# Chapter 15

# Computational Identification of Protein Kinases and Kinase-Specific Substrates in Plants

**Han Cheng, Yongbo Wang, Zexian Liu, and Yu Xue**

## Abstract

The protein phosphorylation catalyzed by protein kinases (PKs) plays an essential role in almost all biological progresses in plants. Thus, the identification of PKs and kinase-specific substrates is fundamental for understanding the regulatory mechanisms of protein phosphorylation especially in controlling plant growth and development. In this chapter, we describe the computational methods and protocols for the identification of PKs and kinase-specific substrates in plants, by using *Vitis vinifera* as an example. First, the proteome sequences and experimentally identified phosphorylation sites (p-sites) in *Vitis vinifera* were downloaded. The potential PKs were computationally identified based on preconstructed Hidden Markov Model (HMM) profiles and ortholog searches, whereas the kinase-specific p-sites, or site-specific kinase–substrate relations (ssKSRs) were initially predicted by the software package of Group-based Prediction System (GPS) and further processed by the iGPS algorithm (in vivo GPS) to filter potentially false positive hits. All primary data sets and prediction results of *Vitis vinifera* are available at: http://ekpd.biocuckoo.org/protocol.php.

**Key words** Protein kinase, Phosphorylation, Kinase-specific substrate, Hidden Markov Model, GPS, Site-specific kinase–substrate relation

## 1 Introduction

As one of the most important and well-studied posttranslational modifications (PTMs), protein phosphorylation takes part in regulating a broad spectrum of biological processes such as signal transduction and environmental response in plants [1, 2]. Protein kinases (PKs), as key regulators responsible for the biochemical reactions, modify their target proteins by chemically adding phosphate groups to specific amino acids, mainly including serine (S), threonine (T), and tyrosine (Y) residues [3–6]. In this regard, the identification of PKs and PK-specific substrates is fundamental for understanding the regulatory mechanisms of protein phosphorylation in controlling plant growth and development.

Although the phosphorylation was discovered nearly sixty years ago [7], the identification and classification of PKs especially

Waltraud X. Schulze (ed.), *Plant Phosphoproteomics: Methods and Protocols*, Methods in Molecular Biology, vol. 1306, DOI 10.1007/978-1-4939-2648-0_15, 

plant PKs is still immature and full of challenge. In 1995, based on the conserved sequence and structural profiles of the catalytic domains, Steven K. Hanks and Tony Hunter performed a seminal study by classifying eukaryotic protein kinases (ePKs) into a hierarchical structure with four levels, including group, family, subfamily and single PK [8]. Using the same rationale, Manning et al. systematically identified 130, 454, 240, and 518 putative PKs in *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Homo sapiens*, and classified these PKs into 10 groups, 134 families, and 201 subfamilies [9]. However, the classification and annotation of PKs at the subfamily level is time-consuming and largely dependent on the manual curation. In this regard, the kinase.com database only contained the annotation information for PKs in 11 species after 10 years of efforts [9]. Recently, we developed a comprehensive database of EKPD (http://ekpd.biocuckoo.org) for PKs and protein phosphatases (PPs) in eukaryotes [10]. First, we collected 1,855 known PKs and 347 known PPs from the scientific literature and several public databases, and then classified these PKs and PPs into 10 groups with 149 families, and 9 groups with 29 families, respectively. At the family level, we totally constructed 139 and 27 Hidden Markov Model (HMM) profiles of PKs and PPs for searching more potential PKs and PPs, separately. Also, orthology searches were performed for PK and PP families without HMM profiles. Totally, EKPD contains 50,433 PKs and 11,296 PPs for 84 eukaryotic species, including 22 plants [10].

