PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2006) **361**, 413–423
doi:10.1098/rstb.2005.1800
Published online 3 February 2006

Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery

Tom L. Blundell[1,2,*], Bancinyane L. Sibanda[1], Rinaldo Wander Montalvão[1], Suzanne Brewerton[1,2], Vijayalakshmi Chelliah[1], Catherine L. Worth[1], Nicholas J. Harmer[1], Owen Davies[1] and David Burke[1]

[1]*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK*
[2]*Astex Technology, 436 Cambridge Science Park, Milton Road, Cambridge CB4 0QA, UK*

Impressive progress in genome sequencing, protein expression and high-throughput crystallography and NMR has radically transformed the opportunities to use protein three-dimensional structures to accelerate drug discovery, but the quantity and complexity of the data have ensured a central place for informatics. Structural biology and bioinformatics have assisted in lead optimization and target identification where they have well established roles; they can now contribute to lead discovery, exploiting high-throughput methods of structure determination that provide powerful approaches to screening of fragment binding.

Keywords: structural biology; structural bioinformatics; structure-based drug design; high-throughput crystallography; virtual screening; multiprotein complexes

1. BACKGROUND

Ideas about the use of X-ray crystallography in drug discovery emerged more than 30 years ago as the first three-dimensional structures of proteins were determined. These ideas included the synthesis of ligands of haemoglobin to decrease sickling (Beddell *et al.* 1976; Goodford *et al.* 1980), the chemical modification of insulins to increase half lives in circulation (Blundell 1972), and the design of inhibitors of serine proteases to control blood clotting. However, apart from an early venture in 1975 by the UK Wellcome Foundation programme (Beddell *et al.* 1976), most pharmaceutical companies considered X-ray crystallography too expensive and time consuming to bring 'in house' and for a time most activity remained in academia.

Within a decade, a radical change in drug design had begun, incorporating the knowledge of the three-dimensional structures of target proteins into the design process. Although structures of the relevant drug targets were usually not available directly from X-ray crystallography, comparative models based on homologues proved useful in defining topographies of the complementary surfaces of ligands and their protein targets, and began to be exploited in lead optimization in the 1980s (Blundell *et al.* 1983; Blundell 1996; Campbell 2000). Eventually crystal structures of key drug targets became available; AIDS drugs such as Agenerase and Viracept were developed using the crystal structure of HIV protease (Lapatto *et al.* 1989; Miller *et al.* 1989); and the influenza drug Relenza was designed using the crystal structure of neuraminidase (Varghese 1999). More than 40 drugs originating from structure-based design approaches have now entered clinical trials (Hardy & Malikayil 2003), and seven of these had achieved regulatory approval and been marketed as drugs by mid-2003.

Protein structure can influence drug discovery at every stage in the design process (figure 1). Classically it has been exploited in lead optimization, a process that uses structure to guide the chemical modification of a lead molecule to give an optimized fit in terms of shape, hydrogen bonds and other non-covalent interactions with the target. Protein structure can also be used in target identification and selection (the assessment of the 'druggability' or tractability of a target). Traditionally, this has involved homology recognition assisted by knowledge of protein structure; but now structural genomics programmes are seeking to define representative structures of all protein families, allowing proposals of binding regions and molecular functions. More recently, X-ray crystallography has been used to assist the identification of hits by virtual screening and more directly in the screening of chemical fragments. The key roles of structural biology and bioinformatics in lead optimization remain as important as ever (Whittle & Blundell 1994; Lombardino & Lowe 2004). Here, we focus on their roles in target identification and lead discovery.

2. TARGET IDENTIFICATION FROM SEQUENCE-STRUCTURE HOMOLOGY RECOGNITION

Protein structures are a rich source of information about membership of families and superfamilies. It is such divergently evolved proteins that need to be

* Author for correspondence (tom@cryst.bioc.cam.ac.uk).

One contribution of 15 to a Discussion Meeting Issue 'Bioinformatics: from molecules to systems'.

