BIOINFORMATICS

Vol. 24 ECCB 2008, pages i14–i20
doi:10.1093/bioinformatics/btn297

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

Chia-Ying Yang[1], Chao-Hui Chang[10], Ya-Ling Yu[4], Tsu-Chun Emma Lin[10], Sheng-An Lee[1], Chueh-Chuan Yen[5], Jinn-Moon Yang[6], Jin-Mei Lai[7], Yi-Ren Hong[8], Tzu-Ling Tseng[9], Kun-Mao Chao[1,2,3,*] and Chi-Ying F. Huang[1,10,11,12,*]

[1]Department of Computer Science and Information Engineering, [2]Graduate Institute of Biomedical Electronics and Bioinformatics, [3]Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 106, [4]Graduate Institute of Life Sciences, National Defense Medical Center University, Taipei 114, [5]Division of Hematology and Oncology, Department of Medicine, Taipei Veterans General Hospital, Taipei, [6]Institute of Bioinformatics, National Chiao-Tung University, Hsinchu 300, [7]Department of Life Science, Fu-Jen Catholic University, Taipei Hsinchuang 242, [8]Department of Biochemistry, Faculty of Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung 807, [9]Department of Proteomics Technology, Industrial Technology Research Institute, 195, Section 4 Chung Hsing Road, Chutung, Hsinchu 310, [10]Institute of Clinical Medicine, [11]Institute of Bio-Pharmaceutical Sciences and [12]Institute of Biotechnology in Medicine, 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, PhosphoPOINT provides robust annotation for kinases, their down-stream substrates and their interaction (phospho)-proteins and this should accelerate the functional characterization of kinome-mediated signaling.

Availability: PhosphoPOINT can be freely accessed in http://kinase.bioinformatics.tw/

Contact: cyhuang5@ym.edu.tw; kmchao@csie.ntu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

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 (Obenauer *et al.*, 2003) offers phosphorylation site prediction for 59 protein kinases. Using manually curated phosphorylation site datasets, numerous computational approaches, such as regular expressions with context-based rules, artificial

*To whom correspondence should be addressed.

© The Author 2008. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014

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 (Blom *et al.*, 1999), NetPhosK (Blom *et al.*, 2004), KinasePhos (Huang *et al.*, 2005), KinasePhos 2.0 (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 (Zanzoni *et al.*, 2007), 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. Moreover, there has been a recent exponential increase in protein phosphorylation sites identified by mass spectrometry, see Ficarro *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), HPRD (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 up-dated 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/sites/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), HPRD (release 6, total 8992 phosphorylation sites), SwissProt (release 51.5, total 6529 phosphorylation sites), and our manually curated 400 kinase–substrate pairs, which are primarily from review articles. The integrated phosphorylation sites are blasted against human protein sequences. Together, PhosphoPOINT collected 4195 phospho-proteins corresponding to 15 738 phosphorylation sites, including 10 937 serine (S), 2425 threonine (T), and 2376 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 kinases 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.

