PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database

Chia-Ying Yang1, Chao-Hui Chang10, Ya-Ling Yu4, Tsu-Chun Emma Lin10, Sheng-An Lee1, Chueh-Chuan Yen7, Jinn-Moon Yang6, Jin-Mei Lai7, Yi-Ren Hong8, Tzu-Ling Tseng3, Kun-Mao Chao12,3,9, and Chi-Ying F. Huang1,10,11,12,∗

1Department of Computer Science and Information Engineering, 2Graduate Institute of Biomedical Electronics and Bioinformatics, 3Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 106, 4Graduate Institute of Life Sciences, National Defense Medical Center University, Taipei 114, 5Division of Hematology and Oncology, Department of Medicine, Taipei Veterans General Hospital, Taipei, 6Institute of Bioinformatics, National Chiao-Tung University, Hsinchu 300, 7Department of Life Science, Fu-Jen Catholic University, Taipei Hsinchuang 242, 8Department of Biochemistry, Faculty of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung 807, 9Department of Proteomics Technology, Industrial Technology Research Institute, 195, Section 4 Chung Hsing Road, Chutuing, Hsinchu 310, 10Institute of Clinical Medicine, 11Institute of Bio-Pharmaceutical Sciences and Technology, National Yang-Ming University, Taipei 112, Taiwan, Republic of China

ABSTRACT

Motivation: To fully understand how a protein kinase regulates biological processes, it is imperative to first identify its substrate(s) and interacting protein(s). However, of the 518 known human serine/threonine/tyrosine kinases, 35% of these have known substrates, while 14% of the kinases have identified substrate recognition motifs. In contrast, 85% of the kinases have protein–protein interaction (PPI) datasets, raising the possibility that we might reveal potential kinase–substrate pairs from these PPIs.

Results: PhosphoPOINT, a comprehensive human kinase interactome and phospho-protein database, is a collection of 4195 phospho-proteins with a total of 15738 phosphorylation sites. PhosphoPOINT annotates the interactions among kinases, with their down-stream substrates and with interacting (phospho)-proteins to modulate the kinase–substrate pairs. PhosphoPOINT implements various gene expression profiles and Gene Ontology cellular component information to evaluate each kinase and their interacting (phospho)-proteins/substrates. Integration of cSNPs that cause amino acids change with the proteins with the phospho-protein dataset reveals that 64 phosphorylation sites result in a disease phenotypes when changed; the linked phenotypes include schizophrenia and hypertension. PhosphoPOINT also provides a search function for all phospho-peptides using about 300 known kinase/phosphatase substrate/binding motifs. Altogether, 35% of these have known substrates, while 14% of the kinases have identified substrate recognition motifs.

Conclusions: The PhosphoPOINT database can be used to identify potential kinase–substrate pairs and to predict the functional characteristics of phospho-proteins.

Supplementary information: Supplementary data are available at Bioinformatics online.

∗To whom correspondence should be addressed.

1 INTRODUCTION

Protein kinase-mediated protein phosphorylation plays an essential role in a variety of cellular signaling. It has been estimated that between 30% and 50% of eukaryotic proteins may undergo phosphorylation (Cohen, 2002). Therefore, improper functioning of these kinases and their down-stream substrates is often manifest in various human diseases. Several databases such as the Protein Kinase Resource (Niedner et al., 2006), http://www.kinase.com, and KinG (Krupa et al., 2004) have provided online resources for protein kinases across various different species. In this study, we have focused on human protein kinases, which are made up of the 518 human serine/threonine/tyrosine protein kinases previously identified by Hidden Markov Model (HMM) and PSI-BLAST computational approaches (Manning et al., 2002).

It is generally accepted that protein kinases exert their functions primarily through their interacting proteins, which are referred to as physical interactions, and/or phosphorylating their down-stream substrates, which are referred to as biochemical interactions. One unique feature of kinase–substrate pairs is that the kinase does not always stably associate with its corresponding down-stream substrate because the biochemical phosphorylation reaction is transient. Moreover, it is imperative to identify the one or more phosphorylation sites of a given substrate, which is then characterized as a phospho-protein, because this facilitates studies using phosphorylation site mutants. This, in turn, allows investigation of the signals forming the kinase relay. However, the ability to distinguish between different phosphorylation sites, which may be catalyzed by various kinases, within a given protein, can be quite difficult and complex.

