Amino Acids (2014) 46:1919–1928
DOI 10.1007/s00726-014-1739-6

ORIGINAL ARTICLE

Prediction of substrate sites for protein phosphatases 1B, SHP-1, and SHP-2 based on sequence features

Zheng Wu · Ming Lu · Tingting Li

Received: 22 October 2013 / Accepted: 31 March 2014 / Published online: 24 April 2014

Abstract Tyrosine phosphorylation plays crucial roles in numerous physiological processes. The level of phosphorylation state depends on the combined action of protein tyrosine kinases and protein tyrosine phosphatases. Detection of possible phosphorylation and dephosphorylation sites can provide useful information to the functional studies of relevant proteins. Several studies have focused on the identification of protein tyrosine kinase substrates. However, compared with protein tyrosine kinases, the prediction of protein tyrosine phosphatase substrates involved in the balance of protein phosphorylation level falls behind. This paper described a method that utilized the k-nearest neighbor algorithm to identity the substrate sites of three protein tyrosine phosphatases based on the sequence features of manually collected dephosphorylation sites. In the performance evaluation, both sensitivities and specificities could reach above 75 % for all three protein tyrosine phosphatases. Finally, the method was applied on a set of known tyrosine phosphorylation sites to search for candidate substrates.

Keywords Protein tyrosine phosphatase · Substrate specificity · Sequence feature · Prediction · k-Nearest Neighbor algorithm

Abbreviations

PTK	Protein tyrosine kinase
PTP	Protein tyrosine phosphatase
PTP1B	Protein tyrosine phosphatase 1B
SHP-1	Src homology 2 domain tyrosine phosphatase 1 (also known as SH-PTP1, src homology 2 domain protein tyrosine phosphatase 1)
SHP-2	Src homology 2 domain tyrosine phosphatase 2 (also known as SH-PTP2, src homology 2 domain protein tyrosine phosphatase 2)

Introduction

Protein phosphorylation is the major post-translational modification in physiological processes and mainly occurs on serine (S), threonine (T), and tyrosine (Y) residues by adding phosphates. Dephosphorylation is the reverse process used to remove phosphates from phosphorylated amino acids. The level of phosphorylation state could influence protein activities and consequently, regulate the signal propagation in cells. Growing evidence suggests that protein phosphorylation participates in various cellular processes, such as migration, proliferation, apoptosis, differentiation, metabolism, and intracellular communication (Graves and Krebs 1999; Manning et al. 2002a, b).

Tyrosine phosphorylation is extensively used for cell communication, cell motility, proliferation, and differentiation (Hunter 1987; Mustelin et al. 2002). Protein tyrosine

Electronic supplementary material The online version of this article (doi:10.1007/s00726-014-1739-6) contains supplementary material, which is available to authorized users.

Z. Wu · M. Lu · T. Li (✉)
Department of Biomedical Informatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China
e-mail: litt@hsc.pku.edu.cn

T. Li
Institute of Systems Biomedicine, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

phosphatases (PTPs) collaborate with protein tyrosine kinases (PTKs) to control the level of phosphorylation state. Abnormal tyrosine phosphorylation could result in various diseases, such as cancer (Zhao et al. 2011), insulin resistance (Andersen et al. 2004), and Noonan syndrome (Tartaglia et al. 2001). Detection of relevant substrates involved in biochemical reactions could help in acquiring a better understanding of specific signal pathways and finding out new targets for therapy. Unfortunately, the mass-spectroscopy method cannot provide exact information of acting enzymes, whereas the p-Tyr antibody technique is time consuming if applied in an unbiased fashion. In silico prediction shows the advantage of providing useful information of candidate substrates and narrowing down experimental efforts.

During the past decades, several focused studies have contributed to the computational prediction methods for PTKs. These algorithms included artificial neural network (Blom et al. 1999), logistic regression (Iakoucheva et al. 2004), support vector machine (Kim et al. 2004), and conditional random field (Dang et al. 2008). Most of them predicted phosphorylation sites based on primary sequences around these sites, whereas the information of high-level structures, functional domains, and subcellular location was gradually considered (Li et al. 2010). Under physiological conditions, PTPs, together with PTKs, maintain the balance of protein phosphorylation levels. Identification of the corresponding dephosphorylated proteins would provide useful information of candidate substrates, as well as contribute to the development of phosphorylation studies.

In contrast to studies on various methods to predict phosphorylation, studies on the prediction of PTPs are few. Ferrari et al. (2011) integrated chip technology and "closeness" in the protein interaction network to identify new substrates of the phosphatase PTP1B, whereas the prediction field of other PTPs remained unexplored. PTP1B is the most intensively studied tyrosine phosphatase and the first enzyme of its class to be purified (Tonks et al. 1988). By decreasing the phosphorylation level of substrates, PTP1B is involved in various biological processes that contain cell differentiation (LaMontagne et al. 1998; Fuentes et al. 2012), cell migration (Stuible et al. 2008; Cortesio et al. 2008), development, and morphology (Lanahan et al. 2010; Chacon et al. 2010). Meanwhile, other well-studied PTPs, such as SHP-1 and SHP-2, were also reported to participate in cell proliferation (Lopez-Ruiz et al. 2011; Mahmood et al. 2012; Kozlowski et al. 1998), cytoskeleton organization (Langdon et al. 2012; Timmerman et al. 2012; Stebbins et al. 2003; Draber et al. 2012), cell communication (Hebeisen et al. 2013; Pani et al. 1995), cell motion (Neel et al. 2003), and so on. However, numerous substrates of these PTPs in relevant pathways still remain to be identified; they are helpful in acquiring a better understanding of how they function to achieve positive and negative signals.

