

The scientific impact of the Structural Genomics Consortium: a protein family and ligand-centered approach to medically-relevant human proteins

Opher Gileadi · Stefan Knapp · Wen Hwa Lee · Brian D. Marsden ·
Susanne Müller · Frank H. Niesen · Kathryn L. Kavanagh · Linda J. Ball ·
Frank von Delft · Declan A. Doyle · Udo C. T. Oppermann · Michael Sundström

Received: 20 May 2007 / Accepted: 22 September 2007 / Published online: 12 October 2007
© Springer Science+Business Media B.V. 2007

Abstract As many of the structural genomics centers have ended their first phase of operation, it is a good point to evaluate the scientific impact of this endeavour. The Structural Genomics Consortium (SGC), operating from three centers across the Atlantic, investigates human proteins involved in disease processes and proteins from *Plasmodium falciparum* and related organisms. We present here some of the scientific output of the Oxford node of the SGC, where the target areas include protein kinases, phosphatases, oxidoreductases and other metabolic enzymes, as well as signal transduction proteins. The SGC has aimed to achieve extensive coverage of human gene families with a focus on protein–ligand interactions. The methods employed for effective protein expression, crystallization and structure determination by X-ray crystallography are summarized. In addition to the cumulative impact of accelerated delivery of protein structures, we demonstrate how family coverage, generic screening methodology, and the availability of abundant purified protein samples, allow a level of discovery that is difficult to achieve otherwise. The contribution of NMR to structure determination and protein characterization is

discussed. To make this information available to a wide scientific audience, a new tool for disseminating annotated structural information was created that also represents an interactive platform allowing for a continuous update of the annotation by the scientific community.

Keywords High-throughput · Protein kinase · Dehydrogenase · Reductase · PDZ · 14-3-3 · Binding specificity · Protein crystallography

Introduction

The long-term goal of structural genomics (SG) has been ambitiously defined as “to make three-dimensional atomic level structures of most proteins easily obtainable from knowledge of their corresponding DNA sequences” (<http://www.nigms.nih.gov/Initiatives/PSI.htm>). Long before this goal is achieved, the multiple specialized SG projects are expected to have a significant impact on many aspects of the biological sciences.

The most readily apparent contribution of SG is the rapid expansion in the number of available protein structures, derived at a reduced cost because of the efficiency of specialized centers. Proper target selection is critical to ensure that the structures solved by SG centers are indeed valuable to the research and industrial community, either because of the intrinsic interest of the proteins investigated, or because of the improved mapping of the protein structure universe, providing homologous structural models.

A second important contribution of SG projects for the scientific community is the development of methods for efficient protein production and structure determination, which could be adopted in smaller research laboratories to improve productivity.

O. Gileadi (✉) · S. Knapp · W. H. Lee ·
B. D. Marsden · S. Müller · F. H. Niesen ·
K. L. Kavanagh · L. J. Ball · F. von Delft ·
D. A. Doyle · U. C. T. Oppermann · M. Sundström
Structural Genomics Consortium, Botnar Research Centre,
University of Oxford, Oxford OX3 7LD, UK
e-mail: opher.gileadi@sgc.ox.ac.uk

Present Address:

L. J. Ball
F.M.P., Campus Berlin-Buch, Robert-Roessle-Str. 10, 13125
Berlin, Germany

Other scientific deliverables of structural genomics derive from the scale and nature of the operations, and include comparative studies on members of protein families, identifying determinants of specificity, deriving general rules, and improving the capability to predict protein structure and function from gene sequences.

The Structural Genomics Consortium (SGC), operating in the Universities of Oxford and Toronto and the Karolinska Institute, was initiated in 2003 to address needs of industrial and academic pharmaceutical research. The SGC investigates human and apicomplexan proteins; the targets are selected based on their potential as drug targets or involvement in disease processes. Technologically, the SGC focuses on interaction of proteins with small molecules (ligands, inhibitors, substrates and co-factors), and on coverage of protein families. This report provides several examples of the impact of research undertaken at the Oxford node of the SGC, including methodology for high-throughput structure determination, generic means for ligand screening, selected examples of insight from specific structures, insights from family coverage, and the possibilities resulting from the availability of large numbers of purified protein samples. The other SGC nodes share the core technologies but investigate non-overlapping target areas.

Finally, the scientific impact depends on dissemination of structural data. We describe a new platform for

distribution of annotated protein structures, which aims at making this data more meaningful to an audience beyond the usual users of the PDB.

Methodology

Protein production

Method adaptation and development for structural genomics involved a change of mindset, no less than developments in instrumentation, chemistry and computer software. Industrialization of protein production—applied to a huge variety of proteins with very divergent chemical properties—is not straightforward. Yet, extensive work in several SG centres have led to a convergence to core procedures, which are widely applicable, and often sufficient to generate purified proteins, crystals and structures (Table 1). Where the core protocol fails, additional steps (e.g., further purification, crystal optimization), or alternative methods (e.g., different cloning vectors) are applied.

Several features of this protocol have been optimized to capture a large portion of target proteins. Gene clones have been predominantly obtained from public and commercial cDNA libraries. However, gene synthesis may become the method of choice, allowing to optimize codon frequency,

Table 1 Core protocols employed at the SGC

1. Source of DNA	1. Sequence-verified cDNA clone collections. 2. Synthetic DNA. 3. RT-PCR, site-directed mutagenesis. 4. Genomic (microbial).
2. Cloning	Ligation-independent cloning. Recombinase-based cloning (e.g., Gateway, InFusion).
3. Expression vectors and hosts	T7 promoters, controlled by Lac repressor.
Bacterial vectors	N-terminal hexahistidine tag, cleavable by specific proteases (TEV, Thrombin, C3). Host strains based on BL21(DE3), often expressing rare-codon tRNAs or chaperone proteins.
4. Eukaryotic expression	Baculovirus-infected insect cells.
5. Protein expression	Rich media, grow at 37°C to mid-log, then induce at low temperature with IPTG. OR: Similar protocol using minimal medium for Selenomethionine or isotopic labelling.
6. Purification	Two-step purification: Affinity chromatography, Gel filtration, all in high-salt buffers (0.5 M NaCl). Optional: tag cleavage and re-purification.
7. Ligand and buffer screening	Thermal denaturation assays are used to screen purified proteins against 1–10 ³ small molecules and several buffer compositions, to identify stabilizing conditions and potential ligands.
8. Crystalliation	Initial coarse screens (2–4 × 96 conditions; 3 protein concentrations each). Vapour diffusion, sitting drops, imaged by robots but scoring done by humans. Include ligands identified from screening or biochemical knowledge to promote crystallization. Follow-up screens and crystal optimization.
9. Data collection and structure determination	Manual or robotic screening of crystals for diffraction properties; data collection in rotating anode or synchrotron sources. Phasing: Molecular replacement (95%), experimental phasing using SeMet derivatives, and MIR.

restriction sites, and mRNA structure and to introduce site-directed mutations. Ligation-independent cloning is a generic, high-throughput process that can be uniformly applied regardless of the target gene or the cloning vector. Short N-terminal fusion tags, including a hexahistidine sequence and a specific protease cleavage site, are almost universally used. It has been widely documented, that larger fusion tags (e.g., GST, thioredoxin, MBP) can enhance solubility of proteins that are not soluble when expressed with a short peptide tag. However, such fusion proteins have not been widely used in the SGC, since removal of the tag often leads to loss of solubility.

