SUPPLEMENTARY MATERIALS

pocketZebra: a web-server for automated selection and classification of subfamily-specific binding sites by bioinformatic analysis of diverse protein families

Dmitry Suplatov, Eugeny Kirilin, Mikhail Arbatsky, Vakil Takhaveev and Vytas Švedas*

Lomonosov Moscow State University, Belozersky Institute of Physicochemical Biology and Faculty of Bioengineering and Bioinformatics, Vorobjev hills 1-73, Moscow 119991, Russia

* Corresponding author: vytas@belozersky.msu.ru

Evaluation dataset

Prediction accuracy of pocketZebra was illustrated on a set of proteins that are known to possess at least two topographically independent binding sites which can be classified as primary or secondary to the main function.

The following steps were performed to build the dataset. First, all entries were automatically retrieved from the ASD database of allosteric proteins and modulators (1) if they were associated with PDB entries that had primary sites annotated in the Catalytic site atlas database (2). In addition, both primary and secondary sites discovered in such a manner had to be associated with crystallographic protein-ligand complexes in the PDB database. We have also manually searched the PDB databank using patterns [allosteric OR secondary] AND [regulator OR inhibitor OR substrate OR antibiotic] to retrieve protein-ligand complexes that were not present in the ASD database. On the second step we removed all proteins if their biological assembly contained more than 10000 atoms. This step was necessary because most of the web-servers from our representative list (Table S4) do not accept large structures for input. The remaining sites were further classified as primary or secondary to the main function. Annotation of primary and secondary sites was retrieved from the CSA and ASD databases or from the corresponding literature. Finally, we removed proteins if their primary and secondary sites were not structurally distinct. If primary and secondary sites are rather sub-sites of one large binding cavity, they are usually identified as a single pocket by the algorithms from the representative list. This results in inability to benchmark programs independently on different types of sites. The final non-redundant set includes 23 primary functional sites and 22 secondary functional sites (Table S1).

Bioinformatic analysis

For each protein chain involved in a ligand binding the corresponding multiple sequence alignments were taken from Pfam seed alignments database (3). A Pfam alignment was rejected in one of the following cases: if it covered only a small part of the protein chain; if it did not include the binding site; if the protein of interest was less than 70% identical to any sequence in the alignment as it means that the alignment does not contain the functional subfamily of the protein of interest. In these cases, the following protocol was used to build multiple alignments of diverse protein families with reasonable computational effort.

The protein of interest was used as a query for PDBeFold algorithm (4) to select structurally similar proteins with matching of at least 30% of secondary structure elements, but not more than 300 structures. Then, a pairwise sequence similarity threshold was chosen in a range from 30% to 50% to select a non-redundant representative set of not more than 20 protein structures. These template proteins were structurally superimposed using Matt (5) to be used as a core alignment of a corresponding protein family. Then, the multiple structure-guided-sequence alignment was built as previously discussed (6). Every template protein was independently used as a query for 2 iterations of PSI-BLAST search (7) to collect homologous sequences from Swissprot database. Only one sequence was retained from a group with more than 95% pairwise identity to remove redundant sequences. Sequences sharing less than 0.25 bits score per column with the corresponding template protein were removed (8). If no similar proteins were found in the Swissprot database the template protein was removed. Incomplete sequences and those that introduced significant insertions to the template protein were removed to reduce amount of columns overpopulated by gaps. Sequence sets were aligned to the corresponding template proteins resulting in a final multiple structure-guided-sequence alignment.

The bioinformatic analysis of the obtained protein family alignments was performed using Zebra algorithm (9) implemented in pocketZebra with default setup and 10000 random shuffles for every column. Columns with more than 5% of gaps were not considered. The most significant functional subfamily classification was automatically selected. The global P-value minimum was used as a threshold to select the most significant subfamily-specific positions.

Evaluation protocol

We evaluated a representative set of web-based methods – Fpocket, POCASA, GHECOM, SiteHound, DogSiteScorer and LIGSITE^{csc} – in their ability to correctly rank known functionally important binding sites against our web-implementation pocketZebra. Biological units of the corresponding proteins from the test set were used for structural analysis. Full-size biological

units were created from the original PDB files using MakeMultimer (http://watcut.uwaterloo.ca/cgi-bin/makemultimer). All heteroatoms were removed. All web-implementations were used with the default set-up. LIGSITE^{csc} was set to identify 20 sites for each protein structure and the "Re-rank by conservation" option was activated.

For each program a list of predicted pockets was obtained with the original ranking. To assess whether a functionally important binding site known from the literature was found or not the following criteria were used similarly to what has been previously described (10): the geometric center of a pocket has to be within 4.5Å from any atom of the ligand; at least 30% of the ligand atoms have to lie within 4.5Å from any atom of the pocket; and at least 20% of the pocket atoms have to lie within 4.5Å from any atom of the ligand. These conditions ensure that the pocket and the ligand have comparable dimensions and that a significant portion of the ligand is covered by a significant part of the pocket. At the same time they allow a pocket to be larger than the ligand but not too large and also accommodate frequent cases when multiple independent pockets are predicted to bind different pharmacophores of a single ligand.

POCASA, GHECOM and SiteHound returned coordinates of probes filling a pocket while LIGSITE^{csc} returned coordinates of a pocket center. Thus, in order to obtain a list of amino acid residues actually forming the corresponding pockets we benchmarked those programs by taking all residues within a certain radius from the identified probes. Radiuses within a range 1Å to 6Å were tested (1Å to 10Å for LIGSITE^{csc}) and values that received the highest Precision-recall AUC score (see METHODS in the main article) were selected for each program – 3Å, 2Å, 5Å and 8Å respectively.

pocketZebra, by default, implements the Fpocket algorithm to actually detect the sites which are then re-ranked by the bioinformatic analysis (see METHODS in the main article). In this work we also used the pocketZebra scoring function to re-rank the binding sites predicted by other methods from our representative list of web-servers. Results showed different trends among the programs (data not shown) which can be explained by varying definitions (exact residue content) of pockets produced by each algorithm. In terms of shape and dimensions compared to the bound ligand the Fpocket, to our opinion, suggested the most accurate predictions for the known sites from the set. Eventually, the default combination Fpocket+pocketZebra showed the best performance (the highest Precision-recall AUC scores).

Table S1. The evaluation dataset.