The present study utilized the information of peptide sequences in the proximity of known dephosphorylation sites to predict putative substrate sites of three PTPs, namely, PTP1B, SHP-1 (also known as SH-PTP1), and SHP-2 (also known as SH-PTP2). Based on manually collected data, the k-nearest neighbor (k-NN) algorithm was used to identify substrate sites. The sensitivities and specificities of the predictive method could both reach above 75 % in the performance evaluation. A web server was available at http://cmbi.bjmu.edu.cn/ptpsite/. Finally, this method was applied to scan the substrate sites of three PTPs from a set of known phosphorylation sites acquired by a mass spectrometer. This study only focused on the prediction for PTPs because the combinatorial subunit principle of serine/threonine protein phosphatases could generate much more diversity and flexibility (Alonso et al. 2004), which means that the specificity of these enzymes was affected by binding to different co-workers. Information about the relevant combined proteins that assign the substrate selectivity should also be considered for serine/threonine protein phosphatases (McConnell and Wadzinski 2009).

Materials and methods

Data preparation

Positive dataset

Protein substrates of PTPs were collected by searching the literature with key words "protein tyrosine phosphatase* AND dephospho*" in PubMed. After reading all those papers and related references, the ones with experimentally verified dephosphorylated site and PTP information were picked out. The dephosphorylated proteins were extracted and mapped to UniProt Database for 21-mer sequences of relevant sites, including the central tyrosine and residues from -10 to $+10$ surrounding it. Given that some substrate sites were at the beginning or the end, the vacant position is filled with a symbol "–" (e.g., EDYFTSTEPQYQPGENL—) during the collection. The exact UniProt ID and accession number of each dephosphorylated protein were also retrieved at the same time. Each dephosphorylated site was carefully checked to ensure the position is exactly the one mentioned in articles. Among the obtained data, three PTPs that contained the most sites, namely, PTP1B, SHP-1, and SHP-2, were selected in this study (positive data are available in Table S1). To avoid high similarity among sequences, only sequence fragments that shared <70 % identity were kept for the

positive dataset during the performance evaluation. Given that previous studies of phosphorylation prediction selected 70 % as the threshold, the same identity was adopted in the present study (Kim et al. 2004; Menor et al. 2012; Dang et al. 2008). If 70 % or more than 70 % residues on corresponding positions of two isometric sequences were similar, only one of them was kept in the positive dataset and the other was discarded. Meanwhile, sequences with "−" were also abandoned because of the difficulty in similarity calculation. The method utilized in this study calculated sequence similarity based on the BLOSUM 62 matrix, which does not have scores for "–". Thus, similarity calculation cannot be applied on the sequences with "–", thereby limiting the terminal sites. In a previous study, Kiemer et al. (2005) constructed predictors for N-terminal acetylation sites using sequences that contained residues on the C-terminal side of the modified residue. Given that most of the positive data in the study contained enough residues on both sides, the few sequences with "−" were discarded and the information on both sides was fully used.

Meanwhile, some dephosphorylated substrate sites were not proved by accurate experiments during the collection (Table S2). These substrate sites were proposed with uncertain words, such as "possible" and "maybe". To guarantee the prediction exactitude, these "unqualified data" were not contained in the positive dataset.

Negative dataset

All the human proteins (Release 2013_08) were downloaded from UniProtKB at http://uniprot.org/pub/databases/uniprot/current_release/knowledgebase/proteomes/ and the corresponding sequences of centered tyrosines were extracted. The negative dataset was obtained by eliminating the positive dataset (Table S1) along with "unqualified data" (Table S2) from the entire tyrosine-centered sequences of human species. Given that our knowledge about the entire protein modification process is limited, collecting a set of peptide sequences that can be safely regarded as non-substrates of any kind of PTPs is difficult. However, the non-substrate sites of a specific PTP should still dominate the negative dataset in consideration of the numerous peptides that contain tyrosines in human species.

Performance evaluation with k-NN algorithm

The k-NN algorithm is one of the simplest machine learning algorithms. In this algorithm, an object is classified by a majority vote of its neighbors based on its similarity to the trained samples. The details of the predictive algorithm are described as follows. Initially, similarity scores between the query peptide and each peptide in positive and negative training sets were calculated according to the BLOSUM 62 matrix. To deal with the imbalanced sample size between two training sets, each score in the positive training set was multiplied by a weight (wi) as the final score. Thereafter, all those scores were mixed and ranked from high to low. If most of the scores that ranked in the top k positions belonged to the positive training set, which means that the given peptide is more similar to the positive samples than the negative ones, the query peptide can be dephosphorylated by this PTP. Otherwise, the query peptide is considered to be a non-substrate sequence. In this case, the value of k is set to 5.

Meanwhile, 1,000 negative training sequences and 100 negative test sequences were selected from the negative dataset. Thus, each time the number of test sample was 101: 1 positive sample and 100 negative samples. The 70 % identity was used for both the positive and negative sequences to avoid over-estimation. During the selection, each of the 100 negative test peptides was guaranteed to share less than 70 % sequence identity with the 1,000 samples in the negative training set. The tests were performed 1,000 times, and the final evaluation was the average value.

Table 1 The numbers of collected sites of three PTPs

PTPs	PTP1B	SHP-1	SHP-2
All sites collected	57	47	48
All known human dephosphorylation sites	51	42	44
Positive dataset after removal of sequences with "–" and highly homologous peptides (over 70 % identity)	50	36	41

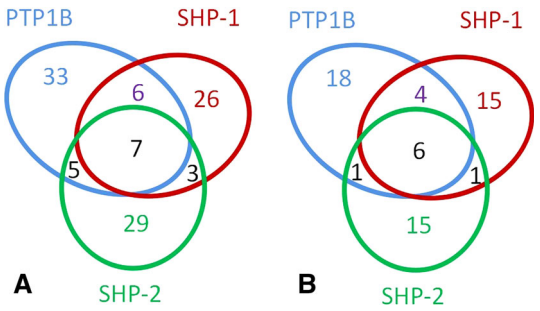

Fig. 1 Venn diagrams representing overlaps of substrate sites and substrate proteins of PTP1B, SHP-1, and SHP-2

Results and discussion

Sequence and functional features of collected data

The dephosphorylation sites were manually collected from the public literature. A total of 57 sites were discovered for PTP1B, 47 for SHP-1, and 48 for SHP-2 (Table S1). Most

