J Struct Funct Genomics (2007) 8:121–140
DOI 10.1007/s10969-007-9036-1

Structural genomics of protein phosphatases

Steven C. Almo · Jeffrey B. Bonanno · J. Michael Sauder · Spencer Emtage · Teresa P. Dilorenzo · Vladimir Malashkevich · Steven R. Wasserman · S. Swaminathan · Subramaniam Eswaramoorthy · Rakhi Agarwal · Desigan Kumaran · Mahendra Madegowda · Sugadev Ragumani · Yury Patskovsky · Johnjeff Alvarado · Udupi A. Ramagopal · Joana Faber-Barata · Mark R. Chance · Andrej Sali · Andras Fiser · Zhong-yin Zhang · David S. Lawrence · Stephen K. Burley

Received: 31 August 2007 / Accepted: 6 November 2007 / Published online: 5 December 2007

Abstract The New York SGX Research Center for Structural Genomics (NYSGXRC) of the NIGMS Protein Structure Initiative (PSI) has applied its high-throughput X-ray crystallographic structure determination platform to systematic studies of all human protein phosphatases and protein phosphatases from biomedically-relevant pathogens. To date, the NYSGXRC has determined structures of 21 distinct protein phosphatases: 14 from human, 2 from mouse, 2 from the pathogen *Toxoplasma gondii*, 1 from *Trypanosoma brucei*, the parasite responsible for African sleeping sickness, and 2 from the principal mosquito vector of malaria in Africa, *Anopheles gambiae*. These structures provide insights into both normal and pathophysiologic processes, including transcriptional regulation, regulation of major signaling pathways, neural development, and type 1 diabetes. In conjunction with the contributions of other international structural genomics consortia, these efforts promise to provide an unprecedented database and materials repository for structure-guided experimental and computational discovery of inhibitors for all classes of protein phosphatases.

Keywords Structural genomics · Phosphatase · NYSGXRC · X-ray crystallography

S. C. Almo (✉) · J. B. Bonanno · T. P. Dilorenzo ·
V. Malashkevich · Y. Patskovsky · J. Alvarado ·
U. A. Ramagopal · J. Faber-Barata · A. Fiser
Albert Einstein College of Medicine, Bronx, NY, USA
e-mail: almo@aecom.yu.edu

J. M. Sauder · S. Emtage · S. R. Wasserman · S. K. Burley (✉)
SGX Pharmaceuticals, Inc., San Diego, CA, USA
e-mail: sburley@sgxpharma.com

S. Swaminathan · S. Eswaramoorthy · R. Agarwal ·
D. Kumaran · M. Madegowda · S. Ragumani
Brookhaven National Laboratory, Upton, NY, USA

M. R. Chance
Case Western Reserve University, Cleveland, OH, USA

A. Sali
University of California at San Francisco, San Francisco, CA, USA

Z.-y. Zhang
Indiana University School of Medicine, Indianapolis, IN, USA

D. S. Lawrence
University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Introduction

In 2000, the National Institute of General Medical Sciences (NIGMS) established the Protein Structure Initiative (PSI) with the goal to "make the three-dimensional atomic-level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences" to support biological and biomedical research (http://www.nigms.nih.gov/Initiatives/PSI.htm). This initial, pilot phase demonstrated the feasibility of the program, and Phase II of the program, PSI-II, was launched in 2005, supporting four large-scale production centers to continue high throughput structure determination efforts and six specialized centers to focus on specific bottlenecks such as membrane proteins and multi-component assemblies (Table 1). More recently, these experimental efforts were supplemented by addition of two centers focused on enhancing comparative protein structure modeling, a PSI materials repository for centralized archiving and distribution of reagents, and a PSI knowledge base for data sharing (Table 1).

Table 1 NIGMS protein structure initiative centers

Large-Scale Production Centers
Joint Center for Structural Genomics http://www.jcsg.org
Midwest Center for Structural Genomics http://www.mcsg.anl.gov
New York SGX Research Center for Structural Genomics http://www.nysgxrc.org/
Northeast Structural Genomics Consortium http://www.nesg.org
Specialized Technology Development Centers
Accelerated Technologies Center for Gene to 3D Structure http://www.atcg3d.org
Center for Eukaryotic Structural Genomics http://www.uwstructuralgenomics.org
Center for High-Throughput Structural Biology http://www.chtsb.org
Center for Structures of Membrane Proteins http://csmp.ucsf.edu
Integrated Center for Structure and Function Innovation http://techcenter.mbi.ucla.edu/
New York Consortium on Membrane Protein Structure http://www.nycomps.org
Homology Modeling Centers
Joint Center for Molecular Modeling
New Methods for High-Resolution Comparative Modeling
Resource Centers
PSI Materials Repository http://www.hip.harvard.edu/
PSI Knowledgebase http://kb-test.psi-structuralgenomics.org/KB/

Target selection represents a critical first step in the structural genomics pipeline, as it dictates the value of the ensuing structures. PSI-II employs a balanced target selection strategy that continues to emphasize the importance of large-scale structure determination and homology model generation, while exploiting the underlying infrastructure to address significant problems of biomedical relevance and to respond to the needs of the larger research community. About 70% of PSI-II efforts focus on the determination of structures with less than 30% amino acid sequence identity to an existing structure. This constraint is central to the overall goals of the PSI, as it is at approximately this level of sequence identity that homology modeling begins to fail due to difficulties in obtaining accurate primary sequence alignments. About 15% of PSI-II activities are committed to projects nominated by the greater scientific community, with the remaining 15% devoted to a biomedically relevant theme developed by each of the four large-scale centers.

NYSGXRC

The New York SGX Research Center for Structural Genomics (NYSGXRC; www.nysgxrc.org) has established a cost-effective, high-throughput X-ray crystallography platform for de novo determination of protein structures. NYSGXRC member organizations include SGX Pharmaceuticals, Inc. (www.sgxpharma.com), the Albert Einstein College of Medicine (www.aecom.yu.edu), Brookhaven National Laboratory (www.bnl.gov), Case Western Reserve University (www.cwru.edu), and the University of California at San Francisco (www.ucsf.edu). Together, scientists from these industrial and academic organizations support all aspects of PSI-II, including Family Classification and Target Selection, Generation of Protein for Biophysical Analyses, Sample Preparation for Structural Studies, Structure Determination, and Analyses and Dissemination of Results. Current NYSGXRC production metrics during the past 12 months (July 1st 2006–June 30th 2007) are as follows: generation of ~2,060 target protein expression clones, ~1,400 successful target protein purifications (all characterized by Matrix-Assisted Laser Desorption Ionization and ElectroSpray Ionization—Mass Spectrometry, and Analytical Gel Filtration), > 360,000 initial crystallization experiments, > 106,000 crystallization optimization experiments, ~3,100 crystals harvested, > 600 X-ray diffraction datasets recorded, and 158 structures deposited in the Protein Data Bank (PDB; www.pdb.org). We average ~110 successful protein purifications per month and one structure deposition every 2–3 days. As mandated by PSI-II, approximately 15% of NYSGXRC resources are devoted to structure determination of its Biomedical Theme targets, protein phosphatases from human and various pathogens.

Motivation

Protein kinases and phosphatases act in counterpoint to control the phosphorylation states of proteins that regulate virtually every aspect of eukaryotic cell and molecular biology. Protein phosphorylation is a dynamic post-translational modification, which allows for processing and integration of extra- and intra-cellular signals. In vivo, protein kinases and phosphatases play antagonistic roles, controlling phosphorylation of specific protein substrates on tyrosine, serine, and threonine sidechains. These reversible phosphorylation events modulate protein function in various ways, including generation of "docking sites" that direct formation of multi-component protein assemblies, alteration of protein localization, modulation of protein stability, and regulation of enzymatic activity. Such molecular events modulate signal transduction pathways responsible for controlling cell cycle progression, differentiation, cell–cell and cell–substrate interactions, cell motility, the immune response, channel and transporter activities, gene transcription, mRNA translation, and basic metabolism.

Aberrant regulation of protein phosphorylation results in significant perturbations of associated signaling pathways and is directly linked to a wide range of human diseases (see [1] for a recent review). PTEN, a phosphoinositide 3-phosphatase and the first member of the greater protein phosphatase family identified as a tumor suppressor, is inactivated by mutations in several neoplasias, including brain, breast, and prostate cancers. Cdc25A and cdc25B are potential oncogenes. Over-expression of PRL-1 and PRL-2 results in cellular transformation and PRL-3 is implicated as a metastasis factor in colorectal cancer. PTP1B is a primary target for therapeutic intervention in diabetes and obesity. CD45 is a target for graft rejection and autoimmunity. Mutations in EPM2A are responsible for a form of epilepsy, characterized by neurological degeneration and seizures.

The importance of protein phosphatases in mammalian physiology is underscored by strategies found in several pathogens, including *Yersinia*, *Salmonella*, and vaccinia viruses, in which pathogen encoded protein phosphatases disrupt host-signaling pathways and are essential for virulence. Systematic structural analysis of protein phosphatases provides an opportunity to make significant progress towards (i) understanding and treating the underlying mechanisms of human diseases, (ii) treating a wide range of opportunistic and infectious microorganisms, and (iii) generating reagents that permit experimentation to uncover new principles in cellular and molecular biology.

Our progress in this endeavor is shown in Fig. 1, wherein the number of distinct phosphatases that have progressed to each experimental stage is shown. We have observed greater attrition for the pathogen phosphatases, due in part, we believe, to the fact that many of the sequences are gene predictions that have not been experimentally verified. To compare our work on the human versus pathogen proteins, of the 62 human/mouse proteins that we successfully purified, 15 yielded structures, whereas of the 55 pathogen phosphatases that we purified, only 5 have yielded structures thus far.

Families of protein phosphatases

The protein phosphatases encompass a range of structural families, mechanistic strategies and substrate specificities. The protein tyrosine phosphatases (PTPs) represent one of the largest families in the human genome with four distinct subfamilies, including (i) the classic PTPs that recognize phosphotyrosine residues (112 human proteins), which are further divided into several subclasses of receptor-like and intracellular cytosolic PTPs, (ii) the promiscuous dual-specificity phosphatases (DSPs), which recognize both phosphotyrosine and phosphoserine/phosphothreonine (33 human proteins) and include subfamiles of the phosphoinositide phosphatases (PTEN and myotubularin) and the mRNA 5′-triphosphatases (BVP and Mce1), (iii) the low molecular weight phosphatases that recognize phosphotyrosine residues, and (iv) the dual-specificity cdc25 phosphatases.

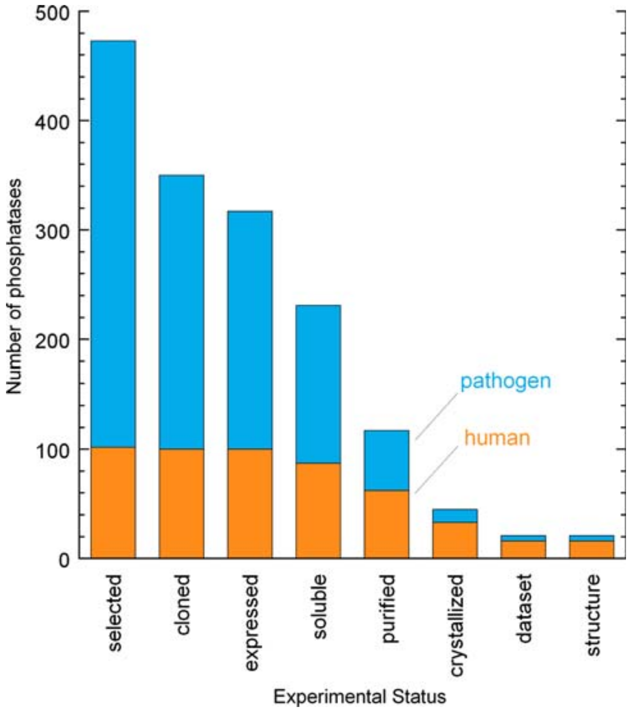


Fig. 1 Progress on protein phosphatase structural studies. Number of protein phosphatase targets at each experimental stage of the NYSGXRC structural genomics pipeline. Human phosphatases (or mammalian orthologs) are shown in orange and pathogen phosphatases are shown in blue

All members of the PTP family catalyze metal-independent dephosphorylation of phospho-amino acids, using a covalent phospho-cysteine intermediate to facilitate hydrolysis. The amino acid sequence hallmark of the PTP family is the *H*C*X*X*G*X*X*R(S/T) motif, which contains the cysteine nucleophile. A sequence alignment showing the family conservation in the 14 amino acids surrounding this motif is shown in Fig. 2a. It is remarkable that this active site feature represents the only amino acid sequence motif that is common to all PTP subfamily members.

