Published online 2 November 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D695–D699***
doi:10.1093/nar/gkm902

NetworKIN: a resource for exploring cellular phosphorylation networks

Rune Linding[1,2,3,*], Lars Juhl Jensen[4], Adrian Pasculescu[1], Marina Olhovsky[1], Karen Colwill[1], Peer Bork[4,5], Michael B. Yaffe[2] and Tony Pawson[1]

[1]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Canada, [2]Center for Cancer Research, Massachusetts Institute of Technology, Cambridge, USA, [3]The Institute of Cancer Research, London, UK, [4]European Molecular Biology Laboratory, Heidelberg and [5]Max-Delbrück-Centre for Molecular Medicine, Berlin, Germany

Received August 13, 2007; Revised October 3, 2007; Accepted October 4, 2007

ABSTRACT

Protein kinases control cellular responses by phosphorylating specific substrates. Recent proteome-wide mapping of protein phosphorylation sites by mass spectrometry has discovered thousands of *in vivo* sites. Systematically assigning all 518 human kinases to all these sites is a challenging problem. The NetworKIN database (http://networkin.info) integrates consensus substrate motifs with context modelling for improved prediction of cellular kinase–substrate relations. Based on the latest human phosphoproteome from the Phospho.ELM and PhosphoSite databases, the resource offers insight into phosphorylation-modulated interaction networks. Here, we describe how NetworKIN can be used for both global and targeted molecular studies. Via the web interface users can query the database of precomputed kinase–substrate relations or obtain predictions on novel phosphoproteins. The database currently contains a predicted phosphorylation network with 20 224 site-specific interactions involving 3978 phosphoproteins and 73 human kinases from 20 families.

INTRODUCTION

Dynamical protein phosphorylation governs many cell biological processes (1). Decades of targeted studies and recent progress in phosphoproteomics has resulted in a large body of protein phosphorylation data (2). Determining how these phosphorylation sites change through time, for example during the cell cycle or following exposure to extracellular stimuli is now possible with techniques such as quantitative mass spectrometry (3). However, it remains difficult to determine which of the 518 human kinases is responsible for the phosphorylation of an observed site; a glance at the Phospho.ELM database reveals that only about a quarter of known *in vivo* phosphorylation sites have been assigned as substrates of a specific kinase, and this fraction is constantly decreasing (2).

This has motivated the development of numerous computational methods for predicting kinase–substrate relations, for example, Scansite (4), NetphosK (5,6), PREDIKIN (7), PredPhospho (8) GPS (9), PPSP (10) and KinasePhos (11). These methods all rely on consensus sequence motifs recognized by the active site of the enzymes, represented by either position-specific scoring matrices (PSSMs), neural networks, support vector machines or other machine-learning representations. However, kinase specificity is known to also depend on other factors, such as auxiliary protein interactions, scaffolds, coexpression and colocalization (collectively referred to as 'context'). We recently introduced a computational framework, NetworKIN, which uses a probabilistic protein association network [STRING (12)] to model the context of kinases and substrates; combined with consensus sequence motifs, this gave a 2.5-fold leap in prediction accuracy over previous methods (13).

Here, we present a database of predicted kinase–substrate relations based on the latest human phosphoproteome and protein association network from the Phospho.ELM (2), PhosphoSite (14) and STRING (12) databases. This database is available via a web interface at http://networkin.info, which enables the user to query the database for any kinases or substrates of interest, to submit new substrates and to explore the evidence underlying a prediction.

