Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

First steps towards effective methods in exploiting high-throughput technologies for the determination of human protein structures of high biomedical value

L. Banci,[a] I. Bertini,[a] S. Cusack,[b] R. N. de Jong,[c] U. Heinemann,[d] E. Y. Jones,[e] F. Kozielski,[f] K. Maskos,[g] A. Messerschmidt,[g] R. Owens,[e] A. Perrakis,[h] A. Poterszman,[i] G. Schneider,[j] C. Siebold,[e] I. Silman,[k] T. Sixma,[h] G. Stewart-Jones,[e] J. L. Sussman,[k] J.-C. Thierry[i] and Dino Moras[i]*

[a]CIRMMP, CERM, Via Sacconi 6, 50019 Sesto Fiorentino, Italy, [b]EMBL-Grenoble, c/o ILL, 6 Rue Jules Horowitz, 38042 Grenoble CEDEX 9, France, [c]Bijvoet Center for Biomolecular Research, NMR Spectrometry, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands, [d]Max-Delbrück-Center for Molecular Medicine, Department of Crystallography, Robert-Rössle-Strasse 10, D-13125 Berlin, Germany, [e]Division of Structural Biology, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, England, [f]Laboratoire des Moteurs Moléculaires (LMM), Institut de Biologie Structurale (CEA-CNRS-UJF), 41 Rue Jules Horowitz, 38027 Grenoble CEDEX 01, France, [g]Max-Planck Insitute of Biochemistry, Department of Proteomics and Signal Transduction, Am Klopferspitz 18, 82152 Martinsried, Germany, [h]Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, [i]Institut de Génétique et de Biologie Moléculaire et Cellulaire, 1 Rue Laurent Fries, BP 163, 67404 Illkirch CEDEX, France, [j]Department of Medical Biochemistry and Biophysics, Karonlinska Institutet, SE 109 51, Stockholm, Sweden, and [k]The Israel Structural Proteomics Centre, The Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Correspondence e-mail: moras@igbmc.u-strasbg.fr

Received 29 November 2005
Accepted 28 July 2006

The EC 'Structural Proteomics In Europe' contract is aimed specifically at the atomic resolution structure determination of human protein targets closely linked to health, with a focus on cancer (kinesins, kinases, proteins from the ubiquitin pathway), neurological development and neurodegenerative diseases and immune recognition. Despite the challenging nature of the analysis of such targets, ∼170 structures have been determined to date. Here, the impact of high-throughput technologies, such as parallel expression of multiple constructs, the use of standardized refolding protocols and optimized crystallization screens or the use of mass spectrometry to assist sample preparation, on the structural biology of mammalian protein targets is illustrated through selected examples.

1. Introduction

Structural Proteomics In Europe (SPINE) is a large proteomics project whose aim is to develop new methods and technologies for high-throughput structural biology and to tackle difficult problems in human health and disease. SPINE is driven by the choice of 'high-value human health targets' rather than 'filling fold space'. These include proteins derived from human pathogens of both bacterial and viral origin associated with diseases such as tuberculosis, anthrax, SARS and herpes, as well as human proteins associated with cancer, immune defence mechanisms and neuronal development plus neurodegenerative diseases. More than 2000 targets were selected, 60% of which were of bacterial or viral origin and 40% of which were from mammalian sources. After less than 3 years of operation, the SPINE consortium (see the editors' preface to this issue for an overview of the consortium and a description of the 20 partner laboratories) has determined the three-dimensional structures of 260 proteins. This number only takes into account novel structures and primary examples of protein–ligand or protein–protein complexes. When structures of complexes with additional ligands or with metal ions are included, the total number approaches 370. At present, 45% of the structures are from bacterial or viral pathogens and 55% are from human or mammalian sources (Fig. 1). This paper will focus on the latter subset, covering structures that are relevant to cancer (cell-surface proteins, transcription factors, nuclear receptors, kinases and signalling proteins, proteases, DNA-repair factors and proteins from the ubiquitin pathway), to neuronal development or neurodegenerative diseases and to immune recognition (all of which are targets included in workpackages 10 and 11 of SPINE). We first summarize the different methodological approaches used and their impact and success rates (§§2 and 3). The second part of the paper discusses several structures determined by the SPINE consortium as 'selected highlights' (§4).

2. Scoreboard and success rates

2.1. Scoreboard for human targets

In the course of the SPINE project, more than 800 targets from higher eukaryotes were selected. A key objective was to introduce novel, automated and high-throughput (HTP) systematic strategies for structure determination by X-ray crystallography and NMR spectroscopy; in particular, the use of HTP was to screen multiple constructs of key targets rather than to generate large numbers of structures. The availability of a large number of experiments provides an opportunity to assess how the structure-determination pipeline currently performs for the subset of human targets and to identify rate-limiting step(s). The SPINE scoreboard presented in Fig. 2(*a*) shows that the main bottleneck of the pipeline is the production of protein suitable for structural analysis: out of the 375 eukaryotic targets that were analyzed (as of 1 November 2005), almost 24% yielded crystals or protein preparation suitable for NMR analysis. When suitable crystals or NMR samples were obtained, success in structure determination by X-ray or NMR protocols approached 80%. If a high-resolution data set has been collected, structure determination by X-ray crystallography is now largely automated and straightforward, with a success rate above 90%. Nevertheless, classical problems that hamper the process still remain, usually in the case of poor diffraction or when crystals of a SeMet-containing protein cannot be obtained. NMR spectroscopy has made an important contribution to the SPINE project, having yielded three-dimensional structures for 27 human proteins or domains (see AB *et al.*, 2006 for detailed analysis of the role of NMR in structural proteomics in general and SPINE in particular, and Fig. 3 for four highlight structures).

2.2. Expression strategies and success rates

A comparison of the number of constructs or expression trials with the number of crystals or good heteronuclear single quantum coherence (HSQC) data sets collected for the entire SPINE project shows that the percentage of constructs leading to the production of a protein suitable for structural analysis is around 10%. This low percentage arises from the fact that only some 30% of the constructs yield soluble protein and less than 30% of these soluble proteins either crystallize or are suitable for NMR analysis. On average, between two and three constructs per target were tested, but in some cases over 30 were analysed for a single target. This multiple-construct strategy proved particularly effective for structural analyses of multi-domain proteins by both NMR (see AB *et al.*, 2006) and X-ray crystallography (for example, studies, detailed in later sections, on large cytosolic proteins such as Dlg1 and MICAL carried out by the Oxford node or on members of the nuclear receptor family reported by the Strasbourg node). These proteins are composed of distinct structural modules and an optimal definition of the limits of these modules is a key parameter in attempts to obtain a 'well behaved' sample suitable for structure determination.