Traditionally, the identification of phosphorylation sites (p-sites) with the approach of site-directed mutagenesis is labor-intensive, time-consuming, and costly. Recently progresses in high-throughput mass spectrometry (HTP-MS) technology and phosphopeptide enrichment techniques such as immobilized metal ion affinity chromatography (IMAC) have enable the large-scale identification of thousands of p-sites in a single experiment [11]. However, the regulatory PKs of these p-sites are still difficult to be identified. In 2004, we developed a novel algorithm of group-based phosphorylation scoring (GPS) for predicting PK-specific substrates [12]. The GPS 1.0 could predict site-specific kinase–substrate relations (ssKSRs) for 52 PK families [12], where GPS 1.1 could predict ssKSRs for 216 PKs of 71 PK groups [13]. Later, we greatly improved the algorithm by developing the GPS 2.0 and 2.1 software packages, which can predict PK-specific substrates for 408 human PKs in hierarchy. Also, we renamed the GPS algorithm as group-based prediction system [14, 15]. More details on the GPS algorithm can be referred to [16]. Because only sequence profiles around p-sites were considered in GPS algorithms and various contextual factors such as kinase–substrate interaction, co-localization and co-expression information can contribute additional specificity for the phosphorylation in vivo [17], the sites predicted by GPS may be only phosphorylated in vitro but not in vivo.

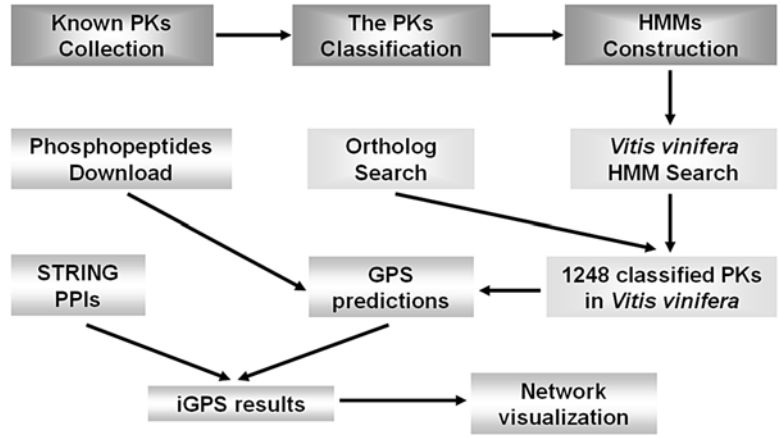


**Fig. 1** The schematic diagram of the computational pipeline for the identification of PKs and kinase-specific substrates in *Vitis vinifera*

Thus, we further developed a new algorithm of iGPS (in vivo GPS), by combining both sequence-based predictions and protein–protein interactions (PPIs) between PKs and substrates [18]. Although GPS and iGPS algorithms were developed mainly for predicting PK-specific substrates in mammalians, they can also be used in plants.

In this chapter, we took *Vitis vinifera* as an example to describe the computational methodologies for identifying PKs and kinase-specific substrates in plants (Fig. 1). First, we summarized how various HMM profiles of catalytic domains were constructed for known and curated PKs at the family level in EKPD database [10]. By downloading the proteome set of *Vitis vinifera*, all potential PKs were identified and classified based on HMM profiles and ortholog search. For the identification of kinase-specific substrates, the GPS software package [14, 15] was used for predicting ssKSRs of experimentally identified p-sites in *Vitis vinifera*. Furthermore, the iGPS algorithm was adopted to reduce false positive hits in predicted ssKSRs by including the PPI information between PKs and substrates [18].

## 2 Materials

### 2.1 Data Resources

1. The kinase.com database (http://kinase.com/kinbase/FastaFiles/), the best annotated database for protein kinases in eukaryotes [9]. All curated kinases were classified into a hierarchical structure with four levels, including group, family, subfamily, and single kinase.

2. The proteome set of *Vitis vinifera* was downloaded from the FTP Server of Ensembl Plants (release version 21, http://plants.ensembl.org/) [19].
3. The information of gene start and end of all proteins in *Vitis vinifera* was obtained from the BioMart service of Ensembl Plants (http://plants.ensembl.org/biomart/martview) [19].
4. The known phosphopeptides and phosphoprotein sequences of *Vitis vinifera were taken* from $P^3$DB (release version 3.0, http://www.p3db.org/), a comprehensive database of phosphoproteomes for 9 plant species from 32 experimental studies [20].
5. The PPI information and corresponding protein sequences of *Vitis vinifera were downloaded from STRING* (release version 9.1, http://string-db.org/), a widely used database containing precalculated PPIs [21].

