transferred the average coordinates of the peptides bound at the catalytic cleft of the eukaryotic STPKs to the models of Mtb STPKs. These modified homology models of Mtb STPKs were used to retrieve the interactable kinase residues for the different amino-acid positions of the bound peptide used for the calculation of the various features/variables to be used in building machine learning based models. The eukaryotic STPKs residue sets corresponding to different peptide binding pockets were also retrieved in the same manner.

2.3 Generation of residue contact matrices between substrate and kinases

The site of phosphorylation on the substrate peptide is referred to as P0, while the three residues flanking the phosphorylation site on the N- and C-termini are referred to as P−3, P−2, P−1 and P+1, P+2 and P+3, respectively. The contacting residue-pairs between the kinase and its bound-peptide were determined using four different distance cut-off criteria - the C^{α} atoms of the substrate peptide and any atom of the kinase residues being at a distance ≤5 Å, ≤6 Å, ≤7 Å and ≤8 Å respectively. Similar analysis was done using the C^{β} atoms of the substrate peptide's residues and any atom of the kinase residues. This information from the 'peptide interacting kinase residues' vs 'bound peptide residue position' matrices was then used for calculating the amino-acid pair compatibility based data features[23,24]. Similarly, 'peptide interacting kinase residues' vs 'bound peptide residue position' matrices using different distance cut-offs were made using the homology models of Mtb STPKs.

2.4 Retrieval of peptide sequences for generation of data features

Rich data are available on the sequences that are phosphorylated by eukaryotic kinases either through high-throughput studies, or through individual kinases-substrate interactions. A search was therefore conducted through standard search engines and in publicly available databases to retrieve information about the protein substrates of the eukaryotic kinases and their respective phosphosites. Only those STPKs were considered whose X-ray structures have been solved in complex with a substrate peptide bound at the active site cleft of the STPK. Various online databases such as PhosphositePlus[25], UniProtKb[26] and Phospho.ELM[27] were used to retrieve a total of 2064 non-redundant eukaryotic phosphosites for the ten STPKs for which kinase-peptide complex X-ray structures are available in PDB. Most of these phosphosite sequences were 15 amino acid residues long and centered on Serine/Threonine. From each such phosphosite, the central hepta-peptide stretch was extracted for consideration of positive training data. Another search strategy employed to retrieve the known kinase-specific phosphosites of Mtb STPKs through UniProtKb[26] and published literature resulted in the identification of 221 phosphosites which were pooled with the 2064 phosphosites of eukaryotic STPKs making the total number of phosphosites used to 2285. The

hepta-peptide sequences to be used for negative data generation were compiled by randomly selecting Ser/Thr-centered hepta-peptides, other than known eukaryotic STPK phosphosite/s in the respective protein sequences. The Ser/Thr-centered hepta-peptide sequence stretches from all the proteins of the Mtb proteome[26] were extracted to develop the input features set. The information related to the known phosphosites used in this study is reported in Supplementary file S1.

2.5 Generation of training dataset (positive and negative instances) and prediction input datasets

Using amino-acid statistical pair potentials, amino-acid pair compatibility matrices and the 'peptide interacting kinase residues' vs 'bound peptide residue position' matrix for each eukaryotic STPK, the kinase residue/s-peptide residue pair-potential and compatibility values for each peptide position of the known phosphosite sequences were derived as described below.

The method for determining contact residues between STPKs and substrate peptide residues is already described above. Three distance cut-off ranges (≤6 Å, ≤7 Å and ≤8 Å) were finally used for each of the six residues of the substrate peptide thereby resulting in 6*3 *i.e.* 18 different data feature values for each hepta-peptide. Similarly, using C^{β} atoms of the substrate peptide residues as a reference, another 18 different data feature values for each hepta-peptide were derived. Combining both, 36 data feature values were derived for each type of amino-acid pair-potentials and amino-acid residue pairs compatibility matrices for every hepta-peptide sequence. This resulted in a total of 144 data feature values for each hepta-peptide to be used as positive dataset features in training the Machine Learning model. Sequences for the negative training dataset were derived by randomly choosing Ser/Thr-centered hepta-peptides, other than the known phosphosite/s, from the substrate proteins. The features for the negative dataset and prediction input dataset (Ser/Thr-centered hepta-peptides from the sequences of Mtb proteins) were derived in the same way using the hepta-peptide sequences from the respective datasets.

2.6 Calculation of the intrinsic disorder

To calculate the intrinsic disorder, the sequences of all Mtb proteins and those of substrate proteins of each of the various eukaryotic kinases under study were downloaded from UniProtKb. These sequences were individually input into a stand-alone version of IUPRED 1.0[28] to derive Intrinsic Disorder (IDR) value for each amino acid residue of a given sequence. Further, using locally written Python programs, 'Average Intrinsic Disorder' values for seven-residue long peptide segments over the entire sequence were derived by iteratively sliding the seven-residue window by one position. The average IDR values of the phosphosites of eukaryotic STPKs were then retrieved and used as a positive data feature. The negative data feature comprised of the average IDR values corresponding

to the sequences of the Negative dataset. Also, the average IDR values of the prediction input sequences dataset from *M. tuberculosis* were retrieved which was used as feature in the input data for predictions.

2.7 Features selection and model optimization

All the training features were compiled in a single file and scaled in a -1 to +1 range. To select the most predictive features, feature selection was carried out in LibSVM using Fischer score (F-score) which is a simple and effective criterion to measure the discrimination between a feature and the label. This method assigns a weight to each feature and ranks the features accordingly. All of the 145 features were found to have predictive value after feature selection process. Grid optimization method of LibSVM was used for deriving optimal values of cost and gamma parameters to build optimal performance SVM models. A generalized model was built to predict the phosphosites of Mtb STPKs. Separate models were built for deriving the phosphosite predictions for PknA and PknK due to lack of kinase interactable residues data at one of the peptide positions. These final models were used in the machine learning based predictions of phosphosites for the 11 STPKs from Mtb.