#	SCOP families	Protein name and source	Primary site		Secondary site		Ref.
			Substrate name	**PDB code**	**Substrate name**	**PDB code**	
1	Caspase catalytic domain	Caspase-7 from *Homo sapiens*	(3S)-3-amino-4-hydroxybutanoic acid (ASJ)	4HQR	2-(2,4-dichloro-phenoxy)-n-(2-mercapto-ethyl)-acetamide (NXN)	1SHJ	11, 12
2	Arginine methyltransferase	Arginine N-methyltransferase 3 from *Homo sapiens*	S-adenosyl-l-homocysteine (SAH)	2FYT	1-(1,2,3-benzothiadiazol-6-yl)-3-[2-(cyclohex- 1-en-1-yl)ethyl]urea (TDU)	3SMQ	13
3	Motor proteins	Kinesin-like protein KIF11 from *Homo sapiens*	Adenosine 5′-diphosphate (ADP)	1X88	Ethyl 4-(3-hydroxyphenyl)-6-methyl-2-thioxo-1,2,3,4-tetrahydropyrimidine-5-carboxylate (NAT)	1X88	-
4	Hexokinase	Hexokinase type I from *Homo sapiens*	Alpha-D-glucose-6-phosphate (G6P), Alpha-D-glucose (GLC)	1CZA	Adenosine 5′-diphosphate (ADP)	1CZA	14
5	Lactate & malate dehydrogenases (C-terminal domain)	L-lactate dehydrogenase from *Bifidobacterium longum*	Nicotinamide-adenine-dinucleotide (NAD)	1LTH	Beta-fructose-1,6-diphosphate (FBP)	1LTH	15
6	Tryptophan synthase beta subunit-like PLP-dependent enzymes	O-acetylserine sulfhydrylase from *Salmonella enterica*	Pyridoxal-5'-phosphate (PLP)	1FCJ	Chloride ion (CL)	1FCJ	16
7	R1 subunit of ribonucleotide reductase (N-terminal domain), R1 subunit of ribonucleotide reductase (C-terminal domain)	Ribonucleotide reductase R1 protein from *Escherichia coli*	Guanosine-5'-diphosphate (GDP)	4R1R	Thymidine-5'-triphosphate (TTP)	4R1R	17
8	Ribonuclease A-like	Seminal ribonuclease from *Bos taurus*	Uridylyl-2'-5'-phospho-guanosine (U2G 130, 132)	11BG	Uridylyl-2'-5'-phospho-guanosine (U2G 131)	11BG	18

9	Phosphotriesterase-like	Parathion hydrolase from *Brevundimonas diminuta*	Cobalt ion (CO)	1QW7	Diethyl 4-methylbenzylphosphonate (EBP)	1QW7	19
10	Multidomain cupredoxins	Copper-containing nitrite reductase from *Alcaligenes faecalis*	Acetamide (ACM 2503)	1ZDS	(Methylsulfanyl)methane (MSM)	1ZDQ	20
11	* Flavin containing amine oxidoreductase	Amine oxidase B from *Homo sapiens*	[[(2r,3s,4s)-5-[(4as)-7,8-dimethyl-2,4-dioxo- 4a,5-dihydrobenzo[g]pteridin-10-yl]-2,3,4- trihydroxy-pentoxy]-hydroxy-phosphoryl] [(2r,3s,4r,5r)-5-(6-aminopurin-9-yl)-3,4-dihydroxy-oxolan- 2-yl]methyl hydrogen phosphate (FA8) 3-phenylpropanal (3PL)	2XCG	2-(2-benzofuranyl)-2-imidazoline (XCG)	2XCG	21
12	* D-ala D-ala ligase N-terminus, D-ala D-ala ligase C-terminus	D-alanine-D-alanine ligase from *Staphylococcus aureus*	Adenosine 5′-diphosphate (ADP)	2I8C	3-chloro-2,2-dimethyl-n-[4-(trifluoromethyl)phenyl]propanamide (G1L)	2I80	22
13	Glucose-1-phosphate thymidylyltransferase	Glucose-1-phosphate thymidylyltransferase from *Pseudomonas aeruginosa*	Thymidine-5'-triphosphate bound to the active site (TTP)	1G2V	N-(6-amino-1-benzyl-2,4-dioxo-1,2,3,4-tetrahydropyrimidin- 5-yl) benzamide (BZ0)	4B2W	23, 24, 25
14	G proteins, Elongation factors, EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain	Elongation factor Tu-A from *Thermus thermophilus HB8*	Pulvomycin (PUL) Phosphoaminophosphonic acid-guanylate ester (GNP)	2C78 2C78	-	-	26,27
15	APH phosphotransferases	Aminoglycoside 3'-phosphotransferase from *Acinetobacter baumannii*	1-tert-butyl-3-(naphthalen-1-ylmethyl)-1H- pyrazolo[3,4-d]pyrimidin-4-amine (0JN)	4GKI	Kanamycin A (KAN)	4GKI	28
16	GABA-aminotransferase-like	5-aminolevulinate synthase from *Rhodobacter capsulatus*	Pyridoxal-5'-phosphate (PLP) Adenine and ribose moieties of Succinyl-coenzyme A (SCA)	2BWO 2BWO	-	-	29
17	Higher-molecular-weight phosphotyrosine protein phosphatases	Protein-tyrosine phosphatase-1B from *Homo sapiens*	4-phosphonooxyphenyl-methyl-[4-phosphonooxy]benzen (BPM)	1AAX	3-(3,5-Dibromo-4-Hydroxy-Benzoyl)-2-Ethyl-Benzofuran-6-Sulfonic Acid (4-Sulfamoyl-Phenyl)-	1T49	30,31, 32

					Amide (892)		
18	Transducin (alpha subunit, insertion domain), G proteins, Adenylyl and guanylyl cyclase catalytic domain,	Adenylate cyclase from *Canis lupus familiaris*	5'-guanosine-diphosphate-monothiophosphate (GSP)	2GVD	Methylpiperazinoforskolin (FKP) in chains A and B	2GVD	33, 34
			Spiro(2,4,6-Trinitrobenzene[1,2a]-2o',3o'-Methylene-Adenine-Triphosphate (128) in chains A and B	2GVD			
19	Protein kinases (catalytic subunit)	Proto-oncogene tyrosine-protein kinase ABL1 from *Mus musculus*	4-[(4-methylpiperazin-1-yl)methyl]-N-(4-methyl-3-{[4-(pyridin-3-yl)pyrimidin-2-yl]amino}phenyl)benzamide (STI)	3K5V	3-(6-{[4-(trifluoromethoxy)phenyl]amino}pyrimidin- 4-yl)benzamide (STJ)	3K5V	35, 36, 37, 38
		3-Phosphoinositide-dependent protein kinase 1 from *Homo sapiens*	-	-	(3S)-4-(5-chloro-1H-benzimidazol-2-yl)-3- (4-chlorophenyl)butanoic acid (A06)	4A06	39, 40
		RAC-alpha serine/threonine-protein kinase from *Homo sapiens*	-	-	1-(1-(4-(7-phenyl-1H-imidazo[4,5-g]quinoxalin- 6-yl)benzyl)piperidin-4-yl)-1H-benzo[d]imidazol- 2(3H)-one (IQO)	3O96	41, 42
		CHK1 Checkpoint kinase from *Homo sapiens*	-	-	(1S)-1-(1H-benzimidazol-2-yl)ethyl (3,4-dichlorophenyl) carbamate (AGX)	3JVR	43, 44, 45
		Mitogen-activated protein kinase 8 from *Homo sapiens*	-	-	Glycerol (GOL)	2H96	46

SCOP families are shown for all chains that are involved in a ligand binding.