### 2.2 Tools

1. MUSCLE (version 3.8.31, http://www.drive5.com/muscle/), an extensively used tool for multiple sequence alignment [22].
2. The HMMER software package (version 3.0, http://hmmer.janelia.org/) [23]. Two programs including hmmbuild and hmmsearch were used in this study. The former can construct HMM profiles from the result of multiple sequence alignments, whereas the latter can search an HMM profile against the target sequence database for finding matches [23].
3. CD-HIT (http://weizhong-lab.ucsd.edu/cd-hit/), a useful tool for clustering similar sequences [24].
4. The blastall program in the stand-alone package of NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [25].
5. The GPS software package (version 2.1.2, http://gps.biocuckoo.org/), mainly for the prediction of kinase-specific p-sites or ssKSRs [14].
6. The iGPS algorithm, which can be used for reducing false positive hits for potentially ssKSRs predicted by GPS [18].
7. Cytoscape (version 2.8.3, http://www.cytoscape.org/), an integrative platform designed for the analysis and visualization of complex networks [26].

## 3 Methods

### 3.1 The Construction of HMM Profiles for the PKs at the Family Level

1. The protein sequences and corresponding kinase catalytic domain sequences of 1,855 curated and pre-classified PKs of *S. cerevisiae*, *C. elegans*, *D. melanogaster*, *Mus musculus*, and *H. sapi*ens were directly downloaded from the kinase.com database [9].

2. Based on the previously established rationales for PK classification [9, 8], we manually classified all collected PKs into 10 groups with 149 families.
3. For 139 families with at least three PKs, we used MUSCLE [22] to align the kinase domain sequences for each family (*see* **Note 1**).
4. Then with the multiple sequence alignment results, we used the hmmbuild program of HMMER [23] to construct 139 HMM profiles for the PK families (*see* **Note 2**).
5. For further characterization of more PKs with constructed HMM profiles, the program of hmmsearch [23] was used. To balance the specificity and sensitivity of the PK prediction, we manually selected a cutoff value each family on the basis of the log-odds likelihood score calculated by hmmsearch (*see* **Note 3**).

### 3.2 Computational Identification and Classification of PKs in Vitis vinifera

1. We downloaded the protein sequences of *Vitis vinifera* from Ensembl Plants [19], and removed low-quality sequences (*see* **Note 4**).
2. For the purpose of eliminating the redundancy, we clustered proteins with a threshold of 100 % identity by CD-HIT [24]. If the identity of multiple proteins in a cluster was 100 %, CD-HIT only retained one sequence of them, while other protein sequences were discarded and not used for any further analysis.
3. Then we applied the hmmsearch program [23] to search the nonredundant protein sequences of *Vitis vinifera* against all PK HMM profiles. If at least one log-odds likelihood score was ≥ the cutoff value of a PK HMM profile, the protein was identified as a PK.
4. For the classification, the calculated log-odds likelihood scores of multiple HMM profiles were compared, and a predicted PK was classified into the family with the highest score.
5. To avoid any redundancy of predicted and classified PKs, we used the Ensembl Gene ID as the unique accession and the transcript with the most significant E-value was represented for its corresponding gene.
6. Because a single gene may generate multiple variant proteins with different Ensembl Gene IDs, we further obtained the chromosomal localization information of genes in *Vitis vinifera* from the Ensembl BioMart service [19], by selecting the "Protein stable ID", "Gene start (bp)", and "Gene end (bp)" of Gene Attributes. If the gene coordinates were identical or overlapped for multiple proteins, we only retained the longest one (*see* **Note 5**).

