Supplementary Data:

EKPD: a hierarchical database of eukaryotic protein kinases and protein phosphatases

Yongbo Wang[1,†], Zexian Liu[1,†], Han Cheng [1], Zhicheng Pan[1], Qing Yang[1], Anyuan Guo[1], and Yu Xue[1,*]

[1]Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

***Running title*:** Eukaryotic PK and PP database

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

*To whom correspondence should be addressed.

Yu Xue, Tel: +86-27-87793903, Fax: +86-27-87793172, E-mail: xueyu@hust.edu.cn.

Supplementary methods

The hypergeometric test

The pre-calculated annotations of Pfam domains (20) for proteins in 84 eukaryotes were downloaded from the BioMart service (http://www.ensembl.org/biomart/martview, Ensembl Genes 70) (18). Totally, there were 1,341,189 proteins annotated with at least one Pfam domain. In our database, there were 48,953 protein kinases (PKs) and 10,605 protein phosphatases (PPs) annotated with at least one Pfam domain. The specific measurements were defined as follows:

N = the number of proteins in 84 eukaryotes annotated by at least one Pfam domain (1,341,189 in this study)

n = the number of proteins in 84 eukaryotes annotated by the Pfam domain *d*

M = the number of proteins in PKs or PPs of 84 eukaryotes annotated by at least one Pfam domain (48,953 for PKs and 10,605 for PPs)

m = the number of proteins in PKs or PPs of 84 eukaryotes annotated by the Pfam domain *d*

The enrichment ratio (E-ratio) of the Pfam domain *d* was calculated, while the hypergeometric distribution equation was used to calculate the *p*-value as follows (27):

$$E-ratio=\frac{\frac{m}{M}}{\frac{n}{N}},$$

$$p-value=\sum_{m'=m}^{n}\frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad \left(E-ratio \geqslant 1\right), \text{ or}$$

$$p-value=\sum_{m'=0}^{m}\frac{\binom{M}{m'}\binom{N-M}{n-m'}}{\binom{N}{n}} \quad \left(E-ratio<1\right)$$

In this work, both over- and under-represented Pfam domains in PKs or PPs were calculated. Because too many hits were generated, a stringent threshold (*p*-value < 10^{-8}) was adopted.

The Yates' chi-squared (χ^2) test

To compare the preferences of Pfam domains in PKs or PPs between animals and plants,

the Yates's corrected version of Pearson's chi-squared test was adopted (From Wikipedia) (27). Given a Pfam domain *d*, the entries in the 2×2 table were defined as below:

a = number of PKs or PPs with the Pfam domain *d* in animals;

b = number of PKs or PPs with the Pfam domain *d* in plants;

c = number of PKs or PPs without the Pfam domain *d* in animals;

d = number of PKs or PPs without the Pfam domain *d* in plants;

$N_l = a + c$, total PKs or PPs in animals (22,874 for PKs and 7,169 for PPs);

$N_w = b + d$, total PKs or PPs in plants (25,954 for PKs and 3,339 for PPs);

$N_y = a + b$; $N_n = c + d$; $N = N_l + N_w = N_y + N_n$.

	Animals	Plants	
Number of PKs or PPs with the Pfam domain *d*	*a*	*b*	N_y
Number of PKs or PPs without the Pfam domain *d*	*c*	*d*	N_n
	N_l	N_w	*N*

The enrichment ratio of number of PKs or PPs with the Pfam domain *d* in animals against plants was calculated as below:

$$E - ratio = \frac{\frac{a}{N_l}}{\frac{b}{N_w}}$$

E-ratio > 1 means the Pfam domain *d* to be over-represented in animals, whereas E-ratio < 1 means the Pfam domain *d* to be over-represented in plants. The χ^2 was calculated as below:

$$\chi^2_{Yates} = \frac{N(\max(0,|ad - bc| - N/2))^2}{N_l N_w N_y N_n}$$

Then the *p*-value (< 10^{-8}) was calculated by the function of CHIDIST(χ^2, degrees_freedom) in Excel. The degrees_freedom is equal to 1 for the 2×2 table.