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014

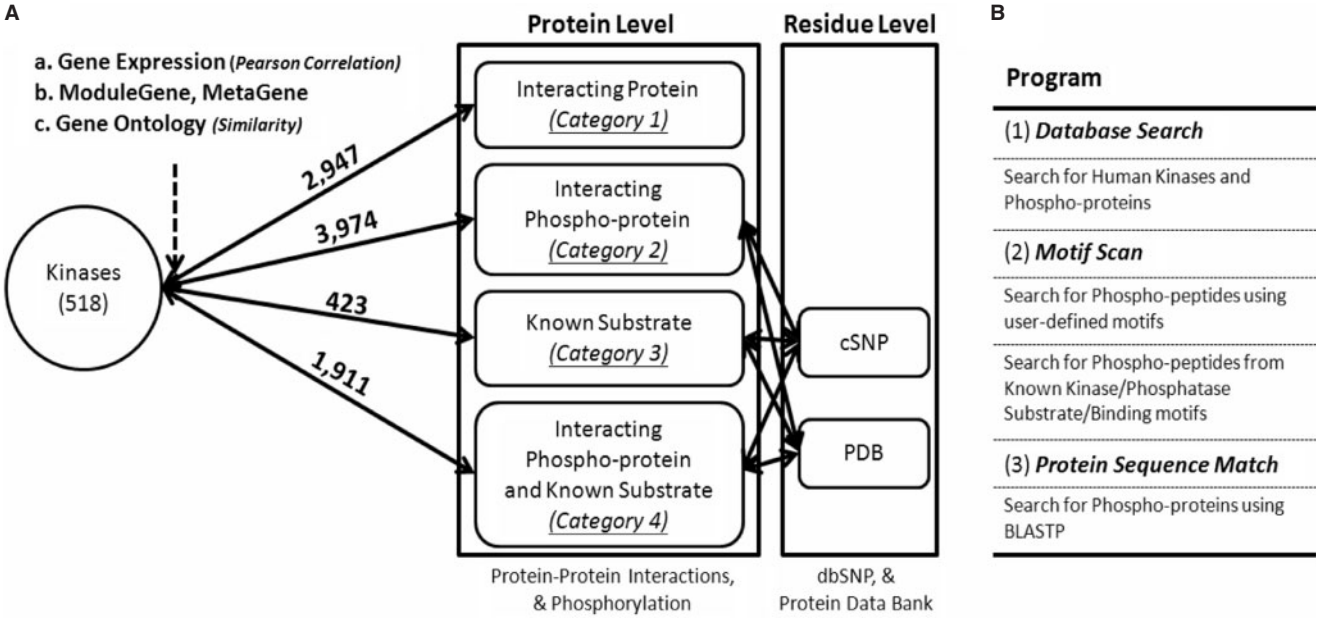

Fig. 1. The system architecture of PhosphoPOINT, a human kinase interactome resource. PhosphoPOINT has integrated human protein kinases, phospho-proteins, and PPI datasets with the goal to delineate four kinds of links among kinases. These include their interacting proteins (2947 links in Category 1), 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 led 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

Category	PhosphoPOINT	Phospho.ELM	HPRD	SwissProt
No. of Phosphosp Proteins	4195	2691	2945	1935
No. of S/T PhosphoProteins	3117	2047	2348	1414
No. of Y PhosphoProteins	480	324	318	346
No. of S/T/Y PhosphoProteins	598	320	279	175
Phosphorylated (S;T;Y) residues	10 937; 2425; 2376	6580; 1343; 1313	6041; 1459; 1492	4357; 1007; 1165
Total Phosphorylation Sites	15 738	9236	8992	6529
Phosphorylated (S;T;Y) residues from High-Throughput (HTP)	6152; 995; 696	4659; 773; 543	1146; 273; 245	–
Phosphorylated (S;T;Y) residues from Low-Throughput (LTP)	3828; 1152; 1349	1769; 542; 683	5837; 1388; 1511	–
Intersected (S;T;Y) residues from both HTP und LTP	420; 97;162	179; 31; 85	302; 87; 47	–

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), 2425 threonine (T), 2376 tyrosine (Y) residues. Among these phosphorylation sites, 7843 (6152+995+696) are from high-throughput (HTP) screening, 6329 (3828+1152+1349) 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 (Niehrs 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 http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014

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 (Kreegipuu *et al.*, 1998). A total collection of 293 known kinase/phosphatase substrate/binding motifs can be obtained from HPRD 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]P) 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 (X[pS/pT]P) and the *GRK1* substrate motif (X[pS/pT]XXX[A/P/S/T]) map 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* (Kufer *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 ($r \geq 0.7$) 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 $r \geq 0.7$ 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 $r \geq 0.7$), 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