* - PFAM families are shown where SCOP classification was not available

Table S2. Contingency table for distribution of the subfamily-specific positions (SSP) in the primary binding sites.

	SSP	**Non-SSP**	**Total**
Pocket	123	263	386
Non-Pocket	1264	6534	7798
Total	1387	6797	8184

Category "Pocket" corresponds to residues within 5Å from ligands bound to the primary sites. Positions were calculated from a non-redundant set of chains that participate in binding of experimentally confirmed primary ligands. Bioinformatic analysis of the subfamily-specific positions was performed as described in the "Bioinformatic analysis" section of Supplementary materials. Each cell describes the actual number of positions in the corresponding category.

Of the positions located within the primary pockets, $\frac{123}{386} = 32\%$ are SSPs; of the positions located outside the primary pockets, $\frac{1264}{7798} = 16\%$ are SSPs. Is there a relationship between localization (pocket or non-pocket) and specificity (SSP or non-SSP) of a position? We carried out a Chi-squared test for independence to assess whether paired observations of the two variables are independent of each other. The null hypothesis H_0 assumes that there is no relationship between localization and specificity of a position (47).

χ^2 test gives a p-value $2.1*10^{-15}$. Consequently, we would reject the H_0 and conclude that for this sample, position being an SSP is positively associated with being located in a primary site.

Table S3. Contingency table for distribution of the subfamily-specific positions (SSP) in the secondary binding sites.

	SSP	**Non-SSP**	**Total**
Pocket	89	197	286
Non-Pocket	1141	6077	7218
Total	1230	6274	7504

Category "Pocket" corresponds to residues within 5Å from ligands bound to the secondary sites. Positions were calculated from a non-redundant set of chains that participate in binding of experimentally confirmed secondary ligands. Bioinformatic analysis of the subfamily-specific positions was performed as described in the "Bioinformatic analysis" section of Supplementary materials. Each cell describes the actual number of positions in the corresponding category.

χ^2 test for independence gives a p-value $1.2*10^{-11}$ (see detailed description for Table S2). Consequently, we would reject the hypothesis of independence and conclude that for this sample, position being an SSP is positively associated with being located in a secondary site.

Table S4. Web-servers that attempt to detect and rank binding sites by implementing different algorithmic strategies and were used for testing and comparison in this publication.

Name	Web-server address	Strategy used to identify the binding sites	Strategy used to rank the binding sites	Ref
Fpocket	http://mobyle.rpbs.univ-paris-diderot.fr/ > Programs > Structure > Pockets > fpocket	Geometric search method based on Voronoi tessellation and alpha sphere detection	By a scoring function calibrated on a training set of known protein-ligand complexes that describes putative capacity of a pocket to bind small molecules	10, 48
POCASA	http://altair.sci.hokudai.ac.jp/g6/service/pocasa/	Geometric search method based on a 3D grid representation of proteins and a probe sphere rolling	By volume depth which includes pocket volume and position	49
GHECOM	http://strcomp.protein.osaka-u.ac.jp/ghecom/	Geometric search method using based on a 3D grid representation, probes and the theory of mathematical morphology	By size and depth of a pocket	50
SiteHound	http://scbx.mssm.edu/sitehound/sitehound-web/Input.html	Energetic search method using the interaction energy between the protein and a chemical probe that can be carbon- or phosphate-like	By energy of binding a model substrate (by default – a carbon probe) to the protein	51
DogSiteScorer	http://dogsite.zbh.uni-hamburg.de/	Grid-based function prediction method which uses a Difference of Gaussian filter	By a set of descriptors representing size, compactness and physicochemical properties of a pocket	52
LIGSITE^{csc}	http://projects.biotec.tu-dresden.de/pocket/	Geometric search method based on 3D grid representation and the Connolly surface	By degree of conservation of surface residues in a pocket	53