RESOURCE OVERVIEW

The foundation of the NetworKIN algorithm is the fact that signalling proteins are modular in nature, that is

*To whom correspondence should be addressed. Tel: + 1 416 586 4800; Fax: + 1 416 586 8869; Email: linding@mshri.on.ca
Correspondence may also be addressed to Tony Pawson. Tel: + 1 416 586 8262; Fax: + 1 416 586 8869; Email: pawson@mshri.on.ca

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

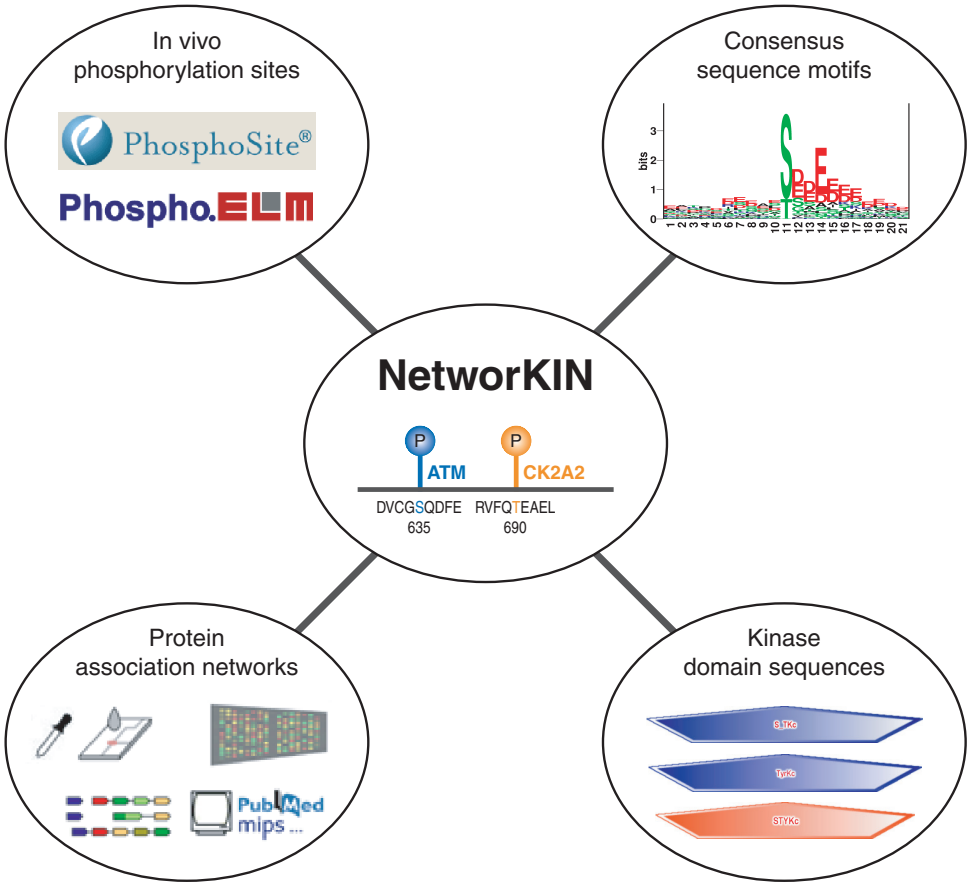


Figure 1. Overview of the NetworKIN resource. NetworKIN starts from a set of known *in vivo* phosphorylation sites, which are obtained and kept synchronized with the Phospho.ELM and PhosphoSite databases to facilitate reciprocal database cross-links. We first compare these phosphorylation sites to consensus sequence motifs from the Scansite and NetPhosK resources in order to predict possible kinase families responsible for the phosphorylation. Second, we capture the kinase and substrate context using a probabilistic functional association network from the STRING database (12). The resulting site-specific kinase–substrate interaction network is stored in a SQL database, which is accessible via a web interface.

they consist of discrete functional modules, such as protein kinase domains and the linear peptide motifs they recognize and phosphorylate. This makes it possible to model the behaviour of such proteins by coupling the prediction of linear motifs to that of identifying the corresponding binding module in a network context. Due to the improved capability of mass spectrometry to identify phosphorylation sites and other post-translational modifications, the scope of modelling these events has changed from predicting what could get phosphorylated to predicting what kinase phosphorylates which sites.