The serine/threonine protein phosphatases are represented by two families which are distinguished by sequence homology and catalytic metal ion dependence. The PPP family members are Zn/Fe-dependent enzymes including PP1, PP2A, and PP2B (calcineurin) (~15 human proteins). The PPM or PP2C-like family members are Mn/Mg-dependent enzymes (~16 human proteins). Despite sharing essentially no sequence similarity, members of

Fig. 2 Human phosphatome phylogenetic tree. (**a**) Sequence logo [2] depicting the conservation of active site residues in the protein tyrosine and dual-specificity phosphatases. (**b**) Dendrogram of protein tyrosine (red branch) and dual-specificity (blue branch) phosphatases based on variation in the active site motif. (**c**) Dendrogram of all other human protein phosphatases based on alignment of the entire catalytic domain, including the metal dependent phosphatases (e.g. PPMs and PPPs) and the members of the haloacid dehalogenase (HAD) superfamily (e.g. CTDSPs, EYAs, MDP-1 and PDXP). Phosphatases with structures in the PDB are indicated by black circles

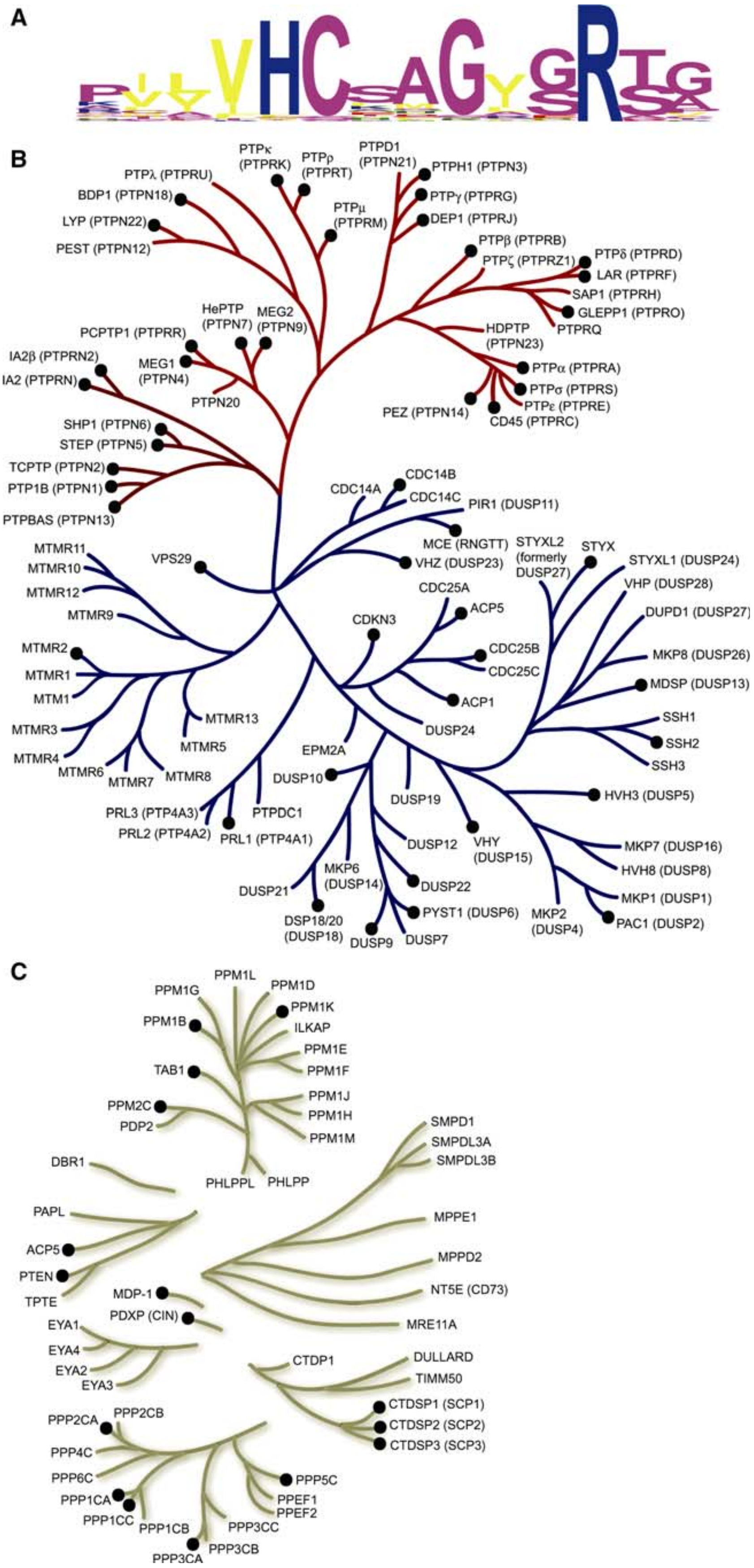


Table 2 NYSGXRC protein phosphatase structures

Gene	Synonym	NYSGX	Species	NCBI	UniProt	Pfam domain	PDB ID
Acp1	ACP1	8663b	Mouse	AAH39744	Q561M1	LMWPc	2p4u
CTDSP2	SCP2	8717a	Human	NP_005721	Q53ZR2	NIF	2q5e
CTDSPL	SCP3	8718a	Human	NP_001008393	Q3ZTU0	NIF	2hhl
DUSP23	LDP3	8673a	Human	NP_060293	Q9BVJ7	DSPc	2img
DUSP28	Dusp28	8736b	Mouse	NP_780327	Q8BTR5	DSPc	2hcm
DUSP9	MKP4	8638a	Human	NP_001386	Q99956	DSPc	2hxp (C290S)
PDXP	CIN	8744a	Human	NP_064711	Q96GD0	Hydrolase	2oyc, 2p27, 2p69
PPM1B	PP2CB	8702a	Human	NP_808907	Q461Q2	PP2C	2p8e
PPM1K	PPM1K	8700a	Human	NP_689755	Q8N3J5	PP2C	2iq1
PTP4A1	PRL1	8648a	Human	NP_003454	Q93096	Y_phosphatase	1rxd
PTPRD	PTPdelta	8613c	Human	NP_002830	P23468	Y_phosphatase	2nv5 rat = human
PTPRG	PTPgamma	8615a	Human	NP_002832	P23470	Y_phosphatase	2pbn, 2hy3
PTPRN	IA2	8620a	Human	NP_002837	Q16849	Y_phosphatase	2i1y
PTPRO	GLEPP1	8635a	Human	NP_109592	A0AV39	Y_phosphatase	2g59
PTPRS	PTPsigma	8623a	Human	NP_002841	Q13332	Y_phosphatase (tandem)	2fh7 (D1–D2)
STYX	STYX	8698a	Human	NP_660294	Q8WUJ0	DSPc	2r0b
Tab1	TAB1	8880z	Mosquito	EAA07598	Q7QD46	PP2C	2irm
PPM1G	Ppm1g	8886z	Mosquito	EAA11252	Q7PP01	PP2C	2i0o (Δ125–398)
PPM1	PPM1	8828z	*T. gondii*	N/A	N/A	PP2C	2isn
PP2C	N/A	8817z	*T. gondii*	CAC86553	Q8WPN9	PP2C	2i44
apaH	N/A	9095b	*T. brucei*	AAX70877	Q57U41	Metallophos	2qjc

both families utilize catalytic mechanisms involving a water nucleophile activated by a bi-nuclear metal center [3]. The haloacid dehalogenase (HAD) superfamily contains a large number of magnesium-dependent phosphohydrolases, which operate through a covalent phospho-aspartic acid intermediate. Recently, a small number of HAD family members have been demonstrated to be protein phosphatases and have been implicated in a range of biological processes [4, 5, 6, 7].

Dendrograms encompassing the recognized human phosphatases are shown in Fig. 2 (experimental 3D public domain structures are denoted therein with black circles). The active site motif (Fig. 2a) compared to a database of human phosphatases was used to construct the PTP/DSP tree (Fig. 2b), whereas a multiple sequence alignment of the catalytic domain sequences was used to characterize homology among the remaining phosphatases (e.g., PPM and PPP families; Fig. 2c). There are over 225 mammalian phosphatase structures in the PDB, providing coverage with either an experimental structure or a high-quality homology model for at least 64 human phosphatases (or ~45% of the human phosphatome). Some human phosphatases have many structural representatives, like PTP1B with over 90 structures, whereas many others have only a single structure in the public domain. Unlike the protein kinases, which can be largely defined by a single Pfam entry, Pkinase (http://pfam.janelia.org/), the phosphatases are more diverse and require at least 12 Pfam entries to describe their varied families (see Table 2 for 7 distinct Pfam entries corresponding to NYSGXRC structures).

NYSGXRC target selection and progress

After the start of PSI-II, the NYSGXRC established a target list of human protein phosphatases for which there was no representative in the PDB. In addition, we selected structurally uncharacterized protein phosphatases from a number of pathogens for which sequence information was available. The coding sequences of most human phosphatases were cloned from cDNA libraries, some were purchased, and 16 were synthesized. All pathogen phosphatases were codon-optimized and synthesized (Codon Devices, Inc., Cambridge, MA). Work on the first group of 93 pathogen phosphatases began at the end of 2005 (*Anopheles gambiae*, *Toxoplasma gondii*, and *Plasmodium falciparum*). In early 2007, work on an additional ~170 pathogen phosphatases was initiated, with targets selected from *Candida albicans*, *Encephalitozoon cuniculi*, *Filobasidiella neoformans*, *Gibberella zeae*, *Cryptosporidium parvum*, *Fusarium graminearum*, *Trichomonas vaginalis*, *Trypanosoma brucei*, *Aspergillus nidulans*, *Cryptococcus neoformans*, *Entamoeba histolytica*, and *Giardia lamblia*.

Two years after the start of PSI-II, we have produced viable expression vectors for 304 phosphatases, and purified crystallization-grade protein for 107 of these NYSGXRC Biomedical Theme targets. We have deposited 24 X-ray crystal structures of 21 distinct protein phosphatases into the PDB (see Table 2 for PDB IDs). Other structural biology groups and structural genomics centers have also determined significant numbers of protein phosphatase structures. Among the most productive are the SGC (Structural Genomics Consortium; http://www.sgc.utoronto.ca) and KRIBB (Korea Research Institute of Bioscience and Biotechnology; https://www.kribb.re.kr/eng/index.asp), which have deposited structures of at least 19 and 7 distinct human phosphatases, respectively, within the past 2 years. There is relatively little overlap of newly deposited structures from competing consortia/efforts.

To help minimize structure overlap among research groups, the NYSGXRC publishes its target list in the TargetDB database (http://targetdb.pdb.org/) on a weekly basis, which includes the experimental status of each target. We compare all of our targets against the contents of the PDB on a weekly basis and typically stop work on those that have been deposited by other groups. We publish experimental protocols for every trial of every target in the PepcDB database (http://pepcdb.pdb.org/); this includes not only detailed general protocols but also information about each clone (DNA sequence and predicted protein sequence, mutations, whether it has been codon-optimized, and small scale expression/solubility results), fermentation (media, volume, induction time and temperature, and resulting pellet weight), protein purification (yield, concentration), purified protein quality as judged by mass spectrometry (pass/fail, exact molecular weight), and crystallization conditions. We encourage this level of transparency for all structural genomics centers. Moreover, we make all of our reagents, such as expression clones, freely available. Sometime in 2008 we anticipate that all NYSGXRC expression clones will be distributed by the centralized PSI Material Repository, located within the Harvard Institute of Proteomics (HIP; http://www.hip.harvard.edu/).

Selected examples

The NYSGXRC Biomedical Theme project has yielded a number of important structures, which have already provided unique insights into a wide range of biological processes with direct relevance to human disease. Illustrative examples are highlighted below.

Protein tyrosine phosphatase sigma (PTPσ)

The 21 human receptor protein tyrosine phosphatases share a common organization with extracellular ligand binding domains, a single transmembrane segment, and intracellular phosphatase catalytic domains that function in concert to regulate signaling through ligand-mediated tyrosine dephosphorylation. Twelve of these receptor PTPs possess two tandem-phosphatase domains with a catalytically active membrane proximal domain (D1) and a membrane distal domain (D2) that is thought to be inactive in most family members. PTPσ belongs to the type 2A sub-family, which possess extracellular ligand binding domains composed of three immunoglobulin-like (Ig) domains and four to nine fibronectin type III-like (FNIII) domains [1, 8, 9]. Additional members of this sub-family include the human leukocyte common antigen-related PTP (LAR) and PTP-delta (PTPδ) and the invertebrate orthologs Dlar and DPTP69D in *Drosophila*, PTP-3 in *Caenorhabditis elegans*, and HmLAR1 and HmLAR2 in *Hirudo medicinalis*. Expression of human receptor PTPs has been detected in all tissues examined, with the majority of PTPσ and PTPδ expression being detected in the brain [9].

PTPσ and other members of the type 2A sub-family play roles in regulating the central and peripheral nervous systems by providing and responding to cues for axon growth and guidance, synaptic function, and nerve repair. These complex functions appear to utilize cell–autonomous and non-cell–autonomous mechanisms, involving signals originating from both the cytoplasmic phosphatase domains and the ligand binding properties of the ectodomain [10]. Using brain lysates from PTPσ-deficient mice, in combination with substrate trapping experiments, N-cadherin and β-catenin were identified as substrates of PTPσ [10]. These findings led to a model of PTPσ-regulated axon growth involving a cadherin/catenin-dependent pathway. In this model, PTPσ directs the dephosphorylation of N-cadherin, which allows for the recruitment of β-catenin. In addition, PTPσ-mediated dephosphorylatin of β-catenin allows for the subsequent linkage to the actin cytoskeleton, resulting in increased adhesion and reduced axon growth. In PTPσ–deficient mice, the resulting hyperphosphorylation of N-cadherin and β-catenin prevents the linkage between the cytoskeleton and the plasma membrane, resulting in reduced adhesion and enhanced axon growth. Further support for this model is provided by observations that dorsal root ganglion axon growth is accelerated in PTPσ-deficient mice. Of particular note, is the enhanced rate of nerve regeneration after trauma (e.g., crush or transection) in PTPσ-deficient mice [11]. In addition to enhanced rates of regeneration, PTPσ-deficient mice show an increased rate of errors in directional nerve growth, suggesting a role in both growth rates and the directional persistence or guidance of advancing neurons. The PTPσ ectodomain has been implicated in non-cell–autonomous functions related to both optimal growth and guidance in regenerating neurons. These contributions to axon growth and regeneration make PTPσ an interesting potential target for therapeutic intervention.