There have been several large-scale in vitro experiments using approaches such as oriented peptide library that have allowed the rapid generation of a kinase substrate motifs knowledge base (Hutti et al., 2004). Using the result of oriented peptide library screening, ScanSite 2.0 (Oenbauer et al., 2003) offers phosphorylation site prediction for 59 protein kinases. Using manually curated phosphorylation site datasets, numerical computational approaches, such as regular expressions with context-based rules, artificial
neural networks, support vector machines (SVMs) and HMMs, have also been developed to predict potential phosphorylation sites for several extensively studied kinases, such as cdc2. These resources include ELM Server (Puntervoll et al., 2003), GPS (Xue et al., 2005), NetPhos (Blin et al., 1999), NetPhosK (Blin et al., 2004), KinasePhos (Huang et al., 2005), KinasePhos2.0 (Wong et al., 2007) and PredPhospho (Kim et al., 2004). In addition, RLIMS-P (Yuan et al., 2006) uses a rule-based text mining approach for online mining of protein phosphorylation information from MEDLINE abstracts. This contrasts with Phospho3D (Zanenzo et al., 2005), which uses enriched structural information on phospho-proteins and diverse annotations at the residue level. The Kinase Pathway Database (Koike et al., 2003) integrates protein-protein, protein-gene, and protein-compound interaction data obtained by automatic literature extraction processes. However, there has been a recent exponential increase in protein phosphorylation sites identified by mass spectrometry, see Piccirillo et al. (2002) for example, and these results have significantly increased the amount of phosphorylation site information available in phosphorylation databases such as Phospho.ELM (Diella et al., 2008), HPREDB (Mishra et al., 2006), SwissProt (http://www.expasy.ch/), MitoCheck (http://www.mitocheck.org/), and PhosphoSite (http://www.phosphosite.org/).

One of the challenges in the kinase field is to reveal which kinase may phosphorylate these newly identified phospho-proteins. One possible approach is to delineate the protein-protein interaction (PPI) for kinases and phospho-proteins. Although Phospho.ELM has utilized MINT (Chatr-aryamontri et al., 2007), a PPI database, to point out potential human kinase–substrate pairs, the integration of phospho-protein information into these kinase–substrate pairs remains incomplete.

In this study, we establish PhosphoPOINT, a comprehensive human kinase interactome and phospho-protein database. PhosphoPOINT integrates 4195 phospho-proteins, 518 serine/threonine/tyrosine kinases, and their corresponding PPI datasets with the goal of delineating the interactions among kinases, their potential substrates and their interacting (phospho)-proteins (Fig. 1A). In order to uncover novel substrates for specific kinases, we have incorporated various gene expression datasets and cellular component information from Gene Ontology (GO) allowing annotation of the kinases and their interacting proteins. Finally, PhosphoPOINT also annotates any amino acids near the phosphorylation sites where a cSNP may cause a phosphorylation site to be lost, and at the same time identifies how such alteration of the phosphorylation site may lead to human disease.

2 MATERIALS AND METHODS

2.1 Collections of human protein kinases and their corresponding PPI datasets

PhosphoPOINT aims to provide a human kinase interactome resource. We have previously integrated several publicly accessible PPI datasets and systematically re-organized these datasets to establish a PPI database, POINT (http://point.bioinformatics.tw/) (Huang et al., 2004). In this study, we have first updated the experimental PPI datasets in POINT to include a total of 44,356 human PPIs corresponding to 9963 proteins. The system architecture (Fig. 1) is primarily based on the previously identified 518 human serine/threonine/tyrosine protein kinases. In addition, we have included in PhosphoPOINT an additional 149 protein kinases that have kinase domains in their protein sequences as determined by a search of the NCBI CDD database (http://www.ncbi.nlm.nih.gov/sitem/entrez?db=cdd) such as PIK3C3 kinase, a lipid kinase.