The standard purification protocol is designed to be widely applicable, and experience has shown that it results in effective purification of a large fraction of proteins solubly expressed in *E. coli*. A protein presented for crystallization must be homogeneous in composition, post-translational modification and oligomeric state; the presence of protein aggregates may be especially detrimental to subsequent crystallization. Affinity purification of highly-expressed proteins eliminates most other proteins, while gel filtration effectively separates different oligomeric forms of the protein and removes protein aggregates, which may otherwise promote irreversible aggregation of the protein preparation. The use of high salt concentration (typically, 0.5 M NaCl) throughout the purification process seems to reduce protein aggregation and non-specific binding of protein contaminants. Tag cleavage followed by another passage through the affinity column provides a further generic and highly effective purification step, which removes other proteins that bind adventitiously to the first affinity column. The generic purification procedure has provided in the majority of cases protein of sufficient purity to achieve crystallization. In most other cases, the generic procedure could be followed by polishing and protein modification steps to achieve homogeneous preparations.

The greatest barrier to production of human proteins in bacteria is recovery of soluble protein. Less than 15% of protein targets yielded detectable levels of soluble protein when tested as full-length constructs in the SGC, while more than 80% were expressed as insoluble aggregates. The key to achieving higher success rates has been the parallel production of large numbers of truncated constructs, often containing a compact protein domain. Construct design is initially based on domain boundary analysis, using a number of bioinformatic tools; 3–4 endpoints are designated around each of the predicted termini of the domain, resulting in 9–16 constructs. We have consistently found that this approach results in a 4-fold increase in the number of targets that can be produced as soluble proteins; a similar impact has been seen on the production of diffracting crystals, which can be dramatically affected by minute changes in protein termini.

Although not rigorously tested, it is presumed that a protein construct that is inherently well-behaved (little tendency to aggregate or denature) will be less dependent on specialized conditions for expression and purification, and may crystallize in a wider range of conditions.

Crystallization, crystal screening and data collection

For successful crystallization of a given target, the SGC's phase I operation appears to have confirmed that the most important driver for success is to explore protein diversity at the crystallization stage. One major form of variation was discussed above, namely testing multiple constructs of the target. Equally effective has been setting up co-crystallization with multiple ligands, along with varying protein concentration in the primary crystallization screens.

At the same time, it appears not to be vital to explore chemical space extensively for any given protein preparation; instead, the primary goal of the initial (coarse) screen can be to identify which preparations are “crystallizable”, and a limited set of coarse screen conditions (~200) generally seems sufficient. Practically, this requires only two 96-well crystallization plates, and by setting up three drops per condition, at different protein-well ratios (in Greiner 3-drop plates), the protein concentration is simultaneously varied. The conditions themselves are derived from those found to be most successful in other high-throughput initiatives [1–3], although according to this “crystallizability” philosophy, the exact composition is probably not important. Naturally, coarse screens do not always yield high-quality crystals that can produce a dataset; however, the SGC operation does not rely on these crystals showing up in coarse screens, and a good optimization infrastructure is in place.

In practice, this diversity exploration leads to large numbers of parallel crystallization experiments, presenting a logistical challenge which, at this scale, can only be met with an efficient robotics and IT infrastructure. For the automation, the SGC has been able to exploit the devices developed on the back of the first wave of structural genomics initiatives, and our investment has been less in developing the machines, than in integrating them and implementing experimental best practices. Particular examples: by minimizing sample requirements with nanolitre crystallization, the available protein can be used in more experiments. The large numbers of drops thereby produced (1.5 million/year) would be practically impossible to view by eye under the microscope, whereas automatic drop imaging on a fixed schedule allows images to be reviewed at leisure at the desk.

Automation has also played an important role in crystal characterization. An automatic sample changer has been

used for initial characterization of diffraction quality of a vast number of crystals. This allows to rank the crystals for more careful data collection, especially at the synchrotron, and to direct further efforts at crystal optimization.

A significant saver of upstream efforts has been to exploit each crystal's diffraction as efficiently as possible, even those traditionally considered to be marginal or problematic. Marginal diffractors would include crystals that are "very small" (<40 μm in longest dimension), twinned, or have streaky or anisotropic diffraction. The latter cases generally require the undivided attention of experienced crystallographers.

Small crystals require an excellent X-ray beam: the PXII beamline of the Swiss Light Source synchrotron provides a beam which is reliably small but also well-aligned and very stable. Most efficient use of the beamline relied on pre-screening all crystals at the laboratory source for thorough work prioritization; real-time data processing during data collection; and close attention to radiation damage of crystals. It has been crucial to have experienced crystallographers on site. Adherence to these good practices has been highly productive: of datasets collected on 24-hour trips to SLS, 66% were used for final structures, while 90% of all depositions relied on synchrotron data. The ability to extract useful data from marginal crystals has been especially productive in combination with the protein/ligand diversity approach of the SGC, as a significant fraction of structures (>50%) could be derived from crystals emerging from the primary screens, saving the need for further optimization.

Phasing and structure solution

Due to the family-based approach, for most SGC targets a homologous structure is already known, and most structures (>95%) can be phased by molecular replacement (MR). While this saves significant experimental efforts upstream compared to experimental phasing, by eliminating the need for selenomethionine-derived protein or heavy atom soaks, we find this does not actually save time overall, because starting phases from MR are heavily phase biased. Removing the bias has required many iterations of careful and incremental model building and refinement by experienced crystallographers who can see the danger signs of a poorly-refined model, and know how to deal with it [4, 5].

The final step, namely finalizing and depositing the model, is in fact a frequent stalling point, not only in high-throughput contexts. The reason is that the final model is not merely a result that can be trivially read off a few measurements, but instead is an interpretation of often

rather noisy data, with a lot of detail that is easy to miss, where individual errors influence the clarity in all areas. Moreover, poor model definition affects biologically interesting parts of a structure, and interpreting it becomes a matter of judgment and using in orthogonal information. Indeed, the "final" model is as much scientific hypothesis as result, and depositing the model means signing off on the hypothesis—which is why it has traditionally been a bottleneck in structural genomics efforts.

The SGC has used a peer proofreading system combined with strict timelines to counteract the problem: before deposition, the structure is reviewed by another crystallographer for errors or alternative interpretations, and comments passed back to the original refiner. The intention is threefold: First, to introduce quality control on the final output. Second, the refiner does not feel compelled to spend excessive time on the model to flush out the final errors, since she knows it will be checked. Third, by mixing up refiners and proofreaders, over time this should lead to common interpretations of marginal modeling decisions. The timelines depend on situation and difficulty, but typically allow two weeks for refinement, a day for proofreading, and two further days for deposition.

This approach has made it possible to deposit novel structures at a considerable rate (6 each month from a team of 6 dedicated and 4–5 occasional crystallographers) without compromising quality.

Information infrastructure

An efficient laboratory information management system (LIMS) has been vital to manage not only target tracking, but also capturing and integrating where possible information generated from robotics, as well as capturing human assessments of experimental outcomes, where these could be entered via a client (e.g., scoring of crystallization images).

Fortuitously, the solution we settled on, BeeHive from Molsoft (<http://www.molsoft.com/bee-hive.html>), is in essence an extremely intuitive database query tool that enables even inexperienced users to extract information relevant to their current work—including the simplification of data entry. This is a weak point of many LIMS solutions, whose focus often evolves around data entry but have very inflexible retrieval mechanisms. This has proved to be a powerful means of communication between all persons involved in a project, allowing immediate and error-free retrieval of "hard" information (e.g., protein sequence, ligand and buffer conditions and project history), as well as evaluation and prioritization of crystals and of concurrent projects.