Supplementary results

The classification of PKs and PPs

As previously described (1, 2), we classified PKs into 10 groups and 149 families: (i) The AGC group has 16 families, such as Akt, DMPK, GRK, MAST, NDR, PDK1, PKA, PKC, PKG, PKN, RSK, RSKL, RSKR, SGK, YANK, and Unique; (ii) The CAMK group contains 18 families, such as CAMK1, CAMK2, CAMKL, CASK, DAPK, DCAMKL, MAPKAPK, MLCK, PHK, PIM, PKD, PSK, RAD53, RSKb, Trbl, Trio, TSSK, and Unique; (iii) The CMGC group has 9 families such as CDK, CDKL, CLK, DYRK, GSK, MAPK, RCK, SRPK, and Unique; (iv) The CK1 group contains 11 families as CK1, Dual, TTBK, TTBKL, VRK, Worm10, Worm6, Worm7, Worm8, Worm9, and Unique; (v) The RGC group has only one family as RGC; (vi) The STE group has 4 families as STE11, STE20, STE7, and Unique; (vii) The TK group contains 31 families, such as Abl, Ack, Alk, Axl, CCK4, Csk, DDR, EGFR, Eph, FAK, Fer, FGFR, InsR, Jak, KIN16, KIN6, Lmr, Met, Musk, PDGFR, Ret, Ror, Ryk, Sev, Src, Syk, Tec, Tie, Trk, VEGFR, and Unique; (viii) The TKL group has 8 families as IRAK, LISK, LRRK, MLK, RAF, RIPK, STKR, and Unique; (ix) The Atypical group contains atypical PKs with 14 families such as ABC1, Alpha, PDHK, PIKK, RIO, BCR, BRD, FAST, G11, H11, TAF1, TIF1, Hisk and FAM20C; (x) The Other group with 37 families contains PKs that can not be classified into the above 9 groups. The Unique family in each group holds PKs that can not be classified into other families.

Based on previously established rationales (12-16), we first classified PSPs into 3 groups and 14 families: (i) The PPP group has 10 families such as PP1, PP2A, PP2B, PP4, PP5, PP6, PP7, Kelch, SLP and PPP-Unique; (ii) The PPM group has 2 families such as PP2C and PDP; (iii) The PSP-other group contains unclassified PSPs with 2 families. The PTPs were also classified into 7 groups with 19 families, while CDC25, LMWPTP and PTPLA have only one family in each group. The other 4 groups are: (i) The Classical PTP group has 2 families such as RPTP and NRPTP; (ii) The DSP group contains 7 families such as aDSP, MKP, PRL, CDC14, SSH, Myotubularins, and PTEN; (iii) The Asp-Based PTP group with 4 families contains atypical PPs, whose activity sites are aspartic acids, whereas other PPs use cysteine residues for catalysis; (iv) The PTP-other group contains unclassified PTPs with 3 families.

Performance evaluation of the HMM-based identification

To evaluate the prediction accuracy and robustness of the HMM identification, our

benchmark data set with 1,855 PKs and 347 PPs was used for testing. For one PK or PP family, the annotated proteins were regarded as positive data (*P*), while all other sequences were taken as negative data (*N*). Two measurements of sensitivity (*Sn*) and specificity (*Sp*) were defined and calculated as shown below:

$$Sn = \frac{TP}{TP + FN}, \text{ and } Sp = \frac{TN}{TN + FP},$$

First, the self-consistency validation was performed directly with the positive data and negative data to represent the prediction accuracy. To further evaluate the prediction robustness, the leave-one-out (LOO) validation was also carried out. The Receiver Operating Characteristic (ROC) curves were drawn, and AROC (area under ROC) values were calculated for 8 PKs (Figure 1A) and 8 PPs (Figure 1B) families, respectively. The results suggested that the HMM predictions are accurate and robust (Figure 1A & 1B). To promise that all curated PKs and PPs can be correctly identified and classified (*Sn* = 100%), we selected different cut-off values for all families (Table S1).