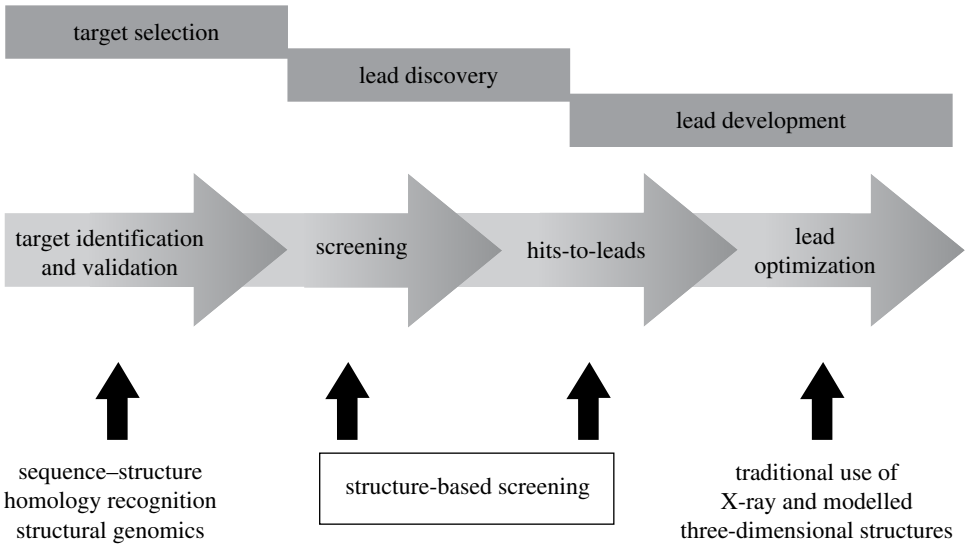

Figure 1. Drug discovery classically follows the path from target selection, through lead discovery to lead development. Although structural biology has historically had a role in the final stages during lead optimization, it is now having an effect at all stages. Homology recognition and structural genomics can aid target selection, while structure-based screening assists the lead discovery and development processes.

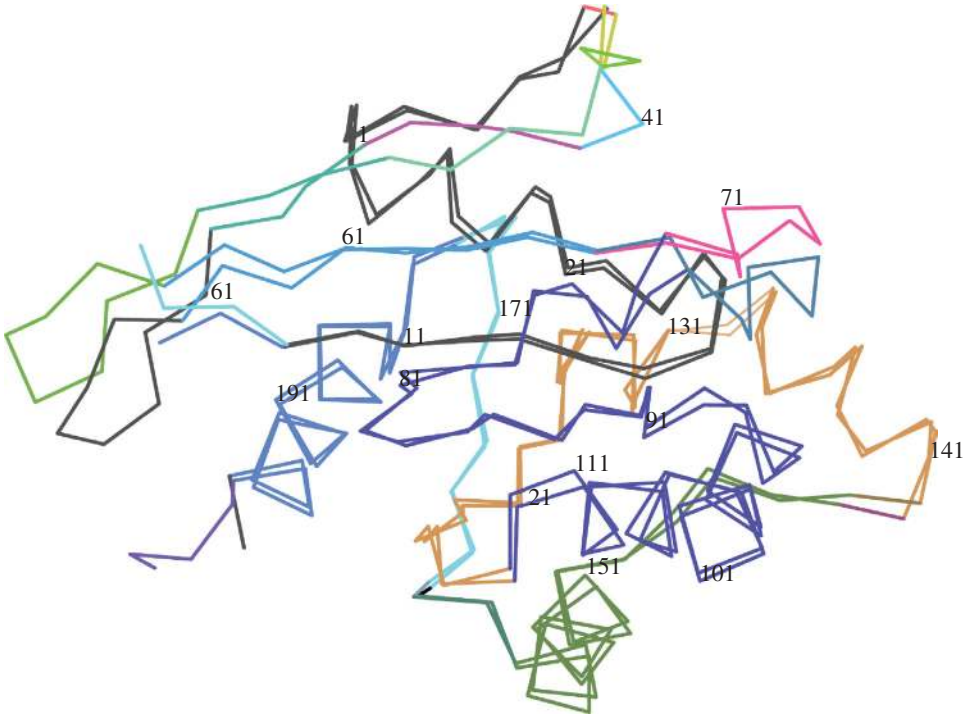

Figure 2. An example of structurally conserved clusters for an ensemble of superposed structures. Two members of the GTP-binding protein family are shown (Pdb codes: 1ftn; 1a4r). Regions with the same colours belong to structurally conserved clusters.

recognized as they are most likely to exhibit similar structure and function. Thus, we are interested in 'homology recognition' rather than 'fold recognition'. A classical example of this process was the recognition of HIV proteinase as a distant member of the pepsin/renin superfamily and the subsequent modelling of its three-dimensional structure and the design of inhibitors (Pearl & Taylor 1987; Blundell 1988). In general, putative relatives are identified, the sequences aligned, and the three-dimensional structures modelled. This is usually helpful in proposing binding sites and molecular functions if key residues are conserved.