2.2 Collections of human phospho-proteins

We have integrated three existing databases, including Phospho.ELM (release 6.0, total 9236 phosphorylation sites), HPREDB (release 6, total 8992 phosphorylation sites), SwissProt (release 51.5, total 6529 phosphorylation sites), and our manually curated 406 kinase–substrate pairs, which are primarily from review articles. The integrated phosphorylation sites corresponded to 1958 human protein sequences. Together, PhosphoPOINT collected 4195 phospho-proteins corresponding to 15,738 phosphorylation sites, including 10,937 serine (S), 2425 threonine (T), and 2576 tyrosine (Y) non-redundant phosphorylation sites (Table 1).

2.3 Data Analysis

PhosphoPOINT focuses on dissecting the relationships among the human 518 kinases and their interacting (phospho)-proteins and known substrates using two different levels of annotation, namely, protein and residue (Fig. 1A). At the protein level, kinases exhibit four kinds of links with their interacting (phospho)-proteins and/or substrates, namely, physically interacting proteins (Category 1), interacting phospho-proteins (Category 2), substrates for a biochemical interaction (Category 3), and substrates as well as interacting phospho-proteins (Category 4). Data analysis revealed that most of these 518 protein kinases (85%) have their interaction proteins. In contrast, only 180 kinases (35%) have their corresponding substrates in PhosphoPOINT and these relationships are an over-estimated because many kinases in the same kinase group (e.g. the CK2 group) have similar properties and the limited studies available cannot distinguish whether a particular substrate is phosphorylated by, for example, CSNK2A1 or CSNK2A2. In addition, of the remaining 338 kinases (65%), 261 kinases interact with one or more phospho-proteins in Category 2, raising the possibility that we might be able to reveal potential kinase–substrate pairs from these kinase interacting phospho-proteins.

In addition, PhosphoPOINT collected 4195 phospho-proteins corresponding to 15,738 phosphorylation sites (Table 1). After analyzing these phosphorylation sites, several interesting observations were revealed. First, ~40% of these phosphorylation sites were obtained from low-throughput studies and these contain 6329 phosphorylation sites and 5094 (about 80%) of them can be phosphorylated by at least one kinase. Interestingly, only 679 (4.3%) of the phosphorylation sites were found by both high-throughput and low-throughput studies, which suggests that there is a need to rapidly identify the potential up-stream kinases for these uncharacterized phospho-proteins. Second, most of the 4195 phospho-proteins (81%) contain only one to five phosphorylation sites, but, nevertheless, 18 phospho-proteins are highly phosphorylated with more than 30 phosphorylation sites (Supplementary Fig. 1). Third, 295 phospho-proteins, corresponding to 872 phosphorylation sites (Nousiainen et al., 2006), have been identified as related to mitosis and this percentage is an under-estimated simply because these are no detailed annotation available from most of the low-throughput studies. Fourthly, an analysis of the number of interacting phospho-proteins for each kinase group indicates that the RGC group of kinases, one out of ten kinase groups (Manning et al., 2002), has the lowest number of interacting phospho-proteins (Supplementary Fig. 2).

2.4 Annotation of the kinase–substrate pairs by matching syn-expression

To detect potential kinase–substrate pairs, we asked whether we could annotate kinases and their interacting (phospho)-proteins by incorporating gene expression datasets and GO cellular component information to pinpoint potential substrates for each kinase. Of the available microarray analysis tools, cluster analysis (Eisen et al., 1998) has been used the most to group together genes with similar gene expression patterns.
A B

**Fig. 1.** The system architecture of PhosphoPOINT, a human kinase interactome resource. PhosphoPOINT has integrated human protein kinases, phospho-proteins, and PPI datasets with the goal to delineate four kinds of links among kinases. These include their interacting proteins (2947 links in Category 1), substrates (Category 3), and substrates as well as interacting phospho-proteins (Category 4). Some of these interacting proteins for kinases are phospho-proteins (3974 links in Category 2), which might have the potential to serve as substrates for the interacting kinases. (A) By calculating the Pearson correlation from the NCI60 and GNF gene expression data or the coverage of these four categories in the same module gene set or meta gene set, and analyzing the similarity of cellular component terms from Gene Ontology (GO) for these four categories, we are able to provide possible functional links for each kinase-substrate pair. Annotation at the residue level includes the addition of Protein Data Bank (PDB) information for these phospho-peptides. Integration of cSNPs, which cause amino acid alteration, further reveals whether the reported phosphorylation sites may be lost, which may lead to functional alteration. (B) List of search programs in PhosphoPOINT. PhosphoPOINT has implemented three different programs, including (1) a database search by kinase gene symbol (e.g. AURKA), Entrez GeneID (e.g. 6790 for AURKA) and SwissProt terms (e.g. STK6_Human), (2) the inputting of user-defined or any of about 300 known kinase/phosphatase substrate/binding motifs to search for possible phospho-peptides, and (3) a search involving user-defined protein sequences as an input to search for human phospho-proteins.