Protein characterization and ligand screening

One of the major challenges in structural genomics is identifying the function and evaluating the functional integrity of the proteins. Examining the physical state of a protein—by methods such as analytical ultracentrifugation, chromatography or dynamic light scattering—is valuable in assessing the prospects for crystallization. In contrast, specific activity assays need to be tailored for each protein class, and may be impractical or impossible when the activity of the protein is not known. We have implemented a generic screen, based on the increase in thermal stability of a protein upon ligand binding. The fluorescent readout is based on monitoring of protein unfolding using a hydrophobicity-sensing dye. Differential Scanning Fluorimetry (DSF) assays [6–9] are ideal for screening a large number of compounds for binding to each target protein. Significantly, the shift in T_m (the unfolding transition midpoint) measured by this method is comparable to measurements obtained by differential scanning calorimetry (DSC), the well-established standard method for thermal shift measurements. In selected cases, a direct correlation between T_m shift and binding constants has been observed [8, 10].

Several advantages have been derived from this capability: First, the identification of relatively strong interacting molecules out of several hundreds of candidates. As detailed below, the compounds discovered in this manner are then included in crystallization experiments; in many cases, only protein–ligand complexes yielded diffracting crystals. Secondly, the reactivity profiles provide data on binding selectivity of the protein active site, which is the most crucial information for drug design; we have often followed up the results from ligand screens by analyzing the structures of several protein–ligand complexes. In parallel, the properties of the protein–ligand interactions are studied by biophysical methods and by enzyme inhibition studies. Third, such screens have allowed us to identify ligands or substrates of proteins with unknown function (sometimes termed “de-orphanizing”). Finally, DSF-based screens can be expanded to explore other conditions, such as buffer composition that enhance the stability of a protein. These conditions may then be introduced to improve the outcome of protein purification and crystallization [8].

The limited scale of protein production and other limitations on resources do not allow a full-scale screen as done in the pharmaceutical industry (10^5 compounds). Rather, we have assembled smaller family-specific compound libraries (10 – 10^3 compounds each), which can reasonably be tested against available amounts of protein (~ 200 μg for 100 assays). The compound libraries are based on the scientific and patent literature; the chemical structure of prospective compounds is used to search an in-house

compilation of vendor databases to identify potential sources. Acquisition of desired compounds is not trivial: not all published compounds, even those appearing in vendor catalogues, are actually available when required; alternative vendors, or collaborative sources may then be accessed. With continuous updating based on current literature and our own experimental results, these libraries have allowed to derive binding profiles and new insights on ligand specificity.

SGC target and biology area selection: relevance for the treatment of human diseases

For any structural genomic organisation target selection is an important consideration as it can have a major impact on the procedures that are implemented during the process of structure determination. There are a number of approaches applied by different structural genomics projects to select targets for structural analysis such as blanket coverage of an organism’s genome, targets with potential novel folds, percentage cut off based on sequence identity or total coverage of selected protein families. The SGC has opted for the family-based approach with an emphasis on protein families whose members are important in human health, disease and are potentially druggable. From our point of view, the main advantages of this approach are 2-fold. Firstly, the methods and procedures identified for one family member can be applied to another family member improving everything from expression, solubility, stability, and purification, to crystallisation and structure determination. Secondly, analysis of the structures from all family members can reveal additional significant information such as ligand binding site specificity, conformational dynamics, understanding of aberrant behaviour of specific family members or the converse revealing common structural properties within all family members.

The availability of high resolution structures constitutes the foundation for structure-guided drug discovery projects. In recent years SG has significantly increased the number of human protein structures available for structure-based design projects [11]. In particular, protein family focused efforts originating from high-throughput structural biology projects have contributed to the structural description of a number of members from human protein families and thus provided valuable structural and chemical information for the design of bioactive compounds. In addition, established expression and crystallization conditions have been used to generate essential reagents, methodologies and technologies which have facilitated research projects in academia and drug discovery programs in industry.

The SGC has focused on providing protein structures to support drug development and understanding of the

structural determinants for human disease. Of 160 unique targets deposited by the SGC (in phase 1), clear disease relevance has been established for 70% and a further 18% are likely to be involved in at least one disease. This pattern holds true for all the human protein families the SGC is working on. The following sections provide an overview of the three distinct biological areas selected at the Oxford site of the SGC.

Biology area I: Structural Genomics of human metabolic enzymes

Selection of metabolic enzymes as biological target area at the SGC was based on two distinct features: they are fundamentally involved in a multitude of human diseases, including cardiovascular, metabolic diseases or cancer, and in addition several enzymes constitute possible drug targets. Emphasis has been given to certain metabolic enzyme families such as oxidoreductases (mostly short-chain dehydrogenases/reductases (SDR), medium-chain dehydrogenases/reductases (MDR), long-chain dehydrogenases/reductases, aldehyde dehydrogenases (ALDH), aldo keto reductases (AKR) and 2'-oxoglutarate dependent oxygenases (2OGs). In addition, pathways of importance, e.g., in lipid or amino acid metabolism were selected with a distribution of about 1:1 between oxidoreductases and other metabolic enzymes. The target list comprises about 300 metabolic enzymes, and after three years of operation, >60 unique novel structures have been solved. Three points of importance are highlighted in this review: structural characterization of enzymes shown to be causative of metabolic inherited diseases, structure determination of drug discovery targets in metabolic diseases such as metabolic syndrome or osteoporosis, and structure-guided "de-orphanization" of insufficiently characterized human gene products or even entire pathways.

Structural basis of inherited metabolic diseases

Genetic defects in enzymes involved in metabolic pathways such as amino acid or lipid catabolism are causative of a whole spectrum of symptoms, including dysmorphologies, mental retardation, neuropathies or life threatening situations like fasting induced hypoglycemia [12, 13]. Understanding of molecular causes and possible interventions of inherited metabolic diseases requires besides biochemical and clinical management a structural template for explanation of mutational effects.

Thus far the focus has been to a large extent on oxidoreductases in the area of metabolic diseases. Associated disorders comprise electron transfer reactions for energy

production (e.g., mitochondrial myopathies), oxidative and reductive roles in the metabolism of amino acids (e.g., hyperprolinemia or branched-chain hydroxyacyl CoA dehydrogenase defects), fatty acids (e.g., inborn errors in α - and β -oxidation of short-, medium- or long-chain fatty acid metabolites), cofactors (e.g., phenylketonuria type 2), hormones (e.g., male pseudohermaphroditism or adrenal hyperplasia), mediators (e.g., congestive heart failure) and lipids (e.g., inborn errors in cholesterol synthesis, CHILD syndrome, Smitz-Opitz Laemml syndrome as examples). The impact of the structural approach is illustrated by the successful structure determination of phytanoyl-CoA hydroxylase [14], the major molecular cause of Refsum disease, a peroxisomal disorder with severe neurological symptoms. The structure provides a framework to interpret the majority of the disease causing polymorphic alleles, and we were able to map those to changes in the active site, around the Fe^{2+} and 2-oxoglutarate binding sites in this 2OG enzyme [14].