The NetworKIN algorithm is designed to work from a set of experimentally identified *in vivo* phosphorylation sites (although the algorithm can also be used *ab initio*). The precomputed results in the database are based on the latest human phosphoproteome from the Phospho.ELM and PhosphoSite databases (2,14) (Figure 1). The release cycle of the database is approximately every 3 months due to the high throughput of mass-spectrometry-driven proteomics, and we intend to keep NetworKIN up-to-date with future releases of Phospho.ELM and PhosphoSite.

These data are processed through the NetworKIN algorithm, which is implemented in Python and C. The sites are classified by matching them to a motif collection (Figure 1) based on the position-specific scoring matrices from Scansite (4) (http://scansite.mit.edu) and the artificial neural networks from NetPhosK (5) (http://www.cbs.dtu.dk/services/NetPhosK). Each consensus sequence motif is considered to be a representative for a family of closely related kinases; for example, the NetPhosK Cdk5 predictor is used for predicting possible phosphorylation sites for all cyclin-dependent kinases. Within a proteome, kinases are identified and assigned to these families based on their best hit in a BLASTP (15) sequence similarity search against a set of 82 representative kinase domain sequences, which have been manually assigned to families. Only hits with an *E*-value better than 10^{-40} and with at least 50% sequence identity are considered.

To capture the biological context of a substrate, we use a probabilistic network of functional associations extracted from the STRING database (12) (http://string.embl.de, Figure 1). This network is based on four fundamentally different types of evidence: genomic context (gene fusion, gene neighbourhood and phylogentic profiles), primary experimental evidence (physical protein interactions and gene co-expression), manually curated

pathway databases, and automatic literature mining. We showed that the three latter evidence types are of comparable importance, whereas genomic context methods contribute very little towards the predictions made by NetworKIN (13). As the curated pathway databases generally contain few errors, a confidence score of 0.9 is assigned to this type of evidence. The best candidate kinases within the appropriate kinase families are identified from a protein network of functional associations [generated using the STRING database (12)] by calculating the proximity to the substrate for all kinases, defined as the probability of the most probable path connecting them (Floyd–Warshall algorithm). The context is thus used as a filter that eliminates many of the false-positive predictions obtained from the sequence motifs and hence improves the prediction accuracy. However, the current algorithm is unable to recover sites that are missed by the sequence motifs (i.e. false-negative predictions).

The resulting predicted kinase–substrate relations are stored in a MySQL relational database. The database also contains cross-references to the Phospho.ELM (2), PhosphoSite (14) and STRING (12) databases. This database can be accessed via a web interface, which consists of a collection of CGI scripts, that query the database backend and format the results as XHTML for display in a web browser.

USING NetworKIN

The NetworKIN database can be accessed in several different ways. In the following, we will explain the various features of the web interface, using the tumour suppressor 53BP1 as an example. For large-scale analysis or visualization, most users will probably prefer to download the complete set of predictions for human phosphoproteins, which is available in tab-separated and Cytoscape format.

For all other users, the primary entry point to NetworKIN is its search interface shown in Figure 2A. The user can select a specific substrate and/or kinase to view the corresponding subset of predictions; in our example, we query for 53BP1 as the substrate and use the wildcard * to obtain predictions for all kinases. The web interface also offers an advanced search form, which enables the user to pose much more refined queries. In either case, the search results will be presented as a table in which each row shows a predicted relation between a kinase and a specific phosphorylation site in a substrate. In case of 53BP1, we get a list of 78 predictions for 39 sites and 12 kinases; the first 10 of these predictions are shown in Figure 2B. For each prediction we list two scores, namely the context score and the motif score, both of which should preferably be high. It should be noted that the motif scores for different kinase families are not comparable; in particular, motif scores from NetPhosK should not be compared with motif scores from Scansite. For this reason, the predictions for a given phosphorylation site are sorted by their context score. As the results of a single query may be extensive, the results can also be downloaded in the formats mentioned previously.