We have determined the structure of the tandem phosphatase domains of human PTPσ (apoPTPσ–D1–D2; PDB ID: 2FH7) (Fig. 3). As observed in the structures of the LAR (type 2A) and CD45 (type 1/6) tandem phosphatases, the D1 and D2 domains of PTPσ are very similar (root-mean-square-deviation or RMSD $\sim$1.0 Å; this and all subsequent RMSD calculations are based on structurally equivalent Cα atoms) [12, 13]. The overall organization of the PTPσ tandem phosphatase domains is very similar to that observed in CD45, LAR, and PTPγ. This similarity is best appreciated by superimposing the D1 domain and examining the distribution of D2 positions (Fig. 4). The relative domain organization is dictated by the residues that contribute to the D1–D2 interface, which are highly conserved among PTPσ, LAR, and PTPγ.

Defining the oligomeric state of the receptor PTPs is central to understanding ligand binding, activation, and underlying regulatory mechanisms. PTPσ–D1–D2 is a monomer in the crystalline state and also behaves as a monomer in solution as judged by analytical gel filtration chromatography (unpublished data). It is remarkable that PTPα and CD45 have been reported to be negatively regulated by homodimerization [14, 15, 16], although this remains an area of intense interest and some controversy [12, 13]. Recently, it has been suggested that PTPσ forms homodimers in the cell and that dimerization is required for ligand binding [17]. The apparent discrepancy between these cell-based results and our biophysical studies may be resolved by demonstrations that dimerization depends, at least in part, on interactions involving the transmembrane segment [17], which is absent from the D1–D2 construct used for our crystallographic and biophysical studies.

Both phosphatase domains in PTPσ possess the characteristic CX$_5$R catalytic site motif and are capable of binding the phosphate analog tungstate (refinement in progress, Fig. 3). However, as observed in most other receptor tyrosine phosphatases, the D2 domain of PTPσ appears to be catalytically inactive [18]. All catalytically active D1 domains possess WPD and KNRY loops, which contain the catalytic acid (D) and participate in phosphotyrosine recognition, respectively. The lack of activity in the PTPσ D2 domain is almost certainly due, at least in part, to the replacement of the WPD Asp with Glu and the KNRY Tyr with Leu (Fig. 5). Similar changes are present in the catalytically defective D2 domains of LAR and PTPα [12, 19], and restoration of the WPD and KNRY sequences in these domains results in a substantial enhancement of catalytic activity. The D2 domain is thought to play an important regulatory role as intermolecular and intramolecular binding interactions between D1 and D2 domains from the same and heterologous PTPs have been shown to modulate the catalytic activity of D1 [14, 18, 20, 21].

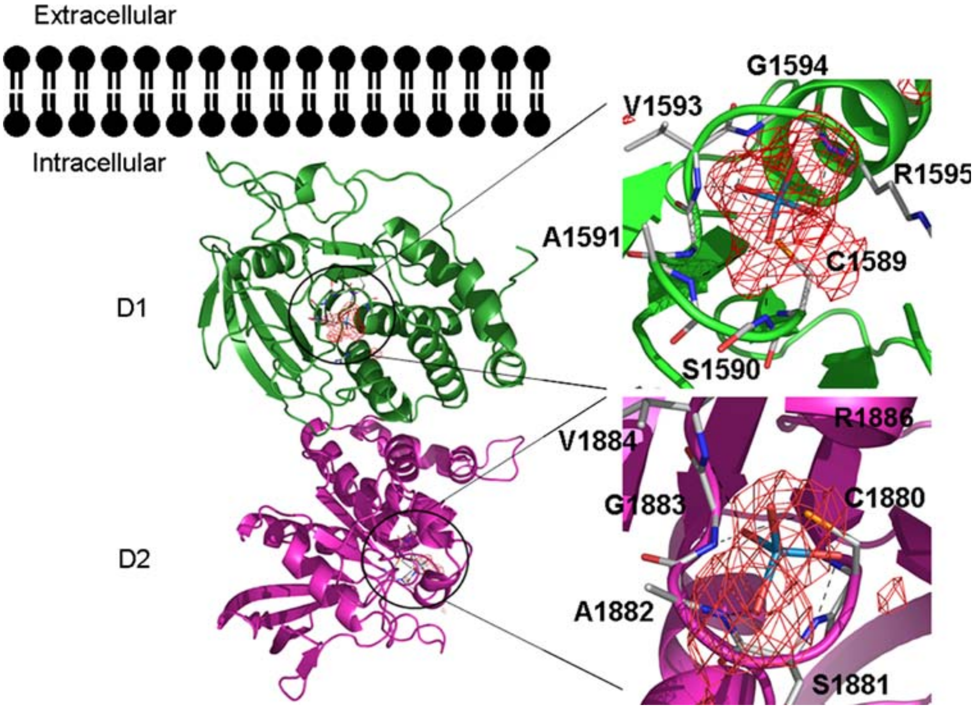
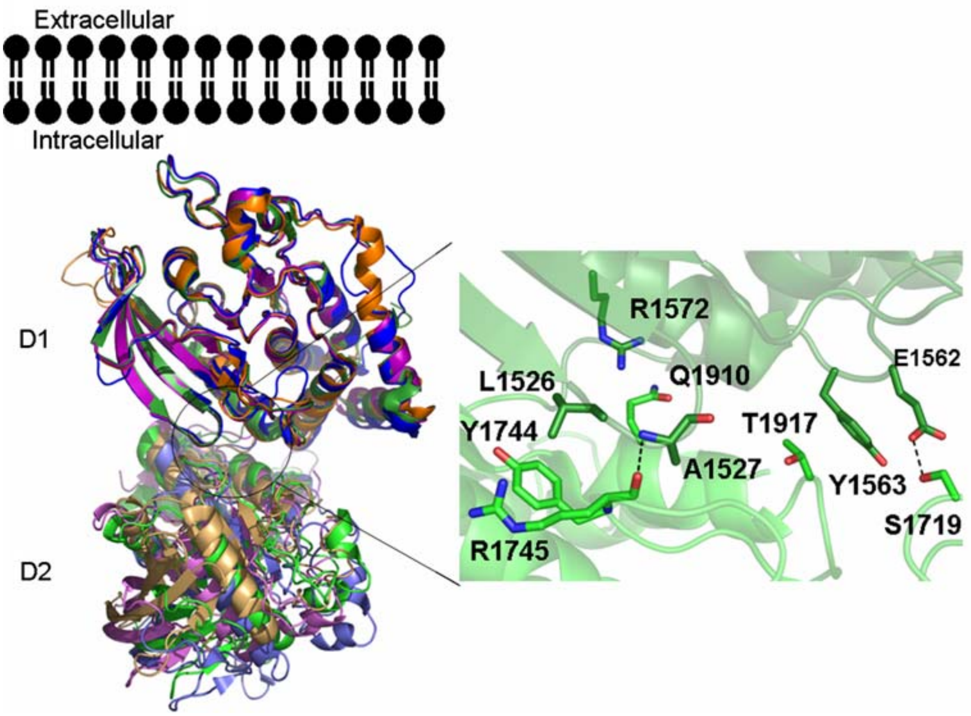


Fig. 3 Structure of the human PTPσ tandem phosphatase domains. The structure of the PTPσ tandem phosphatase domains D1 and D2 is shown as a ribbon diagram with bound tungstate ions as stick and overlapping anomalous difference electron density in red. Domain D1 is shown in dark green and D2 in magenta. Interactions with the tungstate ion in the D1 and D2 active sites are magnified with hydrogen bonds represented as black dashes

Fig. 4 Structural comparison of tandem phosphatase domains of RPTPs. Superposition of the structures of the tandem phosphatase domains of PTPσ (green), LAR (purple), CD45 (blue), and PTPγ (orange). Amino acids involved in interdomain interactions for PTPσ are shown to the right of the D1–D2 structures. For all structures the D1 and D2 domains are shown in dark and light shades of color, respectively



Insulinoma-associated protein 2 (IA-2)

Insulinoma-associated protein 2 (IA-2) is a member of the receptor-type protein tyrosine phosphatase family [22]. It is enriched in the secretory granules of neuroendocrine cells, including pancreatic islet β cells, peptidergic neurons, pituitary cells, and adrenal chromaffin cells [23]. The IA-2 protein is predicted to have a lumenal domain, a single transmembrane helix, and a cytoplasmic tail containing a protein tyrosine phosphatase-like domain. However, enzymatic activity has not been demonstrated for IA-2, and several substitutions within the phosphatase-like domain appear to be responsible for its apparent inactivity against substrates tested thus far [24]. Despite its apparent lack of phosphatase activity, the localization of IA-2 to the membrane of insulin secretory granules suggests that it may be involved in granule trafficking and/or maturation. Indeed, IA-2-deficient mice exhibit defects in glucose-stimulated insulin secretion [25].

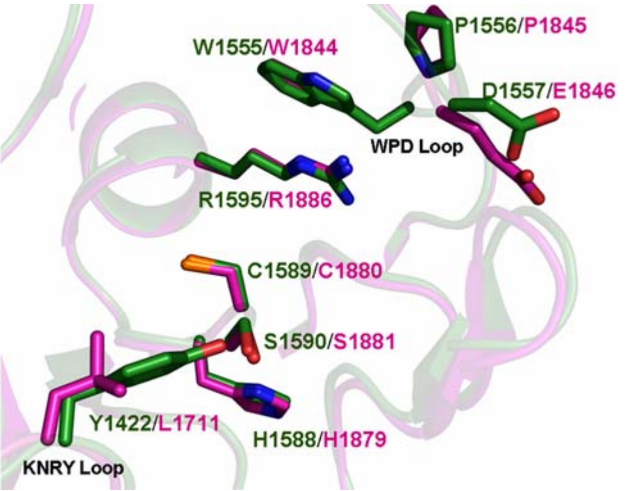


Fig. 5 Comparison of the PTPσ D1 and D2 domain active sites. Superposition of the PTPσ D1 (green) and D2 (magenta) domain active sites. Active site residues and residues making up the WPD and KNRY loops are shown as stick figures. Root-mean-square deviation of D1/D2 superposition for 254 structurally equivalent Cα atoms is ~1.0 Å

IA-2 represents a major autoantigen in type 1 diabetes, with greater than 50% of patients demonstrating circulating antibodies to the protein [26, 27]. Processing of the ~100-kD IA-2 protein involves proteolytic cleavage within the lumenal domain, resulting in a ~64-kD mature form that is immunoprecipitated by insulin-dependent diabetes mellitus patient sera [28]. Autoantibodies to IA-2 can also be detected during the prediabetic period. Measurement of autoantibodies to IA-2, insulin, and glutamic acid decarboxylase (GAD65) enables prediction of type 1 diabetes in at-risk individuals, with the presence of two or more reactivities being highly predictive of future disease [29]. The humoral immune response in diabetes is primarily directed against conformational epitopes located within the cytoplasmic portion of the protein [26, 30, 31]. Distinct B cell epitopes are contained within polypeptide chain segments 605–620, 605–682, 687–979, and 777–937 [26]. As recently reviewed [32], type 1 diabetes patients and at-risk individuals also exhibit $CD4^+$ T cell responses to IA-2 peptides derived from these regions of the protein. $CD8^+$ T cells specific for IA-2 have also recently been reported in type 1 diabetes patients [33].

We have determined the structure of the IA-2 phosphatase-like domain (amino acids 681–979; PDB ID: 2I1Y), which reveals a classic protein tyrosine phosphatase architecture that is most similar to PTP1B (RMSD ~1.4 Å; Fig. 6). This structure also highlights residues responsible for the apparent lack of enzymatic activity. Although the CX_5R sequence is present, several other critical residues are absent, including the catalytic acid (D) in the WPD loop and a major determinant (Y) in the phosphotyrosine-recognition loop. Lack of enzyme activity is thought to be essential for the biological function of IA-2, as it can heterodimerize with both PTPα and PTPε and down-regulate the activity of PTPα [34]. Our structure is particularly noteworthy for defining the relationship between the B cell and T cell epitopes, and thereby providing insights into the pathogenesis of type 1 diabetes. We suggest that T cell response development is facilitated by antibodies to IA-2. The presence of surface antibody would allow B cells to capture fragments of the protein and present IA-2-derived peptides on MHC molecules for recognition by T cells [35]. Antibody-bound protein fragments could also be directed to antigen-presenting cells via Fc receptors, a targeting that allows antigens to be efficiently processed and presented to T cells [36]. A bound antibody can also modulate processing of T cell determinants, suppressing the presentation of some epitopes and enhancing the generation of others, thus influencing the process of epitope spreading [37–39].