The search and advance options in EKPD

The search option provides an interface for querying the EKPD database with one or several keywords or accession numbers. For example, if the keyword of 'AKT1' is inputted and submitted, the results will be shown in a tabular format, with the features of EKPD ID, organism, and protein/gene names/aliases (Figure S2A). Moreover, we provided four additional advance options, such as (i) batch search, (ii) advance search, (iii) protein kinase & protein phosphatase classification, and (iv) BLAST search. (i) Batch search. Users can input multiple keywords (≤ 100) in a line-by-line format for querying multiple entries in EKPD (Figure S2B). (ii) Advance search. In this option, users could use relatively complex and combined keywords to locate the precise information. The interface of the search-engine permits querying by adding or removing searching conditions and linking queries through three operators of "and", "or" and "not" (Figure S2C). (iii) Protein kinase & protein phosphatase classification. Users can input one or multiple protein sequences (≤ 10) in FASTA format, by searching 139 and 27 constructed HMM profiles for PK and PP families respectively. If the protein is determined as a PK or PP, the classification and detection information will be provided (Figure S2D). (iv) BLAST search. This option was designed for finding the related information in EKPD database quickly. The blastall program of NCBI BLAST packages was included in EKPD database. Users can input protein sequences (≤ 10) in FASTA format for searching identical or

homologous proteins (Figure S2E).

Supplementary references

1. Manning, G., Whyte, D.B., Martinez, R., Hunter, T. and Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912-1934.

2. Hanks, S.K. and Hunter, T. (1995) Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.*, **9**, 576-596.

12. Alonso, A., Sasin, J., Bottini, N., Friedberg, I., Osterman, A., Godzik, A., Hunter, T., Dixon, J. and Mustelin, T. (2004) Protein tyrosine phosphatases in the human genome. *Cell*, **117**, 699-711.

13. Andersen, J.N., Del Vecchio, R.L., Kannan, N., Gergel, J., Neuwald, A.F. and Tonks, N.K. (2005) Computational analysis of protein tyrosine phosphatases: practical guide to bioinformatics and data resources. *Methods*, **35**, 90-114.

14. Andersen, J.N., Jansen, P.G., Echwald, S.M., Mortensen, O.H., Fukada, T., Del Vecchio, R., Tonks, N.K. and Moller, N.P. (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. *FASEB J.*, **18**, 8-30.

15. Peng, A. and Maller, J.L. (2010) Serine/threonine phosphatases in the DNA damage response and cancer. *Oncogene*, **29**, 5977-5988.