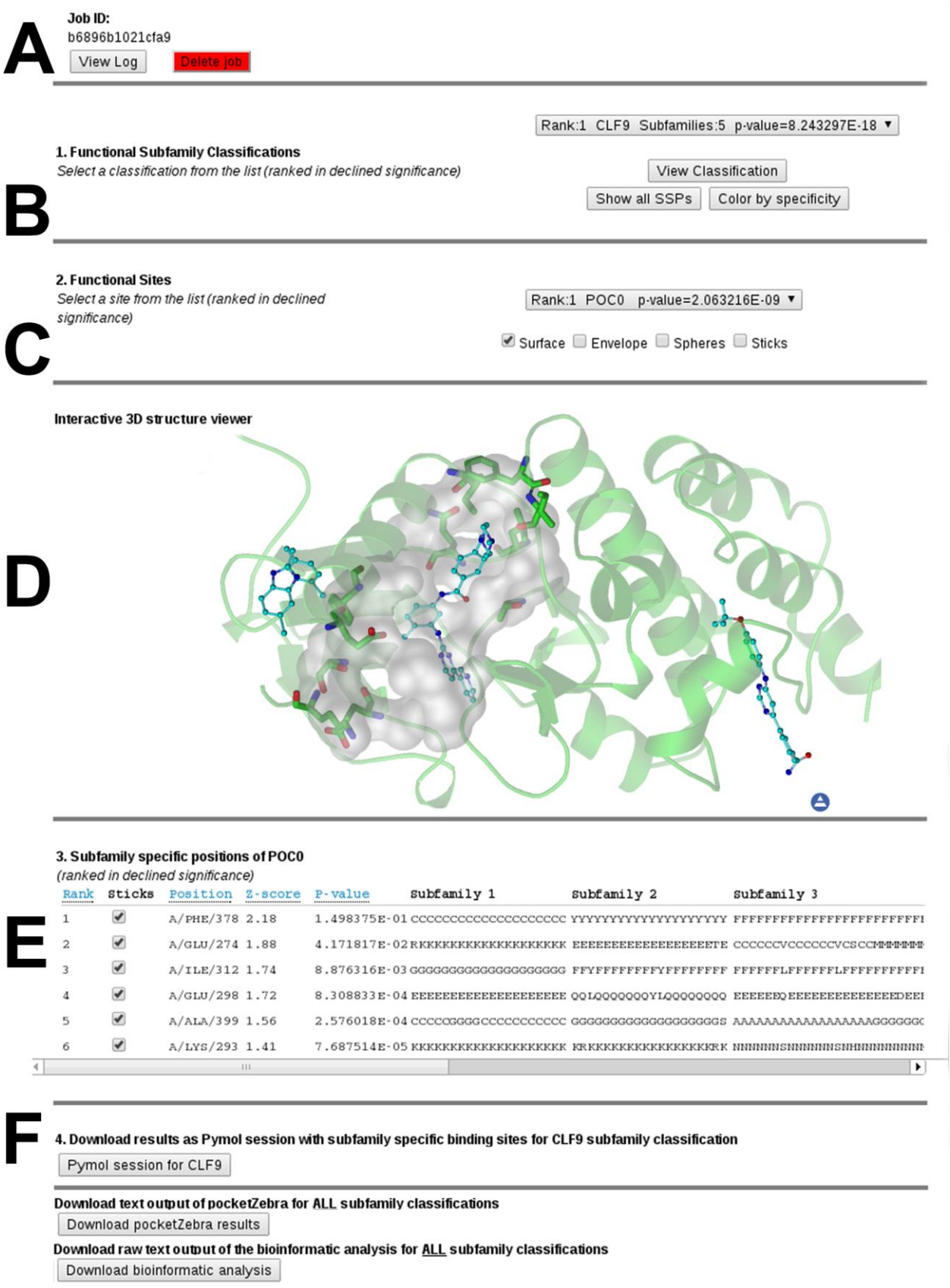

Fig. S1. pocketZebra web-server results page. Heteroatoms are shown as ball-and-sticks and colored in cyan. Subfamily-specific positions are shown as sticks. The original screenshot has been modified in a graphical editor to fit all the necessary information to a single image.

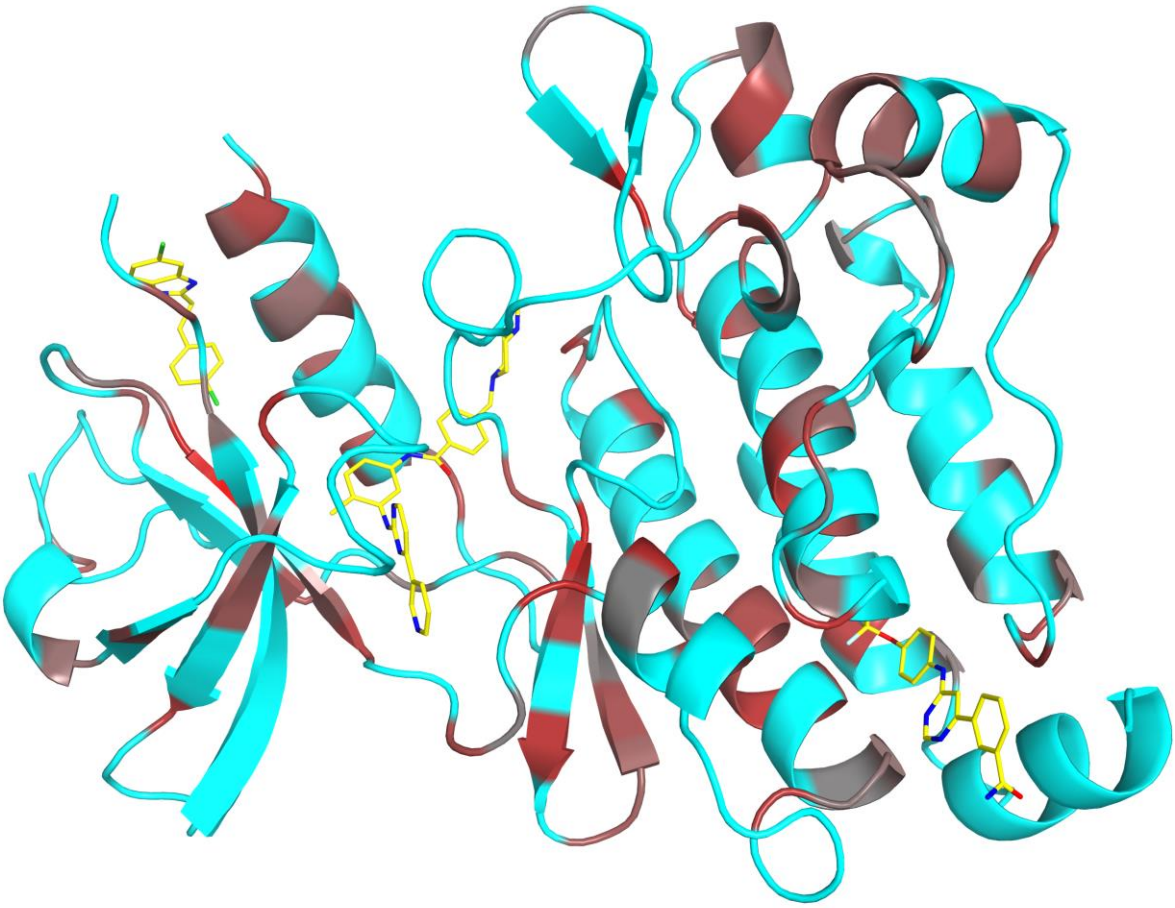

Fig. S2. Subfamily-specific positions in the Protein kinases family. Gradient paint of the Cα-atoms corresponds to estimated specificity of corresponding residues: red stands for highly significant subfamily-specific positions, cyan – for non-specific positions. Ligands that mark functionally important sites were taken from PDBs with codes 3K5V and 4A06 and colored in yellow. This figure was created from a PyMol session with structural representation of pocketZebra results that was automatically produced by the server.