Small C-terminal domain phosphatase 3

The small C-terminal domain phosphatases (SCP) comprise a family of Ser/Thr-specific phosphatases that play a central role in mRNA biogenesis via regulation of RNA polymerase II (RNAP II). RNAP II is a complex enzyme containing 12 protein subunits [40], the largest of which bears a unique C-terminal domain (CTD) that is flexibly linked to a region of the macromolecular machine near the RNA exit pore. The CTD consists of multiple repeats of the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 [41], with the number of repeats being species dependent (e.g., 26 are found in yeast versus 52 in human). Reversible phosphorylation of the CTD plays a crucial role in RNAP II progression through the transcription cycle, controlling both transcriptional initiation and elongation [42, 43]. The CTD phosphorylation status also affects RNA processing events, such as 5′-capping and 3′-processing [44–47]. The CTD of RNAP II is predominantly phosphorylated at Ser2 and Ser5 within its heptapeptide repeats by members of the cyclin-dependent kinase (CDK) family (e.g., CDK7, CDK8, and CDK9) [48, 49], while members of the SCP family work in opposition to restore the unphosphorylated

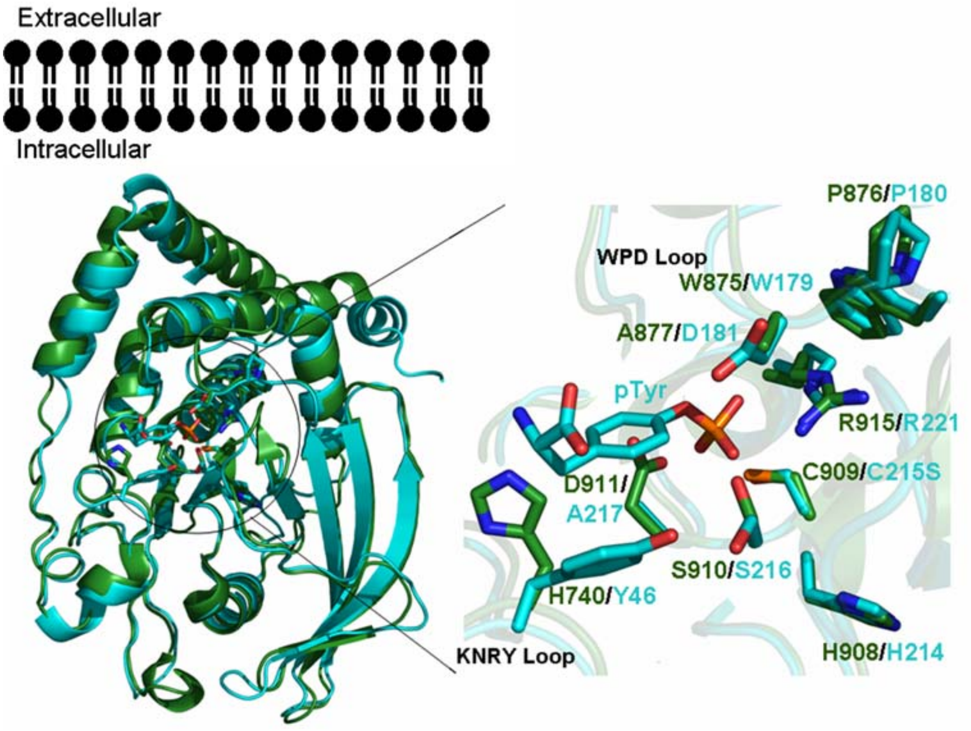


Fig. 6 Comparison of the structures of IA-2 and PTP1B. Superposition of the structures of IA2 (green) and PTP1B (cyan), with active site residues shown as stick figures. Active site residues of IA2 and PTP1B bound to phosphotyrosine are magnified, highlighting differences responsible for the lack of catalytic activity of IA-2

state. It is remarkable that SCP family members also modulate the function of SMAD transcriptional regulators by dephosphorylating SMAD residues in the C-terminus and the linker region connecting two conserved domains [50, 51].

As a consequence of their role in transcriptional regulation, SCP family phosphatases are linked to a wide range of physiological responses and pathologic processes. SCP3 has been identified as a tumor suppressor. The *SCP3* gene is frequently (>90%) deleted or its expression is drastically reduced in lung and other major human carcinomas [52]. In contrast, a related family member, SCP2, was initially identified in a genomic region frequently amplified in sarcomas and brain tumors [53]. Members of the SCP family act as evolutionarily conserved transcriptional regulators that globally silence neuronal genes [54]. SCP2 interacts with the androgen receptor (AR) and appears to control promoter activity by RNAP II clearance during steroid-responsive transcriptional events [55]. FCP1, the first SCP related protein identified, interacts directly with HIV-1 Tat through its non-catalytic domain and is essential for TAT-mediated transcriptional transactivation [56].

To date, the structures of three SCP family members have been determined, including the NYSGXRC structures of SCP2 (PDB ID: 2Q5E) and SCP3 (PDB ID: 2HHL). SCP3 is monomeric both in solution and the crystalline state, and shares high sequence identity (~83%) and significant structural similarity (RMSD ~0.6 Å) with both SCP1 (PDB ID: 2GHQ; [57, 58]) and SCP2. These proteins belong to the haloacid dehalogenase (HAD) superfamily, which encompasses a large number of magnesium-dependent phosphohydrolases characterized by the presence of a conserved DXDX motif (Fig. 7). This signature sequence contributes to the catalytic site, and is responsible for coordination of the catalytically essential magnesium cation, with the first aspartic acid serving as the nucleophile and phoshoryl acceptor [57]. Most HAD family members catalyze phosphoryl transfer reactions involving small molecule metabolites (e.g., phosphoserine). Structures of two HAD protein serine phosphatases from human (PDB ID: 1L8L, [59]) and *Methanococcus jannaschii* (PDB ID: 1F5S; [60]) have also been determined. Despite rather low sequence identity between the SCPs and small molecule phosphatases (14% between 2HHL and 1F5S; RMSD ~2.7 Å), they share significant similarities in overall topology and active site architecture (Fig. 8).

SCP1, SCP2, and SCP3 are composed of a central five-stranded parallel β-sheet flanked by three α-helices on one side and a substantial loop-containing segment on the opposite face. An additional anti-parallel three-stranded β-sheet is formed within the extended loop connecting β-strands 2 and 3 of the central β-sheet. In contrast, the *M. jannaschii* phosphoserine phosphatase (1F5S) contains both the core domain and a capping domain that impinges on the active site. In the case of the tetrameric *Haemophilus influenzae* deoxy-D-mannose-octulosonate 8-phosphatase (PDB IDs: 1K1E and 1J8D), an adjacent monomer packs on top of the core domain and serves as the "cap" (Fig. 8).

As previously discussed by Allen, Dunaway-Mariano, and colleagues [4], the overall architecture of these active sites appear to be related to the type of substrate recognized. For example, the phosphoserine phosphatases, which recognize small molecules, typically possess small catalytic sites that are relatively sequestered from solvent. Such sequestration is generally provided by the additional "capping" domain present in the structure, or by the formation of a higher order oligomeric species that occludes the catalytic site. In contrast, the SCP catalytic sites, which recognize CTD heptad repeats are larger and more accessible to solvent. This architectural variation may be of wider import. For example, in MDP-1, another putative

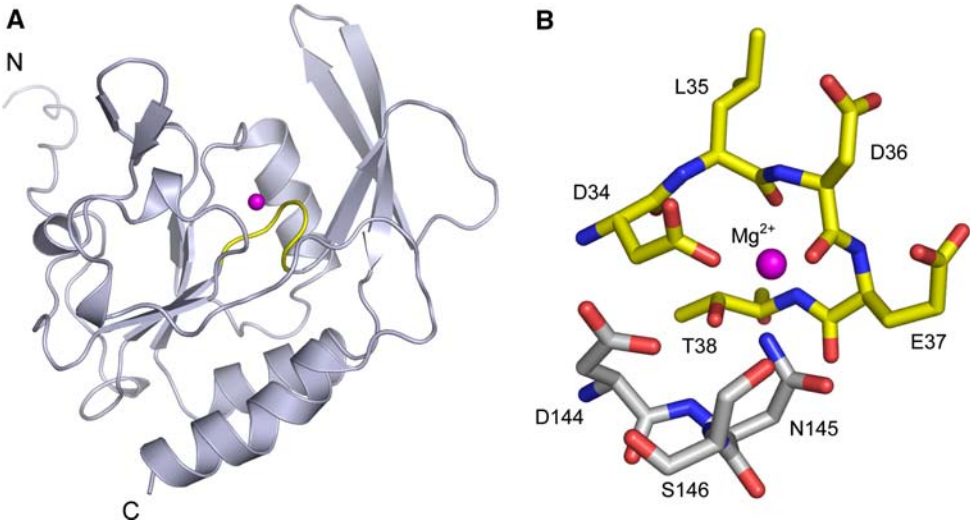


Fig. 7 Structure of SCP3. (**a**) Ribbon diagram of SCP3 showing the DXDX catalytic loop (yellow) and the catalytic Mg^{2+} ion modeled from SCP1 (magenta). (**b**) Atomic details of the SCP3 catalytic site, again with the Mg^{2+} ion modeled from SCP1

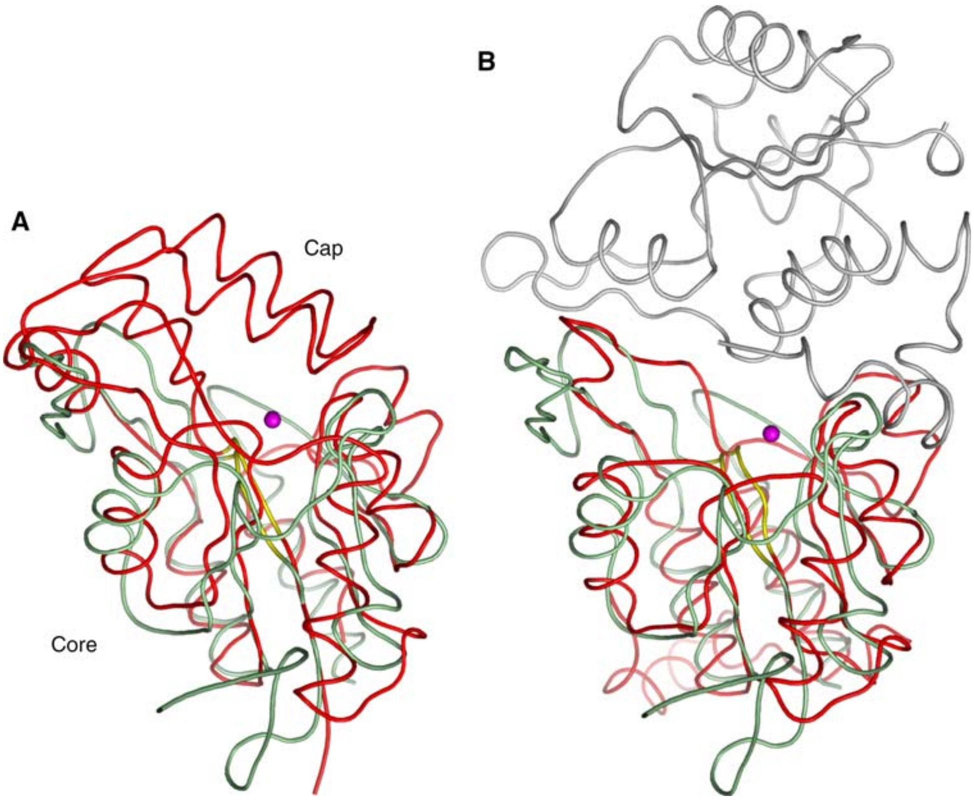


Fig. 8 Structural comparisons of SCP3. (**a**) Superposition of SCP3 (green) with *Methanococcus jannaschii* phosphoserine phosphatase (red, PDB ID: 1F5S). The SCP3 catalytic site is freely accessible to solvent, whereas the alpha-helical capping domain in phosphoserine phosphatase shields its active site. (**b**) Superposition of SCP3 (green) with a dimer of the tetrameric *Haemophilus influenzae* deoxy-D-mannose-oculosonate 8-phosphatase (red and grey, PDB ID: 1K1E). Mg^{2+} ions are shown as pink spheres, and conserved phosphate-binding loops are shown in yellow. The capping domain of 1F5S occludes the active site entrance. In 1K1E, the second subunit of the dimer plays a similar role

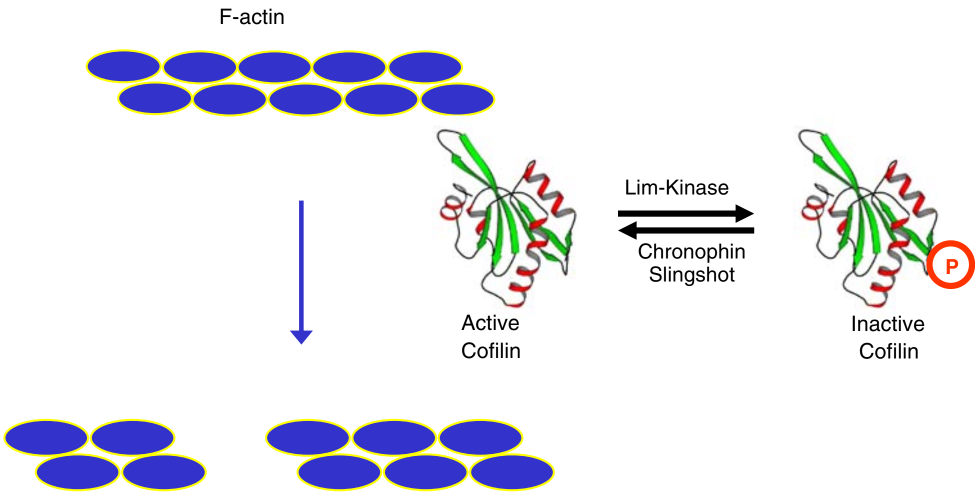
HAD superfamily protein phosphatase that dephosphorylates phosphotyrosine [4], the catalytic site also appears to be highly solvent accessible. Recently published work suggests that MDP-1 might recognize post translationally-encoded protein sugar phosphates, which would also require a substantially more accessible catalytic site [61]. This architectural type is also present in the phosphatase domain of T4 polynucleotide kinase, a HAD superfamily member that utilizes polynucleotide substrates [62]. It is remarkable that all of these polymer-specific HAD phosphatases recognize their substrates and perform catalysis at sites present at termini or linker regions. It, therefore, appears that a combination of catalytic site accessibility and substrate dynamics are required to support the biological activity of these polymer-specific phosphatases.