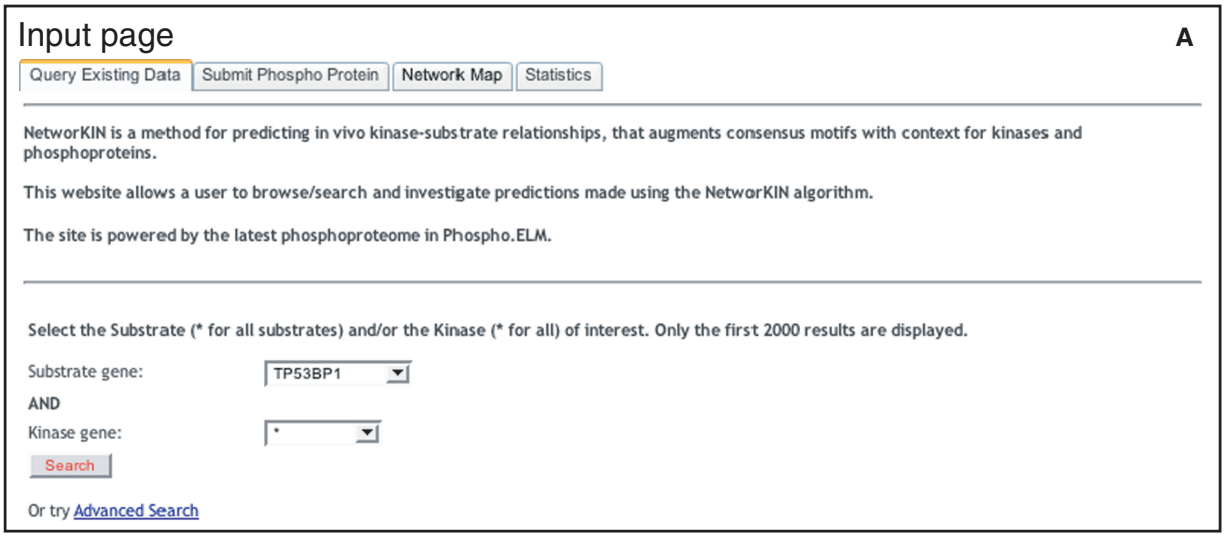
Furthermore, the user can investigate the predictions in greater detail via the web interface. For each substrate, we link to Phospho.ELM or PhosphoSite where the user can find manually curated information on *in vivo* phosphorylation sites including, when known, the kinase(s) involved (Figure 2C). To allow the user to investigate how a specific prediction was made by NetworKIN, we provide a link to the STRING network viewer, in which the most probable path connecting the kinase and the substrate will be highlighted (Figure 2D). Alternatively, the user can select multiple predictions and display the network context for all the proteins involved. From the network viewer, the evidence underlying each individual association can be inspected in further detail. This ability to thoroughly investigate individual predictions is particularly useful for interpreting non-obvious cases, which are often based on indirect links between the kinase and the substrate.