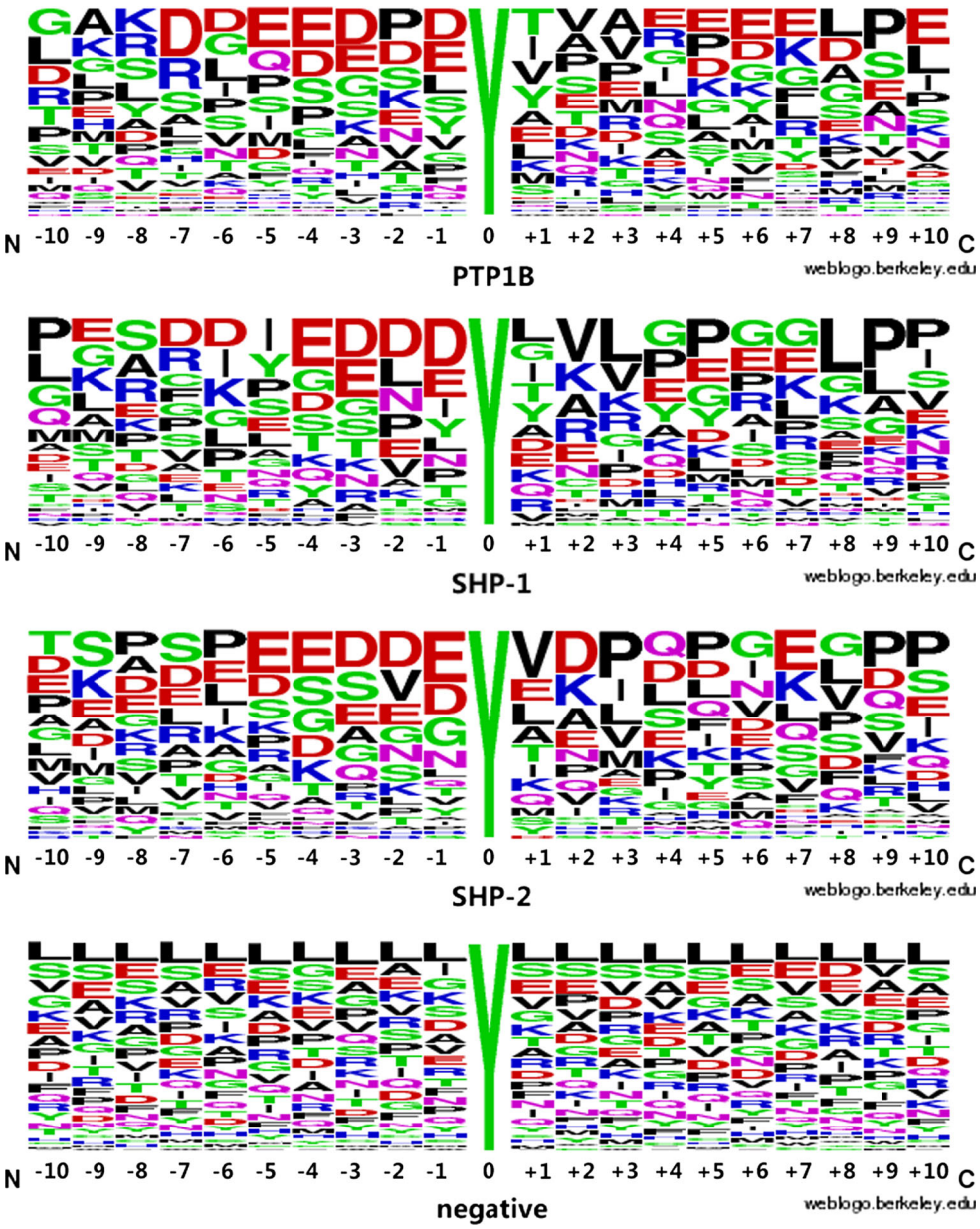

Fig. 2 Sequence frequency analysis of dephosphorylation peptides for PTP1B, SHP-1, and SHP-2. Sequence *logo plots* represent amino acid frequencies for 10 amino acids from both sides of the dephosphorylation site. *N* amino side, *C* carboxyl side

of the dephosphorylation sites were from human, whereas the rest belonged to the mouse species. The following analysis only focused on the dephosphorylation sites of human, which contained 51, 42, and 44 sites for PTP1B, SHP-1, and SHP-2, respectively (Table 1). As shown in Fig. 1, the three PTPs shared some substrate proteins and sites. The overlaps between every two PTPs made up approximately a quarter of the total sum.

The substrate specificity of each PTP was first characterized using all the known 21-mer peptide sequences with dephosphorylation tyrosines surrounded by 10 residues on both sides. From the frequency analysis (Fig. 2) by WebLogo (Crooks et al. 2004), PTP1B preferred Glu (E) at positions −5 and −4 and a large fraction of the C-terminal positions. Asp (D) was the most common residue found at position −7 and several other positions on the proximal N-terminal. Taken together, PTP1B showed more preference for acidic amino acids that flank the phosphorylated tyrosine, which is consistent with the previous studies for PTP1B substrate specificity (Vetter et al. 2000; Pellegrini et al. 1998). Likewise, SHP-1 and SHP-2 favored acidic residues on the N-terminal side, as shown in Fig. 2. However, hydrophobic residues Leu (L), Val (V), and Pro (P) were more preferred than acidic residues in the upstream of SHP-1 targets. As for SHP-2, the C-terminal showed a slightly broader specificity. For example, Val (V), Asp (D), and Pro (P) dominated at positions +1 to +3, respectively. Meanwhile, position +7 showed more

preference for the acidic residue Glu (E). This observation mostly agrees with the careful analysis of substrate specificity for these PTPs (Ren et al. 2011). By contrast, the WebLogo analysis of 1,000 negative sequences showed obvious differences with the positive sequences (Fig. 2). We can see that each PTP recognizes the specific motif around dephosphorylation sites, but the patterns were not so conserved, at least less pronounced than kinases.

