SUPPLEMENTARY INFORMATION

BIOPHYSICAL PREDICTION OF PROTEIN-PEPTIDE INTERACTIONS AND SIGNALING NETWORKS USING MACHINE LEARNING

Joseph M. Cunningham, Grigoriy Koytiger, Peter K. Sorger, Mohammed AlQuraishi

Supplementary Data Tables

Supplementary Table 1 | Experimental sources of PBD-peptide interaction data.

PBD	Number of measured interactions	Sources of high-throughput data (PMID)	Number of domains in family (*Homo sapiens*)	Superfamily accession	Modeled binding mechanism	Affinity ranges[12]
Src Homology 2 (SH2)	396,359	23545499, 23358503, 24728074, 22974441	120	55550	Phospho-tyrosine binding	nM - μM
Phosphotyrosine binding (PTB) / Phosphotyrosine interaction domain (PID)	29,331	23358503, 16982700	33	50729	Phospho-tyrosine binding	nM - μM
Tyrosine kinase (TK)	167,231	23545499	94	56112	Phospho-tyrosine binding / transferase	nM - μM
(Classical) Protein Tyrosine phosphatase (PTP)	70,684	23545499	49	52799	Phospho-tyrosine binding / transferase	nM - μM
Src Homology 3 (SH3)	702,839	23545499, 19841731, 23549480	297	50044	Poly-proline binding	nM - μM
WW	154,741	23545499	92	51045	Poly-proline binding	μM
WASP Homology 1 (WH1) / enabled VASP Homology 1 (EVH1)	36,182	23545499	11	50729	Poly-proline binding	μM

PDZ	505,158	18711339, 15465056, 18828675, 22069443	270	50156	C-terminus binding	nM - µM

Supplementary Table 2 | External models.

Model Name	**Modeled PBD**	**Number of Models**	**Number of Modeled Domains (Fraction of total)**
SH2-PepInt	SH2	1	51 (0.43)
NetPhorest	SH2	11	88 (0.73)
NetPhorest	PTB	3	6 (0.18)
NetPhorest	TK	11	55 (0.59)
NetPhorest	PTP	6	23 (0.47)
PDZ-PepInt	PDZ	9	105 (0.39)

Supplementary Table 3 | Area under the receiver-operating characteristic (AUROC) and p-values (compared to HSM/D; DeLong test) for all PBD modeling methods.

PBD	**PSSM**	**HSM/ID**	**HSM/D**	**Other**
SH2 (N=120)	0.60 (<3.5E-17)	0.87 (3.5E-17)	0.89	SH2-PepInt:0.73 (<3.5E-17) NetPhorest: 0.77 (<3.5E-17)
PTB (N=33)	0.74 (2.4E-45)	0.89 (2.8E-3)	0.92	NetPhorest:0.80 (6.35E-12)
TK (N=94)	0.66 (<2.0E-6)	0.87 (2.0E-6)	0.88	NetPhorest:0.64 (<2.0E-6)
PTP (N=49)	0.75 (6.2E-169)	0.86 (1.4E-11)	0.90	NetPhorest : 0.75 (5.3E-122)
SH3 (N=297)	0.57 (<1.4E-37)	0.90 (1.4E-37)	0.92	-
WW (N=92)	0.60 (<4.7E-10)	0.90 (4.7E-10)	0.92	-
WH1 (N=11)	0.69 (4.9E-56)	0.86 (2.4E-2)	0.88	-
PDZ (N=270)	0.67 (<2.1E-152)	0.92 (2.1E-152)	0.97	PDZ-PepInt:0.80 (4.3E-290)

Supplementary Table 4 | Distribution of inferred peptide sites in the human proteome.

Peptide Type	Number of sites (proteins)	Mean (STD) per protein	Maximum number in a single protein
Phosphosites	17,002 (6,976)	2.4 (+/2.6)	46
C-terminus	2,457 (2,457)	1	1
Other	24,469 (10,378)	5.8 (+/- 7.0)	41

Supplementary Table 5 | High-throughput methods of PPI detection.

Source (PMID)	Method name (assigned)	Detection type	Reported interactions	Constituent proteins	False-discovery rate (FDR)
28514442, 26186194	HT-GYGI (BioPlex)[13,14]	Affinity purification/ mass spectrometry (AP/MS)	~56,000	~11,000	0.01
26496610	HT-MANN[15]	AP/MS	~28,000	~5,500	0.05
26496610	HT-MANN HC[15]	AP/MS	~14,000	~4,300	0.01
25416956	HT-VIDAL[16]	Yeast two-hybrid (Y2H)	~14,000	~4,200	0.01