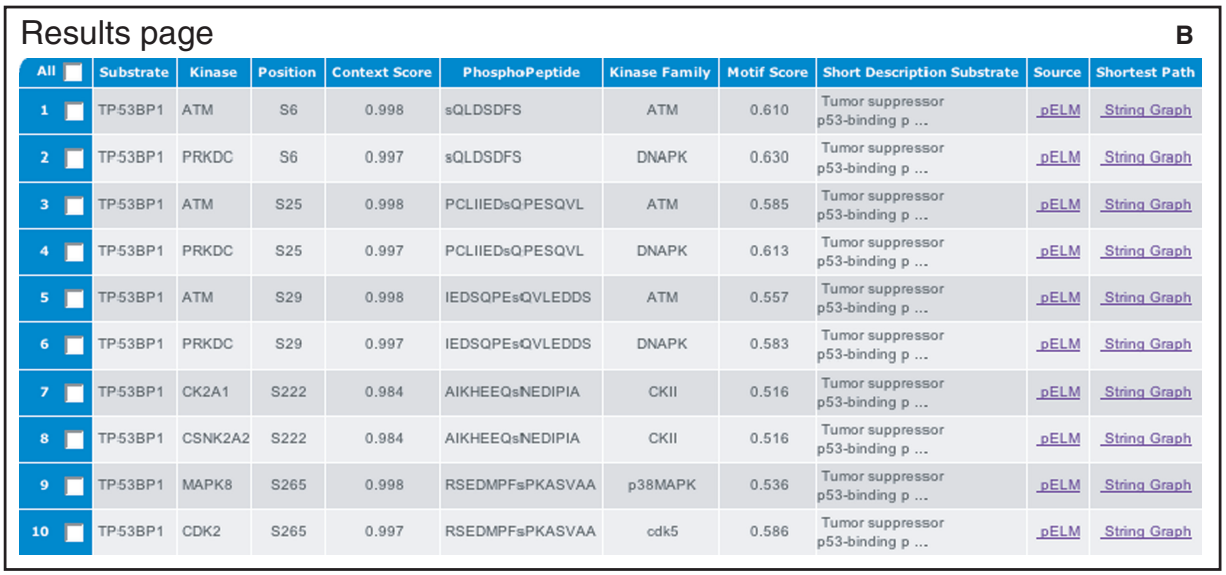
Although Phospho.ELM, PhosphoSite and hence NetworKIN are kept up-to-date with new published phosphorylation sites, many researchers will be interested in predictions for their own, unpublished sites. We thus allow users to submit protein sequences and a corresponding set of phosphorylation sites for analysis; although possible, we discourage submitting sequences without prior knowledge on phosphorylation. After uploading the data, the user will be presented with a confirmation page where potential data entry errors can be detected and fixed. The final predictions will be presented in a tabular format similar to the one used when querying the precomputed results in the database.

Many users are interested in specific kinases or substrates; however, others may want to get an overview of the complete phosphorylation network. To facilitate this, the resource offers a global map of all predictions currently in the database. All kinases and substrates are shown using a colour scale to signify their connectivity, namely the number of substrates for a given kinase or the number of kinases for a given substrate. By selecting one or more kinases, all corresponding substrates are highlighted and vice versa. Deselecting one kinase will deselect only the substrates specific for that kinase, keeping the other ones. We find this approach to be an intuitive way to gain insight into pleiotropic properties of kinases. Similar to the search interface, map selections can be visualized in their network context.

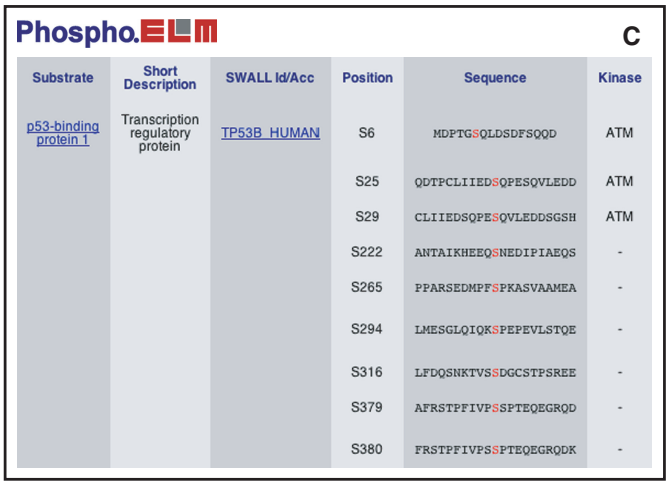
OUTLOOK

In the future, we intend to keep NetworKIN up-to-date with the latest data on phosphorylation sites from Phospho.ELM and PhosphoSite, functional associations from STRING and consensus sequence motifs from Scansite and other sources. Furthermore, the algorithm will be extended to take into account docking motifs (e.g. for MAP kinases) and phosphorylation-dependent binding modules (e.g. SH2, PTB and BRCT domains), which is expected to both improve the prediction accuracy and facilitate more comprehensive modelling of signalling networks. We also intend to extend the method to include