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014

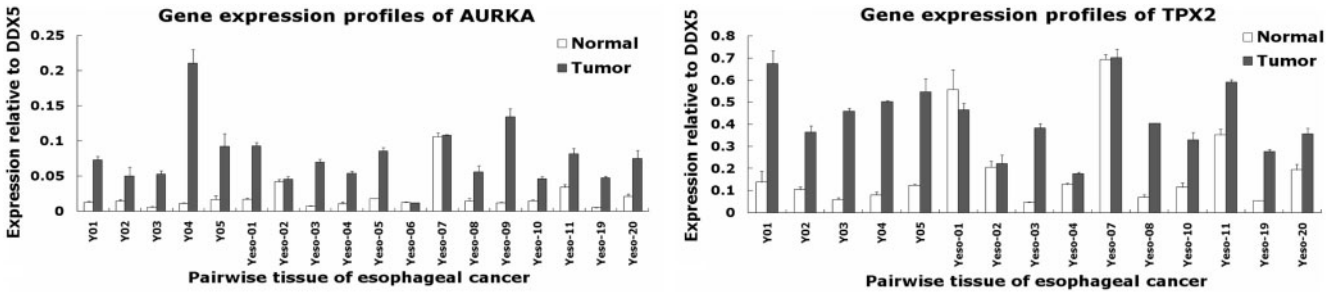

Fig. 2. Similar expression patterns for *AURKA* and *TPX2* from an analysis of 15 adjacent normal-tumor matched esophageal squamous cell carcinoma (ESCC) specimens. Q-RT-PCR was used to examine the gene expression patterns of *AURKA* (A) and *TPX2* (B), which were found to be both up-regulated, in ESCC patient specimens.

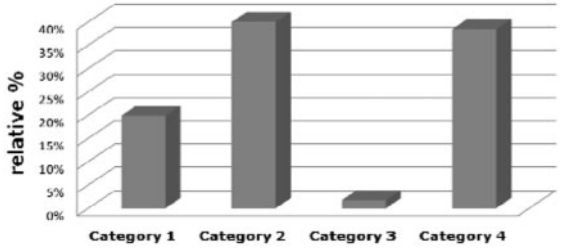

Fig. 3. Prioritized kinase-substrate pairs by implementation of gene expression and GO cellular component datasets. A total 458 kinase pairs belong to the same module gene, or meta gene, or highly correlated (Pearson cor relation $r \geq 0.7$) in NCI60 or GNF datasets and with highly similarity of cellular 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.

rigorously to explore the potential kinase–substrates relationships. In fact, several kinase–substrate pairs, i.e. CDC2-BRCA1 from Category 2, have already been demonstrated in literature (Ruffner *et al.*, 1999), but were not recorded in our database. Together, this data-mining method (syn-expression) is just an initial attempt to answer a number of complex biological questions. Nonetheless, the current approaches may ultimately allow implementation of available biochemical methods to facilitate a greater understanding of down-stream target prioritization and a delineation of the cellular pathways governing kinase–substrate pairs.

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 cSNPs 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 down-stream effecter of neurotrophins action (Porton *et al.*, 2004). In another example, *SCNN1A* expression products, namely epithelial sodium channels (ENaCs), are expressed in principal cells. ENaCs are functionally associated with blood pressure. PKC selectively regulates the functional and surface expression of the polymorphism T663A in hENaCs. Such differential regulation may alter hENaC function and raise the risk for developing hypertension (Yan *et al.*, 2006). Second, changes in amino acids near the phosphorylation sites might influence recognition between the kinase and its down-stream substrate. For example, the substrate recognition motif for the ATM kinase is pSQ (Schwartz and Gygi, 2005) and ATM may phosphorylate TP53 at $\underline{S}^{46}Q$. 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.

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014

Table 2. Lists of phosphorylation site alteration caused by cSNP

Symbol	cSNP	Allele	Altered	phosphorylation mutant phenotype	Mutants in paper	Reference
SYN3	rs599S526	A	S470N	The polymorphism in SYN3 may associate with schizophrenia.	S270N	14732590
SCNN1A	rs2228576	G	T663A	ENaC polymorphisms altering functional channel expression may contribute to the development of hypertension.	T663A	16174865
XRCC1	rs2307184	A	S485Y	The mutant ablates the rapid cellular DNA single-strand breaks.	8 Ser mutatnions to Ala.	15066279

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).