Furthermore, the same sequences as WebLogo were used to display the sequence identity distribution of different sets. In the positive and negative sets, the sequences contained were compared with one another from positions −10 to +10 (except the central tyrosines). Meanwhile, each positive sequence was compared with each negative sequence to show the sequence identity between positive and negative sets. The identical residues on 20 positions for each pair were calculated and the number of pairs that shared the same amount of identical residues was counted. The final identity distributions, divided by the total sum, were shown in supplementary Figure S1. The figure shows that the sequence identity of two sets and between them was relatively low even though that of the positive set was slightly higher, which meant that very few sequences shared high similarity with one another. Taken together, the sequence identity between the positive and negative sets was not easily separable.

Ability of prediction during performance evaluation

Although the previous sections showed that acidic residues at the −1 to −5 positions were much more favorable for substrate recognition by PTPs (Hippen et al. 1993), other positions also showed regular patterns, as shown in Fig. 2. To determine the best length of peptide sequences used for prediction, performance tests were conducted on 5–10 residues on both sides of the central tyrosine, and 70 % identity threshold was used to obtain the suitable positive dataset for each length. The curves in Fig. 3 compared the performance of sequences with different lengths. As described in the experimental procedures, to deal with the imbalanced sample size between positive and negative training sets, each score in the positive training set was multiplied by a weight (wi) as the final score. Setting different values of wi would result in different sensitivities and specificities. Each value in receiver operating characteristic (ROC) curves was the average level of 1,000 performances. The results showed that the best-performing ROC for PTP1B was the pink curve (with nine residues on both sides). Meanwhile, the black line (with eight residues on both sides) of SHP-1 and the red line (five residues on both sides) of SHP-2 showed the best performance compared with other lengths. In this case, the positive dataset for PTP1B, SHP-1, and SHP-2 substrate sites was used with 9, 8, and 5 residues on both sides. By contrast, the performance of the blue curve (with six residues on both sides) for SHP-1 and the black curve (with eight residues on both sides) for SHP-2 was obviously worse than the other lengths. This observation could be attributed to the difference in sequence patterns. For example, the −8 and +8 positions of SHP-2 targets showed less preference for any kind of amino acids (Fig. 2). Even though the sequences with 9 and 10 residues on both sides for SHP-2 targets were also not as good as the other lengths, the more regular patterns (especially the Pro on the C-terminal) at corresponding positions still slightly improved their performance.

Next, the best results in performance evaluation of PTP1B, SHP-1, and SHP-2 were adopted and the sensitivities and specificities at two cutoffs were provided, with details shown in Table 2. The corresponding positions were marked with blue and red in ROC curves (Fig. 4). The horizontal axis represented false positives (1-specificity), whereas the vertical axis represented true positives (sensitivity). For PTP1B, both the sensitivity and specificity could reach above 75 % at the low-stringency cutoff, whereas the specificity could reach as high as 91.0 % at the high-stringency cutoff with a sensitivity of 65.6 %. This result nearly caught up with the previous study of PTP1B prediction, which achieved a sensitivity of 50 % and a specificity of 98 % (Ferrari et al. 2011). However, Ferrari's predicted substrates were proteins rather than exact sites, which made comparing the predicted results impossible, because numerous tyrosines existed in the full sequence of one protein. The other two PTPs also performed as well as PTP1B at the low-stringency cutoff, whereas the sensitivity was nearly 10 % lower than PTP1B at the level of high-stringency cutoff. A highly stringent threshold will improve the specificity but decrease the sensitivity, whereas a less stringent threshold will increase the sensitivity at the price of lower specificity. Choosing the level of cutoffs can predict more candidates or more accurate substrates. To see if a concerned sequence or protein could be dephosphorylated by a given PTP, users can test the query peptide on our predicting website http://cmbi.bjmu.edu.cn/ptpsite/.

To confirm that the performance evaluation in this study was not biased because of lacking independent positive test sequences, the positive dataset was split into two groups: 1/5 and 4/5. Before the division, the 70 % identity threshold was also adopted in the entire positive dataset for different lengths (5–10 residues on both sides of the central tyrosine). The larger group was used to perform leave-one-out tests for sequences with different lengths, as previously mentioned. Then, the best performance for each PTP (15-mer for PTP1B, 17-mer for SHP-1, and 15-mer for SHP-2)

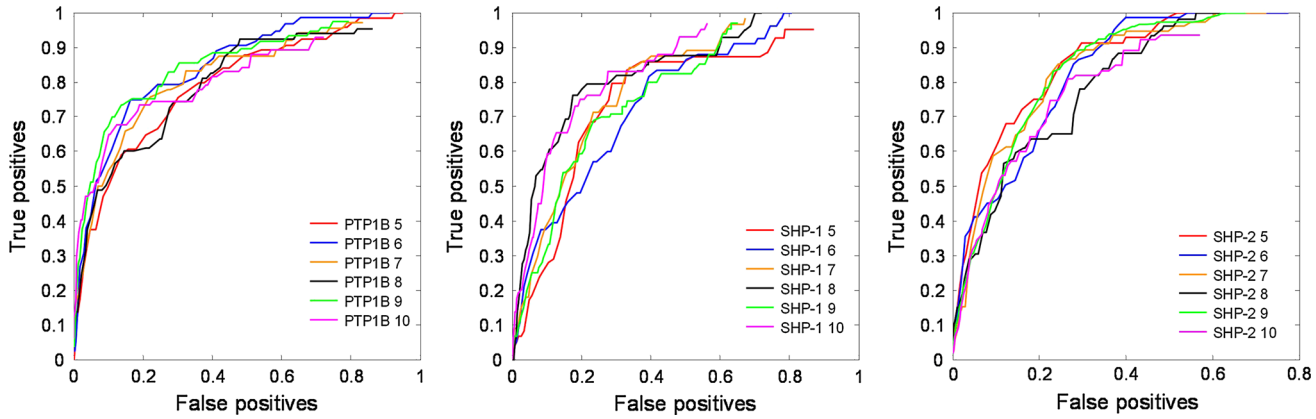

Fig. 3 ROC curves with different *colors* representing different lengths of sequences used for prediction. The sequences that performed best for PTP1B, SHP-1, and SHP-2 are 9, 8, and 5 residues on both sides of central tyrosines, respectively