A global analysis of the expression strategies and results gathered from SPINE partners (Fig. 2*b*) shows that (i) expression trials were performed first in *Escherichia coli*, (ii) strains allowing the co-expression of rare tRNAs with the protein of interest have been widely used and (iii) eukaryotic expression systems were also tested when necessary. On average, 30% of the constructs led to the production of a soluble protein when *E. coli* was the expression system used. Using insect cells or mammalian expression systems, 44 and 75% of the constructs, respectively, yielded soluble protein, leading to a substantial number of interesting structures such as the atypical protein kinase PKC iota (selected by the Munich node as an attractive target for the development of novel therapeutics against cancer) and acid β-glucosylceramidase (Gcase; a molecule that is defective in Gaucher disease, analysed at the Weizmann Institute; see below). In summary, it appears that within SPINE the host of choice for recombinant protein expression is *E. coli*, even for proteins of higher eukaryotes. However, for difficult cases, eukaryotic expression systems such as baculovirus or mammalian cells were often essential.

3. Case studies of the impact of technologies

Selected examples illustrate the impact on the structural biology of mammalian protein targets of SPINE-based HTP technologies such as (i) parallel expression of multiple constructs for a single target using an *E. coli*-based expression pipeline, (ii) screening of targets in baculovirus/insect or mammalian expression systems, (iii) production and crystallization of targets from families of closely related proteins exploiting standardized refolding protocols and optimized crystallization screens, (iv) directed evolution and (v) use of mass spectrometry (MS) to assist sample preparation.

3.1. Parallel expression of multiple constructs

Parallel expression of multiple constructs was found to be very effective when applied to multi-domain proteins, being used by most SPINE partners. For example, the Oxford node used this strategy to generate structural information for the multi-domain scaffold protein Dlg 1, the mammalian homologue of the fly tumour suppressor protein disc large (T. Ose, C. Siebold & E. Y. Jones, manuscript in preparation) and for the cytosolic protein MICAL (**m**olecule **i**nteracting with **CasL**), a putative flavoprotein monooxygenase region required for semaphorin-plexin repulsive axon guidance (Siebold *et al.*, 2005). The production of proteins for the MICAL structural analysis typified the use of the HTP cloning and *E. coli* expression pipeline developed at Oxford (see Aricescu, Asseneberg *et al.*, 2006 and Fig. 4). A total of 54 constructs designed to vary domain(s) length, species and position of affinity tags were selected for generation by means of ligation-independent cloning using Gateway technology.

All constructs encoded either N- or C-terminal 6×His tags to allow nickel-chelate purification of the expressed protein. On the basis of automated small-scale expression screening, three constructs (encoding separate N- and C-terminal portions of the molecule) were selected for full-scale protein production. Purification of the three His-tagged proteins used the standard automated His-affinity-gel filtration program on an ÄKTA Xpress. Quality control using dynamic light scattering indicated that two of the three protein samples were monodisperse and electrospray ionization mass spectrometry confirmed correct molecular weights for all three samples. The three proteins entered crystallization trials with a Cartesian robot dispensing nanolitre-scale drops (100 + 100 nl) using the standard Oxford set of 700 conditions (Walter *et al.*, 2005) at two temperatures (277 and 293 K). One of the constructs, mMICAL(1–489), crystallized under four conditions. These were optimized by applying the 36-well Oxford optimization protocol using the Cartesian robot (Walter *et al.*, 2005). The entire process, from DNA-template preparation to optimized crystals, exploited automated and miniaturized technologies. It was completed within approximately ten weeks. The resultant 1.45 Å resolution crystal structure of the FAD-containing monooxygenase domain of mouse MICAL-1 (residues 1–489) has recently been published (Siebold *et al.*, 2005).

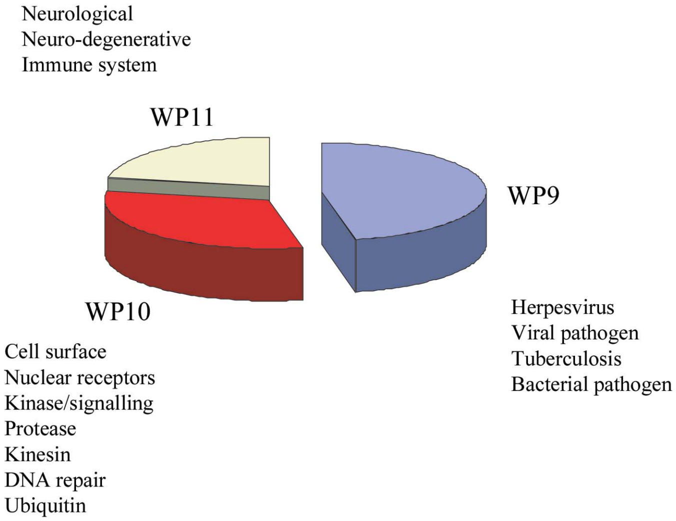

Figure 1
Distribution of the workpackage 9 (WP9), 10 (WP10) and 11 (WP11) structures solved by SPINE.

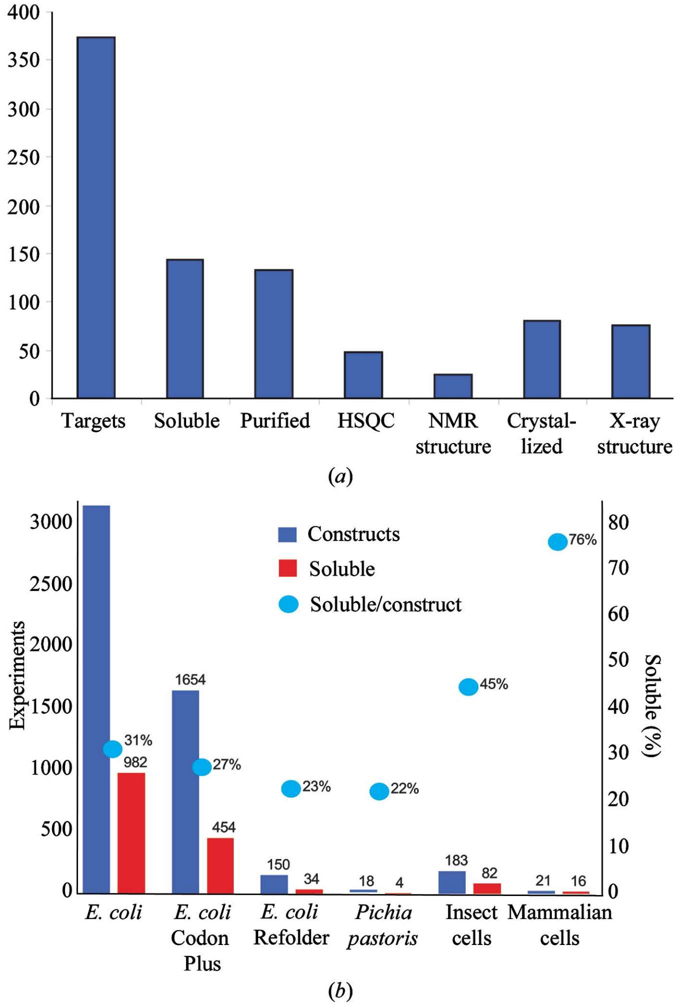

Figure 2
SPINE scoreboards. (*a*) Scoreboard for eukaryotic targets. (*b*) SPINE overall expression statistics as analyzed per expression system