Metabolic enzymes as drug targets

Oxidoreductions at specific positions of lipid hormones such as steroids selectively alter nuclear receptor binding properties. Therefore, inhibition of dehydrogenases/reductases carrying out these reactions selectively influences cellular hormone levels and transcriptional responses. This concept has recently found great attention with the development of specific inhibitors against 11 β -hydroxysteroid dehydrogenase type 1 (11 β -HSD1) as a novel drug target in diabetes and obesity [15–18]. Similar drug development efforts are underway regulating androgen or estrogen levels through specific modulation of distinct hydroxysteroid dehydrogenases (17 β - and 3 α -HSDs in cancer, inflammation, osteoporosis, ageing, and autoimmune diseases). We determined the structure of human 11 β -HSD1 in complex with a clinically relevant inhibitor, carbenoxolone (Wu et al., unpublished) and have provided a platform for drug development efforts. Other hydroxysteroid dehydrogenase structures comprise 17 β -HSDs such as types 4, 8, 10, 11 and a novel type 14 (see below), necessary for determination of off-target activities of compounds directed against type 1 and 3 17 β -HSDs. Other targets of pharmaceutical relevance successfully pursued are farnesyl diphosphate synthase (FDPS) and geranylgeranyl diphosphate synthase (GGPS), which are critical in synthesis of isoprenoids necessary for covalent modification of GTPases involved in cell signalling and survival. Crystal structures of FDPS complexed with nitrogen-containing bisphosphonates currently used for osteoporosis therapy allowed a molecular mechanism of action to be postulated for these drugs [19] (Fig. 1). Furthermore, several

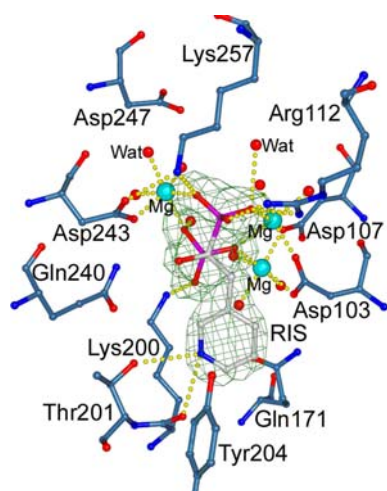


Fig. 1 Bisphosphonate binding to human farnesyl diphosphate synthase. Electron density is shown in green around the clinically used inhibitor risedronate

prokaryotic and parasitic dehydrogenases have been identified as novel targets for antibiotic and antiparasite drug development, and thus allow synchronization with the SGC Toronto efforts, where an apicomplexan/protozoan SG program has been established. Thus, structure determination of related human enzymes will facilitate structure aided drug design and allow virtual and focused screening efforts in this emerging disease area.

Deorphanization of metabolic enzymes and pathways

A significant proportion of the metabolic enzymes targeted were at the time of structure determination devoid of assigned activity or function. High throughput protein production, structure determination and functional characterization allowed “deorphanization” of unknown enzymes. We employed ligand screening, enzyme activity assays, expression and subcellular localization data, as well as structure determination combined with docking analysis to describe novel human enzymes. In the absence of co-crystal structures, interpretation of results from biochemical assays and compound screening was rationalized by *in silico* docking of potential ligands into the active site of the orphan structures. Analysis of the different docking poses was correlated with experimental results, allowing direct visualization of the putative protein–ligand complex. In this manner we determined a novel 17 β -HSD14 [20], possibly involved in cancer, and a novel type-2 R-hydroxybutyrate dehydrogenase, involved in ketone body utilization [21]. Further emphasis was given on novel pathways such as mitochondrial fatty acid synthesis. This recently discovered pathway is important in the synthesis of liponic acid, essential for mitochondrial function. Thus

far we have determined three distinct enzymes of this metabolic route, namely the malonyl transferase (2c2n), ketoacyl synthase (2c9h) and the enoyl-ACP reductase (1zsy). These structures represent the only higher eukaryotic structures thus far available for this pathway. The data will be instrumental to compare to the multidomain type I fatty acid synthase, where we recently solved the structure of the malonyl/acyl transferase domain (2jfk, 2jfd). This cytosolic enzyme is involved in production of endogenous fatty acids and lipids, and is discussed as potential target in metabolic diseases and cancer.

Biology area II: Structural Genomics of transmembrane receptor signalling pathways

Complete coverage of the 14-3-3 protein family

A human protein family that the SGC has completed the structure determination of all members is the 14-3-3 family. This family consists of seven members (β , ϵ , η , γ , σ , τ , and ζ) of which σ [22, 23], τ [24] and ζ [25] structures were previously determined. This protein family plays a central role in many fundamental cellular roles such as cell cycle control, apoptosis, protein trafficking, signal transduction and stress response [26–28].

Before the structural completion of the 14-3-3 family most of the structural studies utilised 14-3-3 ζ which provided details of the conserved peptide binding site [25], the primary peptide interaction [29, 30] and secondary target domain interactions [31]. As all of these structures displayed similar overall conformations, structurally it was proposed that 14-3-3s behaved as “molecular anvils” in that their overall structure remained unchanged whether in the apo-form or bound to their target protein [32]. The structure determination of the remaining members allowed for a family-wide comparative study that revealed another story with a major emphasis on the flexibility of 14-3-3s [33]. This was most obviously with the apo-form of 14-3-3 β in which one of the subunits was in a similar conformation to all other 14-3-3 structure while the opposing monomer displayed a more open conformation for the peptide binding groove (Fig. 2).

Additional flexibility of 14-3-3 proteins was observed when all of the family members were superimposed against one subunit. It became instantly clear that the position of the second subunit varied between the different 14-3-3 isoforms [33]. This is achieved through the N-terminal helices that make up the dimeric interface sliding over one another (Fig. 2). The significance of the interface flexibility is that it allows for the widening or shortening of the distance between the two peptide binding grooves hence allowing a 14-3-3 to accommodate structures of varying

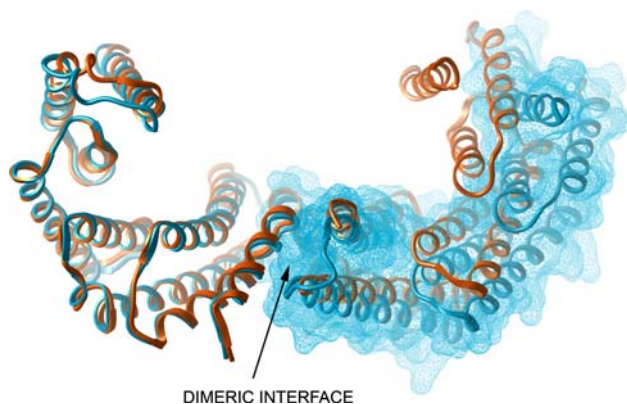


Fig. 2 The flexibility of the 14-3-3 is illustrated by the superimposition of 14-3-3 β (blue) with 14-3-3 η (orange). The monomer conformations of both isoforms are essentially identical on the left hand side. However, the beta monomer on the right side has a more open peptide binding groove and flexibility at the dimeric interface

shapes and sizes. As 14-3-3 are known to have bind hundreds of partners [34–36] this interface flexibility would provide the necessary structural adaptability to accommodate the wide structural range of target proteins.