Methods for the recognition of distant homologues through sequence-structure matching can be classified either as profile methods or threading. The profile methods introduce structural information into traditional sequence comparison algorithms often using structure-dependent propensities (Bowie *et al.* 1991; Rice & Eisenberg 1997). Our approach has been to exploit environment-dependent substitution matrices and gap penalties (Overington *et al.* 1990; Overington *et al.* 1992) in a computer program known as FUGUE (Shi *et al.* 2001). This very effectively improves the recognition of distant homologues that are members of superfamilies. Threading, on the other hand, fits a probe sequence onto the backbone of a known structure, evaluating the compatibility between the sequence and the proposed structure by means of a set

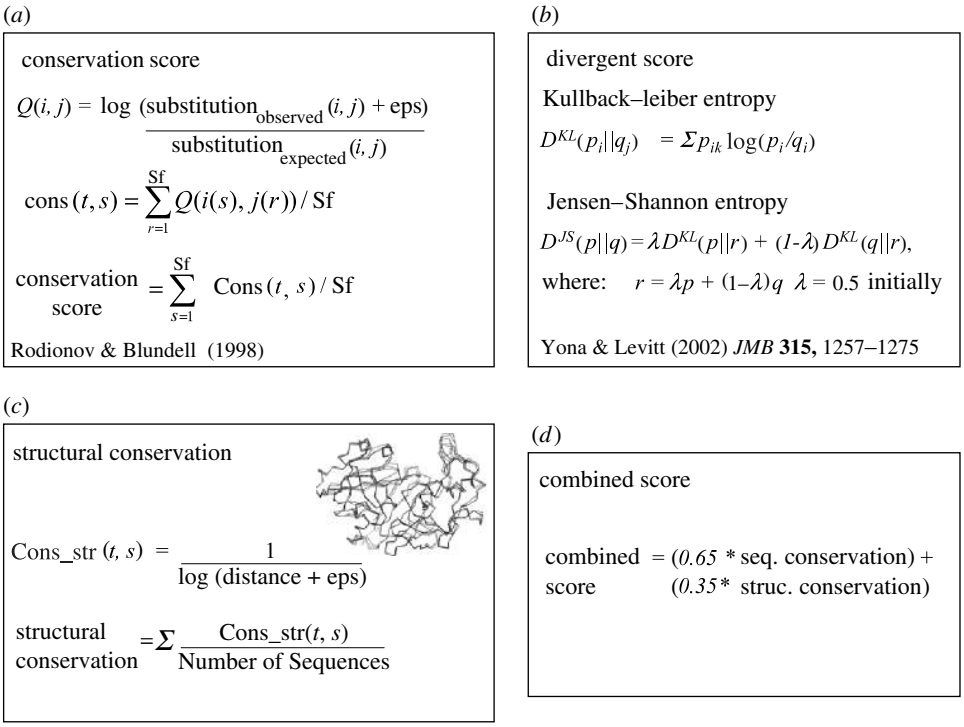

Figure 3. The formula for calculating the two sequence-based scoring systems (*a*) conservation score, (*b*) divergent score and the structure based score (*c*) structural conservation score are shown. (*d*) The empirically determined weights of the sequence based score and structural conservation score, which improves the functional site prediction, are shown.