3.2. The use of parallel methods for expression of target proteins using eukaryotic expression systems

The use of parallel methods for expression of target proteins using eukaryotic expression systems is currently

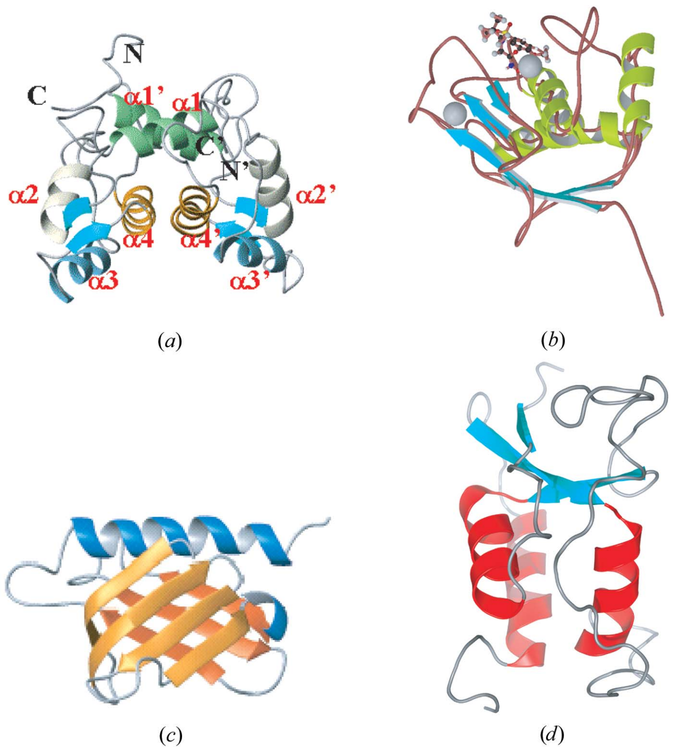

Figure 3
Portfolio of NMR structures from WP10 and WP11. (*a*) The S100A13 Ca^{II}/Co^{II}-binding protein. (*b*) Matrix metalloprotease MMP-12. (*c*) The p62 TFIIH PH domain. (*d*) The USP15 DUSP domain.

emerging as a viable approach for proteins that are not amenable to prokaryotic expression. The structure of the catalytic domain of human atypical protein kinase C iota, determined by the Munich node, illustrates the usefulness of the insect-cell/baculovirus expression system. The availability of mammalian expression systems was decisive for the structure determination of the third leucine-rich repeat domain of slit2, an axon-guidance protein (Grenoble node), and of various ligand-binding regions from the ectodomains of receptor protein tyrosine phosphatases (RPTPμ and RPTPσ) and receptor protein tyrosine kinases (ephrin receptors; Oxford node, see Aricescu, Lu *et al.*, 2006 for a more detailed description of the mammalian transient expression methodology).

3.3. Standardized refolding protocols and optimized crystallization screens

Standardized refolding protocols and optimized crystallization screens have been applied to families of targets such as catalytic domains of matrix metalloproteinases (MMPs) by the Munich or Florence partners, whilst the Oxford node has targeted tumour necrosis factor-like and TNF receptor-like molecules, which in *E. coli* are typically expressed as inclusion bodies. A well developed set of such studies has been conducted in Oxford on the extracellular regions of $\alpha\beta$ T-cell receptors (TCRs). The binding of TCRs to antigen-presenting molecules from the MHC class I- and class II-type families is a key recognition event in the function of the cellular immune response and structural information on these recognition complexes is of direct relevance for the design of new therapeutics and vaccination strategies. TCRs are glycosylated proteins made up of two chains, α and β, each of which can typically be expressed separately in good yields as inclusion bodies in *E. coli*.

Automated small-scale expression screening (varying the induction temperature and host strain) has proved very effective in optimizing this stage (in general, *E. coli* strains B834s and B834pRareLysS were found to be the best for high-level expression). A standardized refolding protocol typically yields 20% of the starting material as heterodimeric TCR. Crystallization screening of TCRs, both unliganded and in complex with specific MHC class I- and MHC class II-type ligands, is focused on a set of 96 optimized conditions selected by mining data on successful crystallization conditions for molecules and complexes of this family (see Berry *et al.*, 2006; Walter *et al.*, 2005). The accumulated experience in Oxford from expression, refolding and crystallization trials on more than 30 TCRs indicates the level of success in going from a pair of α- and β-chain expression constructs to structure is approximately 33%. The major point of attrition is at the crystallization/optimization of diffraction-quality stage and nanolitre-scale crystallization has played a major role in generating a current portfolio of structures which includes nine TCR–MHC class I complexes (Stewart-Jones *et al.*, 2003; Chen *et al.*, 2005).

3.4. Directed evolution

Directed evolution is emerging as a powerful technology to improve soluble expression and has been used in particular by the Grenoble, Weizmann and Oxford nodes. This strategy was applied to members of the serum paraoxonase (PON) family at the Weizmann Institute. PONs have been identified in mammals and other vertebrates and exhibit a wide range of physiologically important hydrolytic activities, including drug metabolism and detoxification of nerve agents. PON1 and PON3 reside on high-density lipoprotein (HDL; 'good cholesterol') and are involved in the prevention of artherosclerosis (Draganov & La Du, 2004). Directed evolution *via* 'family shuffling' and screening resulted in the production of mammalian PON1 variants that express in a soluble and active form in *E. coli* (Aharoni *et al.*, 2004). This in turn permitted purification, subsequent crystallization and solution of the three-dimensional structure of one such variant to 2.2 Å resolution (Fig. 5*a*; Harel *et al.*, 2004). PON1 is a six-bladed β-propeller with a unique active-site lid that is also involved in HDL binding. Two Ca^{2+} ions, one structural and one functional, can be identified within the

???	• Construct planning using the OPPF bioinformatics facilities
2 weeks	• DNA-template preparation and oligo ordering
2 weeks	• PCR and cloning
2 weeks	• Small-scale expression screens
1 week	• Selection and scale-up (three constructs)
1 day	• Purification (ÄKTA Xpress) and protein characterization (DLS and ESI–MS)
2 days	• Nanoscale crystallization using Cartesian technology (three constructs)
2 days	• Crystal optimization (one construct)
2 days	• X-ray data collection
2 weeks	• Structure solution

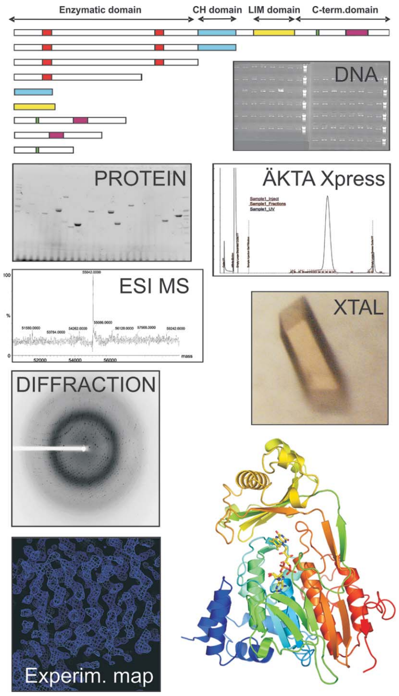

Figure 4
Parellel expression of multiple protein constructs for a single target using an *E. coli*-based expression pipeline: the case of the MICAL-1 protein.

active site. The three-dimensional structure, taken together with the directed evolution studies, permits a detailed description of PON1's active site and catalytic mechanism, which are reminiscent of secreted phospholipase A_2.