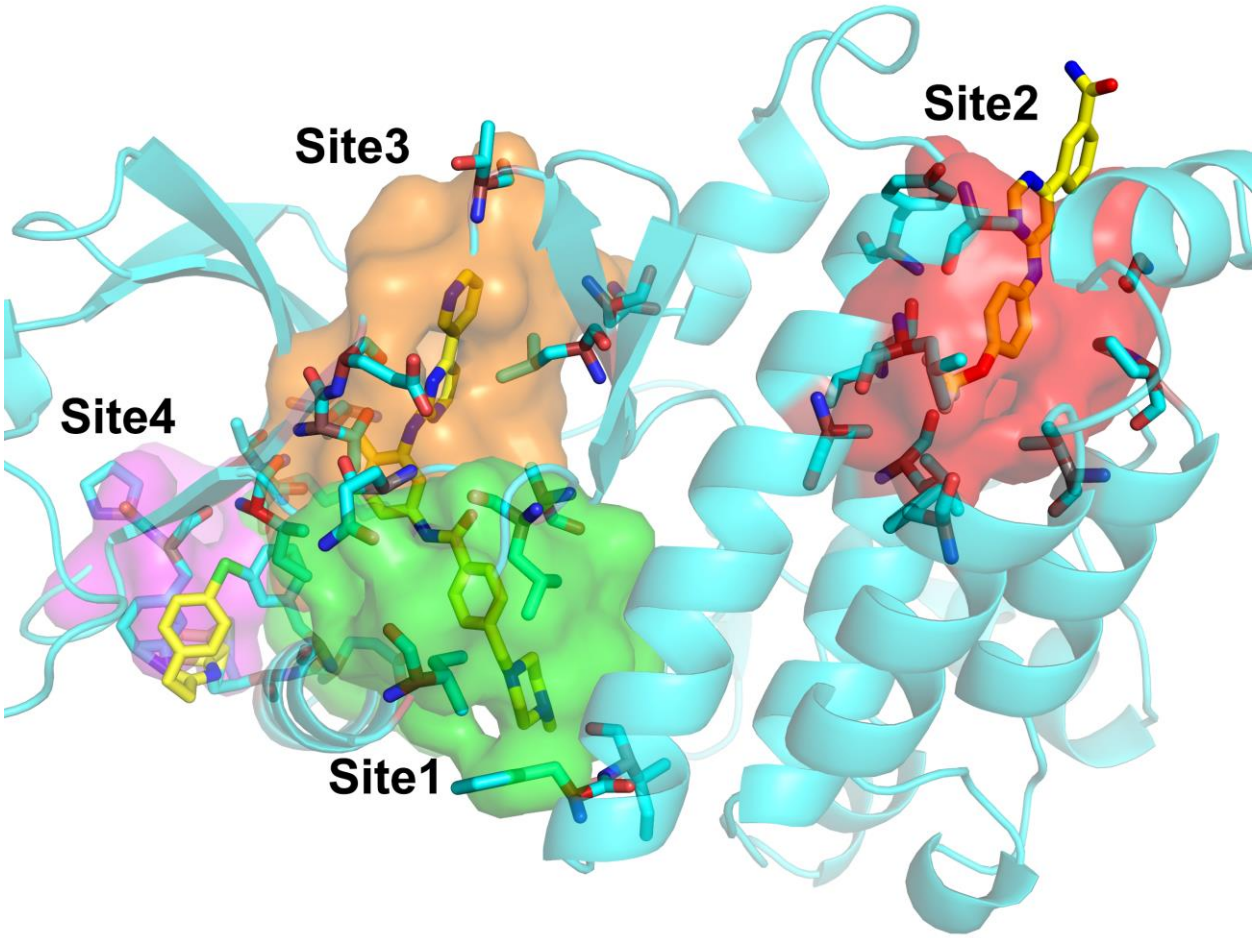

Fig. S3. The top-scoring subfamily-specific binding sites in the Protein kinases family. Ligands that mark the functionally important sites were taken from PDBs with codes 3K5V and 4A06 and colored in yellow. Pockets that were ranked by pocketZebra as 1st and 3rd (Site1 and Site3) bind the primary substrate that occupies the catalytic site, while 2nd and 4th pockets (Site2 and Site4) correspond to topographically independent allosteric sites known from the literature (see Table S1, Protein kinases family). Subfamily-specific positions in the corresponding pockets are shown as sticks. This figure was created from a PyMol session with structural representation of pocketZebra results that was automatically produced by the server.

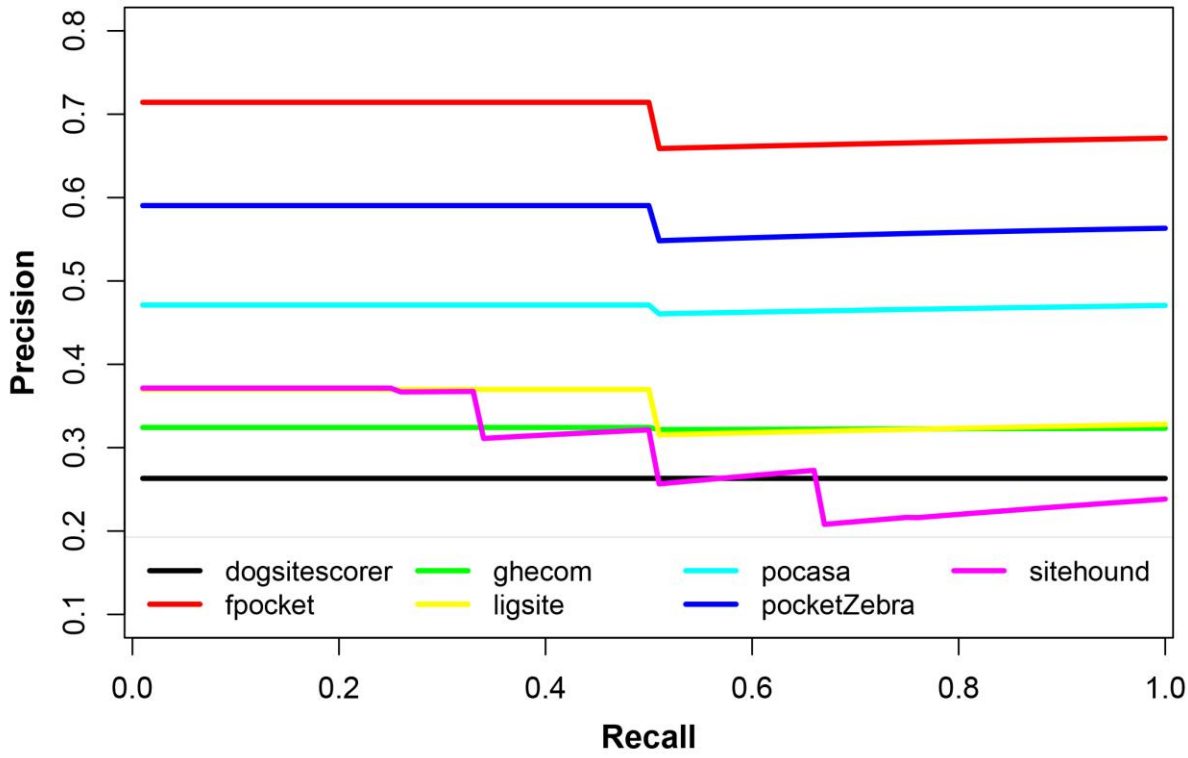

Fig. S4. PR curves for detecting and ranking the primary functional sites by the representative web-servers.