2.8 Evaluation of model performance

The predictive performance of the trained models was evaluated using the following metrics used in machine learning.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall/Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$F1\,score = \frac{2\,TP}{2\,TP + FP + FN}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

where TP, TN, FP, and FN represent the number of true positives, true negatives, false positives and false negatives, respectively. The Receiver-Operating Characteristic (ROC) curves and precision-recall curves were plotted and the Area Under the Curve (AUC) values were calculated as the measures to evaluate and compare the predictive performance of our method with the NetPhos 2.0

and NetPhosBac STPK phosphosites prediction methods.

2.9 Deriving the preferred sequence motifs for Mtb STPKs

Using the large number of predicted sequences we attempted to extract information regarding the compatibility of a substrate peptide to be recognized by the cognate kinase. This was done first by deriving the consensus sequences by calculating the statistical significance for a residue at each aligned position of predicted phosphosites and compared with random unphosphorylated hepta-peptides. We then analysed the consensus sequence in the context of the 3D model of each Mtb STPK. Two kinds of 7-residue motifs were attempted to be understood - the one which each kinase strongly prefers as its substrate, and the other which is strongly disallowed. Using the predicted high-confidence (prediction probability estimate ≥ 0.99) phosphosites and random unphosphorylated hepta-peptides we developed hepta-peptide sequence motifs specific for each Mtb STPK. The predicted consensus sequence motifs specific for each Mtb STPK were developed using the sequence logo tool 'Two Sample Logo'[29]. This tool calculates the statistical significance for occurrence of a residue at each position in the aligned groups of sequences, where the null hypothesis is that the residue is generated according to the same distribution in both positive and negative samples.

2.10 Gene ontology enrichment analysis

The predicted substrates of Mtb STPKs were used to investigate enrichment of cellular processes among them. We used the PANTHER[30,31] web-tool which uses Gene Ontology (GO) terms (cellular processes) of the proteins of Mtb and runs Fischer's Exact test on a query list of proteins to identify the statistically enriched GO terms among the query list. The GO defines concepts/classes used to describe gene/protein function, and relationships between these concepts. It classifies functions along three aspects namely, molecular function (MF), cellular component (CC) and biological process (BP). The gene ontology analysis would shed light upon the physiological changes and the infection stages at which each of the kinases plays an important role.

2.11 Validation using the known phosphosites

To perform a validation of our predicted phosphosites we data-mined and compiled the experimentally known phosphosites of Mtb STPKs from the published literature. For this, we carried out a PubMed search of articles reporting the identification of phosphosites of Mtb STPKs in Mtb proteins. Keywords such as 'Mycobacterium tuberculosis + phosphosite', 'Mycobacterium tuberculosis + phosphorylation', etc were used to retrieve the articles reporting Ser/Thr phosphorylation events in Mtb proteome. The articles included STPK specific sites and also the

phosphoproteomic studies which usually do not pinpoint on the underlying STPKs. We found a total of 221 phosphosites for the ten STPKs of Mtb. We did not find any phosphosites reported for PknI. We also found another 1030 non-kinase specific Mtb phosphosites from three published phosphoproteomic studies[32–34] and one yet unpublished phosphoproteomic study (Vinay Nandicoori, personal communication). After combining the above two sets of phosphosites and removing the redundant sites we obtained 972 known phosphosite sequences in the Mtb proteome which had three residues upstream and downstream the phospho-acceptor serine/threonine residue and these were used to validate our predicted phosphosites.

3 RESULTS

3.1 Construction of machine learning models and features selection

The protein structure-sequence features found to have a predictive value in identification of the phosphosites of each kinase are amino-acid statistical pair potentials[23], amino-acid residue pairs Hydrophobe compatibility, Charge compatibility and Size compatibility values[24]. The phosphorylation site region of the STPK substrate proteins is usually disordered[35,36]. Therefore, we used average intrinsic disorder of the hepta-peptides encompassing the phosphorylation sites as a data feature for the machine learning. A total of 145 features, namely 36 each for four different types of amino-acid pair compatibility matrices and one feature of the average Intrinsic Disorder of the hepta-peptide, were derived for positive dataset (phosphosite hepta-peptide set); negative dataset (randomly chosen Ser/Thr-centered hepta-peptides set) and predictions input datasets i.e. (Ser/Thr-centered hepta-peptides from the sequences of Mtb proteins).

Various kernels of LibSVM were tried out for model building among which the Radial Basis Function (RBF) kernel resulted in SVM models with the highest prediction accuracy after ten-fold cross-validation evaluation of model. After compiling data of all feature types, feature selection procedure was carried out in LibSVM and all the 145 features were found to contribute in enhancing the predictive performance as the highest accuracy model was generated using all the 145 features. We developed an independent test dataset consisting a set of 3582 sequences for training a model and a testing dataset of 988 to evaluate the predictive performance of our methodology by calculating performance metrics such as Accuracy, F1-score, MCC score, Precision, Recall and Specificity for the model as shown in Table 1. This dataset contained 505 phosphosites and 483 non-phosphosites. We evaluated four different machine learning algorithms namely, Random Forest, Logistic Regression, Support Vector Machine (SVM) and K-nearest neighbours by receiver operator characteristic (ROC) curve analysis, to select the most predictive algorithm for our training data features. SVM outperformed all the others with an AUC-ROC value of 0.88 while Random Forest and Logistic Regression both had the next best value AUC-ROC of

0.87 as shown in Figure 3. The final model was built by using 4424 phosphosite sequences (2212 phosphosites and 2212 non-phosphosites) and evaluation of the final SVM model by 10-fold cross-validation approach resulted in 81.6% prediction accuracy. The receiver operator characteristic (ROC) curve analysis as shown in Figure 3 led to an AUC (area under curve) value of 0.88. The precision-recall curve analysis also showed a high AUC value of 0.88 for SVM based model (Supplementary Figure 1). These characteristics therefore suggested that the model has a high value of prediction.