**Table 1.** A summary of four human phospho-protein datasets in PhosphoPOINT

<table>
<thead>
<tr>
<th>Category</th>
<th>PhosphoPOINT</th>
<th>Phospho.ELM</th>
<th>HPRD</th>
<th>SwissProt</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Phosphoproteins</td>
<td>4195</td>
<td>2691</td>
<td>2945</td>
<td>1935</td>
</tr>
<tr>
<td>No. of S/T Phospho-Proteins</td>
<td>3117</td>
<td>2047</td>
<td>2348</td>
<td>1414</td>
</tr>
<tr>
<td>No. of Y Phospho-Proteins</td>
<td>480</td>
<td>324</td>
<td>318</td>
<td>346</td>
</tr>
<tr>
<td>No. of S/T/Y Phospho-Proteins</td>
<td>598</td>
<td>320</td>
<td>279</td>
<td>175</td>
</tr>
<tr>
<td>Phosphorylated (S, T, Y) residues from High-Throughput (HTP)</td>
<td>10,937</td>
<td>2,425,276</td>
<td>6,580,1343,1313</td>
<td>6,041,1459,1492</td>
</tr>
<tr>
<td>Total Phosphorylation Sites</td>
<td>15,738</td>
<td>9236</td>
<td>8992</td>
<td>6,529</td>
</tr>
<tr>
<td>Phosphorylated (S, T, Y) residues from Low-Throughput (LTP)</td>
<td>38,281,152</td>
<td>17,09,542,683</td>
<td>5,837,1388,1511</td>
<td>–</td>
</tr>
<tr>
<td>Intersected (S, T, Y) residues from both HTP and LTP</td>
<td>420,97,162</td>
<td>179,31,85</td>
<td>302,87,47</td>
<td>–</td>
</tr>
</tbody>
</table>

PhosphoPOINT has integrated three existing phospho-protein databases, including Phospho.ELM, HPRD, SwissProt, and our manually curated 400 kinase-substrate pairs. Integration of these four datasets results in a collection of 4195 phospho-proteins with 15,738 phosphorylation sites, consisting of 10,937 serine (S), 2,425 threonine (T), 2,376 tyrosine (Y) residues. Among these phosphorylation sites, 7,846 (6,152+995+696) are from high-throughput (HTP) screening, 6,329 (3,824+1,152+1,349) are from low-throughput (LTP) analysis, and only 679 (420+97+162) are both from HTP and LTP screening. One special note is that there are 887 phosphorylation sites, which do not have annotation from literature in the SwissProt database and it is not possible distinguish whether these are from HTP or LTP.

This permits the exploitation of the overwhelming volume of microarray data and allows an attempt to identify genes that have similar functions or are involved together in the mediation of related biological functions/pathways (e.g. kinase-mediated signaling). This approach is referred to as syn-expression (Nichols and Pollet, 1999). Therefore, we have first calculated the Pearson correlation for the four kinds of pairs (Fig. 1 A, Category 1–4) using gene expression data from the NCI60 and GNF. Briefly, the National Cancer Institute’s Developmental Therapeutics Program has carried out an intensive studies of 60 cancer cell lines, which were derived from tumors of a variety of tissues and organs, namely the NCI60 dataset (Ross et al., 2000); furthermore, Su et al. (2002) conducted gene expression profiling of 79 human tissues and cell lines, namely, the

**Downloaded from** https://academic.oup.com/bioinformatics/article-abstract/24/16/i14/201617 by guest on 23 March 2018
GNF dataset. In addition, to avoid any bias from a single approach, we also adopted a module gene dataset (Segal et al., 2004), an integrated analysis of 1975 published microarrays spanning 22 tumor types and a meta gene dataset (Stuart et al., 2003), which identified pairs of genes that are syn-expression over 3182 microarrays from humans, flies, worms, and yeast and found 22 163 such syn-expression relationships, each of which has been conserved across evolution. The overall aim was to uncover sets of genes that act in concert to carry out a specific function and to annotate the potential kinase-substrate pairs from these groupings.