Table 2 The results in performance evaluation of three PTPs with two cutoffs

PTPs	Low-stringency cutoff			High-stringency cutoff		
	Value of wi	Sensitivity (%)	Specificity (%)	Value of wi	Sensitivity (%)	Specificity (%)
PTP1B	2.35	78.8	75.6	1.70	65.6	91.0
SHP-1	2.60	79.5	78.4	1.90	54.7	91.6
SHP-2	1.75	75.0	79.1	1.45	56.0	91.9

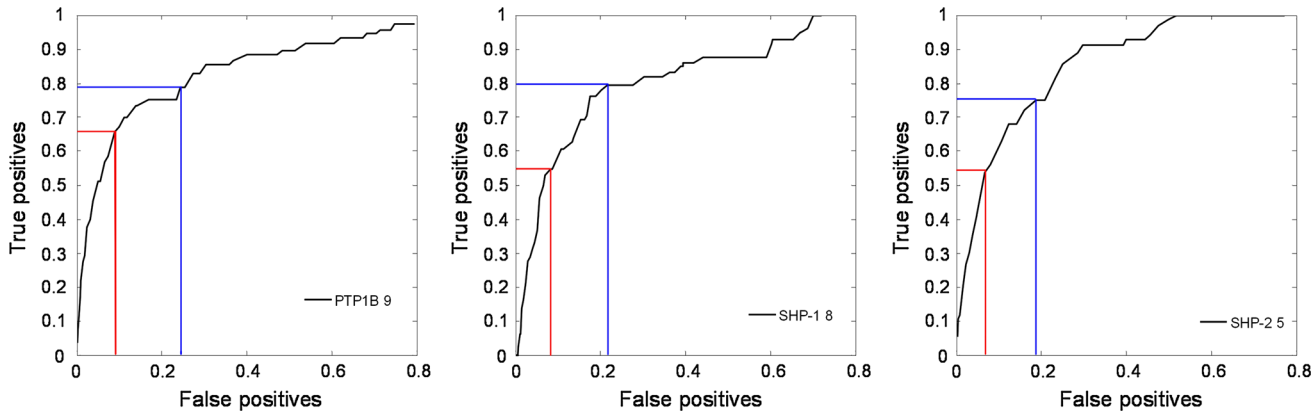

Fig. 4 ROC curves showing predictive results of three PTPs using sequences with the best-performing length. The detailed results at two cutoffs are marked with *blue* and *red*

was picked out and the weights at high cutoffs were adjusted (Table 3). Using the best weights, the rest of the positive sequences were utilized to score the performances independently. Table 3 shows the sensitivities and specificities acquired from the predictive results of 1/5 positive sequences (which were 10 for PTP1B, 8 for SHP-1, and 8 for SHP-2) and 1,000 random negative sequences that shared <70 % identity with the negative training set. At the specificity of around 90 %, each PTP could identify more than 60 % of the 1/5 independent positive sequences. In conclusion, the performance of completely independent positive sequences was still relatively well. The Matthews correlation coefficient and F1 score were also provided as references of the independent tests (Supplementary Table S3). However, both values turned out to be very low because of the large difference between positive and negative sample sizes (about 10:1,000).

The 70 % identity was adopted to avoid high similarity between the training and test sequences. To investigate the effect of identity values on the performance, experiments were conducted with another 4 identity thresholds: 60, 50, 40, and 30 %. Each identity was used not only for the

Table 3 Results of the validation based on independent 1/5 positive sequences

PTPs	Performance using 4/5 of the positive dataset			Tests of the remaining 1/5	
	Value of wi	Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
PTP1B	1.75	58.0	90.4	60.0	89.6
SHP-1	2.25	48.0	91.2	62.5	89.9
SHP-2	1.80	57.0	90.0	62.5	90.2

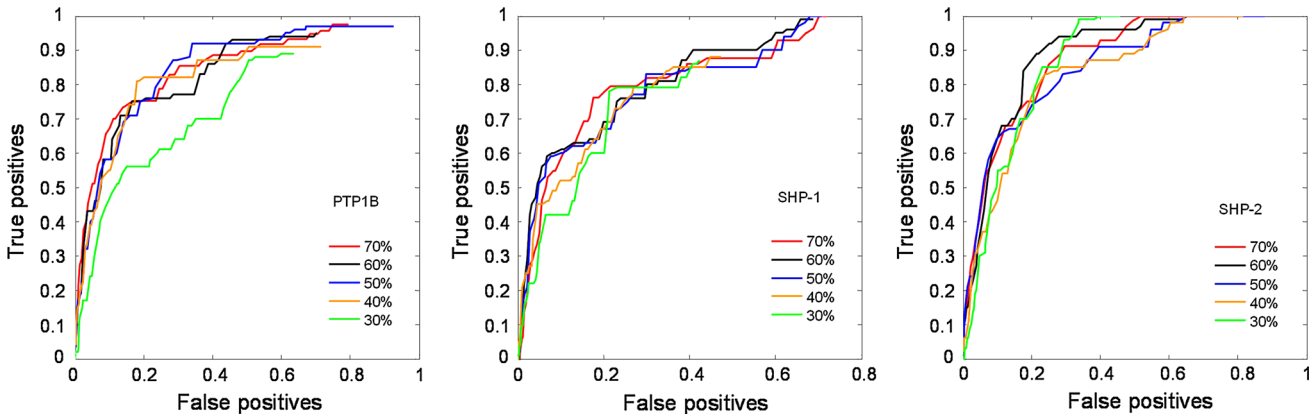

Fig. 5 Comparison of ROC curves using different identity thresholds

positive dataset, but also between the negative test set and negative training set. The best performance acquired from the experiments at different identity levels was compared with one another (Fig. 5). The results in Fig. 5 show that performances did not show large differences at the 70, 60, and 50 % identities because those positive sequences were not largely decreased when the identity threshold dropped from 70 to 50 % (Supplementary Table S4). At the 30 % identity, performances were obviously worse than the others except the one for SHP-2. The 30 % identity was a very strict threshold. For example, in the sequence fragment with five residues on each side, only <3 out of 10 residues could be similar, which leads to the significant reduction of positive sequences.