All	Substrate	Kinase	Position	Context Score	PhosphoPeptide	Kinase Family	Motif Score	Short Description Substrate	Source	Shortest Path
1	TP53BP1	ATM	S6	0.998	sQLDSDFS	ATM	0.610	Tumor suppressor p53-binding p ...	pELM	String Graph
2	TP53BP1	PRKDC	S6	0.997	sQLDSDFS	DNAPK	0.630	Tumor suppressor p53-binding p ...	pELM	String Graph
3	TP53BP1	ATM	S25	0.998	PCLIIEDsQPESQVL	ATM	0.585	Tumor suppressor p53-binding p ...	pELM	String Graph
4	TP53BP1	PRKDC	S25	0.997	PCLIIEDsQPESQVL	DNAPK	0.613	Tumor suppressor p53-binding p ...	pELM	String Graph
5	TP53BP1	ATM	S29	0.998	IEDSQPEsQVLEDDS	ATM	0.557	Tumor suppressor p53-binding p ...	pELM	String Graph
6	TP53BP1	PRKDC	S29	0.997	IEDSQPEsQVLEDDS	DNAPK	0.583	Tumor suppressor p53-binding p ...	pELM	String Graph
7	TP53BP1	CK2A1	S222	0.984	AIKHEEQsNEDIPIA	CKII	0.516	Tumor suppressor p53-binding p ...	pELM	String Graph
8	TP53BP1	CSNK2A2	S222	0.984	AIKHEEQsNEDIPIA	CKII	0.516	Tumor suppressor p53-binding p ...	pELM	String Graph
9	TP53BP1	MAPK8	S265	0.998	RSEDMPFsPKASVAA	p38MAPK	0.536	Tumor suppressor p53-binding p ...	pELM	String Graph
10	TP53BP1	CDK2	S265	0.997	RSEDMPFsPKASVAA	cdk5	0.586	Tumor suppressor p53-binding p ...	pELM	String Graph



Substrate	Short Description	SWALL Id/Acc	Position	Sequence	Kinase
p53-binding protein 1	Transcription regulatory protein	TP53B_HUMAN	S6	MDPTGSQLDSDFSQQD	ATM
			S25	QDTPCLIIEDSQPESQVLEDD	ATM
			S29	CLIIEDSQPESQVLEDDSGSH	ATM
			S222	ANTAIKHEEQSNEDIPIAEQS	-
			S265	PPARSEDMPFSPKASVAAMEA	-
			S294	LMESGLQIQKSPEPEVLSTQE	-
			S316	LFDQSNKTVSSDGCSTPSREE	-
			S379	AFRSTPFIVPSSPTEQEGRQD	-
			S380	FRSTPFIVPSSPTEQEGRQDK	-

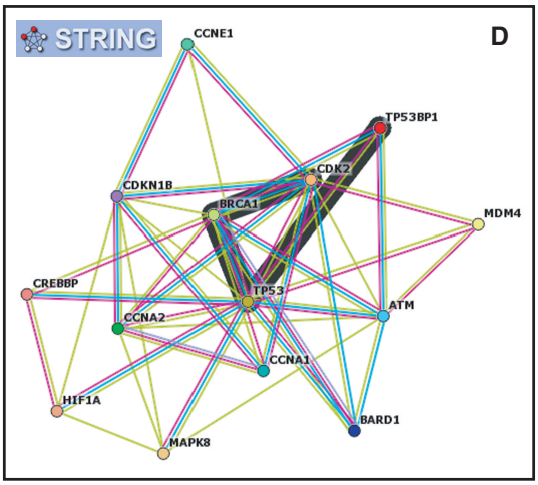


Figure 2. Using NetworKIN. A researcher is interested in the 53BP1 tumour surpressor. From the homepage of NetworKIN, this protein is chosen as the substrate protein to query the database for predictions relating any kinase to specific phosphorylation sites within 53BP1 (**A**). The system returns a total of 78 relations involving 12 different kinases that are predicted to phosphorylated 39 different sites, which are presented in a tabular view (**B**), (only the first 10 predictions are shown). These predictions can now be investigated in further detail by following either the links to the Phospho.ELM and PhosphoSite databases for curated knowledge related to the sites (**C**), or by following the links to the STRING network viewer to visualize the most probable path in the protein assocation network, which connects the kinase and the substrate (**D**).