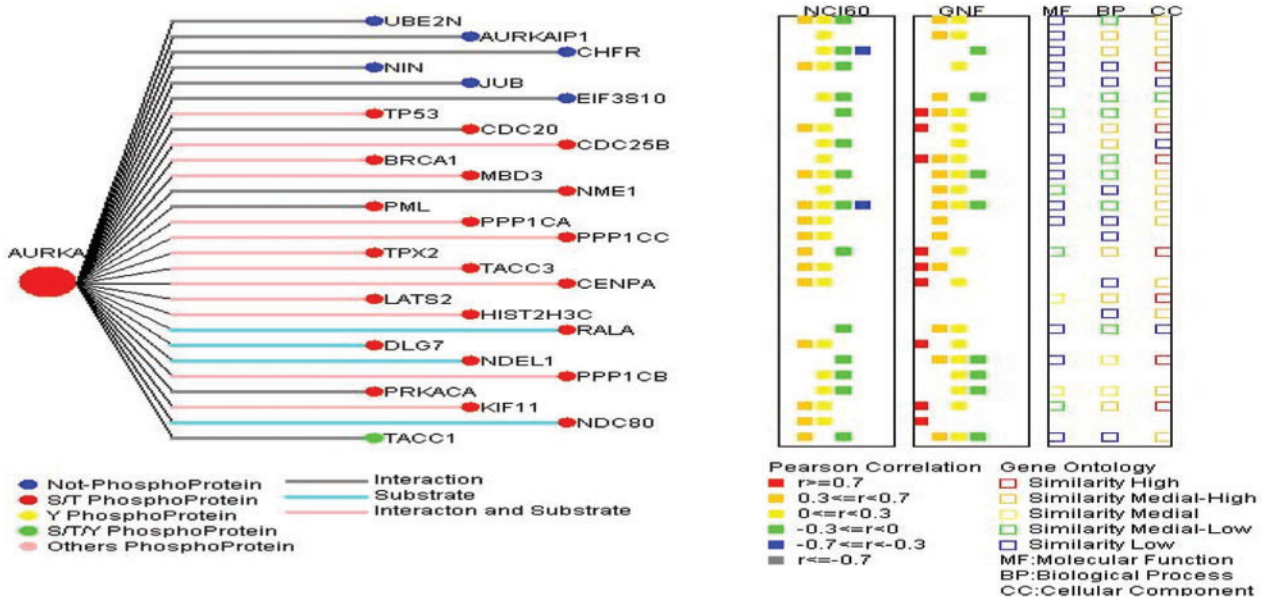

Fig. 4. Annotation and visualization in PhosphoPOINT. To visualize the human kinase (e.g. AURKA) interacting proteins, we have used various colors to represent the different types of nodes (circle)/edges (line). The different node colors represent the phosphorylation status of each protein (e.g. red 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-400-002 and NSC95-2320-B-010-071-MY3) to C.-Y.Huang, NSC95-2627-B-030-001 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

Bhardwaj,N. and Lu,H. (2005) Correlation between gene expression profiles and protein–protein interactions within and across genomes. *Bioinformatics*, **21**, 2730–2738.

Blom,N. *et al.* (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.

Blom,N. *et al.* (2004) Prediction of post-translational glycolysation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Chatr-aryamontri,A. *et al.* (2007) MINT: the Molecular INTeraction database. *Nucl. Acids Res.*, **35**, D572–D574.

Cohen, P. (2002) The origins of protein phosphorylation. *Nat. Cell Biol.*, **4**, E127–E130.