**Table 1**
**The distribution of 49 families in 9 groups for 1,248 classified PKs from *Vitis vinifera***

| PK group | PK family |
|---|---|
| *AGC* | Akt, MAST, NDR, PDK1, PKA, AGC_Unique |
| *CAMK* | CAMK1, CAMKL, CAMK_Unique |
| *CMGC* | CDK, CLK, DYRK, GSK, MAPK, RCK, CMGC_Unique |
| *CK1* | CK1, CK1_Unique |
| *STE* | STE11, STE20, and STE_Unique |
| *TK* | TK_Unique |
| *TKL* | IRAK, MLK, TKL_Unique |
| *Atypical* | ABC1, PDHK, PIKK, RIO, BRD, G11, TAF1, Hisk |
| *Other* | Aur, BUB, Bud32, CDC7, Haspin, IRE, NAK, NEK, SCY, TLK, TTK, ULK, VPS15, WEE, WNK, Other_Unique |

7. Furthermore, we conducted ortholog searches to identify additional PKs for the families without HMM profiles, by the blastall program [25] (*see* **Note 6**).
8. The results of the HMM identifications and ortholog searches were merged together. Totally, we characterized 1,248 PKs with 9 groups and 49 families in *Vitis vinifera* (Table 1).

### *3.3 Prediction of Kinase-Specific Substrates in Vitis vinifera by GPS*

1. We downloaded 927 experimentally identified phosphopeptides of *Vitis vinifera* from $P^3DB$ [20]. Then, we mapped these phosphopeptides to the nonredundant proteome set of *Vitis vinifera* by BLAST [25] and obtained 795 unique p-sites in 539 phosphoprotein sequences.
2. By defining a *phosphorylation site peptide* PSP($m$, $n$) as a phosphorylation residue of S, T, or Y surrounded by $m$ upstream residues and $n$ downstream residues [14, 15], we extracted all PSP(15, 15) items of known p-sites (*see* **Note 7**). And the PSP (15, 15) items were prepared in the FATSA format.
3. Because the GPS tool was mainly developed for the prediction of kinase-specific p-sites in mammalians, the classification information of plant PKs was still not included. Thus, we manually selected predictors in GPS 2.1 for PKs at group and family levels if possible (*see* **Note 8**). Totally, we selected 25 GPS predictors for 1,086 PKs in *Vitis vinifera*.
4. The latest version of the GPS software package was downloaded and directly installed by double-clicking on the icon of the GPS program (*see* **Note 9**).

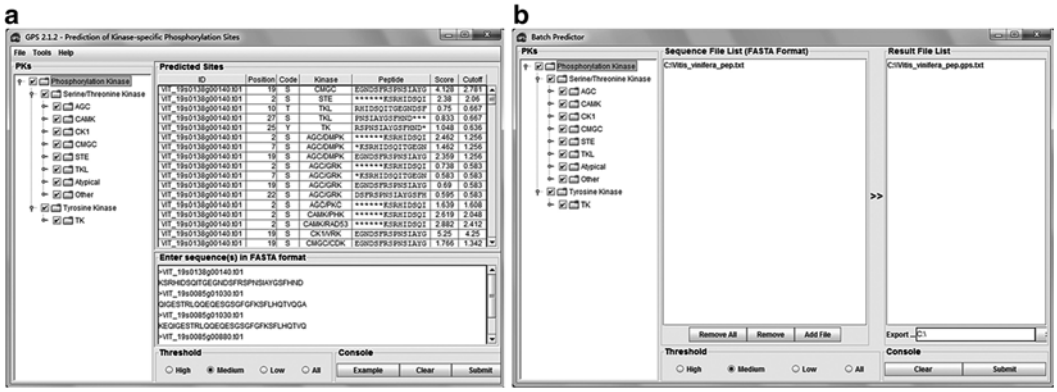


**Fig. 2** The user interface of the GPS software package. (**a**) The PSP (15, 15) items of known p-sites in the FATSA format can be directly inputted, and the predictions will be performed by left-clicking on the "submit" button. (**b**) The option of "Batch Predictor" allows users to import a preprepared file in the FATSA format for the prediction