3 APPLICATION AND RESULTS

3.1 About 92% phospho-peptides can be matched to 128 known substrate motifs corresponding to about 70 kinases

It is generally believed that the substrates of a protein kinase can be phosphorylated at specific sites based on consensus sequences/motifs/functional patterns (Kregepınar et al., 1998). A total collection of 293 known kinase/phosphatase substrate/binding motifs can be obtained from HPGR PhosphoMotif Finder (http://www.hprd.org/PhosphoMotif_finder). Several phospho-protein binding motifs, which are not in HPRD collection, such as PIN1 (Lu et al., 1999), were also included in PhosphoPOINT. Supplementary Figure 3 reports an analysis of about 300 kinase/phosphatase substrate/binding motifs. About 100 kinase/phosphatase binding motifs analyzed, the WW domain binding motif (pS/pT) maps to the greatest number of phospho-peptides (4854). 128 substrate motifs are recognized by about 70 ser/thr and tyr kinases, suggesting that the substrate motif datasets for each ser/thr/tyr kinase are far from complete and further exploration of the substrate specificity for each kinase is required.

When these 128 substrate motifs were used to scan the collected human phospho-peptides, about 92% (14 583) of the phospho-peptides can be matched to one or more substrate motifs (Supplementary Fig. 3). Not surprisingly, the GSK-3, ERK1, ERK2, and CDK5 substrate motif ([pS/pT〈]) maps to 4854 and 4642 phospho-peptides, respectively, suggesting that the specificity of these short substrate motifs may not be sufficient to trace back their corresponding kinases. In addition, only about 70 kinases have their corresponding substrate recognition motifs, implying that we cannot simply employ these motifs to pinpoint kinase-substrate pairs.

3.2 Using gene expression data and GO cellular component information to explore the potential kinase-substrate pairs

PhosphoPOINT aims to provide insights into the interactome of human 518 protein kinases with their potential substrates and their interacting (phospho)-proteins. As described above, we cannot effectively use substrate motifs to pinpoint potential kinase-substrate pairs. In addition, the interacting phospho-proteins for ser/thr kinases exist as both ser/thr and tyr phospho-proteins, or vice versa, suggesting that addition criteria are required to uncover kinase-substrate pairs from the database of kinase and their interacting phospho-proteins. One possible approach is to use genome-wide microarray datasets, which have dramatically expedited comprehensive understanding of the gene expression profiles and provide insights into the molecular mechanisms of gene function. This is despite the fact that analyzing vast numbers of microarray datasets does not necessarily provide a comprehensive understanding of the role of a given gene in a specific biological process.

We have previously shown that the use of the AURKA gene expression profiles to search for and compare transcriptome expression profiles in publicly accessible microarray datasets is able to uncover novel substrate (e.g. HURP) of AURKA (Yu et al., 2005). To further illustrate this point, we used one of the AURKA known substrates, TPX2 (Kofer et al., 2002), as an example to examine the relationship between AURKA and TPX2 in esophageal squamous cell carcinoma (ESCC) specimens by Q-RT-PCR. As shown in Figure 2, both AURKA and TPX2 are up-regulated in ESCC and share similar expression patterns as examined by a Pearson’s correlation coefficient test (r>0.8). This supports the view that commonly expression clusters (or syn-expression) obtained from microarray datasets may be able to pinpoint functionally linked kinase-substrate pairs. To provide further evidence for the use of syn-expression, we have calculated the Pearson correlation for our manually curated 400 kinase-substrate pairs (Category 3 and 4) using gene expression datasets such as NCI60 and GNF as described in the method section. Approximately 25% of the kinase-substrate pairs exhibit high correlation (r>0.7), further supporting the idea that kinase-substrate pairs may, at least in part, be uncovered in this manner.