As all of the human 14-3-3 structures are now known they allow for a detailed bioinformatic analysis of the 14-3-3 family. This approach identified common protein–protein interaction patches at the subunit interfaces plus two additional non-specific protein interaction sites that would attract and bind the globular structured regions of the target protein thus providing a mechanism by which the 14-3-3s can initially attract and then bind a wide range of structurally diverse target proteins [33]. Another more numerous protein–protein interaction family that was targeted by the SGC are the PDZ domains which have been implicated in the regulation of drug transporters [37] and involved in the clustering, targeting and localisation of the target proteins [38]. These domains bind mostly to C-terminal peptides that fall into two classes: class I peptides are $-(\text{Ser/Thr})-\text{X}-\Phi-\text{COO}^-$ while class II peptides are $-\Phi-\text{X}-\Phi-\text{COO}^-$ where X represents any amino acid and Φ represents any hydrophobic residue [39, 40].

PDZ domains

Initial attempts at structure determination of 18 unique human PDZ domains resulted in a successful outcome for only 3 of these targets. To improve our success rate we took advantage of the family based approach and generated new expression clones of the remaining 16 targets with generic class I and II PDZ binding peptides attached to the C-terminus of each domain. The idea was for these peptides to bind adjacent PDZ domains initiating protein–protein interactions and thus crystal nucleation. As such the

linker between the predicted end of the PDZ domain and the C-terminal peptide was varied from 2 to 6 amino acids allowing for flexibility but restraining the distance between adjacent domains [41]. Using this approach we have now solved 11 of the remaining 15 targets many of which have thrown up new details regarding peptide selectivity and structural adaptability of the PDZ domain when bound with a peptide.

As expected for most of these domains the peptide interaction was similar to the standard configuration [42, 43] in that the side-chain of the C-terminal hydrophobic residue (position 0) was bound in a conserved hydrophobic pocket and that the peptide's -2 position Ser/Thr coordinates the His side chain from the αB helix. However, there were a number of surprises of which the biggest was for MPDZ@3 in which a class II mode of binding was observed for a class I peptide which involved a translation of the αB helix (Fig. 6a of [41]).

Biology area III: Structural Genomics of human protein kinases

Kinases play an essential role in most (if not all) signalling pathways and dysregulation has often been linked to disease. Several successful inhibitors developed to target kinases have shown that members of this large protein family are excellent targets for the development of drugs. Currently protein kinases constitute about 25% of presently pursued drug targets in industry [44–47].

There are 518 identified human protein kinases constituting 1.7% of all human genes, which have been grouped into 10 families [48]. Despite the large number of members and their involvement in large variety of pathways, evidence points to a common single ancestral protein. As a result, the structural features as well as key regulatory elements and catalytic mechanism of phosphate transfer are all well conserved. High resolution structures are therefore essential for the rational design of potent and selective inhibitors. Before the contribution of SG efforts, the progression of publicly available kinase structures was linear with only 38 human kinase structures publicly available in 2004 [46]. Currently, 21 novel human kinases structures have been released by the SGC (19 from Oxford), which started to target this protein class in 2004. This increased the number of unique human kinase catalytic domain structures available in the pdb (<http://www.pdb.org/pdb/home/home.do>) to 93 by the end of 2006.

Many structures, released by SG, were only distantly related catalytic domain structures previously known and in some cases provided the first structural information for a subfamily. Thus, these structures significantly enriched the coverage of the three dimensional structure description of

the kinome. Among the structures where the SGC determined the first representative structure of a family were: the NEK (“never in mitosis”/NIMA) family member NEK2, the CDC2 like kinases family member CLK1 and CLK3 as well as the first structure of a NAK (Numb-associated kinases) kinase MPSK1. These kinases are quite diverse in terms of primary structure and it is therefore not surprising that many novel structural features have been discovered. For instance, a novel activation loop architecture characterized by a large helical insert has been discovered in the structure of MPSK1, the structures of CLK1 and CLK3 revealed a family conserved antiparallel beta sheet flanking the kinase hinge region, and the structure of NEK2 identified a short helix following the activation segment DFG motif that may be explored for the development of specific inhibitors [49].

Kinases are extremely flexible proteins that may adopt a number of distinct catalytically active or inactive conformations during their catalytic cycle, upon activation by phosphorylation, or by binding of a regulatory protein, and consequently a number of clinically successful inhibitors have been developed to target specifically the inactive state of kinases [50]. For example the anti-leukaemia drug Imatinib binds selectively to the inactive state cABL characterized by an outward conformation of the DFG motif, a conserved tripeptide motif that ligates Mg^{2+} ions [51, 52]. It is not clear to date how many kinases are able to adopt this conformation, which makes development of these so-called type II inhibitors possible. In general, these are characterized by largely improved specificity.

For the development of conventional inhibitors that target the active state of kinases information about the plasticity of the catalytic domain greatly facilitates the rational design of inhibitors. Consequently it is desirable that several structures of the same target in complex with different ligands are available. Also here the structural information content regarding ligand binding was significantly increased during the last three years by SG. In 2004, only 38 human kinases had a structure available in the public domain and only 12 publicly available structures contained non-adenosine chemotypes [46]. From the 19 structures of kinase catalytic domains released by our laboratory, 16 were determined in the presence of a non-adenosine kinase inhibitor and several structures were determined in complex with more than one inhibitor scaffold (Table 2, and Fig. 3, showing PAK5 apo/inhibitor).

In addition, the SGC has supported development of entirely new inhibitor classes exemplified by co-crystal structures with Ruthenium-half sandwich complexes. These stable organometallic compounds are extremely potent inhibitors for PIM1 kinases [54]. The co-crystal structure of three inhibitors of this class showed that the inert metal centre in this scaffold functions as a hypervalent

carbon, allowing it to occupy the binding pocket efficiently with excellent shape complementarity.

Contributions of NMR to Structural Genomics

NMR as a complementary method to crystallography for protein structure determination

The NMR spectroscopy can play an important role in structural genomics, providing complementary information to that obtained from X-ray crystallography. Importantly for large-scale structural genomics projects, NMR provides an alternative route to solving the high resolution, three-dimensional structures of proteins that prove refractory to crystallization. We were able to use NMR to solve the structures of a number of relatively small protein domains (~ 20 kDa) in which the domain contained at least one flexible region. The RGS domains from the regulator of G-protein signalling proteins, RGS3, RGS10, RGS14, RGS18, RGS20 were all very good examples of this. Multiple constructs of these were designed, which expressed to high yield in stable, highly soluble form yet did not yield high quality crystals despite many months of concerted effort. The domains were therefore expressed as uniformly ^{15}N -labelled proteins using standard growth methods in *E. coli*, and their ^{15}N -HSQC spectra were recorded to assess the feasibility of structure determination by NMR. In all cases, excellent spectral dispersion was observed and we were able to obtain almost complete assignment of the protein resonances. We have since deposited the high resolution NMR structures of three RGS domains in the PDB and the resonance assignments of four RGS domains in the BioMagResBank (BMRB). The structures and assignments of two further non-crystallizing domains (Spred2 EVH1 domain and JARID1CA Bright/ARID domain) have also been deposited, and those of several other non-crystallizing domains ‘rescued’ by NMR are currently underway (Table 3).