3.5. Mass spectrometry

Mass spectrometry (MS) has been widely applied within SPINE for quality assessment of protein quality, analysis of crystal content, domain identification and ligand screening (as discussed by Geerlof *et al.*, 2006). Several nodes including Strasbourg and Oxford used a combination of limited proteolysis and MS to delineate structural domains in cases where the structural determination of the full-length protein had failed. In the case of the p62 subunit from the general transcription/DNA-repair factor TFIIH, a module was identified which turned out to be absolutely required for DNA-repair activity through the nucleotide-excision repair pathway (Jawhari *et al.*, 2004). Its three-dimensional structure revealed an unpredicted pleckstrin homology and phosphotyrosine-binding (PH/PTB) domain, uncovering a new class of activity for this fold (Gervais *et al.*, 2004; Fig. 3*c*). Nondenaturing electrospray MS (ESI–MS) was used by the Strasbourg node to determine the characteristics of ligand and/or co-repressor peptide binding to the ligand-binding domains of nuclear receptors and to probe the stability of the nuclear receptor–co-repressor complexes prior to crystallization in the presence of different types of agonists or antagonists. This approach, based on supramolecular MS, also allowed the detection and identification of fortuitous ligands for the retinoic acid-related orphan receptor beta (RORβ) and for the ultraspiracle protein (USP). These fortuitous ligands were specifically captured from the host cell with the proper stoichiometry. After organic extraction, they were characterized by classical analytical methods and identified as stearic acid for RORβ and phosphatidylethanolamine (PE) for USP, as confirmed by crystallography. These molecules act as 'fillers' and may not be the physiological ligands, but they prove to be essential for stabilization of the active conformation of the LBD, thus permitting its crystallization (Potier *et al.*, 2003). The resulting crystal structures provide a detailed picture of the ligand-binding pocket, allowing the design of highly specific synthetic ligands that can be used to characterize the function of orphan nuclear receptors.

4. Selected highlights

SPINE has led to the determination of a very substantial number of three-dimensional structures of biomedical value, in particular in relation to cancer (kinesins, kinases, proteins from the ubiquitin pathway), to neurological development and neurodegenerative diseases and to immune recognition. Here, we summarize some highlights.

4.1. Cancer

4.1.1. Kinesin. Since the first description of kinesin about 20 years ago, genetic and biochemical methods have led to the discovery of many similar proteins, which now form the kinesin superfamily (Miki *et al.*, 2005). Members of this family are microtubule-associated motor proteins capable of converting the energy generated by ATP hydrolysis into mechanical work. They are involved in intracellular transport and cell division. Human kinesins are currently in the spotlight for the development of new anti-cancer drugs (Jordan & Wilson, 2004). There may be as many as 12 different motor proteins (out of the 41 different predicted kinesin-like motor proteins in the human genome) involved in various aspects of mitotic spindle assembly and function (Zhu *et al.*, 2005), the expression of many of them being restricted to proliferating tissues. Some of these mitotic motors are potential targets for inhibitors leading to mitotic arrest and are therefore interesting candidates for future chemotherapeutic applications (Wood *et al.*, 2001). The most advanced drug candidate targeting a kinesin motor is currently in Phase II clinical trials (Sakowicz *et al.*, 2004).

To provide a better understanding of these molecules, three-dimensional structures have been determined of kinesins relevant to human health or with unusual functional and structural properties, both alone and complexed with inhibitors and/or with associated proteins. The overall goal is to develop and apply interdisciplinary approaches (chemical, cellular, molecular, biochemical and structural approaches) for identifying new and improving existing inhibitors of human mitotic kinesins, with the long-range aim of developing inhibitors of chemotherapeutic value. This was used by the Grenoble partners to exemplify the parallel genome-wide approach to a biomedically important family of proteins. It led to (i) the structure determination of the human kinetochore-associated protein CENP-E (Garcia-Saez, Blot *et al.*, 2004; Garcia-Saez, Yen *et al.*, 2004), essential for some aspects of kinetochore microtubule attachments (Fig. 5*b*), (ii) the pseudo-atomic structure of the MT–CENP-E complex, solved by a combination of X-ray crystallography and three-dimensional image reconstruction (Neumann *et al.*, in preparation), (iii) the structure of human mitotic Eg5 in complex with a new potent antimitotic inhibitor, as well as production of a new crystal form of native Eg5 (I. Garcia-Saez & F. Kozielski, patent pending), and (iv) the structure of a human kinesin-associated protein in two different nucleotide states (Garcia-Saez *et al.*, in preparation).

4.1.2. Kinases. Atypical protein kinase Cs (aPKCs) play critical roles in signalling pathways that control cell growth, differentiation and survival. Therefore, they constitute attractive targets for the development of novel therapeutics against cancer. These molecules were used by the Munich node as a test for the development of more efficient cloning and expression methods in eukaryotic cells (baculovirus technology), leading to the determination of the crystal structure of the catalytic domain of atypical PKCι in complex with the bis(indolyl)maleimide inhibitor BIM1 (Messerschmidt *et al.*, 2005). The overall structure exhibits the classical bilobal kinase fold and is in its fully activated form (Fig. 5*c*). Both phosphorylation sites (Thr403 in the activation loop and Thr555 in the turn motif) are well defined in the structure and form intramolecular ionic contacts that make an important contribution to stabilization of the active conformation of the catalytic subunit. The phosphorylation site in the hydrophobic motif of atypical PKCs is replaced by the phosphorylation mimic glutamate and this is also seen clearly in the structure of PKCι (residue 574). This structure determination provides for the first time the architecture of the turn motif phosphorylation site, which is characteristic of PKCs and PKB/AKT and is completely different to that in PKA. The bound BIM1 inhibitor blocks the ATP-binding site and places the kinase domain in an intermediate open conformation. The PKCι–BIM1 complex is the first kinase domain crystal structure of any atypical PKC and constitutes the basis for rational drug design for selective PKCι inhibitors.