3.2 Benchmarking the performance with other phosphosite prediction tools

We compared the performance of MTB-KSPP with other tools available for prediction of phosphosites of STPKs. Most of the tools available for phosphosite predictions have been developed for specific eukaryotic STPKs or a limited number of families of eukaryotic STPKs and only one tool - NetPhosBac – is available for prediction of phosphosites of bacterial STPKs. Therefore, most of these could not be used to compare the performance of MTB-KSPP, which predicts phosphosites in a kinase independent manner. Consequently, we selected NetPhos 2.0 which is a general predictor of the phosphosites of STPKs and NetPhosBac – a tool developed for the prediction of phosphosites of bacterial STPKs to compare against the predictive performance of MTB-KSPP. The independent test datasets made were used to evaluate the predictive performance of our tool with other tools using ROC curve analysis and precision-recall curve analysis. Since the other tools do not have option for customized model training, we used their pre-trained models available as online web-servers to derive the predictions on the testing set. The ROC curves and the precision-recall curves were plotted in Scikit-learn[37] (Figure 4 and Supplementary Figure 2, respectively) using different thresholds of the scores provided by each method and the AUC (area under the curve) and the average precision values were calculated for the ROC and precision-recall curves. MTB-KSPP performed significantly better than the other two tools used in this benchmarking with the aucROC values of 0.56 NetPhosBac, 0.71 NetPhos 2.0, and 0.88 for MTB-KSPP (Figure 4). Similarly, the precision-recall curve of MTB-KSPP shows better predictive performance with the average-precision (AP) value of 0.88 compared with 0.71 for NetPhos 2.0, and 0.57 for NetPhosBac 1.0 (Supplementary Figure 2). Our method shows an overall better predictive performance compared to these tools based on other evaluation metrics as shown in Table 2.

3.3 Predictions of phosphosites and substrates of Mtb STPKs

A very high number of sites in Mtb proteome were predicted as potential phosphosites for Mtb STPKs as shown in Table 3. To reduce the number of predictions and retain only those the hepta-

peptide sites which have a very high chance of being true phosphosites of each kinase we filtered our predictions based on the probability estimate values of SVM predictions. The number of predicted high-confidence (probability estimate ≥ 0.99) phosphosites and substrates varied for different Mtb STPKs as shown in Table 3. The highest number of phosphosites were predicted for PknD with 4767 phosphosites belonging to 2419 protein substrates. Whereas, PknG has only 09 predicted phosphosites falling into identical number of substrates which is the least among all Mtb STPKs. The reason for this variability in the predicted number of substrates is not clear at this stage. The predicted phosphosites (with probability estimate ≥ 0.80) in Mtb proteins are listed in Supplementary file S2.

3.4 Gene Ontology (GO) enrichment analysis

The GO enrichment analysis on the predicted substrates of four kinases, viz, PknB, PknD, PknF and PknL based on Fisher's exact test with Bonferroni correction $p < 0.05$ shows enrichment of important biological processes vital for the growth and survival of the pathogen. The predicted substrate proteins of other STPKs did not show enrichment of any biological processes. For example, substrates of PknD and PknL show 2.7-fold and 3.3-fold enrichment, respectively in fatty acid biosynthesis and 1.5-fold and 1.7-fold enrichment, respectively in amino-acid metabolism. PknD substrates also show 1.6-fold enrichment enrichment in lipid metabolism and PknL substrates show a 2.5-fold enrichment in DNA metabolic process thus highlighting the plausible roles of these kinases in modulation of activities of these important cellular processes. The predicted protein substrates were also found to be enriched for generic processes including primary metabolic process (GO:0044238), metabolic process (GO:0008152) and cellular amino acid metabolic process (GO:0006520). The details of the enrichment scores and the concerned GO terms of the enriched biological processes of the Mtb STPKs are listed in Supplementary file S3.

3.5 Preferred sequence motifs for Mtb STPKs

The consensus sequence motifs derived from all the predicted substrates of the kinases reflect the preference of specific amino acids by the kinases at various peptide positions. Kinases PknB, PknD, PknE, PknJ and PknL have a large number of predicted substrates, which is explained by a relaxed preference (many amino acids preferred at multiple positions) at many positions, on the other hand, kinases such as PknA, PknG and PknI had only a few predicted phosphosite substrates which is supported by their strict requirement/allowance of only one or two amino acids at various positions of the peptide. For example, consensus motif of PknG (as shown in Figure 5), which has only 09 predicted high confidence phosphosites, shows that it strictly requires Tryptophan at +3 position, and a simultaneous strong depletion or a lack of Arginine at -1 position and lack of Glycine at +2

position. The consensus motif of PknB (Figure 5) shows the N-terminal first position to be highly enriched with the acidic residues Aspartate and Glutamate and depletion of the basic residues Arginine and Lysine. Similarly, in consensus motif of PknJ, the C-terminal end (+3 position) shows high enrichment for the acidic residues Aspartic acid and Glutamic acid and depletion in proportion of basic residues Arginine and Lysine. It is interesting to note that consensus motifs of most STPKs could be derived based on the predicted substrate sequences. We believe therefore, that the predictions of substrates will have a high value in analysing the biological roles of these kinases.

3.6 Structural evidences for specific interactions of consensus peptide residues with Mtb STPKs' residues

Using the sequences of predicted high confidence phosphosites with prediction probability estimate ≥ 0.99 we found the presence of consensus sequence motifs for Mtb STPKs and observed that these motifs have considerable difference with each other as seen from their sequence logos (Figure 5). The consensus sequence as obtained by comparison of predicted substrates can be easily rationalized based on the three-dimensional structures of Mtb STPKs. For example, in the PknB substrate peptide predictions the consensus sequence show strong preference for an acidic residue at the -3 position. Complementary to this peptide site, the binding pocket of the kinase shows presence of the basic amino acid Arg-101 (Figure 6) in the crystal structure of PknB (PDB ID: 1O6Y), which is within a distance of <6 Å of the C^{α} atom of the N-terminal residue (-3 position) of the bound peptide. Similarly, at the same position in the consensus sequence for PknB substrate peptide, depletion of basic residues is also seen. Thus, preference of the PknB is reflected in the consensus motif derived from predicted phosphosites for PknB.

Another illustration of this relates to the PknJ substrate peptide predictions, where the consensus sequence shows strong preference for acidic residues at the +3 position. Complementary to this peptide site, Arginine-45 and Arginine-150 (Figure 6) are present in the binding pocket of PknJ (Figure 6), which may potentially interact with the acidic residue of the bound peptide at its C-terminal end (+3 position). Similarly, this site shows depletion of basic residues in the consensus motif. Therefore, as expected, in the consensus peptide derived from the predicted phosphosites for PknJ, the preferred residues by this binding pocket of PknJ are the acidic amino acids Aspartic acid and Glutamic acid.