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014

Diella,F. *et al.* (2008) Phospho.ELM: a database of phosphorylation sites—update 2008. *Nucl. Acids Res.*, **36**, D240–D244.
Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
Ficarro,S.B. *et al.* (2002) Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. *Nat. Biotechnol.*, **20**, 301–305.
Huang,H.D. *et al.* (2005) KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucl. Acids Res.*, **33**, W226–W229.
Huang,T.W. *et al.* (2004) POINT: a database for the prediction of protein–protein interactions based on the orthologous interactome. *Bioinformatics*, **20**, 3273–3276.
Hutti,J.E. *et al.* (2004) A rapid method for determining protein kinase phosphorylation specificity. *Nat. Methods*, **1**, 27–29.
Kim,J.H. *et al.* (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.
Koike,A. *et al.* (2003) Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource. *Genome Res.*, **13**, 1231–1243.
Kreegipuu,A. *et al.* (1998) Statistical analysis of protein kinase specificity determinants. *FEBS Lett.*, **430**, 45–50.
Krupa,A. *et al.* (2004) KinG: a database of protein kinases in genomes. *Nucl. Acids Res.*, **32**, D153–D155.
Kufer,T.A. *et al.* (2002) Human TPX2 is required for targeting Aurora-A kinase to the spindle. *J. Cell Biol.*, **158**, 617–623.
Lu,P.J. *et al.* (1999) Function of WW domains as phosphoserine- or phosphothreonine-binding modules. *Science*, **283**, 1325–1328.
Manning,G. *et al.* (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
Mishra,G.R. *et al.* (2006) Human protein reference database—2006 update. *Nucl. Acids Res.*, **34**, D411–D414.
Niedner,R.H. *et al.* (2006) Protein kinase resource: an integrated environment for phosphorylation research. *Proteins*, **63**, 78–86.
Niehrs,C. and Pollet,N. (1999) Synexpression groups in eukaryotes. *Nature*, **402**, 483–487.
Nousiainen,M. *et al.* (2006) Phosphoproteome analysis of the human mitotic spindle. *Proc. Natl Acad. Sci. USA*, **103**, 5391–5396.
Obenauer,J.C. *et al.* (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucl. Acids Res.*, **31**, 3635–3641.
Porton,B. *et al.* (2004) A rare polymorphism affects a mitogen-activated protein kinase site in synapsin III: possible relationship to schizophrenia. *Biol. Psychiatry*, **55**, 118–125.
Puntervoll,P. *et al.* (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucl. Acids Res.*, **31**, 3625–3630.
Ross,D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.
Ruffner,H. *et al.* (1999) BRCA1 is phosphorylated at serine 1497 *in vivo* at a cyclin-dependent kinase 2 phosphorylation site. *Mol. Cell Biol.*, **19**, 4843–4854.
Schwartz,D. and Gygi,S.P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol.*, **23**, 1391–1398.
Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.
Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
Su,A.I. *et al.* (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl Acad. Sci. USA*, **99**, 4465–4470.
Wong,Y.H. *et al.* (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucl. Acids Res.*, **35**, W588–W594.
Xue,Y. *et al.* (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucl. Acids Res.*, **33**, W184–187.
Yan,W. *et al.* (2006) Differential modulation of a polymorphism in the COOH terminus of the alpha-subunit of the human epithelial sodium channel by protein kinase Cdelta. *Am. J. Physiol. Renal. Physiol.*, **290**, F279–288.
Yu,C.T. *et al.* (2005) Phosphorylation and stabilization of HURP by Aurora-A: implication of HURP as a transforming target of Aurora-A. *Mol. Cell Biol.*, **25**, 5789–5800.
Yuan,X. *et al.* (2006) An online literature mining tool for protein phosphorylation. *Bioinformatics*, **22**, 1668–1669.
Zanzoni,A. *et al.* (2007) Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucl. Acids Res.*, **35**, D229–D231.

Downloaded from http://bioinformatics.oxfordjournals.org/ at National Chiao Tung University Library on April 25, 2014