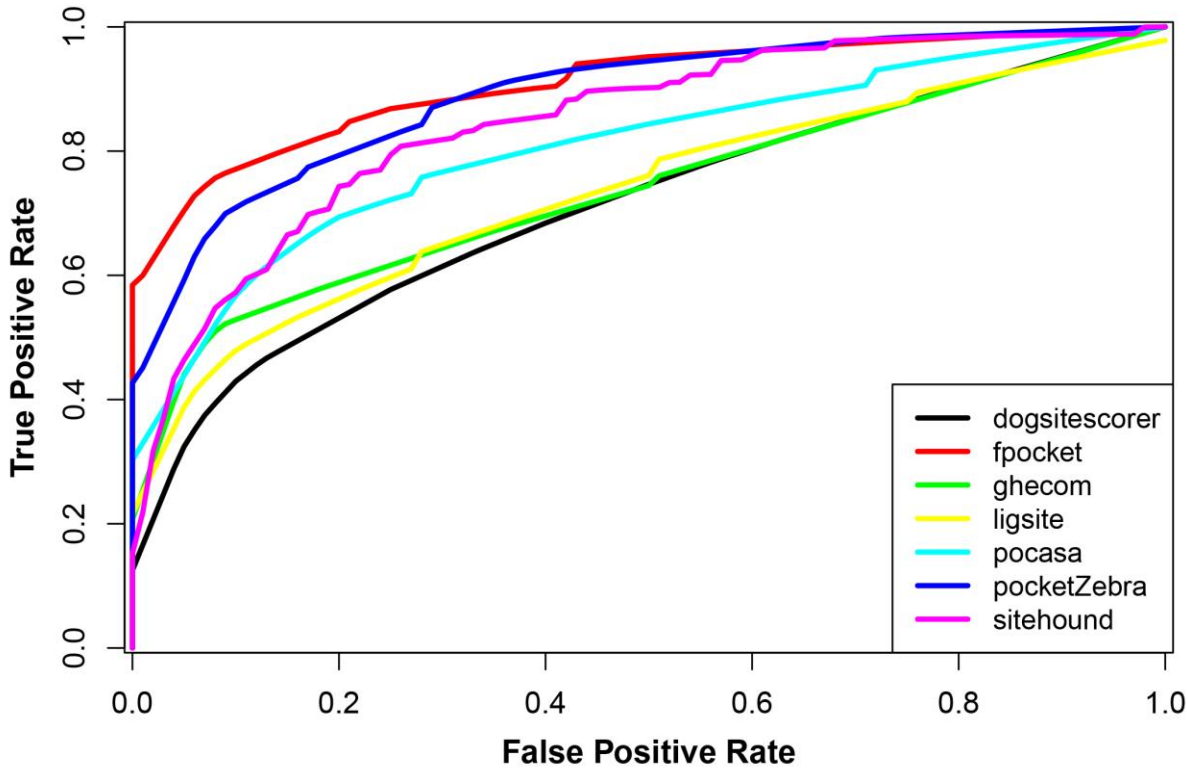

Fig. S5. ROC curves for detecting and ranking the primary functional sites by the representative web-servers.

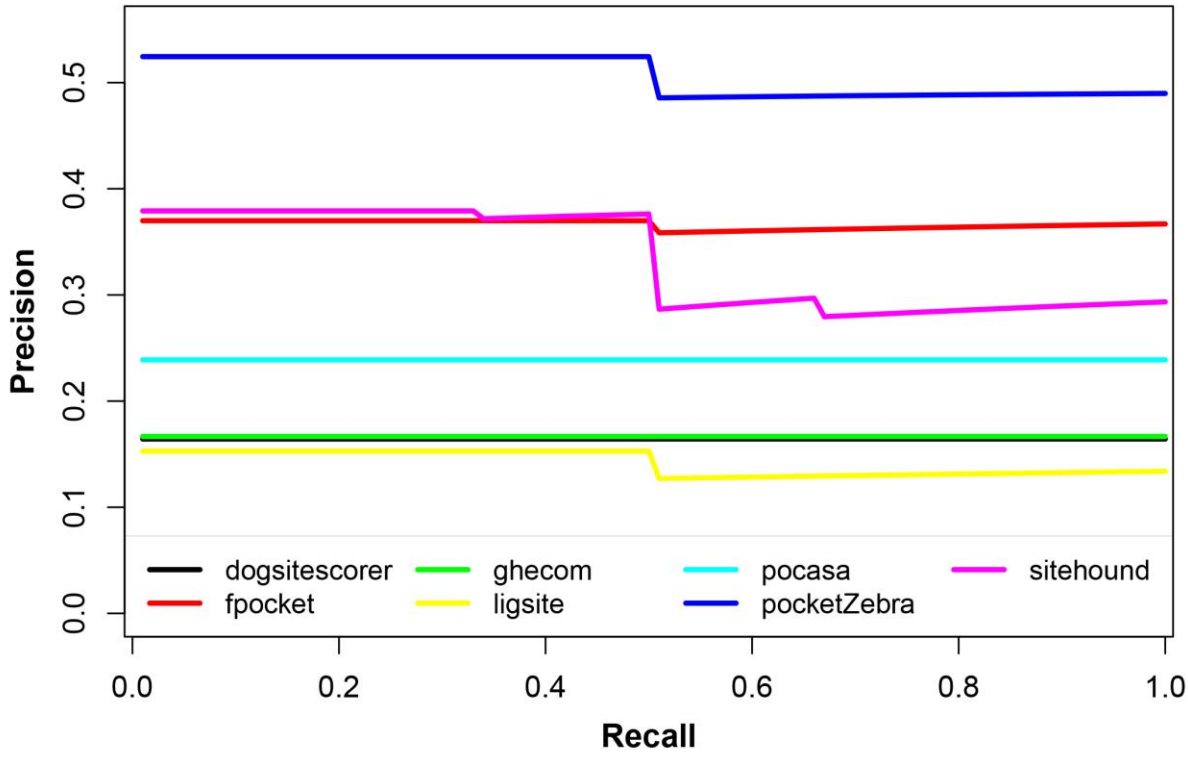

Fig. S6. PR curves for detecting and ranking the secondary functional sites by the representative web-servers.