3.7 Validation of the predicted phosphosites

We validated the high-confidence predicted phosphosites for Mtb STPKs by analysing their overlap with already known Mtb phosphosites compiled from the previous studies on Mtb Ser/Thr protein phosphorylation. Predictions for the kinases PknB, PknD, PknE, PknF and PknL could pick 36, 48,

33, 17 and 43 phosphosites respectively, from the already known set of 1182 phosphosites in the Mtb proteome. The validation results for other STPKs are listed in Table 4.

DISCUSSION

Our analysis shows that amino acids compatibility and pair potentials based features have high predictive value in prediction of phoshorylation sites. Our method not only selects the position-specific features of phosphosite sequence fragments but also intrinsically captures the binding affinity of various peptide-binding pocket positions in the kinase active site. While the phosphosite's average intrinsic disorder feature can capture the structural properties of phosphosite sequence fragments.

The experimental data on phosphorylation of Mtb proteins is inadequate and thus a computational prediction approach used in this study will aid in identification and validation of additional phosphosites in the Mtb proteome. In addition to the known phosphosites of the eukaryotic STPKs used to derive features, we also included the limited number of experimentally identified phosphosites of Mtb STPKs to generate a machine learning model for prediction of phosphosites of Mtb STPKs. Due to bias of experimental techniques of phosphosite identification towards protein which have higher in-vivo abundance, known phosphosites are also potentially biased for proteins of higher abundance and lack information on phosphosites of low abundance proteins or proteins which may express in only specific environmental conditions. We have made use of structural information of Mtb STPKs to predict phosphosites since training the SVM model only on eukaryotic phosphosites could not efficiently predict the limited number of known phosphosites of Mtb STPKs. Our analysis suggests that structural models of kinases can aid in predictions of post-translational modifications (PTMs) especially for the kinases which lack or have limiting amount of experimentally known phosphosites. Therefore, MTB-KSPP can be adopted to predict phosphosites across different species and other types of PTM.

It is tempting to speculate that kinases such as the PknG for which only nine phosphosites were predicted in the Mtb proteome may have a role in phosphorylating only the host proteins, with little effect on the Mtb proteins. We observed that a distinct motif emerges out the predicted phosphosites of each Mtb STPK which signifies the amino acids preferred by these kinases at their different peptide binding pockets. This finding could provide important insights into the mechanism of substrate recognition by these kinases and may also shed light upon the functional modules regulated by these kinases both individually and in combination with other kinases. The results from our gene ontology analysis of the predicted substrates of Mtb STPKs indeed show that cellular processes fatty acid synthesis, lipid metabolism and DNA metabolism which crucial for infection are regulated by few of the Mtb STPKs. A further thorough investigation into the structure and

dynamics of the signalling network and regulatory network in context of gene expression changes across various physiological conditions would unravel more details about the mechanisms employed by these STPKs for efficient modulation of cellular physiology which assists the pathogen in adapting itself in harsh environmental conditions such as, hypoxia and high acidity encountered inside the macrophages. Our phosphosite predictions could efficiently retrieve several of the known phosphosites of Mtb STPKs which highlights the efficiency of our approach to predict new peptide sites in Mtb proteins that can be potentially phophorylated by one or more of the Mtb STPKs. The outcomes from our Mtb STPKs phosphosite prediction and analysis may form a valuable resource for the researchers working on Mtb and may provide starting points in investigating various aspects of Mtb infection and its persistence.

Supporting information: Supporting data file and supplementary data files have been made available at Proteins journal website.

Conflict of Interest: The authors declare that there are no conflicts of interest.

REFERENCES

1. Av-Gay Y, Everett M. The eukaryotic-like Ser/Thr protein kinases of Mycobacterium tuberculosis. *Trends in Microbiology*. 2000;8(5):238-244. doi:10.1016/S0966-842X(00)01734-0

2. Kang C-M, Abbott DW, Park ST, Dascher CC, Cantley LC, Husson RN. The Mycobacterium tuberculosis serine/threonine kinases PknA and PknB: substrate identification and regulation of cell shape. *Genes Dev*. 2005;19(14):1692-1704. doi:10.1101/gad.1311105

3. Ortega C, Liao R, Anderson LN, et al. Mycobacterium tuberculosis Ser/Thr protein kinase B mediates an oxygen-dependent replication switch. *PLoS Biol*. 2014;12(1):e1001746. doi:10.1371/journal.pbio.1001746

4. Dreyfus SE. Artificial neural networks, back propagation, and the Kelley-Bryson gradient procedure. *Journal of Guidance, Control, and Dynamics*. 1990;13(5):926-928. doi:10.2514/3.25422

5. Cortes C, Vapnik V. Support-Vector Networks. *Mach Learn*. 1995;20(3):273-297. doi:10.1023/A:1022627411411

6. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*. 2003;31(13):3635-3641. doi:10.1093/nar/gkg584

7. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*. 1999;294(5):1351-1362. doi:10.1006/jmbi.1999.3310

8. Blom N, Sicheritz-Pontén T, Gupta R, Gammeltoft S, Brunak S. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*. 2004;4(6):1633-1649. doi:10.1002/pmic.200300771

9. Huang H-D, Lee T-Y, Tzeng S-W, Horng J-T. KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic Acids Res*. 2005;33(Web Server issue):W226-229. doi:10.1093/nar/gki471

10. Iakoucheva LM, Radivojac P, Brown CJ, et al. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*. 2004;32(3):1037-1049. doi:10.1093/nar/gkh253

11. Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X. GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res*. 2005;33(Web Server issue):W184-187. doi:10.1093/nar/gki393

12. Miller ML, Soufi B, Jers C, Blom N, Macek B, Mijakovic I. NetPhosBac - a predictor for Ser/Thr phosphorylation sites in bacterial proteins. *Proteomics*. 2009;9(1):116-125. doi:10.1002/pmic.200800285

13. Kim JH, Lee J, Oh B, Kimm K, Koh I. Prediction of phosphorylation sites using SVMs. *Bioinformatics*. 2004;20(17):3179-3184. doi:10.1093/bioinformatics/bth382

14. Gao J, Thelen JJ, Dunker AK, Xu D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Mol Cell Proteomics*. 2010;9(12):2586-2600. doi:10.1074/mcp.M110.001388