NMR as an assessment tool for the feasibility of structure determination

Further examples where NMR has proven useful as a rescue strategy include particular families of signalling domains which have a known tendency to be partially unfolded in their unliganded states. Some examples include certain WW domain [55, 56]. We successfully identified peptide binding partners for a WW-tandem construct using the SPOTs screening technique [57, 58] following which the most strongly binding peptides were synthesized on a large scale for NMR measurements. Although the ^{15}N -

Table 2 Protein kinase structures determined by SGC

Name	PDB ID	Resolution [\AA]	Inhibitor name	Disease link	Family
CLK1	1Z57	1.70	Hymenialdisine	Pot. Genetic ^e	CMGC
CLK3 ^c	2EU9	1.53	none	Pot. Genetic ^e	CMGC
CK1 γ 1	2CMW	1.75	Compound 52		CK1
CK1 γ 2	2C47	2.40	5-Iodotubercidin	Genetic	CK1
CK1 γ 3 ^d	2CHL	1.95	Triazolodiamine 1	Cancer	CK1
ERK3	2I6L	2.25	none	Cancer	CMGC
ASK1	2CLQ	2.30	Staurosporine	Inflammation, CV ^f	STE
NEK2	2JAV	2.10	SU11652	Cancer	Other-NEK
PAK4 ^a	2CDZ	2.40	Cdk1 Inhibitor	Cancer	STE
PAK5	2F57	1.80	Cdk1 Inhibitor	Pot. Cancer	STE
PAK6	2C30	1.60	none	Cancer	STE
PIM1 ^b	1XWS	1.80	BIM I, HB1	Cancer, Inflammation	CAMK
PIM2	2IWI	2.80	HB1	Cancer, Inflammation	CAMK
SLK ^c	2J51	2.10	Triazolodiamine 1	Pot. Cancer ^e	STE
MPSK1	2BUJ	2.60	Staurosporine	Pot. Cancer ^e	Other-NAK
STK10	2J7T	2.0	SU11274	Not known	STE
DAPK3	2J90	2.0	Pyridone 6	Cancer, Inflammation	CAMK
CAMK1G	2JAM	1.7	SU11652	Not known	CAMK
CAMK1D	2JC6	2.5	GSK inhibitor XIII	Genetic	CAMK

^a PAK4 also deposited as an apo-structure in two different spacegroups: 2BVA, 2J01

^b Structures with different inhibitors and substrate: PIM1: 2BIK, 2BZH, 2BZI, 2BZJ, 2BZK, 2C3I, 2BIL, 2J2I; SLK: 2JA0

^c CLK3 also deposited as phosphorylated protein: pdb-code: 2EXE. Detailed description of structures of targets solved in Oxford is available in form of “iSee” datapacks freely downloadable on: <http://www.sgc.ox.ac.uk/structures/KIN.html>

^d CK1 γ 3 with different inhibitors: 2CHL, 2IZR, 2IZS, 2IZU, 2IZT

^e A formal link to the disease has not been established so far but is likely

^f Cardiovascular disease

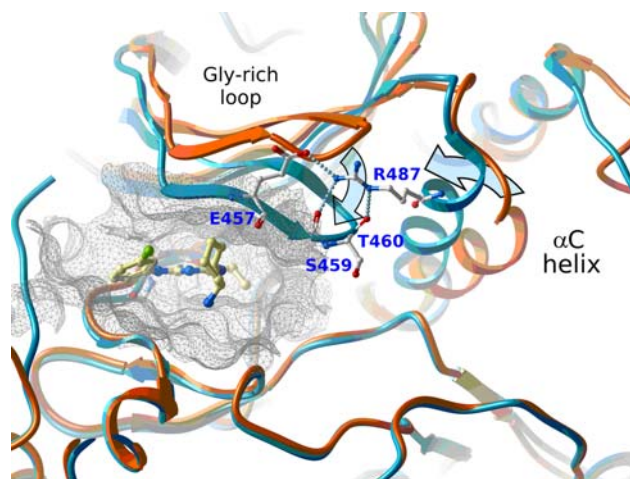


Fig. 3 Superimposition of apo-PAK5 (cyan) and the PAK5 purine complex (orange), highlighting the decomposed movements of the glycine-rich loop (flapping) and the α C helix (swinging) [53]

HSQC spectra of this pair of tandem domains in isolation were very unhelpful, the spectra of their complexes showed significant improvements in signal dispersion,

indicating that in the complexed form, the protein was better folded. At this point, the protein entered our NMR structure determination pipeline. The recording of a quick 15N-HSQC spectrum has in several cases allowed us to rescue protein constructs with promising but borderline behaviour, for example, proteins showing good signal dispersion but low-medium levels of aggregation. Far from abandoning these constructs, we took these constructs ‘back to the drawing board’ and made rational construct improvements with the help of bio-informatic tools. Successfully re-designed constructs were then re-screened for fold quality by 15N-HSQC. After 2–3 iterations of this procedure, it was often possible to refine promising constructs sufficiently for structure determination. For example, Fig. 4 shows the stepwise improvement in the spectral properties of a hopeful, though initially problematic DNA-recognition domain from the oxygenase protein, JARID1CA. The NMR structure is now deposited (PDB code: 2JRZ). In all of the above cases, a quick 15N-HSQC showed immediately whether the structure determination of a protein, having to failed to crystallize, should be

Table 3 Deposited NMR structures and assignments

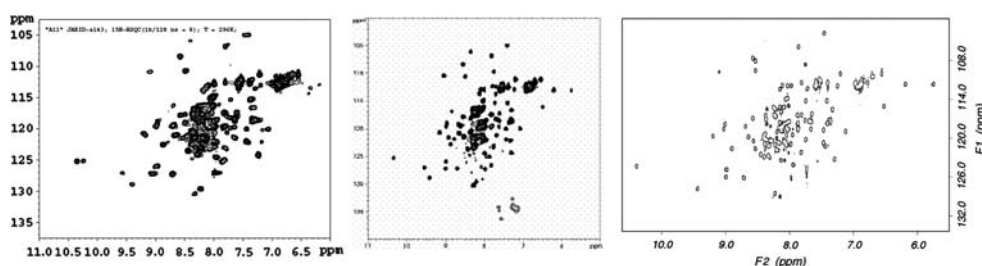
Gene	PDB deposition	Resonance assignment deposition
RGS3	–	BMRB-15178
RGS10	2I59	BMRB-7272
RGS14	2JNU	BMRB-15128
RGS18	2OWI	BMRB-7106
SPRED2	2JP2	BMRB-5939
JARID1CA	2JRZ	BMRB-15348

pursued or abandoned, hence reducing unnecessary attrition in the structure determination pipeline.

The study of protein dynamics by NMR

The use of NMR to study the rotational correlation times and internal dynamics of the proteins offers good explanations as to why crystallization sometimes fails even for well-folded proteins. In all of the proteins we rescued by NMR, ^{15}N heteronuclear NOE and ^{15}N T₁, T₂ relaxation data revealed regions of internal mobility within the proteins, which would have hindered long-range order and impaired or prevented efficient crystal packing. A striking example was the case of the RGS domain from RGS10, in which NMR relaxation data confirmed true local mobility in a region of the domain which not only lacked in NMR restraints, but also showed no electron density in the crystal structure of the complex of RGS10 with G- α -i3 (PDB 2IHB). Comparison of mobility in RGS domains from different branches of the phylogenetic tree leads to clues about their specificity and helps to guide further investigations. In some cases, the ^{15}N T₁ and T₂ data have also identified partial dimerization in proteins that fail to crystallize, thus explaining the latter. NMR relaxation data were in each case confirmed by analytical ultracentrifugation (AUC). The combined information allowed us to decide whether these proteins should be highlighted as candidates for structure determination by NMR and to judge the best conditions under which they should be studied.