4.1.3. Nuclear receptors. Small molecules such as retinoids, steroid hormones, fatty acids, cholesterol metabolites or xenobiotics are involved in the regulation of numerous physiological and patho-physiological processes by binding to and controlling the activity of members of the nuclear receptor (NR) superfamily of transcription factors. In addition to natural ligands, synthetic agonists and antagonists have been identified that in some cases specifically target NR isotypes or elicit tissue-specific, signalling pathway-specific or promoter-selective transcriptional responses. Since the structure determination of apo-retinoid X receptor (RXR), holo-retinoic acid receptor (RAR) and holo-thyroid receptor (TR) ligand-binding domains (LBDs) in the mid-1990s, the major principles of ligand-dependent NR action and determinants of (isotype-) selective ligand binding have been revealed. The application of a coherent set of HTP approaches by the Strasbourg node (see Alzari *et al.*, 2006; Aricescu, Assenberg *et al.*, 2006) led to the structure determination of more than 20 crystal structures (novel structures or primary examples of protein–ligand complexes) from both classical nuclear receptors such as RAR, vitamin D receptor (VDR), oestrogen receptor (ER) and orphan receptors such as RXR, oestrogen-related receptor (ERR) or retinoic acid-related orphan receptor (ROR). ERR is an example of an orphan receptor (a natural ligand not yet having been identified) that is constitutively active. The gamma isotype (ERRγ) is expressed in a variety of foetal cell lines and adult tissues, including brain, lung, kidney, bone marrow and spinal cord. Previous studies identified diethylstilbestrol (DES), an oestrogen-receptor (ER) agonist, as an antagonist for all ERR isotypes. The structure determination of the ERRγ LBD revealed a transcriptionally active LBD conformation in complex with a SRC1 coactivator peptide in the absence of any ligand (Greschik *et al.*, 2002). The ERRγ ligand-binding pocket is the smallest observed so far in the active receptor conformation. Point mutants of ERRγ in which the ligand-binding pocket is either significantly enlarged or filled up by bulkier side chains remain constitutively active. Therefore, the transcriptional activity of ERRγ does not depend on the presence of an endogenous agonist. The crystal structure of the ERRγ–DES complex showed that the binding of DES to ERRγ results in a conformational change of a single phenylalanine residue (Phe435) which partially fills the ligand-binding pocket in the apo-LBD and thus interferes with the position of His12 in the transcriptionally active LBD conformation (Greschik *et al.*, 2004; Fig. 5*d*). Accordingly, the mutation of Phe435 to leucine (the corresponding residue in ERs) abolishes the antagonist action of DES on ERRγ.

NRs function either as homodimers or as heterodimers with retinoid X receptor (RXR). A structure-based sequence analysis aimed at discovering the molecular mechanism that controls the dimeric association of the ligand-binding domain reveals two sets of differentially conserved residues, which partition the entire NR superfamily into two classes related to their oligomeric behaviour (Brelivet *et al.*, 2004). Site-directed mutagenesis confirms the functional importance of these residues for the dimerization process and/or transcriptional activity. All homodimers belong to class I, in which the related residues contribute a communication pathway of two salt bridges linking helix 1 on the cofactor-binding site to the dimer interface (Fig. 5*e*).

4.1.4. Proteases. Human tissue kallikrein 4 (hK4) belongs to a 15-member multigene family of closely related serine proteinases that are up- or down-regulated in certain cancers. hK4 itself is specifically expressed in prostate and ovary. The Munich node has purified and characterized recombinant hK4, identifying active monomers and inactive oligomers and establishing micromolar zinc as an inhibitor of the enzymatic activity. hK4 has been crystallized in the presence of zinc, nickel and cobalt ions in three crystal forms containing cyclic tetramers and octamers. These structures confirm the trypsin-like fold of hK4, but display a novel zinc site between His25 and Glu77 that links the 70–80 loop with the N-terminal segment (Debela *et al.*, 2006). This link suggests a mechanism by which zinc modulates the catalytic activity of hK4 *via* effects on Ile16, which induce a functional active site. Whereas hK4 lacks the typical ‘kallikrein-loop’ sequence, the unusual conformation of its 99-loop creates a groove-like acidic S2 subsite. These findings explain the amino-acid substrate profile of hK4 for the subsites S1 to S4. In addition, the structural analysis reveals that hK4 exhibits negatively charged surface patches suggestive of a putative exosite for prime-side substrate recognition.

Matrix metalloproteinases (MMPs) are a class of proteases which are responsible for the degradation of the extracellular matrix. They are multi-domain proteins in which the catalytic domain requires a zinc ion to provide an active catalytic centre. Some of these proteins are well validated pharmacological targets as they are involved in a variety of diseases such as cancer, inflammation, multiple sclerosis *etc.* Within SPINE, a number of catalytic domains of MMPs (MMP1, MMP3, MMP10, MMP12 and MMP8) were studied by NMR and X-ray crystallography and several structures were determined with both techniques by the Florence and Munich nodes. In the case of MMP12 (Fig. 3*b*), the comparison between the solution and X-ray structures, free or in complex with different inhibitors, provided key information on the protein dynamics and flexibility, molecular characteristics which can be relevant for inhibitor binding (Bertini *et al.*, 2005). The structural characterization of a series of protein–ligand adducts