15. Ruzzene M, Bertacchini J, Toker A, Marmiroli S. Cross-talk between the CK2 and AKT signaling pathways in cancer. *Adv Biol Regul*. 2017;64:1-8. doi:10.1016/j.jbior.2017.03.002

16. Gassaway BM, Petersen MC, Surovtseva YV, et al. PKCε contributes to lipid-induced insulin resistance through cross talk with p70S6K and through previously unknown regulators of insulin signaling. *Proc Natl Acad Sci USA*. 2018;115(38):E8996-E9005. doi:10.1073/pnas.1804379115

17. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235-242. doi:10.1093/nar/28.1.235

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

19. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics*. 2016;54:5.6.1-5.6.37. doi:10.1002/cpbi.3

20. Williams CJ, Headd JJ, Moriarty NW, et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Sci*. 2018;27(1):293-315. doi:10.1002/pro.3330

21. Benkert P, Biasini M, Schwede T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics*. 2011;27(3):343-350. doi:10.1093/bioinformatics/btq662

22. Lowe ED, Noble ME, Skamnaki VT, Oikonomakos NG, Owen DJ, Johnson LN. The crystal structure of a phosphorylase kinase peptide substrate complex: kinase substrate recognition. *EMBO J*. 1997;16(22):6646-6658. doi:10.1093/emboj/16.22.6646

23. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*. 1999;8(2):361-369. doi:10.1110/ps.8.2.361

24. Biro JC. Amino acid size, charge, hydropathy indices and matrices for protein structure analysis. *Theor Biol Med Model*. 2006;3:15. doi:10.1186/1742-4682-3-15

25. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res*. 2015;43(Database issue):D512-520. doi:10.1093/nar/gku1267

26. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45(D1):D158-D169. doi:10.1093/nar/gkw1099

27. Dinkel H, Chica C, Via A, et al. Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res*. 2011;39(Database issue):D261-267. doi:10.1093/nar/gkq1104

28. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*. 2005;21(16):3433-3434. doi:10.1093/bioinformatics/bti541

29. Vacic V, Iakoucheva LM, Radivojac P. Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics*. 2006;22(12):1536-1537. doi:10.1093/bioinformatics/btl151

30. Thomas PD, Campbell MJ, Kejariwal A, et al. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res*. 2003;13(9):2129-2141. doi:10.1101/gr.772403

31. Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res*. 2010;38(suppl_1):D204-D210. doi:10.1093/nar/gkp1019

32. Fortuin S, Tomazella GG, Nagaraj N, et al. Phosphoproteomics analysis of a clinical Mycobacterium tuberculosis Beijing isolate: expanding the mycobacterial phosphoproteome catalog. *Front Microbiol*. 2015;6:6. doi:10.3389/fmicb.2015.00006

33. Verma R, Pinto SM, Patil AH, et al. Quantitative Proteomic and Phosphoproteomic Analysis of H37Ra and H37Rv Strains of Mycobacterium tuberculosis. *J Proteome Res*. 2017;16(4):1632-1645. doi:10.1021/acs.jproteome.6b00983

34. Gil M, Lima A, Rivera B, et al. New substrates and interactors of the mycobacterial Serine/Threonine protein kinase PknG identified by a tailored interactomic approach. *J Proteomics*. 2019;192:321-333. doi:10.1016/j.jprot.2018.09.013

35. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic Disorder and Protein Function. *Biochemistry*. 2002;41(21):6573-6582. doi:10.1021/bi012159+

36. Johnson LN, Lewis RJ. Structural basis for control by phosphorylation. *Chem Rev*. 2001;101(8):2209-2242.

37. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*. 2011;12:2825−2830.

Tables:

Table 1. Evaluation metrics of our SVM based method on the independent testing dataset

Performance evaluation measure	Value
Accuracy	80.3%
Precision	0.82
Recall/Sensitivity	0.79
Specificity	0.82
F1-score	0.803
MCC-score	0.606
AUC-ROC	0.88
Average Precision: Precision-Recall	0.86

Table 2. Evaluation metrics for comparing our method with other phosphosite prediction tools

Method/tool	Accuracy (%)	F1-score	MCC-score	Precision	Recall	Average precision
Mtb-KSPP	80.3	0.803	0.606	0.82	0.79	0.86
NetPhos2.0	65.9	0.69	0.32	0.64	0.75	0.76
NetPhosBac1.0	53.7	0.42	0.093	0.585	0.33	0.63

Table 3. Number of predicted high-confidence (probability >= 0.99) phosphosites and substrates for Mtb STPKs

Mtb STPK	No. of predicted phosphosites	No. of predicted protein substrates
PknA	53	52
PknB	1949	1434
PknD	4767	2419
PknE	1078	892
PknF	951	790
PknG	09	09
PknH	877	746
PknI	123	116
PknJ	1940	1435
PknK	944	800
PknL	4209	2353

Table 4. Overlap of predicted phosphosites of Mtb STPKs with experimentally known phosphosites

Mtb STPK	No. of predicted phosphosites	No. of known phosphosites in predicted phosphosites
PknA	53	0
PknB	1949	36
PknD	4767	48
PknE	1078	33
PknF	951	17
PknG	09	0
PknH	877	12
PknI	123	03
PknJ	1940	12
PknK	944	15
PknL	4209	43

Figure legends:

Figure 1. Framework of the structure-sequence features based phosphosites predictor MTB-KSPP. The input features is a matrix of 4570 * 145 (no. of features) generated by deriving mean value for each peptide-residue position for the hepta-peptide sequences centered at the phosphosites and non-phosphosites (4570 training instances). Residue pairs list for each peptide position were formulated from hepta-peptides and kinase-peptide interacting residues matrix of corresponding STPK for which the amino-acid pair potentials and amino-acid pair compatibility matrices were derived. Also, the average Intrinsic disorder value was calculated for each hepta-peptide training instance.

Figure 2. Comparison of kinase structures from Eukaryotes and Prokaryotes. The eukaryotic serine/threonine protein kinase Phosphorylase kinase (in gray colour) (PDB ID: 2PHK) superposed with PknB (in purple colour) (PDB ID: 1O6Y), a serine/threonine protein kinase from *Mycobacterium tuberculosis*. The Activation Loop part missing in PknB's structure is highlighted in purple colour in the phosphorylase kinase structure.