16. Shi, Y. (2009) Serine/threonine phosphatases: mechanism through structure. *Cell*, **139**, 468-484.

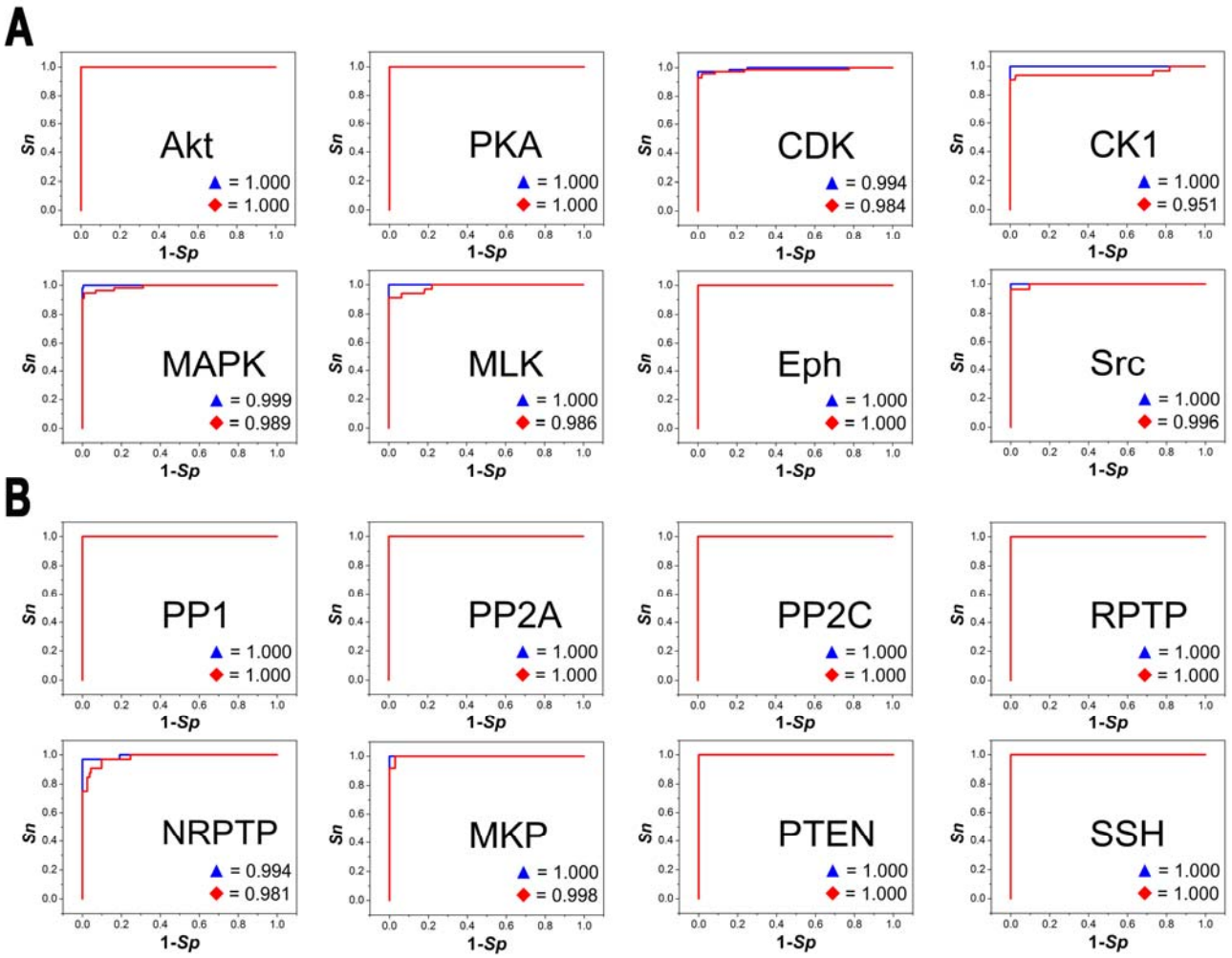
18. Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S. *et al*. (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48-55.

20. Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290-301.

27. Liu, Z., Cao, J., Ma, Q., Gao, X., Ren, J. and Xue, Y. (2011) GPS-YNO2: computational prediction of tyrosine nitration sites in proteins. *Mol. Biosyst.*, **7**, 1197-1204.

Supplementary Figures

Supplementary Figure S1 – The prediction performances of the HMM identifications. The ROC curves and AROC values were carried out for the self-consistency (Curves & triangles in blue) and LOO (Curves & diamonds in red) validations of (A) 8 PK and (B) 8 PP families, respectively.



Supplementary Figure S2 – The search and advance options. (A) The database can be searched with one or multiple keywords; (B) Batch search permits users to input multiple keywords in a line-by-line format for querying (≤ 100); (C) Advance search allows users to query with more than one searching conditions; (D) Protein kinases and protein phosphatases classification option scans protein sequences (≤ 10) in FASTA format against pre-constructed HMM profiles; (E) Blast search option search protein sequences (≤ 10) for detecting identical or homologous sequences.