To our knowledge, only a fraction of tyrosines are phosphorylated, and they are more likely to be within disordered regions (Iakoucheva et al. 2004). Phosphotyrosines might have specific sequence contexts because of the disordered regions. The good performance of the method utilized in this study might be caused by discriminating phosphotyrosines versus non-phosphotyrosines, not the substrate specificity of different PTPs. Recently, numerous phosphorylated sites of several species have been identified by mass spectroscopy (Beltrao et al. 2012). Among those data, more than 8,000 tyrosine sites of human proteins were discovered. To confirm the discrimination ability, the same tests were performed using numerous phosphotyrosines identified by mass spectroscopy as the negative dataset. Figure 6 shows the performance results after replacing the whole tyrosines with identified phosphorylated tyrosines. The predictor could still discriminate substrates although not as well as the results with the whole tyrosines. Table 4 shows the detailed sensitivities and specificities at two cutoff levels. Compared with the results using whole tyrosines as negative dataset (Table 2), PTP1B still performed better than the other two PTPs. At the low-stringency cutoff, the sensitivities and specificities dropped to 60–70 %, whereas at the high-stringency cutoffs with specificities of around 90 %, the sensitivities of three PTPs declined about 20 %. Even though the results were not as good as the tests using whole tyrosines, the predictor could still discriminate different substrates.

To explore whether one PTP could distinguish the substrates of other PTPs, the collected substrate sites of each PTP were treated as three query peptide sets. The three sets were tested for each PTP on our predictive website at the high-stringency cutoff level with a specificity of above 90 %. The results obtained from the analysis are listed in Table 5, which presents the percentage of identified substrates for each PTP in three sets. Table 5 shows that 100 % of the known substrate sites of PTP1B could be recognized by the predictor of PTP1B, because all the collected data were used to train the predictor on the website. By contrast, 34.5 % of the SHP-1 substrate sites were predicted as substrates of PTP1B. Among the three known substrate sets, the overlaps among them made up

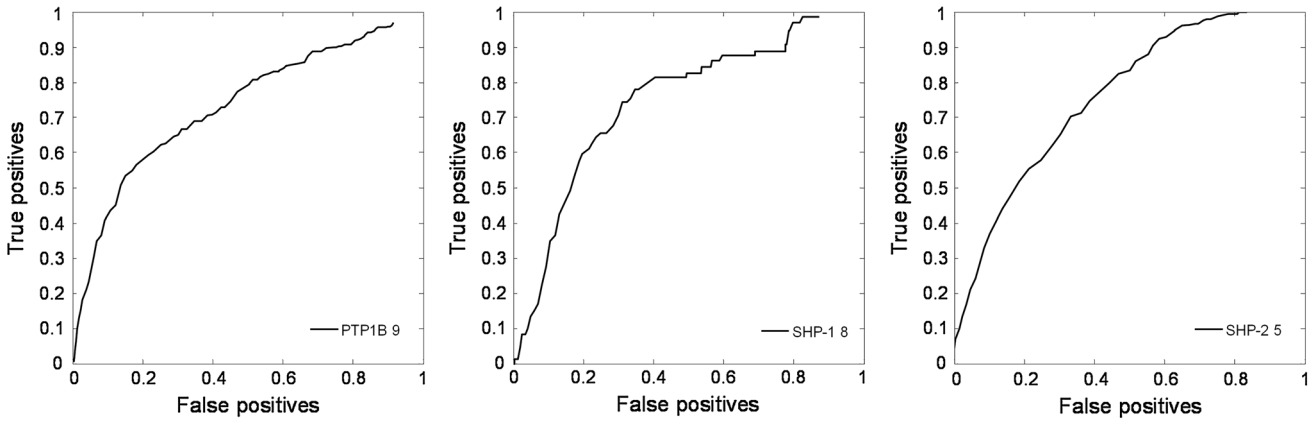

Fig. 6 ROC curves using phosphotyrosines instead of tyrosines from human proteome as negative dataset

Table 4 Performance evaluation using phosphorylated tyrosines as negative dataset

PTPs	Low-stringency cutoff			High-stringency cutoff		
	Value of wi	Sensitivity (%)	Specificity (%)	Value of wi	Sensitivity (%)	Specificity (%)
PTP1B	2.25	64.5	71.2	1.65	40.8	90.7
SHP-1	2.55	67.6	71.5	1.85	34.9	89.5
SHP-2	1.75	62.0	71.9	1.40	32.8	91.4

Table 5 Results in the table presented the percentage of identified substrates of each PTP using three collected sets

	PTP1B (%)	SHP-1 (%)	SHP-2 (%)
Substrate sites of PTP1B	100	21.1	41.0
Substrate sites of SHP-1	34.5	100	31.3
Substrate sites of SHP-2	29.6	51.7	100

Table 6 The numbers of predicted sites of three PTPs

PTPs	PTP1B	SHP-1	SHP-2
The number of predicted dephosphorylation sites	1176	1571	1239

approximately a quarter of the total sum (Fig. 1). In consideration of the overlaps in the collected data, the discrimination ability of different PTPs was acceptable.