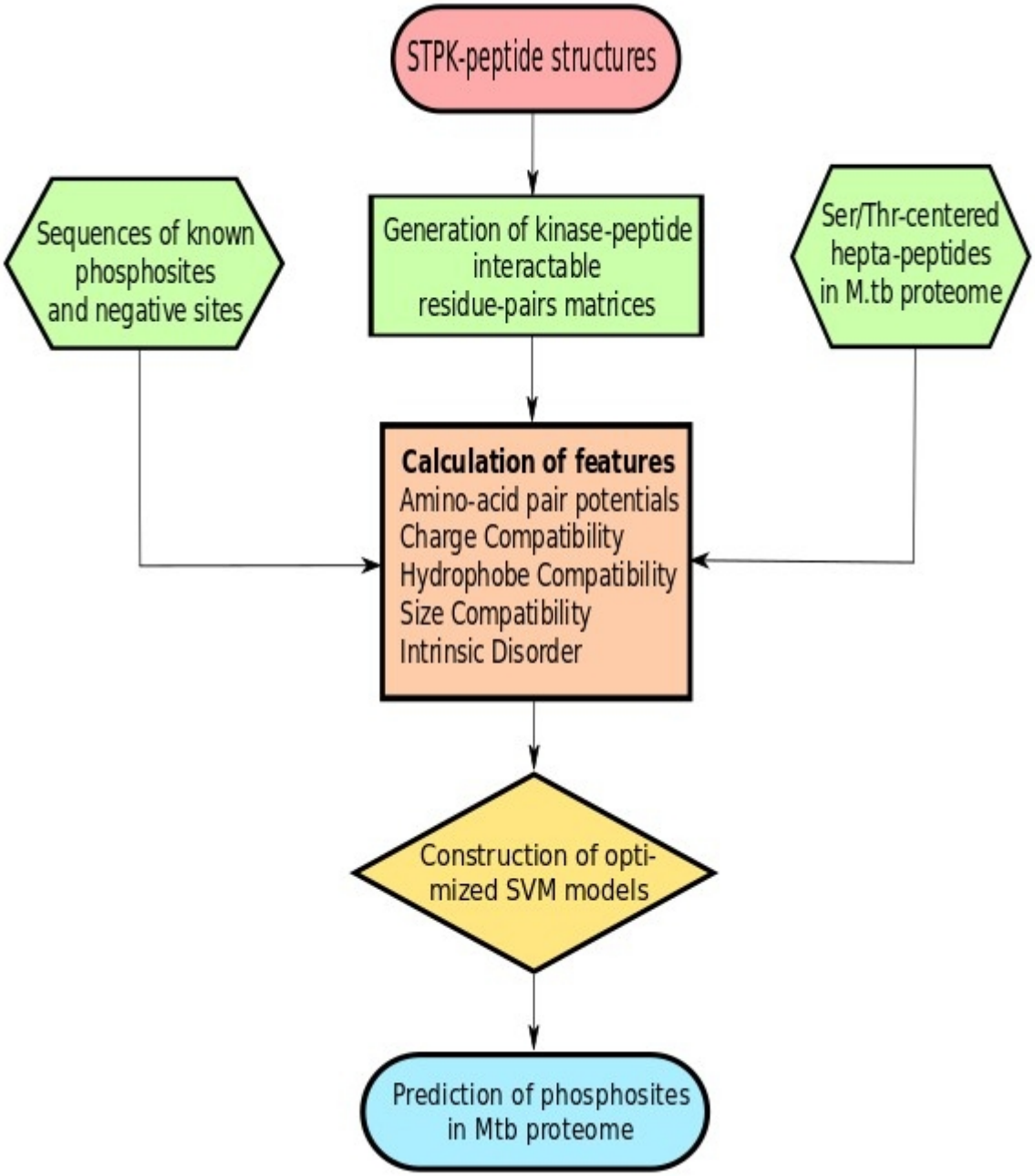
Figure 3. ROC curves plot comparing the predictive performance of our SVM based method with other machine learning algorithms.

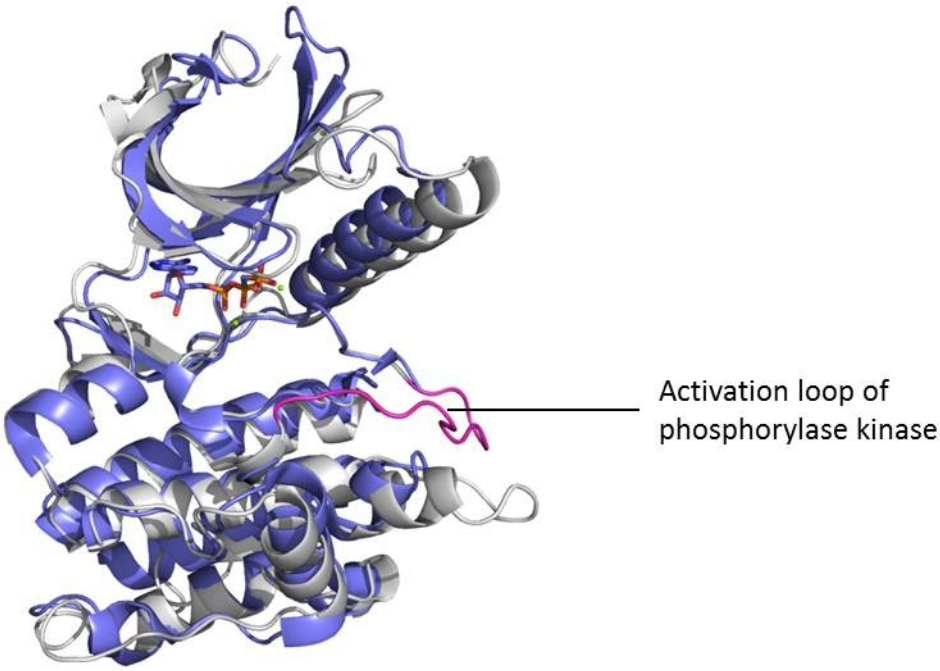
Figure 4. ROC curves plot of benchmarking the predictive performance of MTB-KSPP with other tools available for STPK phosphosite prediction.