5. For the prediction of kinase-specific p-sites, we directly inputted all PSP (15, 15) items in the FATSA format into the text form of GPS, and selected all predictors in GPS with the default threshold (*see* **Note 10**). By left-clicking on the "submit" button, the prediction results can be generated soon (Fig. 2a).
6. For a larger data set, the kinase-specific p-sites can be predicted by the "Batch Predictor" in GPS. The PSP (15, 15) items in the FATSA format can be stored in a text file, and then inputted into the Batch Predictor (Fig. 2b).
7. After the prediction, we only retained the results in which the position of p-sites was 16. Based on the GPS predictor-PK relations of *Vitis vinifera*, we assigned the exact PKs to all predictable p-sites (*see* **Note 11**).
8. Totally, we predicted 171,241 ssKSRs between the 1,072 PKs and 483 substrates for the 674 p-sites, with an average of 254.1 upstream PKs per p-site.

### *3.4 Prediction of In Vivo ssKSRs in Vitis vinifera by iGPS Algorithm*

1. We obtained the PPI information of *Vitis vinifera from the* STRING database [21]. Then using BLAST [25], we mapped the protein sequences of interacting proteins in PPIs to the nonredundant proteome set of *Vitis vinifera*. Totally, we got 2,636,727 nonredundant PPI pairs in 18,070 proteins for *Vitis vinifera*.
2. The iGPS algorithm combined both the sequence-based predictions and contextual factors to reduce false positive predictions [18] (*see* **Note 12**). Here, we reserved the GPS predictions only if the relations of PKs and their substrates were supported by PPIs from STRING.

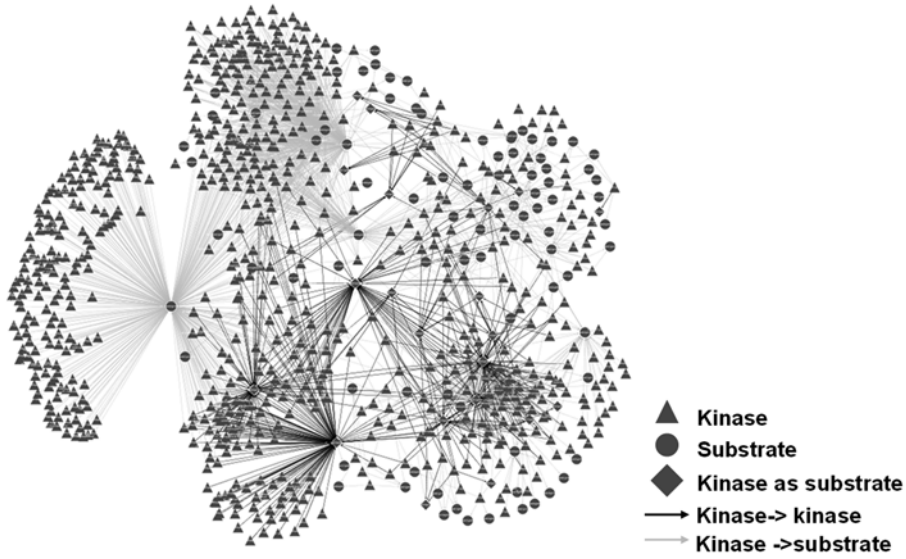


**Fig. 3** The directional KSPN of *Vitis vinifera* can be visualized by Cytoscape [26]. Multiple ssKSRs between a PK and a substrate was regarded as a single KSR [18]

3. By iGPS algorithm, we finally predicted 2,574 ssKSRs between 737 PKs and 110 substrates for the 129 p-sites, with an average of 20.0 regulatory PKs per p-site.
4. For the construction and visualization of kinase-substrate phosphorylation network (KSPN), we only counted multiple ssKSRs of a PK and a substrate as a single kinase–substrate relation (KSR). 6). In the KSPN, the nodes indicated PKs or substrates, while the edges represented KSRs. As previously described [18], the KSPN is directional, and we defined two types of orientations including PK→Substrate (a PK phosphorylates a substrate which is not a PK), PK→PK (a PK phosphorylates a PK). The final KSPN of *Vitis vinifera was visualized by* Cytoscape [26] and *contained 2,204 KSRs for 737 PKs and 110 substrates* (Fig. 3).