Identification of substrates for three PTPs

According to the results of the performance evaluation (Fig. 3), the 19-, 17-, and 11-mer sequences were used for PTP1B, SHP-1, and SHP-2 to perform prediction. The sample size of the positive dataset that resulted from the removal of sequences with "–" and highly homologous peptides (over 70 % identity) is listed in Table 1. The results show that the specificity above 90 % could predict relatively accurate substrates. In the set of human tyrosine phosphorylation sites that contained more than 8,000 targets, possible substrate sites were scanned at the specificity of 91.0 % for PTP1B, 91.6 % for SHP-1, and 91.9 % for SHP-2. The k-NN algorithm was trained using all the known dephosphorylation sites and 1,000 negative samples randomly selected from the negative dataset (the entire negative dataset with the appropriate length for each PTP is shown in Table S5). The numbers of predicted dephosphorylation sites at the high-stringency cutoff are listed in Table 6 (the detailed sites and sequences can be found in Table S6). Reported articles have revealed that PTP1B is exceptionally active toward multiply phosphorylated substrates (Ren et al. 2011), especially for two side-by-side pYs (Myers et al. 2001). This observation was consistent with our collected dataset and predicted results. The predicted sequences that contained Y–Y structures made up about 13.6 % of the total PTP1B substrate sites, which outnumbered the percentage for SHP-1 (11.6 %) and SHP-2 (9.0 %).

Conclusion

This study utilized manually collected substrate sites in PubMed to predict the substrate sites of three PTPs, namely, PTP1B, SHP-1, and SHP-2. Analysis of substrate sequences revealed that specific motifs could be recognized by different enzymes. The performance tests achieved both sensitivities

and specificities of above 75 % using appropriate lengths of peptide sequences. Applying the predictor utilized in this study on the set of proteins that have already been identified as phosphorylated substrates could provide a set of possible candidates for experimental validation (Table S6). To improve the prediction performance of the method used in this study, collecting data and updating the webserver will be maintained. More PTPs can also be included in the system with the availability of more dephosphorylation data.

Acknowledgments We thank Dr. Pufeng Du for the useful discussions. This work was supported by grants from the National Basic Research Program (2011CBA01104), the National High-tech R&D Program (2012AA020401) of China, the National Natural Science Foundation of China (31371337 and 61105003), and Beijing Higher Education Young Elite Teacher Project (YETP0055).

Conflict of interest The authors declare that they have no conflict of interest.

References

Alonso A, Sasin J, Bottini N, Friedberg I, Osterman A, Godzik A, Hunter T, Dixon J, Mustelin T (2004) Protein tyrosine phosphatases in the human genome. Cell 117(6):699–711. doi:10.1016/j.cell.2004.05.018

Andersen JN, Jansen PG, Echwald SM, Mortensen OH, Fukada T, Del Vecchio R, Tonks NK, Moller NP (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. FASEB J 18(1):8–30. doi:10.1096/fj.02-1212rev

Beltrao P, Albanese V, Kenner LR, Swaney DL, Burlingame A, Villen J, Lim WA, Fraser JS, Frydman J, Krogan NJ (2012) Systematic functional prioritization of protein posttranslational modifications. Cell 150(2):413–425. doi:10.1016/j.cell.2012.05.036

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294(5):1351–1362

Chacon PJ, Arevalo MA, Tebar AR (2010) NGF-activated protein tyrosine phosphatase 1B mediates the phosphorylation and degradation of I-kappa-Balpha coupled to NF-kappa-B activation, thereby controlling dendrite morphology. Mol Cell Neurosci 43(4):384–393. doi:10.1016/j.mcn.2010.01.005

Cortesio CL, Chan KT, Perrin BJ, Burton NO, Zhang S, Zhang ZY, Huttenlocher A (2008) Calpain 2 and PTP1B function in a novel pathway with Src to regulate invadopodia dynamics and breast cancer cell invasion. J Cell Biol 180(5):957–971. doi:10.1083/jcb.200708048

Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14(6):1188–1190. doi:10.1101/gr.849004

Dang TH, Van Leemput K, Verschoren A, Laukens K (2008) Prediction of kinase-specific phosphorylation sites using conditional random fields. Bioinformatics 24(24):2857–2864. doi:10.1093/bioinformatics/btn546

Draber P, Sulimenko V, Draberova E (2012) Cytoskeleton in mast cell signaling. Front Immunol 3:130. doi:10.3389/fimmu.2012.00130

Ferrari E, Tinti M, Costa S, Corallino S, Nardozza AP, Chatrary-amontri A, Ceol A, Cesareni G, Castagnoli L (2011) Identification of new substrates of the protein-tyrosine phosphatase PTP1B by Bayesian integration of proteome evidence. J Biol Chem 286(6):4173–4185. doi:10.1074/jbc.M110.157420

Fuentes F, Zimmer D, Atienza M, Schottenfeld J, Penkala I, Bale T, Bence KK, Arregui CO (2012) Protein tyrosine phosphatase PTP1B is involved in hippocampal synapse formation and learning. PLoS ONE 7(7):e41536. doi:10.1371/journal.pone.0041536

Graves JD, Krebs EG (1999) Protein phosphorylation and signal transduction. Pharmacol Ther 82(2–3):111–121

Hebeisen M, Baitsch L, Presotto D, Baumgaertner P, Romero P, Michielin O, Speiser DE, Rufer N (2013) SHP-1 phosphatase activity counteracts increased T cell receptor affinity. J Clin Invest 123(3):1044–1056. doi:10.1172/JCI65325

Hippen KL, Jakes S, Richards J, Jena BP, Beck BL, Tabatabai LB, Ingebritsen TS (1993) Acidic residues are involved in substrate recognition by two soluble protein tyrosine phosphatases, PTP-5 and rrbPTP-1. Biochemistry 32(46):12405–12412

Hunter T (1987) A thousand and one protein kinases. Cell 50(6):823–829

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32(3):1037–1049

Kiemer L, Bendtsen JD, Blom N (2005) NetAcet: prediction of N-terminal acetylation sites. Bioinformatics 21(7):1269–1270. doi:10.1093/bioinformatics/bti130

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

Kozlowski M, Larose L, Lee F, Le DM, Rottapel R, Siminovitch KA (1998) SHP-1 binds and negatively modulates the c-Kit receptor by interaction with tyrosine 569 in the c-Kit juxtamembrane domain. Mol Cell Biol 18(4):2089–2099