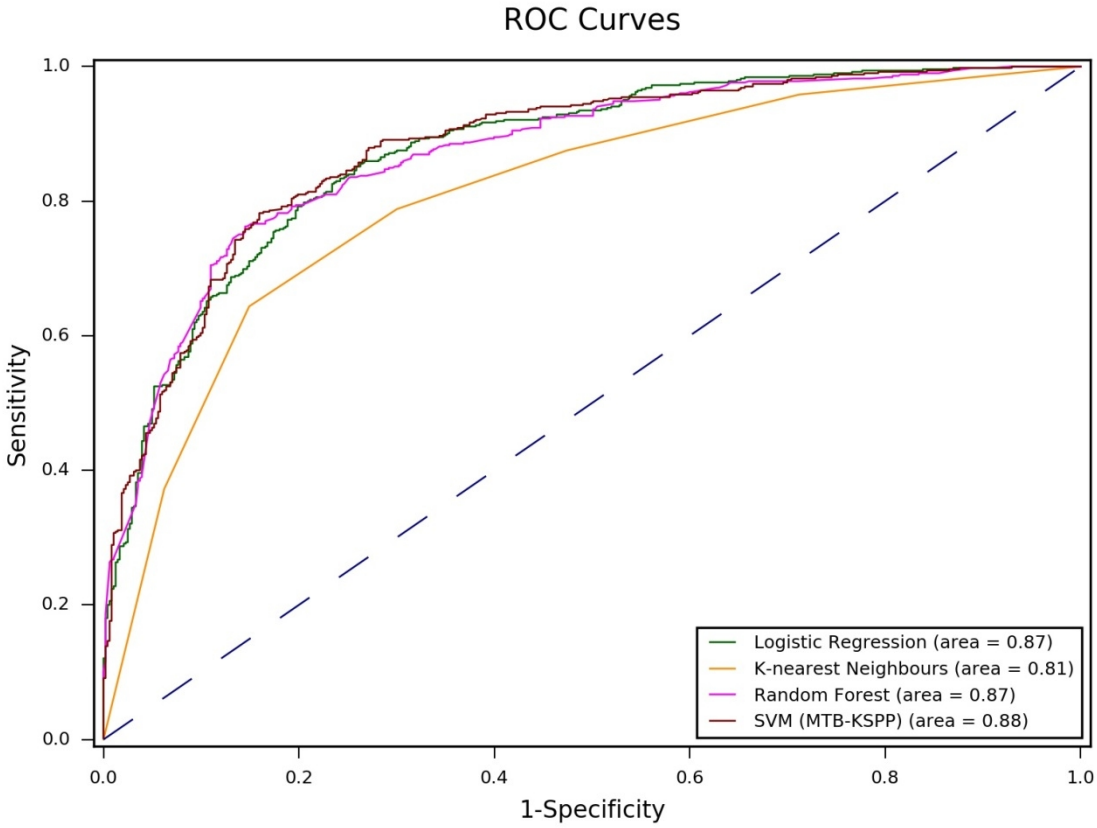
Figure 5. Graphical representations of the predicted sequence motifs recognized by Mtb STPKs PknA, PknB, PknF and PknG. The X-bar indicates the positions of the amino-acid residue of the peptide motif centered on Serine/Threonine (position 4). The numbers on the Y-bar indicate the percentage enrichment of the amino-acid residues at individual positions of the predicted peptide motif recognized by each Mtb STPK.

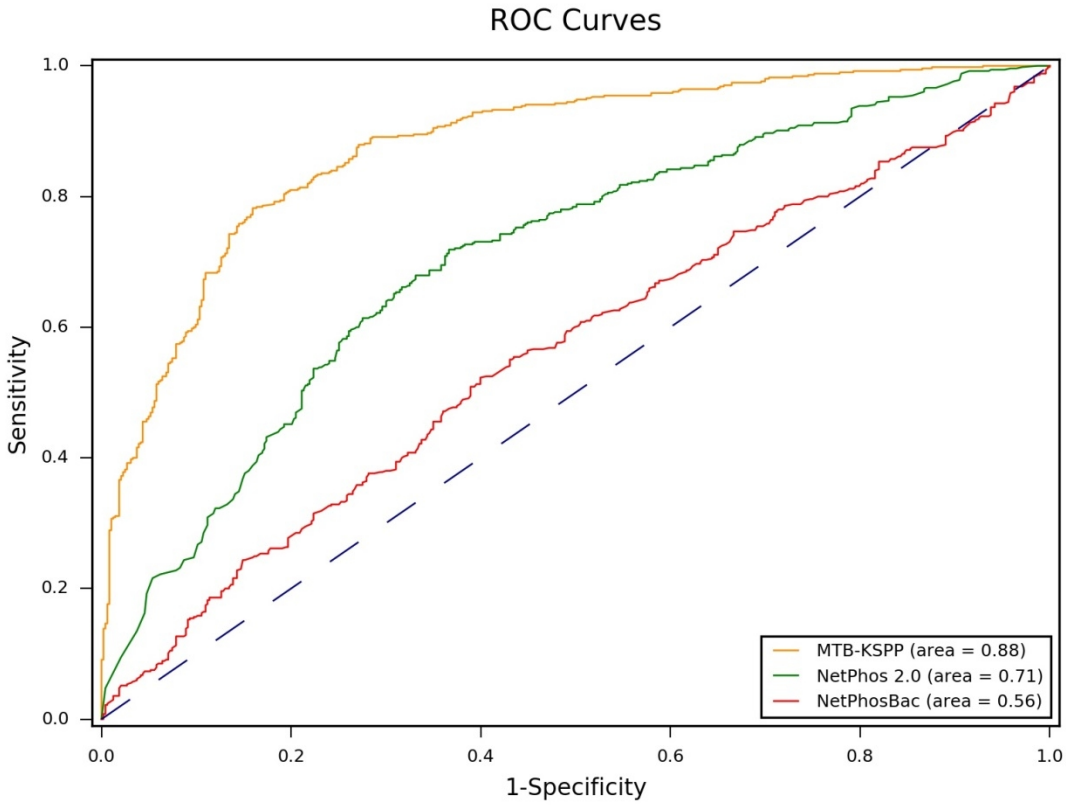
Figure 6. (A) The kinase residue Arg-101 (in Yellow colour) of Mtb PknB kinase (in lightblue colour) is within an interactable distance of the N-terminal end -3 position residue of the bound peptide (in lightpink colour) modelled in the peptide binding pocket of PknB. The predicted consensus motif for PknB substrate peptide shows enrichment of acidic residues Aspartate and Glutamate at -3 position (C^{α} atom is coloured Red). (B) The kinase residues Arg-45 and Arg-150 (in Yellow colour) of Mtb PknJ kinase (in lightblue colour) are within an interactable distance of the C-terminal end +3 position residue of the bound peptide (in lightpink colour) modelled in the peptide binding pocket of PknJ. The predicted consensus motif for PknJ substrate peptide shows enrichment of acidic residues Aspartate and Glutamate at +3 position (C^{α} atom is coloured Red).