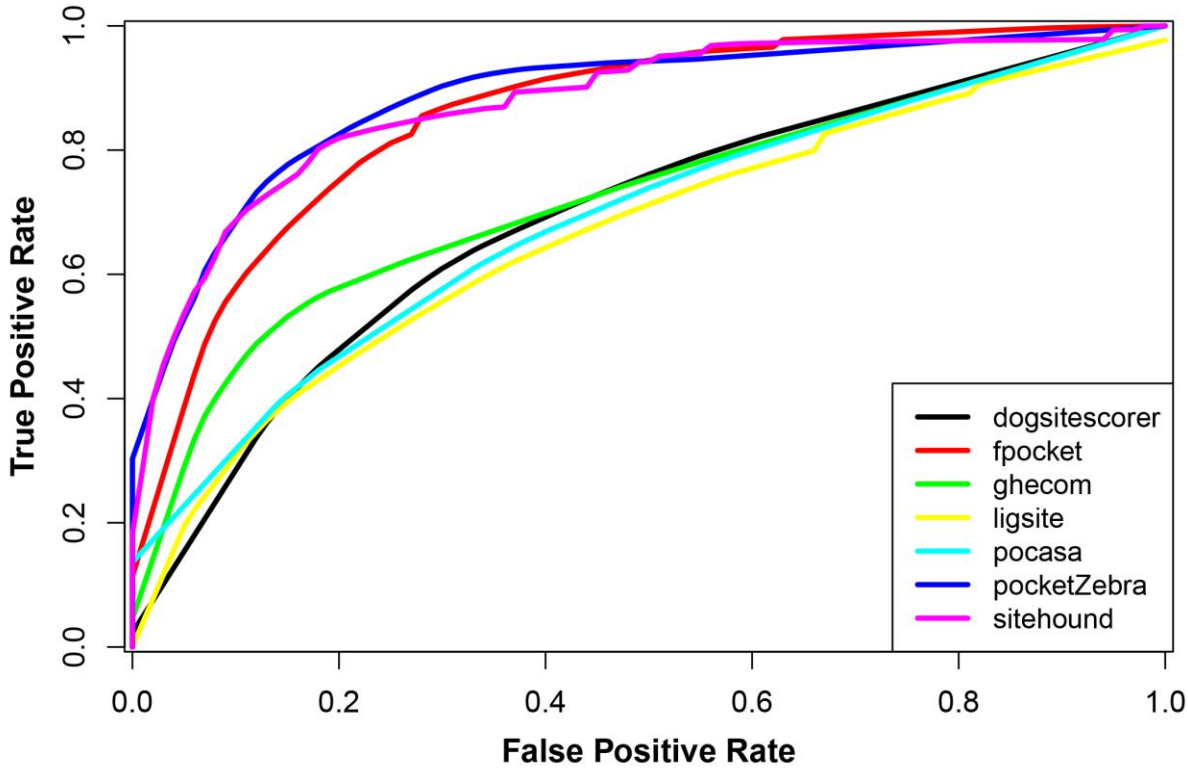

Fig. S7. ROC curves for detecting and ranking the secondary functional sites by the representative web-servers.

REFERENCES

1. Huang,Z., Zhu,L., Cao,Y., Wu,G., Liu,X., Chen,Y., et al. (2011) ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic acids res.*, **39**, D663-D669.

[2] Porter,C.T., Bartlett,G.J., and Thornton,J.M. (2004). The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res*., **32(suppl 1)**, D129-D133.

3. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., et al. (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281-D288.

4. Krissinel,E., and Henrick,K. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallogr D Biol Crystallogr*., **60(12)**, 2256-2268.

5. Menke,M., Berger,B., and Cowen,L. (2008) Matt: local flexibility aids protein multiple structure alignment. *PLoS Comput Biol*., **4(1)**, e10.

6. Suplatov,D., Kirilin,E., Takhaveev,V., and Švedas,V. (2013) Zebra: a web server for bioinformatic analysis of diverse protein families. *J Biomol Struct Dyn.*, doi:10.1080/07391102.2013.834514.

7. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25(17)**, 3389-3402.

8. Fischer, J.D., Mayer, C.E., & Söding J. (2008) Prediction of protein functional residues from sequence by probability density estimation. Bioinformatics. 24(5), 613-20.

9. Suplatov,D., Shalaeva,D., Kirilin,E., Arzhanik,V., and Švedas,V. (2014) Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity. *J Biomol Struct Dyn.*, **32(1)**, 75-87.

10. Le Guilloux,V., Schmidtke,P., and Tuffery,P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC bioinformatics*, **10(1)**, e168.

11. Hardy,J.A., Lam,J., Nguyen,J.T., O'Brien,T., and Wells, J.A. (2004) Discovery of an allosteric site in the caspases. *Nucleic acids res.*, **101(34),** 12461-12466.

12. Kang,H.J., Lee,Y.M., Bae,K.H., Kim,S.J., and Chung,S.J. (2013) Structural asymmetry of procaspase-7 bound to a specific inhibitor. *Acta Crystallogr. Sect. D-Biol. Crystallogr.*, **69(8)**, 1514-1521.

13. Siarheyeva,A., Senisterra,G., Allali-Hassani,A., Dong,A., Dobrovetsky,E., Wasney,G.A., et al. (2012) An allosteric inhibitor of protein arginine methyltransferase 3. *Structure*, **20(8)**, 1425-1435.

14. Aleshin,A.E., Kirby,C., Liu,X., Bourenkov,G.P., Bartunik,H.D., Fromm,H.J., et al. (2000) Crystal structures of mutant monomeric hexokinase I reveal multiple ADP binding sites and conformational changes relevant to allosteric regulation. *J Mol Biol.*, **296(4)**, 1001-1015.

15. Iwata,S., Kamata,K., Yoshida,S., Minowa,T., and Ohta,T. (1994) T and R states in the crystals of bacterial L–lactate dehydrogenase reveal the mechanism for allosteric control. *Nat. Struct. Mol. Biol.*, **1(3)**, 176-185.

16. Burkhard,P., Tai,C.H., Jansonius,J.N., and Cook,P.F. (2000) Identification of an allosteric anion-binding site on O-acetylserine sulfhydrylase: structure of the enzyme with chloride bound. *J Mol Biol.*,**303(2)**, 279-286.

17. Eriksson,M., Uhlin,U., Ramaswamy,S., Ekberg,M., Regnström,K., Sjöberg,B.M., et al. (1997) Binding of allosteric effectors to ribonucleotide reductase protein R1: reduction of active-site cysteines promotes substrate binding. *Structure*, **5(8)**, 1077-1092.

18. Vitagliano,L., Adinolfi,S., Sica,F., Merlino,A., Zagari,A., and Mazzarella,L. (1999) A potential allosteric subsite generated by domain swapping in bovine seminal ribonuclease. *J Mol Biol.*, **293(3)**, 569-577.

19. Grimsley,J.K., Calamini,B., Wild,J.R., and Mesecar,A.D. (2005) Structural and mutational studies of organophosphorus hydrolase reveal a cryptic and functional allosteric-binding site. *Arch. Biochem. Biophys.*,**442(2)**, 169-179.

20. Wijma,H.J., MacPherson,I., Alexandre,M., Diederix,R.E., Canters,G.W., Murphy,M.E., et al. (2006) A rearranging ligand enables allosteric control of catalytic activity in copper-containing nitrite reductase. *J Mol Biol.*, **358(4)**, 1081-1093.

21. Bonivento,D., Milczek,E.M., McDonald,G.R., Binda,C., Holt,A., Edmondson, D.E., et al. (2010) Potentiation of ligand binding through cooperative effects in monoamine oxidase B. *J Biol Chem.*, **285(47)**, 36849-36856.

22. Liu,S., Chang,J.S., Herberg,J.T., Horng,M.M., Tomich,P. K., Lin,A. H., et al. (2006) Allosteric inhibition of Staphylococcus aureusd-alanine: d-alanine ligase revealed by crystallographic studies. *Nucleic acids res.*, **103(41)**, 15178-15183.

23. Alphey,M.S., Pirrie,L., Torrie,L.S., Boulkeroua,W.A., Gardiner,M., Sarkar,A., et al. (2012) Allosteric Competitive Inhibitors of the Glucose-1-phosphate Thymidylyltransferase (RmlA) from Pseudomonas aeruginosa. *ACS Chem Biol.*, **8(2)**, 387-396.

24. Sivaraman,J., Sauvé,V., Matte,A., and Cygler,M. (2002) Crystal structure of Escherichia coli glucose-1-phosphate thymidylyltransferase (RffH) complexed with dTTP and Mg2+. *J Biol Chem.*, **277(46)**, 44214-44219.

25. Blankenfeldt,W., Asuncion,M., Lam,J.S., and Naismith,J.H. (2000) The structural basis of the catalytic mechanism and regulation of glucose-1-phosphate thymidylyltransferase (RmlA). *EMBO J.*, **19(24)**, 6652-6663.