## 4 Notes

1. If not specified, the default parameters were selected for all bioinformatics tools used in this chapter.
2. The HMM profiles for 139 PK families can be available at: http://ekpd.biocuckoo.org/faq.php.
3. For a given protein sequence, the hmmsearch program [23] will compare it with each HMM profile by calculating a log-odds likelihood score and an E-value. Because the E-value

depends on the size of inputted data set and will be not equal when different data sets are used, we choose a realistic constant value of the log-odds likelihood score as the threshold [10].

4. Because the annotation quality of *Vitis vinifera* proteome is poor, we removed the protein sequences that had at least one "X" residue which indicates an unspecified amino acid.
5. By using the PK HMM profile of each family, we totally characterized 1,243 PKs in *Vitis vinifera*.
6. We adopted the computational approach of reciprocal best hit (RBH) [27], and used each member in PK families without HMM profiles to search in the proteome of *Vitis vinifera*. Then the sequence with the highest score was chosen to search in the corresponding proteome of the curated PK. If the selected PK was also the best hit, the predicted sequence was regarded as a PK and classified into the corresponding family. By this method, five additional PKs were identified.
7. For p-sites that locate in N-terminal or C-terminal of protein sequences, we complemented the phosphopeptides to PSP (15, 15) with "*" characters if necessary.
8. The basic hypothesis for assigning GPS predictors to PK groups or families is that similar PKs classified in a same group or family would recognize similar SLMs of substrate modification.
9. The GPS 2.1.2 release was implemented in JAVA, and several installation packages were constructed to support three major Operating Systems including Windows, Mac and Linux/Unix. In this chapter, the file "GPS_2.1.2_windows_20120913.exe" was downloaded.
10. In GPS, the threshold values were selected based on the false positive rates (FPRs), which were estimated from a randomly generated dataset containing 200,000 PSP (7, 7) peptides [14, 15]. For serine/threonine PKs, the high, medium, and low thresholds were chosen with FPRs of 2, 6, and 10 %. For tyrosine PK, the high, medium, and low thresholds were selected with FPRs of 4, 9, and 15 %. The medium thresholds were adopted as the default parameters.
11. Because the real p-sites are only a small proportion of total S/T or Y residues in protein sequences, the authors don't recommend the ab initio prediction of kinase-specific p-sites directly from the primary sequences. Instead, the inclusion of experimentally identified p-sites and the prediction of potential PKs for these real p-sites will greatly reduce the false positive predictions.
12. It was widely adopted that SLMs around the p-sites provide primary specificity for PK recognition [14, 15]; however, a number of additional contextual factors, such co-localization,

coexpression, co-complex, and physical interaction of the PKs with their substrates, contribute additional modification specificity in vivo [17, 18]. The PPI information was considered as a major contextual filter in iGPS algorithm.

## Acknowledgement

This work was supported by grants from the National Basic Research Program (973 project) (2013CB933900 and 2012CB910101), Natural Science Foundation of China (31171263, and 81272578), and International Science & Technology Cooperation Program of China (2014DFB30020).