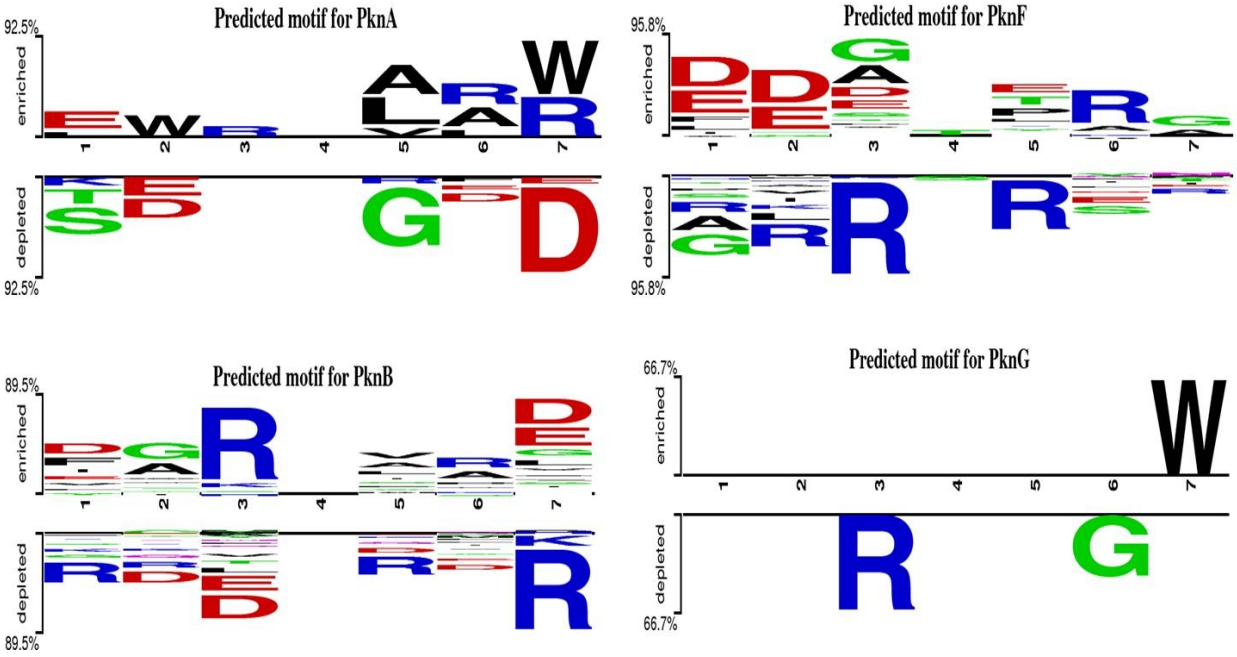
Figures:

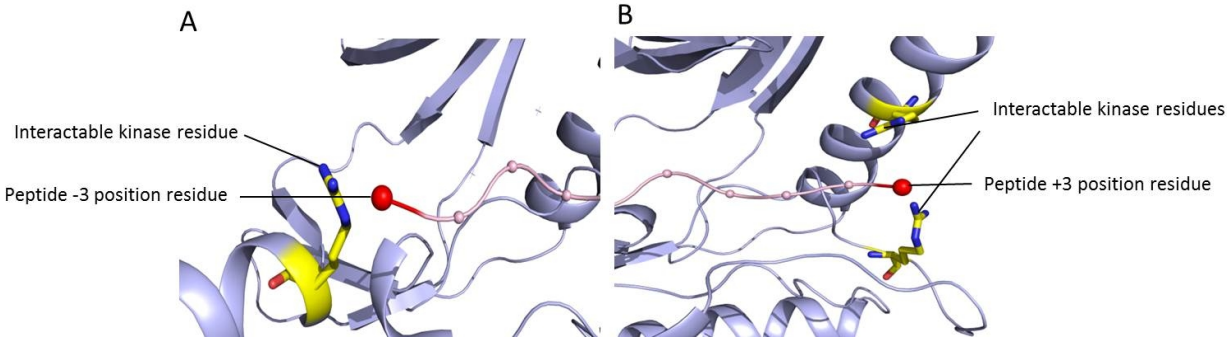

STPK-peptide structures
Sequences of known phosphosites and negative sites
Generation of kinase-peptide interactable residue-pairs matrices
Ser/Thr-centered hepta-peptides in M.tb proteome
Calculation of features
Amino-acid pair potentials
Charge Compatibility
Hydrophobe Compatibility
Size Compatibility
Intrinsic Disorder
Construction of optimized SVM models
Prediction of phosphosites in Mtb proteome


Activation loop of
phosphorylase kinase


ROC Curves
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression (area = 0.87)
K-nearest Neighbours (area = 0.81)
Random Forest (area = 0.87)
SVM (MTB-KSPP) (area = 0.88)


ROC Curves
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
MTB-KSPP (area = 0.88)
NetPhos 2.0 (area = 0.71)
NetPhosBac (area = 0.56)


Predicted motif for PknA
92.5%
enriched
depleted
92.5%
Predicted motif for PknF
95.8%
enriched
depleted
95.8%
Predicted motif for PknB
89.5%
enriched
depleted
89.5%
Predicted motif for PknG
66.7%
enriched
depleted
66.7%


A
B
Interactable kinase residue
Peptide -3 position residue
Interactable kinase residues
Peptide +3 position residue