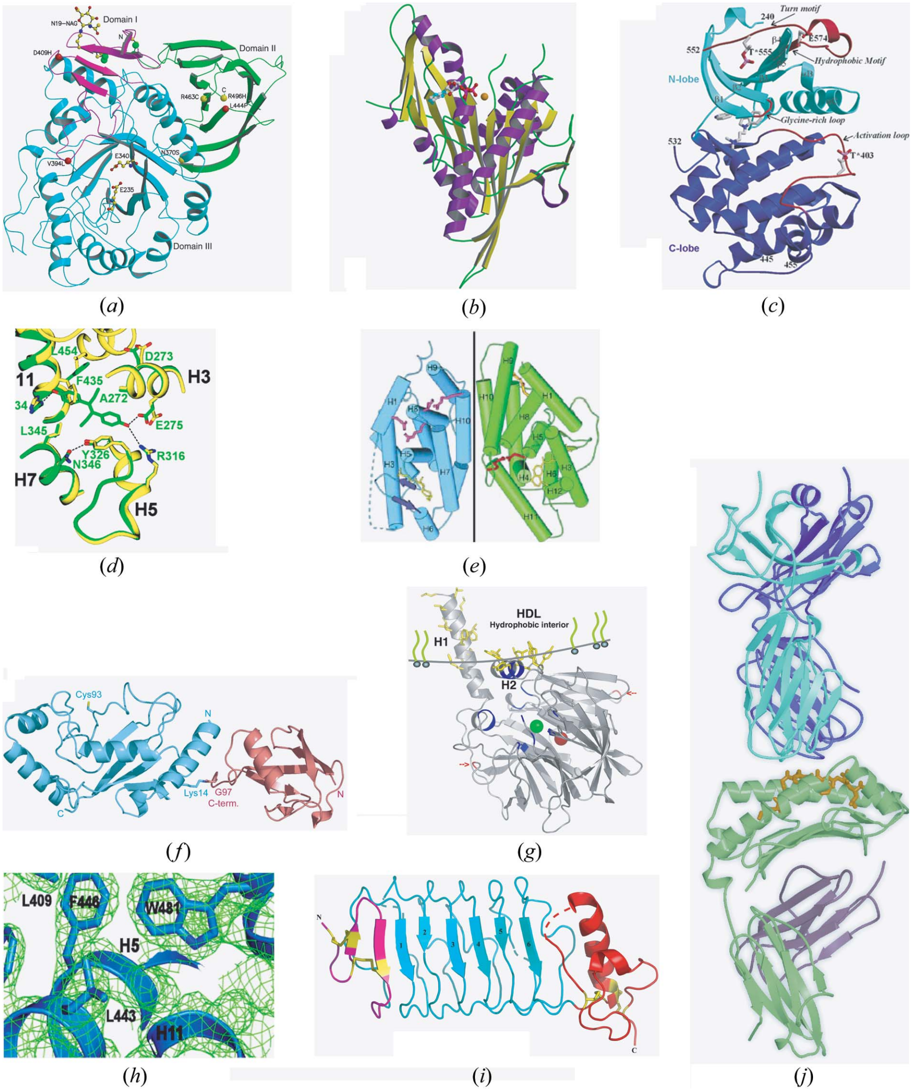

Figure 5
Portfolio of X-ray structures from WP10 and WP11. (*a*) PON-1. (*b*) Human kinesin CENP-E with MgADP in the active site. (*c*) Catalytic domain of the atypical protein kinase C iota. (*d*) ERR LBD–DES complex (green) superimposed with the ERR apoLBD (yellow). (*e*) RXRα/RARα LBD heterodimer showing the communication pathways in nuclear receptors. (*f*) Ubiquitin-conjugating enzyme E2-25K–SUMO complex. (*g*) GlcCerase. (*h*) NGFI-B ligand-binding pocket. (*i*) Slit2 (third leucine-rich repeat domain). (*j*) pMHC–TCR complex.

provided the necessary structural background for rational drug design.

4.1.5. Ubiquitination. Ubiquitin (Ub) conjugation of cellular proteins plays an important role in many biological processes such as cell-cycle progression, DNA repair, transcriptional activation, oncogenesis and intracellular proteolysis. Considerable progress has been made in the understanding of Ub conjugation and its roles in regulating protein modification and degradation. The process of conjugation with ubiquitin and ubiquitin-like molecules is mediated through an E1/E2/E3 cascade of enzymes. To provide better understanding of the specificity and selectivity of this system, a number of enzymes and target complexes in the ubiquitin-(like) cascade have been studied by the Amsterdam node, studies which have showcased the application of HTP technologies (in particular crystallization; see Berry *et al.*, 2006) The crystal structures of the human E2 protein Rad6 and several active-site mutants of Rad6 were determined. The ubiquitin E2 E2-25K was established to be a target for the ubiquitin-like protein SUMO and the crystal structures of native and SUMO-modified E2-25K were determined (Fig. 5*f*). These structures revealed that the modification did not result in structural changes (Pichler *et al.*, 2005). Through a SPINE-based collaboration, the Amsterdam node also established the modification site to be Lys14, a surprising finding since this is a non-consensus site located in the context of several SUMO consensus sites. In the three-dimensional structure this residue is within a helix and it was demonstrated that the same sequence in the context of an unfolded peptide is modified at the traditional consensus site, rather than on Lys14. This showed that the structural context is critical for modification with SUMO.

4.2. Neurological diseases

4.2.1. Copper-binding proteins. Several proteins involved in neurological diseases bind copper ions or their level of expression and/or activity is modulated by copper. As part of the HTP technologies developed in SPINE (see AB *et al.*, 2006), the Florence node has studied the copper ATPase7A and 7B, also called Menkes and Wilson proteins as mutations on them are responsible for the corresponding fatal or serious diseases. These ATPases are membrane-bound proteins which feature six soluble domains, each of them capable of copper binding. The solution structures of four single domains (2, 3, 5 and 6) of ATPase7A and of domains 5–6 of 7B were determined as well as that of the soluble copper chaperone. In particular, this work yielded atomic level insights into the effects of the disease-causing mutation A629P which occurs in the last of the six copper(I)-binding domains (Banci *et al.*, 2005). The A629P mutation reduces the affinity for copper(I) and makes the protein β-sheet more solvent-accessible, possibly resulting in an enhanced susceptibility of ATP7A to proteolytic cleavage and/or in reduced capability of copper(I) translocation.

4.2.2. Structural basis for Gaucher disease. The most common lysosomal storage disease, Gaucher disease, is caused by mutations in the gene coding for acid-β-glucosidase (GlcCerase; Futerman *et al.*, 2004). Type 1 Gaucher disease is characterized by hepatosplenomegaly and types 2 and 3 by early or chronic onset of severe neurological symptoms. No clear correlation exists between the ~200 GlcCerase mutations and disease severity, although homozygosity for the common mutations, N370S and L444P, is associated with non-neuronopathic and neuronopathic disease, respectively.

By a procedure that involved partial deglycosylation and used the HTP microbatch crystallization protocols developed within SPINE (see Berry *et al.*, 2006), the Weizmann team were able to crystallize GlcCerase and subsequently to solve its X-ray structure at 2.0 Å resolution (Dvir *et al.*, 2003; Fig. 5*g*). The catalytic domain consists of a $(\beta/\alpha)_8$ TIM-barrel, as expected for a member of the glucosidase hydrolase A family. The distance between the catalytic residues Glu235 and Glu340 is consistent with a catalytic mechanism of retention. Asn370 is located on the longest α-helix (helix 7), which has several other mutations of residues that point into the TIM-barrel. Helix 7 is at the interface between the TIM-barrel and a separate immunoglobulin-like domain on which Leu444 is located, suggesting an important regulatory or structural role for this non-catalytic domain. More recently, the structure of GlcCerase conjugated with the irreversible inhibitor conduritol-B-epoxide was solved (Premkumar *et al.*, 2005). Two alternative conformations were distinguished for a pair of flexible loops located at the entrance to the active site and analysis of the dynamics suggests that they act as a lid at the entrance to the active site. This is supported by a cluster of mutations in loop 394–399 that cause Gaucher disease by reducing catalytic activity. Thus, the native and complex structures reveal the possibility of engineering improved GlcCerase for enzyme-replacement therapy and for designing structure-based drugs aimed at restoring the activity of defective GlcCerase.

4.2.3. NGFI-B. The human NR4A subfamily of nuclear receptors comprises three members: NGFI-B, Nurr1 and NOR1. NGFI-B is a ligand-independent orphan nuclear receptor expressed in various tissues, notably the brain, and plays a role in multiple cellular events such as cell proliferation, differentiation and apoptosis. This receptor has been implicated in neurodegenerative pathologies such as Parkinson's disease, manic depression and schizophrenia. Nurr1 is almost exclusively expressed in brain, is essential for the development of midbrain dopaminergic neurons and, accordingly, is also linked to Parkinson's disease. NOR-1 has been isolated from cultured forebrain neurons undergoing apoptosis and plays a role in brain development, but also acts in other tissues outside the central nervous system.