Chronophin

Chronophin is an example of the burgeoning family of "moonlighting" proteins [63, 64], as it plays a central role in vitamin B6 metabolism and serves as a major regulator of the actin cytoskeleton [5]. This protein was first identified as pyridoxal phosphatase, which specifically dephosphorylates pyridoxal 5′-phosphate (PLP), the coenzymatically active form of vitamin B6 that participates in a remarkable range of enzymatic transformations [65]. PLP synthesis requires a flavin mononucleotide-dependent pyridoxine 5′-phosphate (PNP) oxidase and an ATP-dependent pyridoxal kinase. Degradative pathways for PLP include the action of one or more pyridoxal phosphatases. The overall mechanism by which PLP levels are regulated is dauntingly complex and includes PLP biosynthetic and degradative pathways, PLP binding proteins, and proteins that regulate availability and/or transport of synthetic precursors. Given its central role in PLP metabolism, it is not surprising that the PLP phosphatase is expressed in all human tissues examined and is particularly abundant in brain, suggesting a specialized role in the CNS [66].

More recently, Bokoch and colleagues demonstrated that chronophin plays a direct role in regulating the actin cytoskeleton [67] (Fig. 9). Under physiological conditions, monomeric actin (G-actin) spontaneously polymerizes to form actin filaments (F-actin) with chemically and structurally distinct ends, termed barbed and pointed [68, 69]. Further assembly of F-actin into an array of higher order assemblies underlies all actin-based dynamic processes, including cell motility, vesicle movement, and cytokinesis. In vivo, actin polymerization is dominated by monomer addition to barbed ends, making barbed end generation a central focus in studies of dynamic actin-based processes [69, 70]. Cofilin is a 15 kD protein that severs the actin filament by disrupting the noncovalent bonds between monomers comprising the filament [71]. Filament severing increases the number of polymerization competent barbed ends, and results in increased rates of actin polymerization.

Fig. 9 Cofilin-mediated F-actin severing. F-actin severing activity of cofilin is regulated by a phosphorylation cycle involving the LIM kinase and the slingshot and chronophin phosphatases. Actin monomers are represented by blue ellipses



Cofilin is also thought to be involved in depolymerization, because under certain conditions cofilin not only increases the number of filaments, but can also increase the rate of monomer dissociation from the newly created pointed ends [71]. Cofilin itself is regulated by a phosphorylation cycle: LIM kinase-mediated phosphorylation of Ser3 inactivates cofilin [72] while the action of the slingshot phosphatases return cofilin to the active state [71, 73]. Bokoch's work exploited a wide range of techniques to demonstrate chronophin specificity. siRNA knockdowns of chronophin activity (reduced phosphatase activity) increases phosphocofilin levels, while overexpression of chronophin decreases phosphocofilin levels. Decreased chronophin activity also results in stabilization of F-actin structures in vivo and causes massive defects in cell division [67].

The NYSGXRC structure of human chronophin (PDB ID: 2OYC) confirmed that this protein is indeed a member of the HAD family. It possesses a typical core domain that contains the catalytic signature sequence DXDX. There is also a very substantial capping domain that abuts the catalytic site (Fig. 10). The structure of the PLP-bound form of the enzyme was obtained by substituting Mg^{2+} (PDB ID: 2P27) with the catalytically inert Ca^{2+} (PDB ID: 2P69). This substitution results in a change of metal ligation from six to seven coordinate via bidentate coordination of the catalytic aspartic acid (Asp25), which results in the near complete loss of activity (Fig. 11). Our structure demonstrates that bound PLP is largely buried, with the phosphate being completely shielded from solvent (Figs. 10 and 12). Immediately after our first structure was deposited, three chronophin structures from Kang et al. [74] were made public via the PDB (PDB IDs: 2CFR, 2CFS, and 2CFT).

Demonstration of cofilin phosphatase activity is remarkable given the structural and functional features associated with all previously characterized polymer-specific HAD family phosphatases. As noted above, two members of the SCP family, MDP-1 and T4 polynucleotide kinase, all lack a capping domain, resulting in a relatively open and solvent-accessible catalytic site. In contrast, chronophin appears unique among the polymer-directed

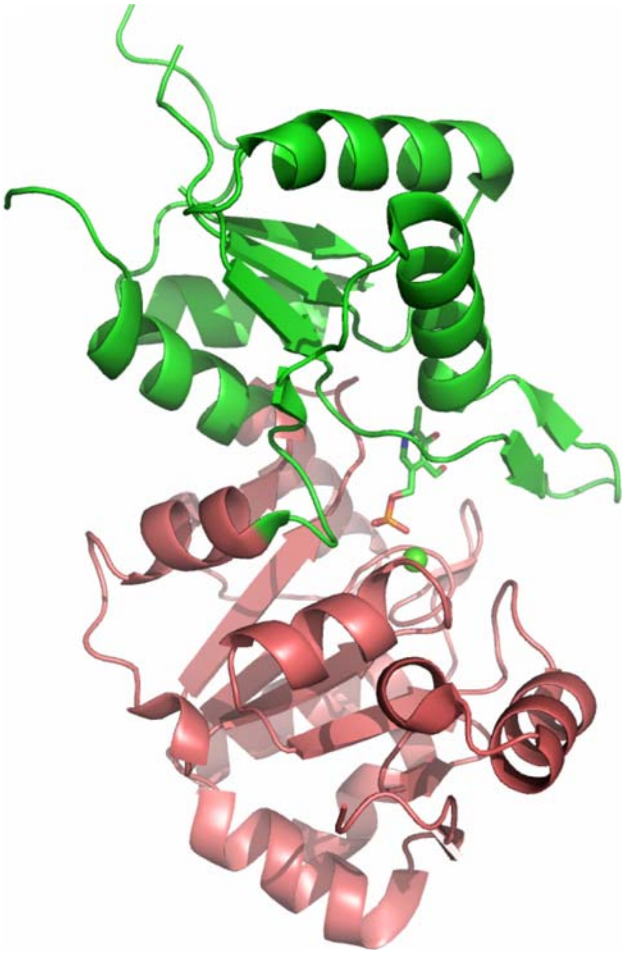


Fig. 10 Chronophin structure. Cartoon representation of chronophin/PLP phosphorylase with bound PLP and Ca^{2+}. The core domain is colored raspberry, the capping domain green, PLP is shown with stick representation and the Ca^{2+} is shown as a green sphere. The catalytic site lies at the interface between the core and capping domains

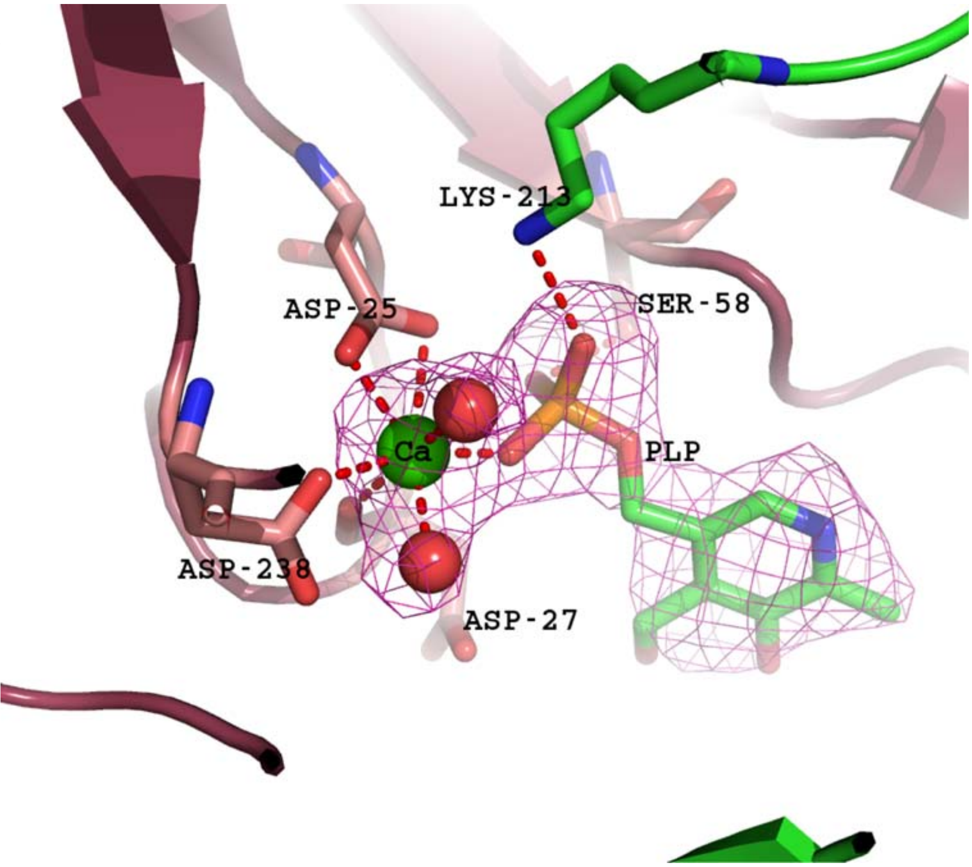


Fig. 11 Chronophin catalytic site. The active site of chronophin with its ligand PLP and inhibitory Ca^{2+}. The Ca^{2+} (green sphere) is hepta-coordinated and participates in a bidentate interaction with the active site nucleophile Asp-25

HAD phosphatases examined to date, because the presence of a substantial capping domain largely occludes the catalytic site (Figs. 12 and 13). Analysis of our chronophin structure suggests that conformational reorganization of the core and capping domains is required to facilitate binding of the phosphorylated N-terminus of cofilin. The N-terminus of cofilin is unstructured in solution [75, 76] and, like the dynamic properties described above for other HAD-associated polymeric substrates, this behavior is almost certainly essential for it to gain access to the chronophin catalytic site. It, therefore, appears that the HAD family has evolved various mechanisms by which to perform chemistry on polymeric substrates. Availability of our structures and protein reagents provides the necessary foundation for a detailed study of chronophin function. These results will be particularly relevant to our fundamental understanding of cancer, as the phosphorylation status of cofilin is directly implicated in the actin-based mechanisms underlying invasion, intravasation, and metastasis of mammary tumors [77].

Other phosphatases

4In addition to the human phosphatase structures described above, the NYSGXRC has determined and made publicly available X-ray structures from *Toxoplasma gondii* (PDB ID: 2ISN), mosquito (TAB1, PDB ID: 2IRM; and PPM1G, PDB ID: 2I0O), and *Trypanosoma brucei* (PDB ID: 2QJC), which will be described elsewhere.

More than a pretty picture

As highlighted by some of the examples described in detail above, a systematic structural characterization of the protein phosphatases provides significant mechanistic and functional insights. Notwithstanding these and other successes, we believe that full exploitation of the growing structural phosphatome database must include efforts to discover new therapeutic phosphatase inhibitors and generate reagents that allow for specific inhibition of signaling pathways in cell culture and whole animal model systems. There is much interest in the pharmaceutical and biotechnology industries in protein phosphatases as drug discovery targets, as evidenced by recent publications on PTP1B and related targets (reviewed in [78]). Below, we briefly highlight two structure-based approaches from academe to the problem of discovering potent phosphatase inhibitors.