Based on the above, we calculated the Pearson correlation of NCI60 and GNF datasets for the four kinds of pairs (Fig. 1A, Category 1–4) to detect potential kinase-substrate pairs. Most (between 46% and 66%) of the kinase-substrate/interacting (phospho)-protein pairs belong to the medium correlation (0<r<0.3) group and only 3.4%, 5.3%, 3.8%, and 13.6% belongs to high correlation (r>0.7) in Category 1–4, respectively. The result of this analysis is similar to a recent study (Bhardwaj and Lu, 2005). Similarly, the coverage of these four categories and our curated dataset in the same module gene set or meta gene set are 10.6%, 12.4%, 5.9%, 20.4% and 31.0%, respectively. Measurement of the similarity of cellular component terms from GO is also evaluated and the highly similarity proportion are 14.4%, 18.4%, 14.9%, 27.2% and 32.9% in Category 1–4 and our curated dataset, respectively. It is likely that the incorporations of gene expression datasets into kinase-substrate/interacting (phospho)-protein pairs using high Pearson’s correlation coefficient (r>0.7) as a criteria may dramatically lose coverage of potential kinase-substrate pairs, as exemplified by our curated dataset. On the other hand, our curated dataset has a higher correlation ratio than Category 1–4, raising the possibility that some of the collected datasets might be not accurate and that, as a result, a setting of r>0.7 might increase accuracy when uncovering novel kinase-substrate pairs.

Finally, in PhosphoPOINT, we have incorporated the gene expression datasets from NCI60 and GNF (Pearson correlation r>0.7), the module gene dataset, the meta gene dataset, and GO cellular component information to analyze kinase-substrate/interacting (phospho)-protein pairs across the four kinds of categories and this resulted in 458 pairs (Fig. 3). Figure 3 reports the proportion of each category in 458 pairs. Among these 458 pairs, 38.4% (176 kinase–substrate pairs) and 40.0% (183 kinase–substrate pairs) are from Category 4 and Category 2, respectively. Moreover, 61 pairs can be found in both Category 4 and our curated dataset. Therefore, 183 pairs in Category 2 should be tested
Fig. 2. Similar expression patterns for \textit{AURKA} and \textit{TPX2} from an analysis of 15 adjacent normal-tumor matched esophageal squamous cell carcinoma (ESCC) specimens. Q-RT-PCR was used to examine the gene expression patterns of \textit{AURKA} (A) and \textit{TPX2} (B), which were found to be both up-regulated, in ESCC patient specimens.

C.-Y. Yang et al.

Fig. 3. Prioritized kinase-substrate pairs by implementation of gene expression and GO cellular component datasets. A total 458 kinase pairs belong to the same module gene, or meta gene, or highly correlated (Pearson correlation $r \geq 0.7$) in NCI60 or GNF datasets and with highly similarity of cellular component terms from GO in PhosphoPOINT. Among these pairs, 40.0\% (183 pairs) and 38.4\% (176 pairs) are from Category 2 and Category 4, respectively.

3.3 Phosphorylation site alteration caused by cSNP

The amino acids around the phosphorylation site play a pivotal role in the recognition of distinct protein kinases. Coding Single Nucleotide Polymorphism (cSNP) is a DNA sequence variation occurring when a single nucleotide differs between members of a species and causes an amino acid residue to change within the protein sequence. To address whether the collected phosphorylation sites may exhibit any alterations, we have gathered cSNP information from NCBI dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/) (download, May 2007). Interestingly, of 109 262 cSNP data points, 1515 cSNPs occur in the region flanking with (position $-7$ to $+7$) the observed phosphorylation site (position 0) (Supplementary Fig. 4) and thus might influence the recognition of a kinase towards its preferred substrate. First, 64 phosphorylation sites were potentially lost and some examples are listed in Table 2. Briefly, S470N of synapsin III has been identified as a polymorphism possibly linking to schizophrenia using 118 schizophrenia individuals and 330 population controls for association analysis. S470 is phosphorylated during early development by MAP kinase, which is a downstream effector of neurotrophins action (Porton et al., 1999). Second, changes in amino acids near the phosphorylation sites might influence recognition between the kinase and its downstream substrate. For example, the substrate recognition motif for the ATM kinase is pS$^{Q(}$Schwartz and Gygi, 2005$^{)}$ and ATM may phosphorylate TP53 at S$^{Q(46)}$. Alteration of the amino acid at Q$^{47}$ to P by cSNP might influence the recognition of ATM toward TP53.