To gain insight into the structural basis for the distinct activation potentials, the Strasbourg team determined the crystal structure of the NGFI-B ligand-binding domain (LBD) at 2.4 Å resolution (Flaig *et al.*, 2005). The NGFI-B LBD adopts the canonical three-layered α-helical sandwich fold with the helix H12 in an active position. The ligand-binding pocket does not contain a ligand, but instead is filled with bulky aromatic residues (Fig. 5*h*). Superimposition with the Nurr1 LBD revealed a significant shift of the position of helix 12 potentially caused by conservative amino-acid exchanges in helix 3 or helix 12. Replacement of the helix 11–12 region of Nurr1 with that of NGFI-B dramatically reduces the transcriptional activity of the Nurr1 LBD. Similarly, mutation of Met414 in helix 3 to leucine or of Leu591 in helix 12 to isoleucine (the corresponding residues found in NGFI-B) significantly affects Nurr1 transactivation. In comparison, swapping the helix 11–12 region of Nurr1 into NGFI-B results in a modest increase of activity. These observations reveal that LBD activity is highly sensitive to changes that influence helix 12 positioning. Furthermore, mutation of hydrophobic surface residues in the helix 11–12 region (outside the canonical co-activator surface constituted by helices 3, 4 and 12) severely affects Nurr1 transactivation. Taken together, these data suggest that a novel co-regulator surface that includes helix 11 and a specifically positioned helix 12 determine the cell-type-dependent activities of the NGFI-B and the Nurr1 LBDs.

4.2.4. SLIT axon-guidance proteins. Developing axons must navigate through the embryo by processing a number of different signals in their immediate environment. Slit and Roundabout (Robo) provide a key ligand–receptor interaction for such a process during neuronal development, especially at the midline of the central nervous system of vertebrate and invertebrates (Stein & Tessier-Lavigne, 2001). Slits are large multi-domain proteins secreted by glial cells and Robos are large multi-domain transmembrane receptors expressed on the axonal growth cones. Slits contain four tandem leucine-rich repeat (LRR) domains at their N-terminus, domains that are well known to mediate protein–protein interactions. Recent studies have revealed that the second LRR domain (D2) is essential for its interaction with Robo and that the fourth LRR domain (D4) may play an important role in dimerization (Howitt *et al.*, 2004).

The Grenoble team determined the structure of the third LLR from mammalian Slit2 (McCarthy *et al.*, in preparation; Fig. 5*i*). This domain is composed of six LRR repeats flanked at the N- and C-termini by cysteine-rich capping domains, often found in secreted LRR proteins. The mammalian Slit2 D3 differs from the recently published *Drosophila* Slit D3 (Howitt *et al.*, 2004) in that it contains an additional LRR repeat. Two potential glycosylation sites were identified in the primary sequence of Slit2 D3. However, only one of these is glycosylated and is located on the convex face. The mammalian Slit D3 sequences are well conserved, with 60% identity over ~220 residues, with the best conservation observed on the concave face. This is not surprising because LRR proteins are well known to generally mediate their protein–protein interactions *via* their concave face. It is thus plausible that the concave face is able to mediate an interaction between mSlits and other as yet unidentified co-receptors. These are quite likely to exist given the growing importance of Slit proteins in neurogenesis, angiogenesis, immune response and carcinogenesis.

4.3. The immune system

4.3.1. TCR–MHC class I complexes. As noted earlier in this report, the extracellular regions of T-cell receptors (TCR) are responsible for a primary event in the human cellular immune response: the recognition of peptide antigens presented by MHC class I and class II molecules. Peptide antigens presented by MHC class I molecules can originate from viral proteins expressed within a cell because it is infected or from proteins distinctive to a tumour cell. The binding of TCRs (on the surface of a cytotoxic T cell) to such peptides displayed in complex with MHC class I molecules (on the surface of a target cell) flags the target cell for destruction by the T cell. A molecular-level understanding of these recognition events is therefore of considerable relevance for the design of therapeutic strategies (*e.g.* T-cell vaccines) which aim to modulate/boost the cellular immune response to combat infectious diseases such as AIDs as well as pathologies such as cancer. In Oxford, exploitation of miniaturized and parallelized approaches to protein production and crystallization developed in SPINE has resulted in crystal structures for a series of human MHC class I–peptide–TCR complexes which includes complexes of TCRs recognizing two different tumour antigens (Chen *et al.*, 2005; Hamer *et al.*, in preparation) and TCRs recognizing three different AIDS-virus antigens (Stewart-Jones *et al.*, in preparation; Fig. 5*j*).

For the first of the tumour-antigen studies the structural analysis is already interwoven into a strategy for vaccine design (Chen *et al.*, 2005). In this study, the crystal structure of the immunodominant HLA-A2 tumour epitope (NY-ESO-$1_{157–165}$; SLLMWITQ**C**) complexed with a specific TCR revealed that TCR binding centres on a prominent pair of hydrophobic side chains (methionine-tryptophan) located at positions 4 and 5 in the peptide. A second structure showed that optimization of cysteine to valine at peptide position 9 results in improved peptide shape complementarity to both HLA-A2 and TCR. Binding analyses confirmed tighter binding of the analogue peptide (SLLMWITQ**V**) to HLA-A2 and a threefold to fourfold improved binding for the soluble TCR, while a series of functional analyses established that the changes in the biophysical characteristics of the recognition system resulted in enhanced immunogenicity. These studies thus provide a basis for the rational optimization of peptides for use in clinical trials to boost anti-tumour immune responses.

5. Conclusion

This first large-scale EC Integrated Project on Structural Proteomics has contributed to a evolution in the way structural biology is now being carried out in Europe through the democratization of the use of new technologies (*e.g.* affordable expression-screening and nano-crystallization robots) and the development of novel strategies (*e.g.* ligation-free cloning, use of directed evolution protocols or eukaryotic expression systems for the production of soluble protein, MS-assisted ligand screening) at various steps of the structure-determination pipeline. By its policy of promoting an open decentralized network and focusing on high-value targets, SPINE has gone beyond the potentially divisive dichotomy between the 'traditional' way of doing structural biology ('one post-doc/one project' with in-depth complementary functional investigations) and 'factory-style' structural genomics (multiple parallel projects, abandonment of failures, targets of often unknown function). As illustrated by the examples presented in this paper, the SPINE mode of work in which HTP techniques have been exploited for high-value targets is having a significant impact on the structural biology of human-related proteins, thereby providing valuable information which can influence complementary biological and clinical studies.

This work forms part of SPINE (Structural Proteomics In Europe), contract No. QLG2-CT-2002-00988, funded by the European Commission under the Integrated Programme 'Quality of Life and Management of Living Resources'.

References

AB, E. *et al.* (2006). *Acta Cryst.* D**62**, 1150–1161.

Aharoni, A., Gaidukov, L., Yagur, S., Toker, L., Silman, I. & Tawfik, D. S. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 482–487.

Alzari, P. M. *et al.* (2006). *Acta Cryst.* D**62**, 1103–1113.