Supplementary Notes

Supplementary Note 1 | Estimate of domain properties in the human proteome

We assume the actual number of PBDs sites in the human proteome to be the "modular binding" domains[12] falling into 22 canonically recognized families having folded, peptide-binding elements: PTB, SH2, PTP, 14-3-3, BRCT, FF, FHA, MH2/DWB, POLO-Box, EVH1/WH1, GYF, SH3, WW, LRR, PHD, Chromo, MBT, Tudor, PWWP, BROMO, PDZ, and WD40. This comprises a total of 1,809 PBDs having seven distinct binding chemistries: phosphotyrosine, phosphoserine/threonine, polyproline, methyl-lysine, acetyl-lysine, C-terminus, and miscellaneous (e.g. WD40, a β-propeller repeat that recognizes many of the previous chemistries). Two PBD families—LRR and WD40—represent repetitive sequence elements that fold and function as single units. Proteins containing these elements are therefore assumed to contain a single functional PBD. In this paper we also model two enzyme families, Tyrosine Kinases (TK) (n = 94; there are 538 protein kinases made up of 11 families[17]) and phosphatases (PTPs) (n = 49; there are 110 protein phosphatases made up of 13 families[18]). For this paper, we combine these estimates (n = 2,396) into a single estimate for the total number of protein-interacting domains. In principle, HSM is capable of representing all human PBDs (or PBDs from other organisms) but the current work is limited to 966 protein interacting domains (~39% of the total; 823 / 1,809 = ~46% of PBDs) in eight domain families having three binding chemistries simply because the amount of data available on the other families is too small for model training or evaluation (<100 binding interactions per family).

Supplementary Note 2 | Analysis of structures and models inform biophysical constraints

The two domain models, HSM/ID and HSM/D are defined by two biophysically-informed constraints: (i) a standardized PBD-peptide coordinate system and (ii) a shared set of energy potentials. These constraints are derived from analysis of both structural and inferred energetic data. The first constraint was derived from structural analysis of PBDs. Domain structures were first aligned using rigid rotations and translations (**Supplementary Fig. 1a**). We then analyzed the 'biophysical' profiles of every residue position within the bound peptide. We define a residue's biophysical profile to be the counts of amino acids of different classes (positively-charged, negatively-charged, hydrophobic, hydrophilic) within a given spatial distance from that residue (*i.e.* a ball of a given radius surrounding a peptidic residue). When comparing two residues, biophysical distance is defined as the Euclidean distance between their respective profiles. We computed pairwise distances between all pairs of peptidic residues at different distance thresholds (10Å, 15Å, 20Å, 25Å). The local context (at 10 - 15Å distance) is strongly correlated with physical distance between residues (**Supplementary Fig. 1b**).

The second constraint was inferred from analysis of the learned energy functions generated from our previous SH2-specific model[19]. When the inferred energy functions were compared using both t-SNE[1] and PCA[2] (**Supplementary Fig. 2**), several displayed strong pairwise similarity, with a number of clusters emerging. This suggested that reuse of a limited set of energy “basis” functions might help a new model generalize better. We formalized this observation in the HSM/D model, which resulted in significantly improved model performance over HSM/ID ($p \leq 2.4 \times 10^{-2}$).

References

1. Maaten, L. van der & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
2. F.R.S, K. P. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
3. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **28**, 321–377 (1936).
4. Zarrinpar, A., Bhattacharyya, R. P. & Lim, W. A. The Structure and Function of Proline Recognition Domains. *Sci STKE* **2003**, re8–re8 (2003).
5. Zarrinpar, A. & Lim, W. A. Converging on proline: the mechanism of WW domain peptide recognition. *Nat. Struct. Mol. Biol.* **7**, 611–613 (2000).
6. Cesareni, G., Gimona, M., Sudol, M. & Yaffe, M. *Modular Protein Domains*. (John Wiley & Sons, 2006).
7. Harris, B. Z. & Lim, W. A. Mechanism and role of PDZ domains in signaling complex assembly. *J. Cell Sci.* **114**, 3219–3231 (2001).
8. Kaneko, T. *et al.* Loops Govern SH2 Domain Specificity by Controlling Access to Binding Pockets. *Sci Signal* **3**, ra34–ra34 (2010).
9. Wagner, M. J., Stacey, M. M., Liu, B. A. & Pawson, T. Molecular Mechanisms of SH2- and PTB-Domain-Containing Proteins in Receptor Tyrosine Kinase Signaling. *Cold Spring Harb. Perspect. Biol.* **5**, a008987 (2013).
10. Superti-Furga, G., Fumagalli, S., Koegl, M., Courtneidge, S. A. & Draetta, G. Csk inhibition of c-Src activity requires both the SH2 and SH3 domains of Src. *EMBO J.* **12**, 2625–2634 (1993).
11. Creixell, P. *et al.* Unmasking Determinants of Specificity in the Human Kinome. *Cell* **163**, 187–201 (2015).
12. Jadwin, J. A., Ogiue-Ikeda, M. & Machida, K. The application of modular protein domains in proteomics. *FEBS Lett.* **586**, 2586–2596 (2012).
13. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425–440 (2015).
14. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).
15. Hein, M. Y. *et al.* A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell* **163**, 712–723 (2015).
16. Rolland, T. *et al.* A Proteome-Scale Map of the Human Interactome Network. *Cell* **159**, 1212–1226 (2014).
17. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **298**, 1912–1934 (2002).
18. Chen, M. J., Dixon, J. E. & Manning, G. Genomics and evolution of protein phosphatases. *Sci Signal* **10**, eaag1796 (2017).
19. AlQuraishi, M., Koytiger, G., Jenney, A., MacBeath, G. & Sorger, P. K. A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nat. Genet.* **46**, 1363–72 (2014).