Fig. 4 Visible improvement in quality of ^{15}N -HSQC spectra over two rounds of iterative construct re-design for the JARID1CA Bright/ARID domain. The leftmost (initial) construct shows potential. The structure of the final construct on the far right was determined by NMR (PDB code: 2JRZ)



Future and outlook

The future role that NMR will play in structural genomics will depend heavily on the continued development and implementation of new, faster methods of data acquisition, processing, resonance- and NOE-assignment and structure determination and refinement. These topics have been covered extensively in other reviews; for a concise summary see [59] and references therein. The potential time gains that could be gained from these methods make high throughput structure determination by NMR a realistic possibility for the future.

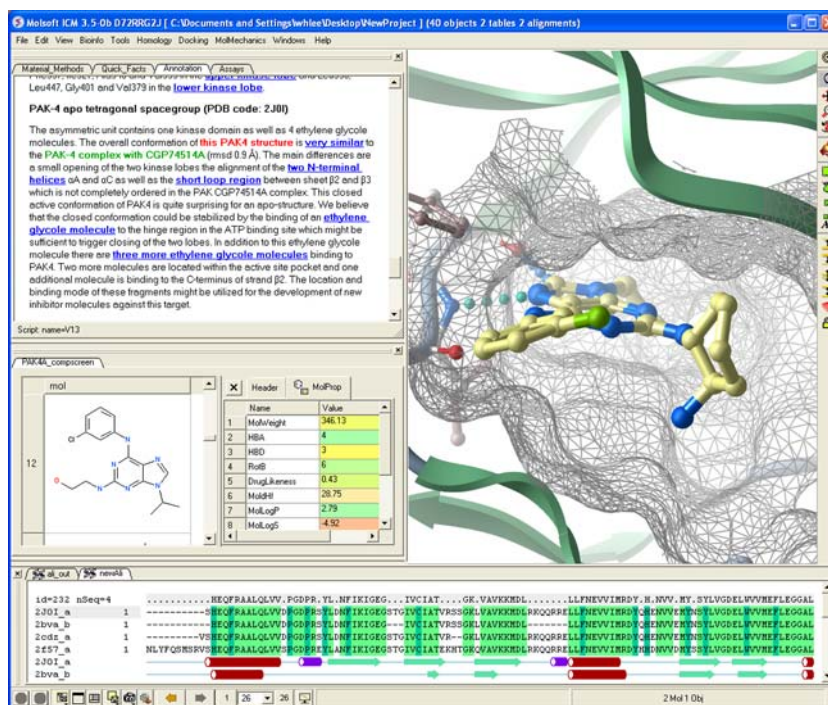
Structural bioinformatics and rationalisation of experimental results

A crystal structure of a protein in absence of ligand or substrate may not always provide insight on reaction mechanisms or specificity. Ideally, such information can be derived from additional structures with bound ligands. In the absence of such co-crystals, interpretation of results from biochemical assays and compound screening is more speculative. However, these results can be rationalised with in silico docking of potential ligands into the active site of unliganded protein structures. An example illustrating this point is the analysis of the DHRS10 structure [20]. Analysis of the different docking poses can be correlated with experimental results, allowing direct visualisation of the putative protein–ligand complex. With these results, further modifications of the enzyme can be suggested more reliably, allowing a faster progress towards the complete elucidation of the mechanistics.

Dissemination of structural genomics data and knowledge

Structural genomics produces a wealth of information of different types: DNA and protein sequences, biochemical information, coordinates of crystal structures, and structural

Fig. 5 Screenshot of iSee datapack. The annotation text (top left panel) includes links (blue text), which lead to structural images focused at areas of interest, simultaneously accessing other types of information (sequence alignment, small molecule formulae, etc.)



annotation. This information is deposited in one or more public databases, predominantly the PDB, in addition to publication in journals. This form of data distribution does not adequately disseminate the full information to a wide scientific audience. The first issue is the fragmentation of data between different formats. A user may have to read text information in a journal paper, which may include a few two-dimensional Figures; then download a PDB structure file and image with a separate application; and then perform analysis and alignment of data from, say, SNP database using alignment software. The second issue is that non-structural biologists do not routinely access PDB files, especially of structures that were not published in pubmed-indexed journals.

We have approached this challenge by developing a new intuitive dissemination concept in conjunction with Molsoft LLC (San Diego, CA) [60]. This concept, (which we denoted *iSee*) integrates all the information associated with any given target solved by SGC into a small, self-contained file, annotated by the authors (Fig. 5). The file not only allows the direct visualisation of text information, but also offers an interactive visualisation feature fully integrated to the structural data being presented. At any stage, the annotation written by the expert can be coupled with an interactive molecular graphics scene. Transition between each annotated viewpoint is fully animated on-the-fly, to convey a sense of three-dimensionality which is vital for the user to grasp the spatial relationship between different features on a structure.

Each of these files (called an *iSee* datapack), as well as the software needed to visualise them (ICM-Browser) are

available for free download from our website (<http://www.sgc.ox.ac.uk/iSee>).

We also maintain and curate each of these files by revising each datapack quarterly to ensure that all the recently disclosed information is added (either by ourselves through follow-up experiments or by external collaborators working on the same targets). Each of the datapacks has a built-in automated updating function that can be executed on user's request.

Acknowledgements We thank all members of the SGC for performing the work reviewed in this paper. The SGC is a registered charity (number 1097737) funded by the Wellcome Trust, Glaxo-SmithKline, Genome Canada, the Canadian Institutes of Health Research, the Ontario Innovation Trust, the Ontario Research and Development Challenge Fund, the Canadian Foundation for Innovation, VINNOVA, The Knut and Alice Wallenberg Foundation, The Swedish Foundation for Strategic Research, and Karolinska Institutet.

References

1. Newman J, Egan D, Walter TS, Meged R, Berry I, Ben Jelloul M, Sussman JL, Stuart DI, Perrakis A (2005) *Acta Crystallogr D Biol Crystallogr* 61:1426–1431
2. Page R, Grzechnik SK, Canaves JM, Spraggon G, Kreuzsch A, Kuhn P, Stevens RC, Lesley SA (2003) *Acta Crystallogr D Biol Crystallogr* 59:1028–1037
3. Cudney R, Patel S, Weisgraber K, Newhouse Y, McPherson A (1994) *Acta Crystallogr D Biol Crystallogr* 50:414–423
4. Kleywegt GJ (2000) *Acta Crystallogr D Biol Crystallogr* 56:249–265
5. Emsley P, Cowtan K (2004). *Acta Crystallogr D Biol Crystallogr* 60:2126–2132