Aricescu, A. R., Assenberg, R. *et al.* (2006). *Acta Cryst.* D**62**, 1114–1124.

Aricescu, A. R., Lu, W. & Jones, E. Y. (2006). *Acta Cryst.* D**62**, 1243–1250.

Banci, L., Bertini, I., Cantini, F., Migliardi, M., Rosato, A. & Wang, S. (2005). *J. Mol. Biol.* **352**, 409–417.

Berry, I. M., Dym, O., Esnouf, R. M., Harlos, K., Meged, R., Perrakis, A., Sussman, J. L., Walter, T. S., Wilson, J. & Messerschmidt, A. (2006). *Acta Cryst.* D**62**, 1137–1149.

Bertini, I., Calderone, V., Cosenza, M., Fragai, M., Lee, Y. M., Luchinat, C., Mangani, S., Terni, B. & Turano, P. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 5334–5339.

Brelivet, Y., Kammerer, S., Rochel, N., Poch, O. & Moras, D. (2004). *EMBO Rep.* **5**, 423–429.

Chen, J. L., Stewart-Jones, G., Bossi, G., Lissin, N. M., Wooldridge, L., Choi, E. M., Held, G., Dunbar, P. R., Esnouf, R. M., Sami, M., Boulter, J. M., Rizkallah, P., Renner, C., Sewell, A., van der Merwe, P. A., Jakobsen, B. K., Griffiths, G., Jones, E. Y. & Cerundolo, V. (2005). *J. Exp. Med.* **201**, 1243–1255.

Debela, M., Magdolein, V., Grimminger, V., Sommerhoff, C., Messerschmidt, A. & Huber, R. (2006). Submitted.

Draganov, D. I. & La Du, B. N. (2004). *Naunyn Schmiedebergs Arch. Pharmacol.* **369**, 78–88.

Dvir, H., Harel, M., McCarthy, A. A., Toker, L., Silman, I., Futerman, A. H. & Sussman, J. L. (2003). *EMBO Rep.* **4**, 704–709.

Flaig, R., Greschik, H., Peluso-Iltis, C. & Moras, D. (2005). *J. Biol. Chem.* **280**, 19250–19258.

Futerman, A. H., Sussman, J. L., Horowitz, M., Silman, I. & Zimran, A. (2004). *Trends Pharmacol. Sci.* **25**, 147–151.

Garcia-Saez, I., Blot, D., Kahn, R. & Kozielski, F. (2004). *Acta Cryst.* D**60**, 1158–1160.

Garcia-Saez, I., Yen, T., Wade, R. H. & Kozielski, F. (2004). *J. Mol. Biol.* **340**, 1107–1116.

Geerlof, A. *et al.* (2006). *Acta Cryst.* D**62**, 1125–1136.

Gervais, V., Lamour, V., Jawhari, A., Frindel, F., Wasielewski, E., Dubaele, S., Egly, J. M., Thierry, J. C., Kieffer, B. & Poterszman, A. (2004). *Nature Struct. Mol. Biol.* **11**, 616–622.

Greschik, H., Flaig, R., Renaud, J. P. & Moras, D. (2004). *J. Biol. Chem.* **279**, 33639–33646.

Greschik, H., Wurtz, J. M., Sanglier, S., Bourguet, W., van Dorsselaer, A., Moras, D. & Renaud, J. P. (2002). *Mol. Cell*, **9**, 303–313.

Harel, M., Aharoni, A., Gaidukov, L., Brumshtein, B., Khersonsky, O., Meged, R., Dvir, H., Ravelli, R. B., McCarthy, A., Toker, L., Silman, I., Sussman, J. L. & Tawfik, D. S. (2004). *Nature Struct. Mol. Biol.* **11**, 412–419.

Howitt, J. A., Clout, N. J. & Hohenester, E. (2004). *EMBO J.* **23**, 4406–4412.

Jawhari, A., Boussert, S., Lamour, V., Atkinson, R. A., Kieffer, B., Poch, O., Potier, N., van Dorsselaer, A., Moras, D. & Poterszman, A. (2004). *Biochemistry*, **43**, 14420–14430.

Jordan, M. A. & Wilson, L. (2004). *Nature Rev. Cancer*, **4**, 253–265.

Messerschmidt, A., Macieira, S., Velarde, M., Badeker, M., Benda, C., Jestel, A., Brandstetter, H., Neuefeind, T. & Blaesse, M. (2005). *J. Mol. Biol.* **352**, 918–931.

Miki, H., Okada, Y. & Hirokawa, N. (2005). *Trends Cell. Biol.* **15**, 467–476.

Pichler, A., Knipscheer, P., Oberhofer, E., van Dijk, W. J., Korner, R., Olsen, J. V., Jentsch, S., Melchior, F. & Sixma, T. K. (2005). *Nature Struct. Mol. Biol.* **12**, 264–269.

Potier, N., Billas, I. M., Steinmetz, A., Schaeffer, C., van Dorsselaer, A., Moras, D. & Renaud, J. P. (2003). *Protein Sci.* **12**, 725–733.

Premkumar, L., Sawkar, A. R., Boldin-Adamsky, S., Toker, L., Silman, I., Kelly, J. W., Futerman, A. H. & Sussman, J. L. (2005). *J. Biol. Chem.* **280**, 23815–23819.

Sakowicz, R., Finer, J. T., Beraud, C., Crompton, A., Lewis, E., Fritsch, A., Lee, Y., Mak, J., Moody, R., Turincio, R., Chabala, J. C., Gonzales, P., Roth, S., Weitman, S. & Wood, K. W. (2004). *Cancer Res.* **64**, 3276–3280.

Siebold, C., Berrow, N., Walter, T. S., Harlos, K., Owens, R. J., Stuart, D. I., Terman, J. R., Kolodkin, A. L., Pasterkamp, R. J. & Jones, E. Y. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 16836–16841.

Stein, E. & Tessier-Lavigne, M. (2001). *Science*, **291**, 1928–1938.

Stewart-Jones, G. B., McMichael, A. J., Bell, J. I., Stuart, D. I. & Jones, E. Y. (2003). *Nature Immunol.* **4**, 657–663.

Walter, T. S., Diprose, J. M., Mayo, C. J., Siebold, C., Pickford, M. G., Carter, L., Sutton, G. C., Berrow, N. S., Brown, J., Berry, I. M., Stewart-Jones, G. B., Grimes, J. M., Stammers, D. K., Esnouf, R. M., Jones, E. Y., Owens, R. J., Stuart, D. I. & Harlos, K. (2005). *Acta Cryst.* D**61**, 651–657.

Wood, K. W., Cornwell, W. D. & Jackson, J. R. (2001). *Curr. Opin. Pharmacol.* **1**, 370–377.

Zhu, C., Zhao, J., Bibikova, M., Leverson, J. D., Bossy-Wetzel, E., Fan, J. B., Abraham, R. T. & Jiang, W. (2005). *Mol. Biol. Cell*, **16**, 3187–3199.