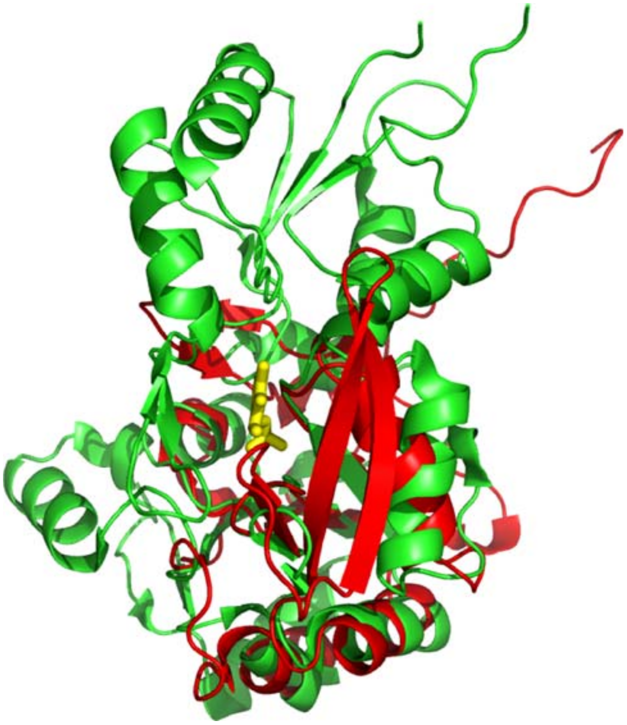


Fig. 12 Chronophin capping domain. Superposition of chronophin (green) and SCP (CTD phosphatase) (red; PDB ID: 2HHL), which lacks a capping domain. The core domains share 11% sequence identity and superimpose with a DALI Z-score of 6.6 and an RMSD of ~2.9 Å for 116 structurally equivalent Cα atoms

Fragment condensation

The concept of inhibitor (or agonist) design via condensation of individual small molecule fragments known to bind proximally in the vicinity of a target protein active site has received growing attention over the past decade [79]. The simple rationale is that the geometrically appropriate covalent linkage of multiple low affinity fragments/functionalities will generate high affinity species. A rigorous thermodynamic treatment of this phenomenon was published by Jencks as early as 1981 [80] and demonstrated that the enhanced affinities of the final species is the consequence of both the additivity of the intrinsic binding energies (ΔG^{i}) of the individual fragments and the favorable entropic contribution associated with the linking of multiple fragments.

Fragment libraries are typically composed of 1,000–10,000 low molecular weight compounds representing a breadth of chemically diverse substructures that possess various chemical functional groups, which can facilitate either target binding or further chemical elaboration. Selection for solubility, lipophilicity, H-bonding donors/acceptors, and known toxicity or ADME (*A*bsorption, *D*istribution, *M*etabolism, and *E*xcretion) properties are typical considerations in building the library.

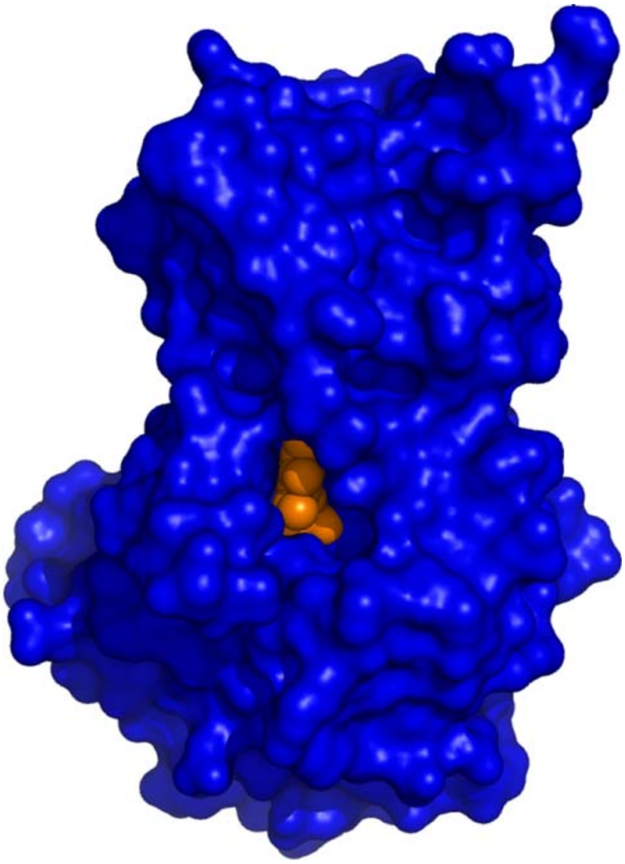


Fig. 13 Inaccessibility of PLP in the Chronophin Catalytic Site. Surface representation of chronophin (blue) with bound ligand PLP (orange) in the same orientation as shown in Fig. 12. The PLP is viewed on the edge and the phosphoryl group is completely buried from solvent

Various approaches to fragment binding detection have been adopted, among them mass spectrometry [81], NMR [82, 83], crystallography [84, 85] and combinations therefrom [86, 87]. Special considerations for fragment based discovery using X-ray crystallography include the necessity for obtaining a well characterized crystal form of the target protein which diffracts well ($d_{min} < 2.5$ Å), possesses a lattice amenable to soaking experiments (i.e., without occlusion of the target site), and is able to withstand exposure to modest quantities of organic solvents (e.g., ethanol and DMSO are popular solvents for chemical fragments). Initial crystal soaking experiments can be conducted on mixtures of the library to speed the initial screening with follow up soaks using individual fragments, where an individual component can not be conclusively identified from the resulting electron density maps. Judicious inclusion of heavy atom containing fragments (Br, I) can provide a significant anomalous scattering signal which can be used for simple deconvolution of fragment mixtures and defining fragment orientation. In addition, the use of halogen-substituted fragments provides a convenient synthetic “handle” for carbon–carbon bond formation.

There are numerous examples in the literature of the success of this method, many coming from small biotechnology companies. Hartshorn et al. [88] have published a useful summary of the initial findings on five very different targets (p38 MAP kinase, CDK2, thrombin, ribonuclease A, and PTP1B). A detailed summary of the approach applied to spleen tyrosine kinase has been published by Blaney et al. [85]. The power of the method is highlighted by the following example for PTP1B, which represents the simplest possible approach to fragment screening with a screening library composed solely of phosphotyrosine.

An analysis of the substrate specificity of PTP1 (the rat ortholog of human PTP1B) revealed that this enzyme catalyzes hydrolysis of a wide variety of low molecular weight phosphate monoesters. One of these substrates (compound **1**; Fig. 14) exhibits poor turnover (6.9 s^{-1}) coupled with an extraordinarily low K_m (16 μM) for a nonpeptidic species. The latter finding suggested that substrate might be binding tightly to the enzyme in a less than optimal fashion with respect to catalysis. Crystallographic analysis of compound **1** bound to a catalytically incompetent C215S mutant form of PTP1B revealed that the substrate can occupy one of two mutually exclusive binding modes: (1) a catalytically competent active site-bound form and (2) a nonproductive peripheral site-bound form (Fig. 15a) [89]. Furthermore, the less sterically demanding phosphotyrosine (“library of one”) was observed to simultaneously bind to both sites (Fig. 15b). Although the active site is highly conserved among all PTPase family members, the peripheral position is not. This latter observation suggested that a bidentate ligand, capable of occupying both positions on the phosphatase, would not only display enhanced affinity but also enhanced selectivity for PTP1B.

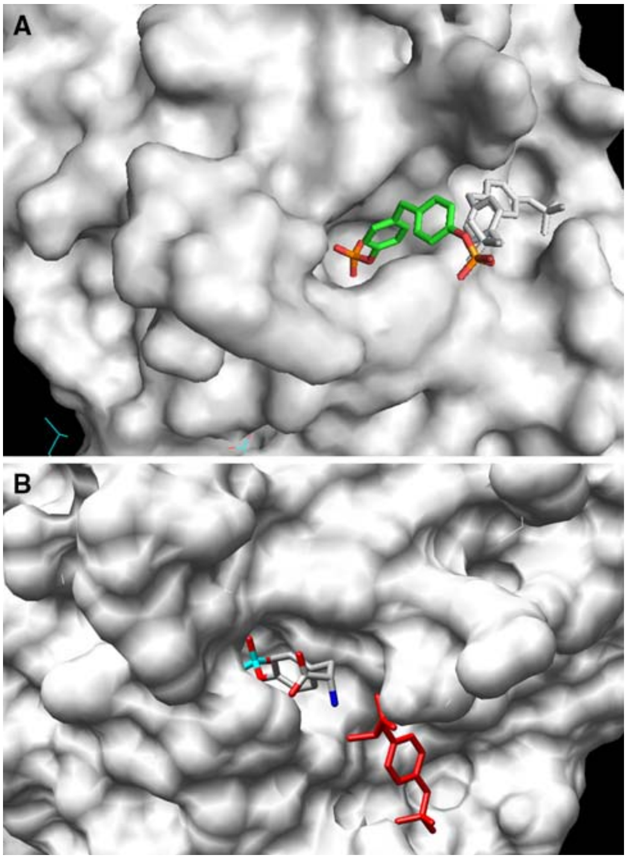


Fig. 15 Surface representation of the PTP1B active site. (**a**) Compound **1** binds to a catalytically incompetent form of PTP1B via one of two mutually incompatible binding modes, which encompass either the active site (multicolor) or a secondary peripheral site (black and white). (**b**) Phosphotyrosine simultaneously binds to the active site (multicolor) and a secondary peripheral site (red)

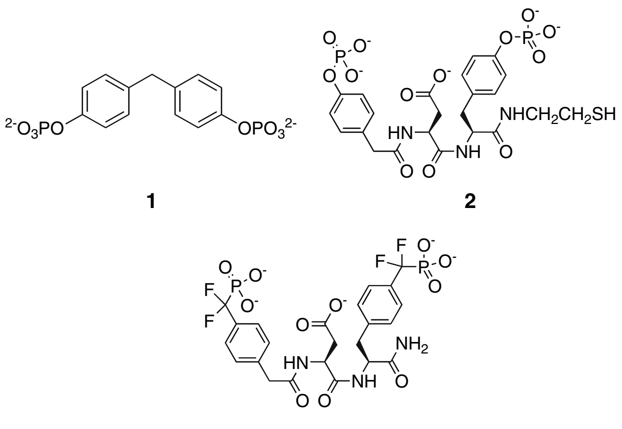

1
2
3

Fig. 14 Compounds that Bind PTP1B. The low K_m, low k_{cat} substrate, compound **1**, compound **2**; and a highly-selective, nonhydrolyzable bidentate inhibitor, compound **3**

Based on these observations, a combinatorial library of 184 compounds was designed containing (1) a fixed phosphotyrosine moiety, (2) eight structurally diverse aromatic acids (terminal elements) to target the unique peripheral site, and (3) twenty-three structurally diverse “linkers” to tether the phosphotyrosine with the array of terminal elements [90]. Compound **2** was identified as the lead species from the library and the corresponding nonhydrolyzable difluorophosphonate (compound **3**) was synthesized. Compound **3** exhibits high affinity ($K_i = 2.4$ nM) and extraordinary selectivity (1,000–10,000-fold versus an array of phosphatases) for PTP1B. Analogs of this compound have recently been shown to serve as insulin sensitizers and mimetics in cell culture as well as appetite suppressors in animal models. These results are consistent with the role of PTP1B as a negative regulator of the insulin and leptin signaling pathways [91].

The structural biology that served as the source of inspiration to prepare the library of bidentate phosphatase inhibitors was based on the surprising finding that PTP1B can simultaneously bind two phosphotyrosine amino acid residues. However, nature had an additional surprise in store with the bidentate ligand compound **3**, designed to coordinate to both the active and peripheral sites. Subsequent crystallographic analysis revealed that although compound **3** binds to the active site in the expected fashion, it does not coordinate to the predicted peripheral site (Fig. 16). Instead, the bisphosphonate compound **3** coordinates in an unanticipated fashion to a completely different secondary site on the surface of the enzyme [92]. This secondary site, however, like the initially identified peripheral site, is not conserved among PTPase family members. The latter appears to provide the structural basis for the extraordinary selectivity displayed by compound **3** and its congeners for PTP1B. This serendipitous finding was the product of structure-guided library-screening and relied on access to purified protein samples; this underscores one of the under-appreciated and under-utilized deliverables of the Protein Structure Initiative: the expression clones and the purified proteins themselves.

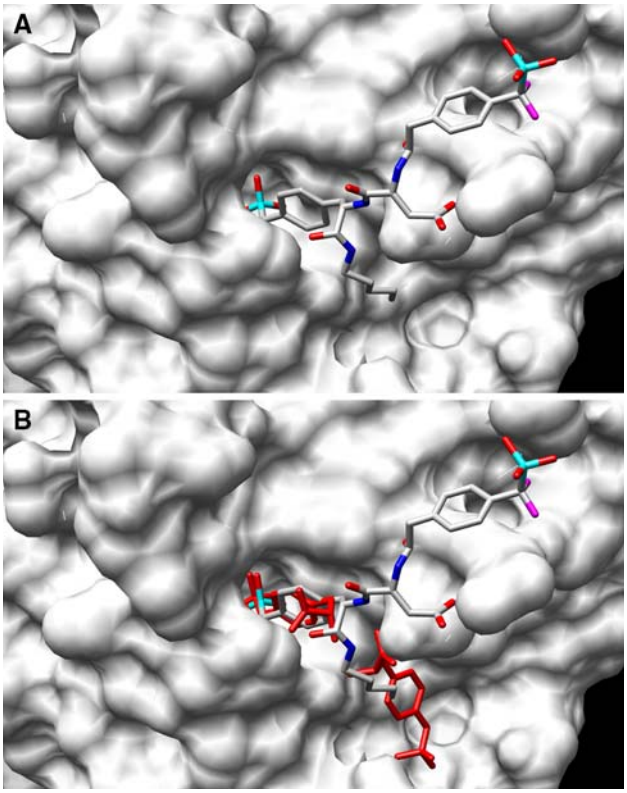


Fig. 16 Structure of the PTP1B/Compound **3** Complex. (**a**) The bidentate ligand, compound **3**, is bound to the active site and a secondary peripheral site different from that observed with compound **1** and phosphotyrosine. (**b**) Overlay of the double binding mode of phosphotyrosine (red) and the bidentate ligand compound **3** (multicolor)

Virtual inhibitor discovery

In silico virtual ligand screening (VS) uses computational approaches to identify small molecules ligands of target macromolecules. The process can be usefully divided into two stages, including docking and scoring. Docking utilizes a structure of the macromolecular target to calculate whether or not a particular compound can fit within a putative binding cleft (e.g., enzyme active site). Scoring methods estimate the free energy of binding for a particular ligand bound to the target in a particular pose, thereby permitting prioritization of predicted target–ligand complexes according to calculated binding energy.