6. Pantoliano MW, Petrella EC, Kwasnoski JD, Lobanov VS, Myslik J, Graf E, Carver T, Asel E, Springer BA, Lane P, Salem FR (2001) *J Biomol Screen* 6:429–440
7. Poklar N, Lah J, Salobir M, Macek P, Vesnaver G (1997) *Biochemistry* 36:14345–14352
8. Vedadi M, Niesen FH, Allali-Hassani A, Fedorov OY, Finerty PJ, Jr Wasney GA, Yeung R, Arrowsmith C, Ball LJ, Berglund H, Hui R, Marsden BD, Nordlund P, Sundstrom M, Weigelt J, Edwards AM (2006) *Proc Natl Acad Sci USA* 103:15835–15840
9. Niesen FH, Berglund H, Vedadi M (2007) *Nat Protoc* 2: 2212–2221
10. Bullock AN, Debreczeni JE, Fedorov OY, Nelson A, Marsden BD, Knapp S (2005) *J Med Chem* 48:7604–7614
11. Chandonia JM, Brenner SE (2006) *Science* 311:347–351
12. Rinaldo P, Matern D, Bennett MJ (2002) *Annu Rev Physiol* 64:477–502
13. Bennett MJ, Rinaldo P, Strauss AW (2000) *Crit Rev Clin Lab Sci* 37:1–44
14. McDonough MA, Kavanagh KL, Butler D, Searls T, Oppermann U, Schofield CJ (2005) *J Biol Chem* 280:41101–41110
15. Oppermann U (2006) *Endocr Metab Immune Disord Drug Targets* 6:259–269
16. Walker BR, Seckl JR (2003) *Expert Opin Ther Targets* 7: 771–783
17. Tomlinson JW, Bujalska I, Stewart PM, Cooper MS (2000) *Endocr Res* 26:711–722
18. Thieringer R, Hermanowski-Vosatka A (2005) *Expert Rev Cardiovasc Ther* 3:911–924
19. Kavanagh KL, Guo K, Dunford JE, Wu X, Knapp S, Ebetino FH, Rogers MJ, Russell RG, Oppermann U (2006) *Proc Natl Acad Sci USA* 103:7829–7834
20. Lukacik P, Keller B, Bunkoczi G, Kavanagh K, Hwa Lee W, Adamski J, Oppermann U (2007) *Biochem J* 402:419–427
21. Guo K, Lukacik P, Papagrigoriou E, Meier M, Lee WH, Adamski J, Oppermann U (2006) *J Biol Chem* 281:10291–10297
22. Wilker EW, Grant RA, Artim SC, Yaffe MB (2005) *J Biol Chem* 280:18891–18898
23. Benzinger A, Popowicz GM, Joy JK, Majumdar S, Holak TA, Hermeking H (2005) *Cell Res* 15:219–227
24. Xiao B, Smerdon SJ, Jones DH, Dodson GG, Soneji Y, Aitken A, Gamblin SJ (1995) *Nature* 376:188–191
25. Liu D, Bienkowska J, Petosa C, Collier RJ, Fu H, Liddington RC (1995) *Nature* 376:191–194
26. Berdichevsky A, Guarente L (2006) *Cell Cycle* 5:2588–2591
27. Hermeking H, Benzinger A (2006) *Semin Cancer Biol* 16:183–192
28. Darling DL, Yingling J, Wynshaw-Boris A (2005) *Curr Top Dev Biol* 68:281–315
29. Yaffe MB, Rittinger K, Volinia S, Caron PR, Aitken A, Leffers H, Gamblin SJ, Smerdon SJ, Cantley LC (1997) *Cell* 91:961–971
30. Petosa C, Masters SC, Bankston LA, Pohl J, Wang B, Fu H, Liddington RC (1998) *J Biol Chem* 273:16305–16310
31. Obsil T, Ghirlardo R, Klein DC, Ganguly S, Dyda F (2001) *Cell* 105:257–267
32. Yaffe MB (2002) *FEBS Lett* 513:53–57
33. Yang X, Lee WH, Sobott F, Papagrigoriou E, Robinson CV, Grossmann JG, Sundstrom M, Doyle DA, Elkins JM (2006) *Proc Natl Acad Sci USA* 103:17237–17242
34. Jin J, Smith FD, Stark C, Wells CD, Fawcett JP, Kulkarni S, Metalnikov P, O'Donnell P, Taylor P, Taylor L, Zougman A, Woodgett JR, Langeberg LK, Scott JD, Pawson T (2004) *Curr Biol* 14:1436–1450
35. Meek SE, Lane WS, Piwnica-Worms H (2004) *J Biol Chem* 279:32046–32054
36. Pozuelo Rubio M, Geraghty KM, Wong BH, Wood NT, Campbell DG, Morrice N, Mackintosh C (2004) *Biochem J* 379: 395–408
37. Kato Y, Watanabe C, Tsuji A (2006) *Eur J Pharm Sci* 27:487–500
38. Kim E, Sheng M (2004) *Nat Rev Neurosci* 5:771–781
39. Songyang Z, Fanning AS, Fu C, Xu J, Marfatia SM, Chishti AH, Crompton A, Chan AC, Anderson JM, Cantley LC (1997) *Science* 275:73–77
40. Stricker NL, Christopherson KS, Yi BA, Schatz PJ, Raab RW, Dawes G, Bassett DEJ, Bredt DS, Li M (1997) *Nat Biotechnol* 15:336–342
41. Elkins JM, Papagrigoriou E, Berridge G, Yang X, Phillips C, Gileadi C, Savitsky P, Doyle DA (2007). *Protein Sci* 16:683–694
42. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R (1996) *Cell* 85:1067–1076
43. Morais Cabral JH, Petosa C, Sutcliffe MJ, Raza S, Byron O, Poy F, Marfatia SM, Chishti AH, Liddington RC (1996) *Nature* 382:649–652
44. Cohen P (2002) *Nat Rev Drug Discov* 1:309–315
45. Faivre S, Djelloul S, Raymond E (2006) *Semin Oncol* 33: 407–420
46. Vieth M, Higgs RE, Robertson DH, Shapiro M, Gragg EA, Hemmerle H (2004) *Biochim Biophys Acta* 1697:243–257
47. Vieth M, Sutherland JJ, Robertson DH, Campbell RM (2005) *Drug Discov Today* 10:839–846
48. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S (2002) *Science* 298:1912–1934
49. Rellos P, Ivins FJ, Baxter JE, Pike A, Nott TJ, Parkinson DM, Das S, Howell S, Fedorov O, Shen QY, Fry AM, Knapp S, Smerdon SJ (2007) *J Biol Chem* 282:6833–6842
50. Noble ME, Endicott JA, Johnson LN (2004) *Science* 303:1800–1805
51. Nagar B, Bornmann WG, Pellicena P, Schindler T, Veach DR, Miller WT, Clarkson B, Kuriyan J (2002) *Cancer Res* 62:4236–4243
52. Nagar B, Hantschel O, Young MA, Scheffzek K, Veach D, Bornmann W, Clarkson B, Superti-Furga G, Kuriyan J (2003) *Cell* 112:859–871
53. Eswaran J, Lee WH, Debreczeni JE, Filippakopoulos P, Turnbull A, Fedorov O, Deacon SW, Peterson JR, Knapp S (2007) *Structure* 15:201–213
54. Bullock AN, Debreczeni JE, Edwards AM, Sundstrom M, Knapp S (2006) *Proc Natl Acad Sci USA* 103:7637–7642
55. Fedoroff OY, Townson SA, Golovanov AP, Baron M, Avis JM (2004) *J Biol Chem* 279:34991–35000
56. Kanelis V, Rotin D, Forman-Kay JD (2001) *Nat Struct Biol* 8:407–412
57. Frank R (1992) *Tetrahedron* 48:9217–9232
58. Frank R (2002) *J Immunol Methods* 267:13–26
59. Yee A, Gutmanas A, Arrowsmith CH (2006) *Curr Opin Struct Biol* 16:611–617
60. Abagyan R, Lee WH, Raush E, Budagyan L, Totrov M, Sundstrom M, Marsden BD (2006) *Trends Biochem Sci* 31:76–78