LaMontagne KR Jr, Flint AJ, Franza BR Jr, Pandergast AM, Tonks NK (1998) Protein tyrosine phosphatase 1B antagonizes signalling by oncoprotein tyrosine kinase p210 bcr-abl in vivo. Mol Cell Biol 18(5):2965–2975

Lanahan AA, Hermans K, Claes F, Kerley-Hamilton JS, Zhuang ZW, Giordano FJ, Carmeliet P, Simons M (2010) VEGF receptor 2 endocytic trafficking regulates arterial morphogenesis. Dev Cell 18(5):713–724. doi:10.1016/j.devcel.2010.02.016

Langdon Y, Tandon P, Paden E, Duddy J, Taylor JM, Conlon FL (2012) SHP-2 acts via ROCK to regulate the cardiac actin cytoskeleton. Development 139(5):948–957. doi:10.1242/dev.067579

Li T, Du P, Xu N (2010) Identifying human kinase-specific protein phosphorylation sites by integrating heterogeneous information from various sources. PLoS ONE 5(11):e15411. doi:10.1371/journal.pone.0015411

Lopez-Ruiz P, Rodriguez-Ubreva J, Cariaga AE, Cortes MA, Colas B (2011) SHP-1 in cell-cycle regulation. Anticancer Agents Med Chem 11(1):89–98

Mahmood S, Kanwar N, Tran J, Zhang ML, Kung SK (2012) SHP-1 phosphatase is a critical regulator in preventing natural killer cell self-killing. PLoS ONE 7(8):e44244. doi:10.1371/journal.pone.0044244

Manning G, Plowman GD, Hunter T, Sudarsanam S (2002a) Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 27(10):514–520

Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002b) The protein kinase complement of the human genome. Science 298(5600):1912–1934

McConnell JL, Wadzinski BE (2009) Targeting protein serine/threonine phosphatases for drug development. Mol Pharmacol 75(6):1249–1261. doi:10.1124/mol.108.053140

Menor M, Baek K, Poisson G (2012) Probabilistic prediction of protein phosphorylation sites using classification relevance units machines. ACM SIGAPP Appl Comput Rev 12(4):8–20. doi:10.1145/2432546.2432547

Mustelin T, Feng GS, Bottini N, Alonso A, Kholod N, Birle D, Merlo J, Huynh H (2002) Protein tyrosine phosphatases. Front Biosci 7:d85–d142

Myers MP, Andersen JN, Cheng A, Tremblay ML, Horvath CM, Parisien JP, Salmeen A, Barford D, Tonks NK (2001) TYK2 and JAK2 are substrates of protein-tyrosine phosphatase 1B. J Biol Chem 276(51):47771–47774. doi:10.1074/jbc.C100583200

Neel BG, Gu H, Pao L (2003) The 'Shp'ing news: SH2 domain-containing tyrosine phosphatases in cell signaling. Trends Biochem Sci 28(6):284–293. doi:10.1016/S0968-0004(03)00091-4

Pani G, Kozlowski M, Cambier JC, Mills GB, Siminovitch KA (1995) Identification of the tyrosine phosphatase PTP1C as a B cell antigen receptor-associated protein involved in the regulation of B cell signaling. J Exp Med 181(6):2077–2084

Pellegrini MC, Liang H, Mandiyan S, Wang K, Yuryev A, Vlattas I, Sytwu T, Li YC, Wennogle LP (1998) Mapping the subsite preferences of protein tyrosine phosphatase PTP-1B using combinatorial chemistry approaches. Biochemistry 37(45):15598–15606. doi:10.1021/bi981427+

Ren L, Chen X, Luechapanichkul R, Selner NG, Meyer TM, Wavreille AS, Chan R, Iorio C, Zhou X, Neel BG, Pei D (2011) Substrate specificity of protein tyrosine phosphatases 1B, RPTPalpha, SHP-1, and SHP-2. Biochemistry 50(12):2339–2356. doi:10.1021/bi1014453

Stebbins CC, Watzl C, Billadeau DD, Leibson PJ, Burshtyn DN, Long EO (2003) Vav1 dephosphorylation by the tyrosine phosphatase SHP-1 as a mechanism for inhibition of cellular cytotoxicity. Mol Cell Biol 23(17):6291–6299

Stuible M, Dube N, Tremblay ML (2008) PTP1B regulates cortactin tyrosine phosphorylation by targeting Tyr446. J Biol Chem 283(23):15740–15746. doi:10.1074/jbc.M710534200

Tartaglia M, Mehler EL, Goldberg R, Zampino G, Brunner HG, Kremer H, van der Burgt I, Crosby AH, Ion A, Jeffery S, Kalidas K, Patton MA, Kucherlapati RS, Gelb BD (2001) Mutations in PTPN11, encoding the protein tyrosine phosphatase SHP-2, cause Noonan syndrome. Nat Genet 29(4):465–468. doi:10.1038/ng772

Timmerman I, Hoogenboezem M, Bennett AM, Geerts D, Hordijk PL, van Buul JD (2012) The tyrosine phosphatase SHP2 regulates recovery of endothelial adherens junctions through control of beta-catenin phosphorylation. Mol Biol Cell 23(21):4212–4225. doi:10.1091/mbc.E12-01-0038

Tonks NK, Diltz CD, Fischer EH (1988) Purification of the major protein-tyrosine-phosphatases of human placenta. J Biol Chem 263(14):6722–6730

Vetter SW, Keng YF, Lawrence DS, Zhang ZY (2000) Assessment of protein-tyrosine phosphatase 1B substrate specificity using "inverse alanine scanning". J Biol Chem 275(4):2265–2268

Zhao H, Lo YH, Ma L, Waltz SE, Gray JK, Hung MC, Wang SC (2011) Targeting tyrosine phosphorylation of PCNA inhibits prostate cancer growth. Mol Cancer Ther 10(1):29–36. doi:10.1158/1535-7163.MCT-10-0778