Compound library selection plays an important role in VS. In principle, vast compound sets of diverse properties (including size, lipophilicity, and chemical substructure) are accessible to VS due to the availability of powerful computational resources. For example, Irwin and Shoichet have compiled a freely available database of ~4.6 million commercially available compounds with multiple defined subsets, useful search functions, literature references, chemical similarity cross-references, vendor information, and atomic coordinates (http://blaster.docking.org/zinc/; [93]). Moreover, the National Cancer Institute maintains a repository of over 140,000 compounds together with associated literature and structural information; 1,900 of these are arrayed on multi-well plates intended for use in ligand discovery (http://dtp.nci.nih.gov/index.html). For a review of the growing availability of chemical databases, see [94].

While there certainly exists a benefit to having a source of the actual compound for follow-up in vitro studies, hypothetical small molecule compounds can also be generated in silico and serve as inputs for VS calculations. In practice, electronic compound libraries are processed [95] and filtered for a variety of properties (e.g., Lipinski's rules) either prior to VS ('forward filtering') in order to generate a smaller test set on which more thorough calculations can be undertaken, or after VS ('backward filtering') to eliminate excessive downstream chemistry on candidates with poor chemical properties [96]. Moreover, in contrast to fragment approaches (see above) for which very small molecules are observed to bind specifically, albeit sometimes quite weakly, to the target protein, VS methods encounter difficulty placing fragment molecules unless the binding landscape is limited to a manageably small area of the target protein surface. Thus, for VS, library components are usually larger and somewhat more elaborated compounds.

Selection of an appropriate "druggable" macromolecular target is also a critical concern. The availability of high resolution structural data not only supports VS in general, but allows for the objective assessment, selection, and accurate boundary definition of the protein surface site to be interrogated (e.g., [97, 98]).

A VS campaign against a phosphatase target for which no known non-phosphate based inhibitor was known has been published recently [99]. The authors selected PP2Cα as a target due to its importance in cell cycle and stress response pathways and the availability of a crystal structure (PDB ID: 1A6Q; [100]). While inspection of the enzyme binding site revealed three putative binding pockets, PP2Cα represents a more difficult case for VS due to the presence of two bound metal ions coordinated by six waters leading to ambiguity concerning the relevant target structure. Using AutoDock 3.0 [101], the authors ran docking and scoring calculations against controls (pSer, pThr, and pNPP docked nicely with phosphate groups overlaying a free phosphate observed in the structure) and the NCI Diversity Set (1,990 compounds). This initial screening exercise was followed by a chemical similarity search of the Open NCI Database (currently over 250,000 small molecule structures) resulting in a second generation library, which was also docked against PP2Cα and the resulting hits scored. Additional docking calculations were run on the most attractive looking hits against the apo form of the protein (deleting metals and waters) and all 64 possible permutations of water deletions (while still retaining the active site metals). Perhaps as an indication of binding to pockets surrounding the metals, there was some agreement among the runs with metal/water deletions when compared to the original run. Post-scoring filtering for solubility (compounds with log P < 6 were kept) yielded some leads, which were then requested from the NCI and run in an inhibition assay against PP2Cα and a panel of three additional Ser/Thr phosphatases. Additional runs testing for inhibition due to aggregation by varying enzyme concentration or adding detergent were conducted. Remarkably, at compound concentrations of 100 μM, many of VS hits demonstrated robust inhibitory activity; one compound (109268) showed ~80% inhibition of PP2Cα and reasonable selectivity for PP2Cα and PP1 over PP2A and PP2B (80% versus 40% inhibition) with no appreciable aggregation effects. No further elaboration of these leads was described.

In light of this result, it is of considerable interest that the NYSGXRC has recently determined the structure of a human PP2Cβ (PPM1B) fragment (PDB ID: 2P8E) and the *Toxoplamsa gondii* ortholog of PPM2C (PDB ID:

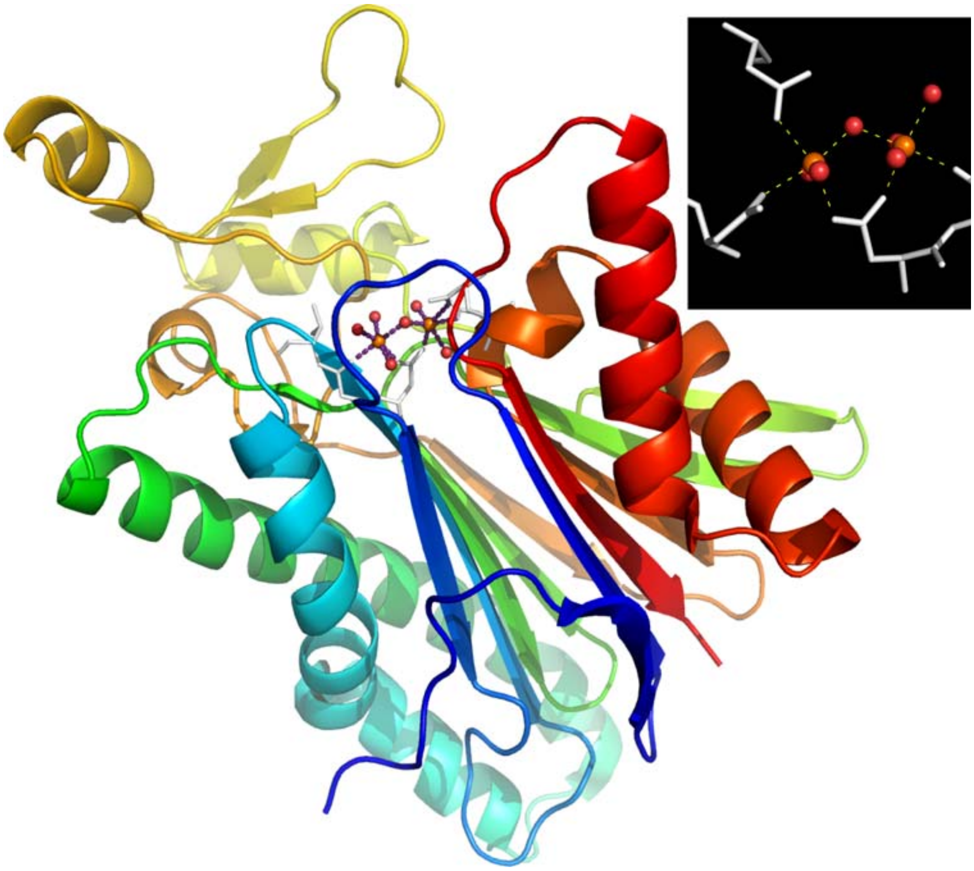


Fig. 17 X-ray structure of *T. gondii* PP2C (PDB ID: 2i44). Inset: two calcium ions supported by conserved aspartate residues and coordinated waters

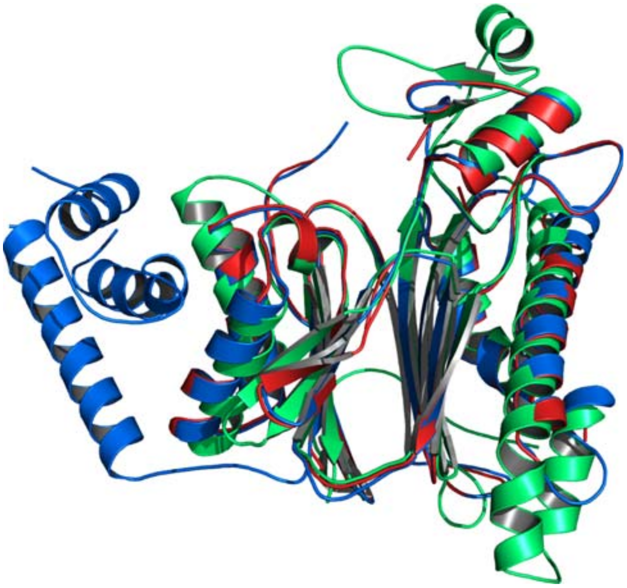


Fig. 18 Structural alignment and overlay of PP2Cs. PP2Ctg (green), PP2Cα (blue), and PP2Cβ (red)

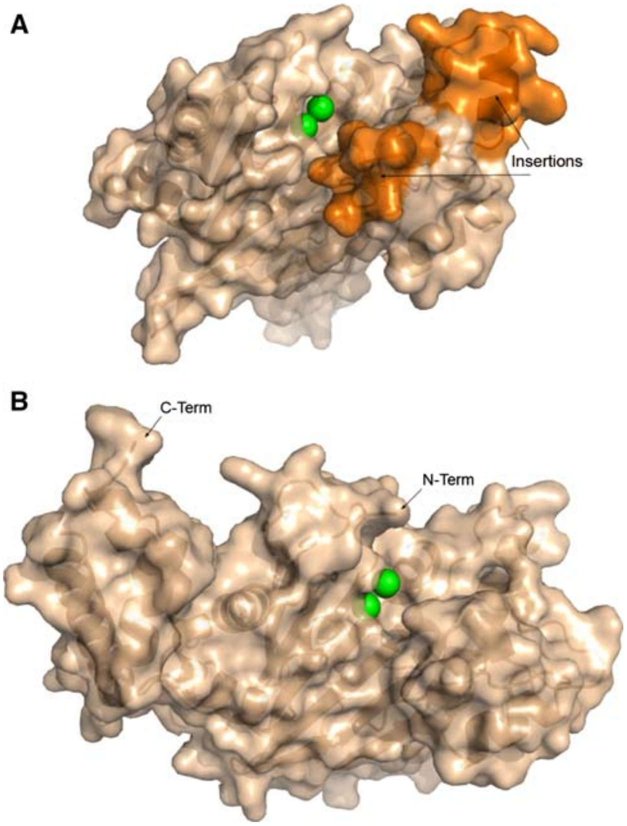


Fig. 19 Surface representations of PP2Cs. (**a**) PP2Ctg: Ca^{2+} shown as green spheres; surface corresponding to amino acid insertions 207–213 and 218–232 colored orange. (**b**) PP2Cα in similar orientation as PP2Ctg: Mn^{2+} shown as green spheres

2I44; herein referred to as PP2Ctg). The structures of PP2Cα and PP2Cβ are nearly identical (RMSD ~0.8 Å; 80% sequence identity), although the PP2Cβ expression construct does not encompass ~90 C-terminal residues, which form a small helical subdomain distal to the active site in PP2Cα. The structure of Ser/Thr phosphatase PP2Ctg (Fig. 17) compares well with human PP2Cα [100] and human PP2Cβ structures. RMSDs calculated by comparing the structure of PP2Ctg to hPP2Cα and hPP2Cβ are ~1.9 Å (23% sequence identity) and ~1.8 Å (25% sequence identity), respectively (Fig. 18). The four-layered αββα architecture, the number of strands, and the location and coordination of the di-metal center are conserved among these structures. Although the core catalytic domain structure is conserved, there are significant differences between the human and the *T. gondii* enzymes. The helices in PP2Ctg are longer than those found in the human enzyme structures in one of the α-layers and PP2Ctg lacks the C-terminal 3 helices present in the human sequences. Importantly, in the vicinity of the active site, PP2Ctg possesses two insertions (residues 207–213 and 218–232 of PP2Ctg) forming an α-helix and a β-sheet, which are not found in the human PP2Cs (Figs. 18 and 19a). Another significant difference affecting the nature of the active site arises from the disposition of the N-terminus. In the human structures, the N-terminus is close to the active site and forms an additional strand along one side of the beta-sandwich (Fig. 19b), whereas in PP2Ctg, the N-terminus folds down away from the metals and forms contacts on either side of the sheets. Several other residues found within the vicinity of the active site and identified as potentially important for ligand binding to human PP2Cα [99] are not conserved: V34→K, E35→H, H62→T, A63→V, and R186→F (PP2Cα numbering). This sequence/structure divergence gives rise to significant variation in the active sites of human versus *T. gondii* PP2Cs (Fig. 19a, b) and may provide the basis for discovery and development of parasite selective phosphatase inhibitors.

Conclusion

The impact of various structural genomics efforts on our structural and functional understanding of the human protein phosphatases and protein phosphatases from a wide range of biomedically-relevant pathogens is already apparent. Current coverage of the human protein phosphatome is ~45%, with the promise of many more structures to come within the next few years. These data will provide insights into both normal and pathophysiologic processes, including transcriptional regulation,

regulation of major signaling pathways, neural development, and type 1 diabetes. They will also help to stimulate and support discovery of specific small molecule inhibitors of phosphatases, for use both in biomedical research and in various clinical settings.

Acknowledgements The NYSGXRC is supported by NIH Grant U54 GM074945 (Principal Investigator: S. K. Burley). We gratefully acknowledge the efforts of all members of the NYSGXRC, past and present. This work was supported by NIH Grant U54 GM074945. Use of the Advanced Photon Source was supported by the U. S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. Use of the SGX Collaborative Access Team (SGX-CAT) beam line facilities at Sector 31 of the Advanced Photon Source was provided by SGX Pharmaceuticals, Inc., who constructed and operates the facility.