4 CONCLUSION AND FUTURE PERSPECTIVES

To fully understand how a protein kinase regulates biological processes, it is imperative to first identify its substrate(s) and interacting protein(s). PhosphoPOINT not only integrates various kinase and phospho-protein datasets, but also provides robust annotation that helps to bridge between the kinase and its potential substrates using available PPI and cSNP datasets. Of the 518 included kinases, about 35\% of the kinases have known substrates and about 14\% of the kinases have substrate recognition motifs. In contrast, 85\% of the kinases interact with one or more phospho-proteins, raising the possibility that it should be possible to reveal many more potential kinase–substrate pairs from these handful interacting phospho-proteins by, for example, incorporating gene expression datasets and cellular component information from GO. This web server may ultimately augment our predictive power and accelerate the functional characterization of those poorly analyzed kinases and their possible regulatory mechanisms. The PhosphoPOINT database will be regularly updated as soon as Phospho.ELM, HPRD or SwissProt datasets are released and can be freely accessed on request. The information related to the use of this database can be downloaded from this website.
Table 2. Lists of phosphorylation site alteration caused by cSNP

<table>
<thead>
<tr>
<th>Symbol</th>
<th>cSNP</th>
<th>Allele</th>
<th>Altered phosphorylation mutant phenotype</th>
<th>Mutants in paper</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYN3</td>
<td>rs5995226</td>
<td>A</td>
<td>The polymorphism in SYN3 may associate with schizophrenia.</td>
<td>S270N</td>
<td>14732590</td>
</tr>
<tr>
<td>SCN11A</td>
<td>rs2228576</td>
<td>G</td>
<td>ENaC polymorphisms altering functional channel expression may contribute to the development of hypertension.</td>
<td>T663A</td>
<td>16174865</td>
</tr>
<tr>
<td>XRCC1</td>
<td>rs2307184</td>
<td>A</td>
<td>The mutant ablates the rapid cellular DNA single-strand breaks.</td>
<td>S485Y</td>
<td>8 Ser mutations to Ala. 15066279</td>
</tr>
</tbody>
</table>

Integration of the phospho-proteins and cSNP databases shows that the listed phospho-proteins may undergo alteration of their original function and this could result in a phenotype change (Detail lists are in Supplementary Table 1).

Fig. 4. Annotation and visualization in PhosphoPOINT. To visualize the human kinase (e.g. AURKA) interacting proteins, we have used various colors to represent the different types of nodes (circle/edges (line)). The different node colors represent the phosphorylation status of each protein (e.g. red stands for this protein containing ser/thr phosphorylation sites). Edges represent the relationship between kinase and its interacting (phospho)-proteins/substrates. (Right) For each kinase and its interacting (phospho)-proteins/substrates pairs, we have calculated the Pearson correlation (solid rectangle) of each pair using the NCI60 and GNF microarray datasets. The red solid rectangle means a highly positive correlation ($r \geq 0.7$) between each pairs. The similarity of cellular component terms from GO for each kinase pairs is also illustrated in the third line with a hollow rectangle.

5 WEBSITE INTERFACE

PhosphoPOINT provides three main search programs (Fig. 1B). First, the Database Search program can be used to optionally input a kinase or phospho-proteins gene symbol, Entrez GeneID and SwissProt terms. Figure 4 uses AURKA kinase as an example to illustrate such a search result. Next, the Motif Scan program allows either user-defined or known kinase/phosphatase substrate/binding motifs as an input for a search of possible phospho-peptides. Finally, the Protein Sequence Match program allows the user to search for possible human phospho-proteins using user-defined protein sequences.

ACKNOWLEDGEMENTS

Funding: This research was supported by grants from NSC (NSC95-2627-B-010-071-MY3) to C.-Y.Huang, NSC95-2627-B-030-003 to J.-M.Lai, NSC95-2627-B-194-001 to F.-S.Wang, NSC95-2627-B-002-011 to C.-Y.Kao, and NSC95-2221-E-002-126-MY3 to K.-M.Chao.

Conflict of Interest: none declared.

REFERENCES

et al.


---