phosphatases as soon as data for this is available to us. Although, NetworKIN is so far specifically aimed at protein phosphorylation, many other post-translational modifications are mediated by enzymes that recognize short linear motifs. We thus expect the same principle of specificity through context to apply. For example, the modifications of histone tails through acetylation, methylation and phosphorylation has been shown to be context dependent (16), and acetylated or methylated sites in turn bind interaction domains, such as bromo- or chromodomains. Extending the resource to cover also other post-translational modifications is thus a long-term goal.

ACKNOWLEDGEMENTS

We thank Christian von Mering for developing the best-path viewer at our request and Sara Quirk, Claus Jørgensen and Ginny I. Chen for feedback on the online resource and comments on this manuscript. Thanks to Chris Tan Soon Hen for technical assistance. We are deeply grateful to Phospho.ELM and PhosphoSite for their continued hard and absolutely essential work on high-quality annotation of the phosphoproteomes.
This project was funded by Genome Canada through Ontario Genomics Institute, by the National Institute of Health (U54-CA112967 and GM60594) as well as through the ADIT Integrated Project, contract number LSHB-CT-2005511065, and through the BioSapiens Network of Excellence, contract number LSHG-CT-2003-503265, both funded by the European Commission FP6 Programme. R.L. is a Human Frontiers Science Research Fellow. Funding to pay the Open Access publication charges for this article was provided by Genome Canada.

Conflict of interest statement. None declared.

REFERENCES

1. Manning,B.D. and Cantley,L.C. (2007) Akt/pkb signaling: navigating downstream. *Cell*, **129**, 1261–1274.
2. Diella,F., Gould,C.M., Chica,C., Via,A. and Gibson,T.J. (2008) Phospho.ELM: a database of phosphorylation sites—update 2008. *Nucleic Acids Res.*, **36**, In Press.
3. Schmelzle,K. and White,F.M. (2006) Phosphoproteomic approaches to elucidate cellular signaling networks. *Curr. Opin. Biotechnol.*, **17**, 406–414.
4. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res.*, **31**, 3635–3641.
5. Hjerrild,M., Stensballe,A., Rasmussen,T.E., Kofoed,C.B., Blom,N., Sicheritz-Pontén,T., Larsen,M.R., Brunak,S., Jensen,O.N. *et al.* (2004) Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J. Proteome Res.*, **3**, 426–433.
6. Blom,N., Sicheritz-Pontén,T., Gupta,R., Gammeltoft,S. and Brunak,S. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.
7. Brinkworth,R.I., Breinl,R.A. and Kobe,B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl Acad. Sci. USA*, 100, 74–79.
8. Kim,J.H., Lee,J., Oh,B., Kimm,K. and Koh,I. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.
9. Xue,Y., Zhou,F., Zhu,M., Ahmed,K., Chen,G. and Yao,X. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res.*, **33**, W184–W187.
10. Xue,Y., Li,A., Wang,L., Feng,H. and Yao,X. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, **7**, 163.
11. Wong,Y.H., Lee,T.Y., Liang,H.K., Huang,C.M., Wang,T.Y., Yang,Y.H., Chu,C.H., Huang,H.D., Ko,M.T. *et al.* (2007) KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res.*, **35**, W588–W594.
12. von Mering,C., Jensen,L.J., Kuhn,M., Chaffron,S., Doerks,T., Krüger,B., Snel,B. and Bork,P. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res.*, **35**, D358–D362.
13. Linding,R., Jensen,L.J., Ostheimer,G.J., van Vugt,M.A., Jørgensen,C., Miron,I.M., Diella,F., Colwill,K., Taylor,L. *et al.* (2007) Systematic discovery of in vivo phosphorylation networks. *Cell*, **129**, 1415–1426.
14. Hornbeck,P.V., Chabra,I., Kornhauser,J.M., Skrzypek,E. and Zhang,B. (2004) PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.