References

1. Tonks NK (2006) Nat Rev Mol Cell Biol 7(11):833–846
2. Schneider TD, Stephens RM (1990) Nucl Acids Res 18:6097–6100
3. McCluskey A, Sim AT, Sakoff JA (2002) J Med Chem 45(6):1151–1175
4. Peisach E, Selengut JD, Dunaway-Mariano D, Allen KN (2004) Biochemistry 43(40):12770–12779
5. Wiggan O, Bernstein BW, Bamburg JR (2005) Nat Cell Biol 7(1):8–9
6. Meinhart A, Kamenski T, Hoeppner S, Baumli S, Cramer P (2005) Genes Dev 19(12):1401–1415
7. Jemc J, Rebay I (2007) Annu Rev Biochem (Epub ahead of print)
8. Chagnon MJ, Uetani N, Tremblay ML (2004) Biochem Cell Biol 82:664–675
9. Pulido R, Serra-Pages C, Tang M Streuli M (1995) Proc Natl Acad Sci USA 92:11686–11690
10. Siu R, Fladd C, Rotin D (2007) Mol Cell Biol 27:208–219
11. Sapieha PS, Duplan L, Uetani N, Joly S, Tremblay ML, Kennedy TE, Di Polo A (2005) Mol Cell Neurosci 28:625–635
12. Nam HJ, Poy F, Krueger NX, Saito H, Frederick CA (1999) Cell 97:449–457
13. Nam HJ, Poy F, Saito H, Frederick CA (2005) J Exp Med 201:441–452
14. Jiang G, den Hertog J, Su J, Noel J, Sap J, Hunter T (1999) Nature 401:606–610
15. Tertoolen LG, Blanchetot C, Jiang G, Overvoorde J, Gadella TW Jr, Hunter T, den Hertog J (2001) BMC Cell Biol 2:8
16. Xu Z, Weiss A (2002) Nat Immunol 3:764–771
17. Lee S, Faux C, Nixon J, Alete D, Chilton J, Hawadle M, Stoker AW (2007) Mol Cell Biol 27:1795–1808
18. Wallace MJ, Fladd C, Batt J, Rotin D (1998) Mol Cell Biol 18:2608–2616
19. Buist A, Zhang YL, Keng YF, Wu L, Zhang ZY, den Hertog J (1999) Biochemistry 38:914–922
20. Blanchetot C, den Hertog J (2000) J Biol Chem 275(17):12446–12452
21. Blanchetot C, Tertoolen LG, Overvoorde J, den Hertog J (2002) J Biol Chem 277(49):47263–47269
22. Lan MS, Lu J, Goto Y, Notkins AL (1994) DNA Cell Biol 13:505–514
23. Solimena M, Dirkx R Jr, Hermel JM, Pleasic-Williams S, Shapiro JA, Caron L, Rabin DU (1996) EMBO J 15:2102–2114
24. Magistrelli G, Toma S, Isacchi A (1996) Biochem Biophys Res Commun 227:581–588
25. Saeki K, Zhu M, Kubosaki A, Xie J, Lan MS, Notkins AL (2002) Diabetes 51:1842–1850
26. Lampasona V, Bearzatto M, Genovese S, Bosi E, Ferrari M, Bonifacio E (1996) J Immunol 157:2707–2711
27. Lan MS, Wasserfall C, Maclaren NK, Notkins AL (1996) Proc Natl Acad Sci USA 93:6367–6370
28. Xie H, Deng YJ, Notkins AL, Lan MS (1998) Clin Exp Immunol 113:367–372
29. Verge CF, Gianani R, Kawasaki E, Yu L, Pietropaolo M, Jackson RA, Chase HP, Eisenbarth GS (1996) Diabetes 45:926–933
30. Xie H, Zhang B, Matsumoto Y, Li Q, Notkins AL, Lan MS (1997) J Immunol 159:3662–3667
31. Zhang B, Lan MS, Notkins AL (1997) Diabetes 46:40–43
32. DiLorenzo TP, Peakman M, Roep BO (2007) Clin Exp Immunol 148:1–16
33. Ouyang Q, Standifer NE, Qin H, Gottlieb P, Verchere CB, Nepom GT, Tan R, Panagiotopoulos C (2006) Diabetes 55:3068–3074
34. Gross S, Blanchetot C, Schepens J, Albet S, Lammers R, den Hertog J, Hendriks WJ (2002) Biol Chem 277(50):48139–48145
35. Ke Y, Kapp JA (1996) J Exp Med 184:1179–1184
36. Villinger F, Mayne AE, Bostik P, Mori K, Jensen PE, Ahmed R, Ansari AA (2003) J Virol 77:10–24
37. Davidson HW, Watts C (1989) J Cell Biol 109:85–92
38. Simitsek PD, Campbell DG, Lanzavecchia A, Fairweather N, Watts C (1995) J Exp Med 181:1957–1963
39. Watts C, Lanzavecchia A (1993) J Exp Med 178:1459–1463
40. Sayre MH, Tschochner H, Kornberg RD (1992) J Biol Chem 267:23376–23382
41. Corden JL, Cadena DL, Ahearn JM Jr, Dahmus ME (1985) Proc Natl Acad Sci USA 82:7934–7938
42. Dahmus ME (1996) J Biol Chem 271(32):19009–19012
43. Goodrich JA, Tjian R (1994) Cell 77:145–156
44. Shatkin AJ, Manley JL (2000) Nat Struct Biol 7:838–842
45. Ahn SH, Kim M, Buratowski S (2004) Mol Cell 13:67–76
46. Ho CK, Shuman S (1999) Mol Cell 3:405–411
47. Rodriguez CR, Cho EJ, Keogh MC, Moore CL, Greenleaf AL, Buratowski S (2000) Mol Cell Biol 20:104–112
48. Hengartner CJ, Myer VE, Liao SM, Wilson CJ, Koh SS, Young RA (1998) Mol Cell 2:43–53
49. Zhou M, Halanski MA, Radonovich MF, Kashanchi F, Peng J, Price DH, Brady JN (2000) Mol Cell Biol 20:5077–5086
50. Sapkota G, Knockaert M, Alarcon C, Montalvo E, Brivanlou AH, Massague J (2006) J Biol Chem 281(52):40412–40419
51. Knockaert M, Sapkota G, Alarcon C, Massague J, Brivanlou AH (2006) Proc Natl Acad Sci USA 103:11940–11945
52. Kashuba VI, Li J, Wang F, Senchenko VN, Protopopov A, Malyukova A, Kutsenko AS, Kadyrova E, Zabarovska VI, Muravenko OV, Zelenin AV, Kisselev LL, Kuzmin I, Minna JD, Winberg G, Ernberg I, Braga E, Lerman MI, Klein G, Zabarovsky ER (2004) Proc Natl Acad Sci USA 101:4906–4911
53. Su YA, Lee MM, Hutter CM, Meltzer PS (1997) Oncogene 15:1289–1294
54. Yeo M, Lee SK, Lee B, Ruiz EC, Pfaff SL, Gill GN (2005) Science 307:596–600
55. Thompson J, Lepikhova T, Teixido-Travesa N, Whitehead MA, Palvimo JJ, Janne OA (2006) EMBO J 25:2757–2767
56. Abbott KL, Archambault J, Xiao H, Nguyen BD, Roeder RG, Greenblatt J, Omichinski JG, Legault P (2005) Biochemistry 44:2716–2731
57. Kamenski T, Heilmeier S, Meinhart A, Cramer P (2004) Mol Cell 15:399–407

58. Zhang Y, Kim Y, Genoud N, Gao J, Kelly JW, Pfaff SL, Gill GN, Dixon JE, Noel JP (2006) Mol Cell 24:759–770
59. Kim HY, Heo YS, Kim JH, Park MH, Moon J, Kim E, Kwon D, Yoon J, Shin D, Jeong EJ, Park SY, Lee TG, Jeon YH, Ro S, Cho JM, Hwang KY (2002) J Biol Chem 277:46651–46658
60. Wang W, Kim R, Jancarik J, Yokota H, Kim SH (2001) Structure 9:65–71
61. Fortpied J, Maliekal P, Vertommen D, Van Schaftingen E (2006) J Biol Chem 281(27):18378–18385
62. Galburt EA, Pelletier J, Wilson G, Stoddard BL (2002) Structure 10:1249–1260
63. Jeffery CJ (2003) Ann Med 35(1):28–35
64. Jeffery CJ (1999) Trends Biochem Sci 24(1):8–11
65. Fonda ML, (1992) J Biol Chem 267:15978–15983
66. Jang YM, Kim DW, Kang TC, Won MH, Baek NI, Moon BJ, Choi SY, Kwon OS (2003) J Biol Chem 278(50):50040–50046
67. Gohla A, Birkenfield J, Bokoch GM (2004) Nat Cell Biol 7:21–29
68. Pollard TD, Cooper JA (1986) Annu Rev Biochem 55:987–1035
69. Schafer DA, Cooper JA (1995) Annu Rev Cell Biol 11:497–518
70. Zigmond S (1996) Current Opin Cell Biol 8:66–73
71. Ono S (2007) Int Rev Cytol 258:1–82
72. Scott RW, Olson MF (2007) J Mol Med (Epub ahead of print)
73. Huang TY, DerMardirossian C, Bokoch GM (2006) Curr Opin Cell Biol 18(1):26–31
74. Kang ME, Dahmus ME (1993) J Biol Chem 268:25033–25040
75. Blanchoin L, Robinson RC, Choe S, Pollard TD (2000) J Mol Biol 295(2):203–211
76. Pope BJ, Zierler-Gould KM, Kuhne R, Weeds AG, Ball LJ (2004) J Biol Chem 279(6):4840–4848
77. Wang W, Mouneimne G, Sidani M, Wyckoff J, Chen X, Makris A, Goswami S, Bresnick AR, Condeelis JS (2006) J Cell Biol 173:395–404
78. Pei Z, Liu G, Lubben TH, Szczepankiewicz BG (2004) Curr Pharm Des 10(28):3481–3504
79. Hajduk PJ, Greer J (2007) Nat Rev Drug Discov 6(3):211–219
80. Jencks WP (1981) Proc Natl Acad Sci USA 78:4046–4050
81. Erlanson DA, Wells JA, Braisted AC (2004) Annu Rev Biophys Biomol Struct 33:199–223
82. Shuker SB, Hajduk PJ, Meadows RP, Fesik SW (1996) Science 274(5292):1531–1534
83. Baurin N, Aboul-Ela F, Barril X, Davis B, Drysdale M, Dymock B, Finch H, Fromont C, Richardson C, Simmonite H, Hubbard RE (2004) J Chem Inf Comput Sci 44(6):2157–2166
84. Mooij WT, Hartshorn MJ, Tickle IJ, Sharff AJ, Verdonk ML, Jhoti H (2006) ChemMedChem 1(8):827–838
85. Blaney J, Nienaber V, Burley SK (2006) In: Jahnke W, Erlanson DA (ed) Fragment-based approaches in drug discovery. WILEY-VCH Verlag GmbH and Co., KGaA, Weinheim, pp 215–248
86. Liu G, Xin Z, Pei Z, Hajduk PJ, Abad-Zapatero C, Hutchins CW, Zhao H, Lubben TH, Ballaron SJ, Haasch DL, Kaszubska W, Rondinone CM, Trevillyan JM, Jirousek MR (2003) J Med Chem 46(20):4232–4235
87. Moy FJ, Haraki K, Mobilio D, Walker G, Powers R, Tabei K, Tong H, Siegel MM (2001) Anal Chem 73(3):571–581
88. Hartshorn MJ, Murray CW, Cleasby A, Frederickson M, Tickle IJ, Jhoti H (2005) J Med Chem 48(2):403–413
89. Puius YA, Zhao Y, Sullivan M, Lawrence DS, Almo SC, Zhang ZY (1997) Proc Natl Acad Sci USA 94(25):13420–13425
90. Shen K, Keng YF, Wu L, Guo XL, Lawrence DS, Zhang ZY (2001) J Biol Chem 276(50):47311–47319
91. Taylor SD, Hill B (2004) Expert Opin Investig Drugs 13(3):199–214
92. Sun JP, Fedorov AA, Lee SY, Guo XL, Shen K, Lawrence DS, Almo SC, Zhang ZY (2003) J Biol Chem 278(14):12406–12414
93. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45(1):177–182
94. Baker M (2006) Nat Rev Drug Discov 5:707–708
95. Cummings MD, Gibbs AC, DesJarlais RL (2007) Med Chem 3(1):107–113
96. Klebe G (2006) Drug Discov Today 11(13/14):580–594
97. An J, Totrov M, Abagyan R (2005) Mol Cell Proteomics 4(6):752–761
98. Vajda S, Guarnieri F (2006) Curr Opin Drug Discov Devel 9(3):354–362
99. Rogers JP, Beuscher AE IV, Flajolet M, McAvoy T, Nairn AC, Olson AJ, Greengard P (2006) J Med Chem 49:1658–1667
100. Das AK, Helps NR, Cohen PT, Barford D (1996) EMBO J 15:6798–6809
101. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ (1998) J Comput Chem 19(14):1639–1662