26. Parmeggiani,A., Krab,I.M., Okamura,S., Nielsen,R.C., Nyborg,J., and Nissen,P. (2006) Structural basis of the action of pulvomycin and GE2270 A on elongation factor Tu. *Biochemistry*, **45(22)**, 6846-6857.

27. Uejima,T., Ihara,K., Goh,T., Ito,E., Sunada,M., Ueda,T., et al. (2010) GDP-bound and Nucleotide-free Intermediates of the Guanine Nucleotide Exchange in the Rab5· Vps9 System. *J Biol Chem.*, **285(47)**, 36689-36697.

28. Peter,J.S., Peter,S., Tushar,S., Nick,T., Alfredo,C., and Gerard,D.W. (2013) Structure-guided optimization of protein kinase inhibitors reverses aminoglycoside antibiotic resistance. *Biochem J.*, **454(2)**, 191-200.

29. Astner,I., Schulze,J.O., van den Heuvel,J., Jahn,D., Schubert,W.D., and Heinz,D.W. (2005) Crystal structure of 5-aminolevulinate synthase, the first enzyme of heme biosynthesis, and its link to XLSA in humans. *EMBO J.*, **24(18)**, 3166-3177.

30. Barford,D., Flint,A.J., and Tonks,N.K. (1994) Crystal structure of human protein tyrosine phosphatase 1B. *Science*, **263(5152)**, 1397-1404.

31. Puius,Y.A., Zhao,Y., Sullivan,M., Lawrence,D.S., Almo,S.C., and Zhang,Z.Y. (1997) Identification of a second aryl phosphate-binding site in protein-tyrosine phosphatase 1B: a paradigm for inhibitor design. *Proc Natl Acad Sci U S A*, **94(25)**, 13420-13425.

32. Wiesmann,C., Barr,K.J., Kung,J., Zhu,J., Erlanson,D.A., Shen,W., et al. (2004) Allosteric inhibition of protein tyrosine phosphatase 1B. *Nat Struct Mol Biol.*, **11(8)**, 730-737.

33. Topal,H., Fulcher,N.B., Bitterman,J., Salazar,E., Buck,J., Levin,L.R., et al. (2012) Crystal Structure and Regulation Mechanisms of the CyaB Adenylyl Cyclase from the Human Pathogen Pseudomonas aeruginosa. *J Mol Biol.*, **416(2)**, 271-286.

34. Mou,T.C., Gille,A., Fancy,D.A., Seifert,R., and Sprang,S.R. (2005) Structural basis for the inhibition of mammalian membrane adenylyl cyclase by 2′(3′)-O-(N-methylanthraniloyl)-guanosine 5′-triphosphate. *J Biol Chem.*, **280(8)**, 7253-7261.

35. Weisberg,E., Manley,P.W., Breitenstein,W., Brüggen,J., Cowan-Jacob,S.W., Ray,A., et al. (2005) Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer cell*, **7(2)**, 129-141.

36. Nagar,B., Bornmann,W.G., Pellicena,P., Schindler,T., Veach,D.R., Miller,W.T., et al. (2002) Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571).*Cancer Research*, **62(15)**, 4236-4243.

37. Cowan-Jacob,S.W., Fendrich,G., Floersheimer,A., Furet,P., Liebetanz,J., Rummel,G., et al. (2006) Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Crystallogr D Biol Crystallogr.*, **63(1)**, 80-93.

38. Yang,J., Campobasso,N., Biju,M.P., Fisher,K., Pan,X.Q., Cottom,J., et al. (2011) Discovery and characterization of a cell-permeable, small-molecule c-Abl kinase activator that binds to the myristoyl binding site. *Chemistry & biology*, **18(2)**, 177-186.

39. Biondi,R.M., Komander,D., Thomas,C.C., Lizcano,J.M., Deak,M., Alessi,D.R., et al. (2002) High resolution crystal structure of the human PDK1 catalytic domain defines the regulatory phosphopeptide docking site. *EMBO J.*, **21(16)**, 4219-4228.

40. Lopez-Garcia,L.A., Schulze,J.O., Fröhner,W., Zhang,H., Süß,E., Weber,N., et al. (2011) Allosteric regulation of protein kinase PKCζ by the N-terminal C1 domain and small compounds to the PIF-pocket. *Chemistry & biology*, **18(11)**, 1463-1473.

41. Wu,W.I., Voegtli,W.C., Sturgis,H.L., Dizon,F.P., Vigers,G.P., and Brandhuber,B.J. (2010) Crystal structure of human AKT1 with an allosteric inhibitor reveals a new mode of kinase inhibition. *PLoS One*, **5(9)**, e12913.

42. Zhu,X., Kim,J.L., Newcomb,J.R., Rose,P.E., Stover,D.R., Toledo,L.M., et al. (1999) Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. *Structure*, **7(6)**, 651-661.

43. Chen,P., Luo,C., Deng,Y., Ryan,K., Register,J., Margosiak,S., et al. (2000) The 1.7 A crystal structure of human cell cycle checkpoint kinase Chk1: implications for Chk1 regulation. *Cell*, **100(6)**, 681-692.

44. Vanderpool,D., Johnson,T.O., Ping,C., Bergqvist,S., Alton,G., Phonephaly,S., et al. (2009) Characterization of the CHK1 allosteric inhibitor binding site. *Biochemistry*, **48(41)**, 9823-9830.

45. Matthews,T.P., Klair,S., Burns,S., Boxall,K., Cherry,M., Fisher,M., et al. (2009) Identification of inhibitors of checkpoint kinase 1 through template screening. *J Med Chem.*, **52(15)**, 4810-4819.

46. Zhao,H., Serby,M.D., Xin,Z., Szczepankiewicz,B.G., Liu,M., Kosogof,C., el al. (2006) Discovery of potent, highly selective, and orally bioavailable pyridine carboxamide c-Jun NH2-terminal kinase inhibitors. *J Med Chem.*, **49(15)**, 4455-4458.

47. Rice,J. (2007). Chapter 13: The analysis of categorical data, pp. 514-541 in Mathematical statistics and data analysis, 3rd Edition. Duxbury Advanced.

48. Schmidtke,P., Le Guilloux,V., Maupetit,J., and Tufféry,P. (2010) Fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.*, **38**, W582-W589.

49. Yu,J., Zhou,Y., Tanaka,I., and Yao,M. (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. *Bioinformatics*, **26(1)**, 46-52.

50. Kawabata,T. (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. *Proteins*, **78(5)**, 1195-1211.

51. Hernandez,M., Ghersi,D., and Sanchez,R. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.*, **37**, W413-W416.

52. Volkamer,A., Kuhn,D., Grombacher,T., Rippmann,F., and Rarey,M. (2012) Combining global and local measures for structure-based druggability predictions. *J Chem Inf Model.*, **52(2)**, 360-372.

53. Huang,B., and Schroeder,M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol.*, **6(1)**, e19.