### References

1. Olsen JV, Blagoev B, Gnad F, Macek B, Kumar C, Mortensen P, Mann M (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. Cell 127(3):635–648. doi:10.1016/j.cell.2006.09.026
2. Meng X, Xu J, He Y, Yang KY, Mordorski B, Liu Y, Zhang S (2013) Phosphorylation of an ERF transcription factor by Arabidopsis MPK3/MPK6 regulates plant defense gene induction and fungal resistance. Plant Cell 25(3):1126–1142. doi:10.1105/tpc.112.109074
3. De Verdier CH (1952) Isolation of phosphothreonine from bovine casein. Nature 170(4332):804–805
4. Levene PA, Alsberg CL (1906) The cleavage products of vitellin. J Biol Chem 2:127–133
5. Lipmann FA, Levene PA (1932) Prokaryotic elongation factor Tu is phosphorylated *in vivo*. J Biol Chem 98:109–114
6. Sutherland EW Jr, Wosilait WD (1955) Inactivation and activation of liver phosphorylase. Nature 175(4447):169–170
7. Fischer EH, Krebs EG (1955) Conversion of phosphorylase b to phosphorylase a in muscle extracts. J Biol Chem 216(1):121–132
8. Hanks SK, Hunter T (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB Journal 9(8):576–596
9. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) The protein kinase complement of the human genome. Science 298(5600):1912–1934. doi:10.1126/science.1075762
10. Wang Y, Liu Z, Cheng H, Gao T, Pan Z, Yang Q, Guo A, Xue Y (2014) EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases. Nucleic Acids Res 42(1):D496–D502. doi:10.1093/nar/gkt1121
11. Nuhse TS, Stensballe A, Jensen ON, Peck SC (2003) Large-scale analysis of in vivo phosphorylated membrane proteins by immobilized metal ion affinity chromatography and mass spectrometry. Mol Cell Proteomics 2(11):1234–1243. doi:10.1074/mcp.T300006-MCP200
12. Zhou FF, Xue Y, Chen GL, Yao X (2004) GPS: a novel group-based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 325(4):1443–1448. doi:10.1016/j.bbrc.2004.11.001
13. Xue Y, Zhou F, Zhu M, Ahmed K, Chen G, Yao X (2005) GPS: a comprehensive www server for phosphorylation sites prediction. Nucleic Acids Res 33(Web Server issue):W184–W187. doi:10.1093/nar/gki393
14. Xue Y, Ren J, Gao X, Jin C, Wen L, Yao X (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy. Mol Cell Proteomics 7(9):1598–1608. doi:10.1074/mcp.M700574-MCP200
15. Xue Y, Liu Z, Cao J, Ma Q, Gao X, Wang Q, Jin C, Zhou Y, Wen L, Ren J (2011) GPS 2.1: enhanced prediction of kinase-specific phosphorylation sites with an algorithm of motif length selection. Protein Eng Des Se 24(3):255–260. doi:10.1093/protein/gzq094
16. Yu Xue ZL, Jun Cao, Jian Ren (2011) Computational prediction of post-translational modification sites in proteins. Systems and computational biology—molecular and cellular experimental systems, Ning-Sun Yang (Ed), ISBN: 978-953-307-280-7, InTech, DOI:105772/18559
17. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, Bork P, Yaffe MB, Pawson T

(2008) NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res 36(Database issue):D695–D699. doi:10.1093/nar/gkm902

18. Song C, Ye M, Liu Z, Cheng H, Jiang X, Han G, Songyang Z, Tan Y, Wang H, Ren J, Xue Y, Zou H (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data. Mol Cell Proteomics 11(10):1070–1083. doi:10.1074/mcp.M111.012625
19. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM (2013) Ensembl 2013. Nucleic Acids Res 41(Database issue):D48–D55. doi:10.1093/nar/gks1236
20. Yao Q, Ge H, Wu S, Zhang N, Chen W, Xu C, Gao J, Thelen JJ, Xu D (2014) P3DB 3.0: from plant phosphorylation sites to protein networks. Nucleic Acids Res 42(1):D1206–D1213. doi:10.1093/nar/gkt1135
21. Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41(Database issue):D808–D815. doi:10.1093/nar/gks1094
22. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797. doi:10.1093/nar/gkh340
23. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform 23(1):205–211
24. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. doi:10.1093/bioinformatics/btl158
25. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36(Web Server issue):W5–W9. doi:10.1093/nar/gkn201
26. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. doi:10.1101/gr.1239303
27. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278(5338):631–637