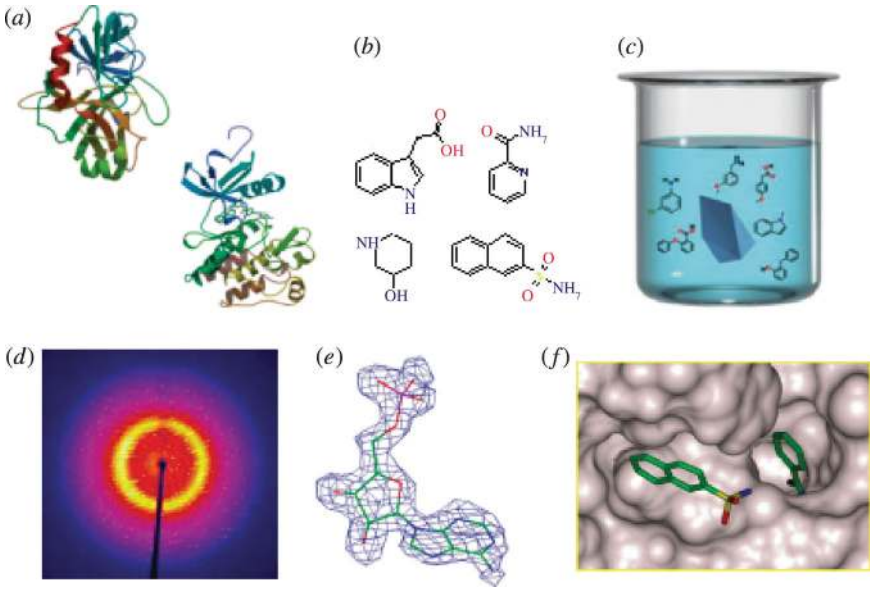

Figure 4. The Pyramid system allows lead discovery through a fragment-based approach of molecular fragment matching and fitting. (*a*) High resolution target structure determination. (*b*) Generation of Astex drug fragment library. Virtual screening used to enrich the library for fragments likely to bind the target. (*c*) Drug fragment cocktails used for protein crystal soaks, 4–8 compounds per cocktail. (*d*) High throughput protein/ligand X-ray crystallography. Automated X-ray data collection and analysis. (*e*) Electron density analysed by AutoSolve in order to identify bound drug fragment. (*f*) Structure-based optimization of hits to leads.

of empirical potentials, which are derived from well-resolved protein structure data (Jones *et al.* 1992). This method remains a powerful tool for fold recognition, but sequence-structure comparison methods using profiles offer better homology recognition performance (Lindahl & Elofsson 2000). Combined algorithms have been reported; for example, GenTHREADER uses the sequence comparison method to generate the sequence-structure alignment and then evaluates the alignment using threading potentials (Jones 1999).

Once a homologue of known structure has been identified it can be modelled using a variety of comparative (homology) modelling procedures. For example, those that use a fragment-assembly approach such as COMPOSER (Sutcliffe *et al.* 1987) or 3D-JIGSAW (Bates & Sternberg 1999) or alternatively a restraint-based approach such as MODELLER

(Sali & Blundell 1993), which uses probability density functions derived from homologous structures and general features, obtained from the statistical analysis of a large numbers of known protein structures. Programs based on the satisfaction of spatial constraints usually produce complete models but with variable quality in different regions. These problems occur due to inadequacies in handling deletions and especially insertions where constraints cannot easily be derived from homologues or other proteins of known structure. Also, these programs normally need more computational power and time in order to process and generate the model. Both approaches give good models if the sequence identity is greater than 30% but the accuracy falls off sharply when it is lower, mainly due to the difficulty in obtaining good alignments, predicting shifts of core residues and building loops (Venclovas *et al.* 2003).

Our recent attempts to improve comparative modelling have centred on the idea of structurally conserved clusters (SCCs; R. W. Montalvão *et al.* 2005, unpublished results). One problem associated with fragment assembly approaches is that structural conserved regions are usually defined as regions where all proteins in the same family show the same conformation for the main chain atoms independently of their classification in a secondary structure element or loop region. This implies that the length of the structural conserved region tends to be proportional to the family percentage sequence identity (PID) and inversely proportional to the number of its members. A further problem is that superposition of Cα atoms often leads to equivalencing regions of quite different conformation. In order to overcome these difficulties, we have developed a computer program called CHORAL, which uses a knowledge-based method comprised of an amalgam of differential geometry and pattern recognition algorithms to identify the conserved structural patterns in homologous protein families (figure 2). CHORAL defines the SCCs as regions with equivalent differential geometry, i.e. curvature and torsion. Propensity tables are used to classify and to select patterns that are most likely to represent the structure of the core for a target protein (R. W. Montalvão *et al.* 2005, unpublished results). Our modelling process is completed by knowledge-based approaches in ANDANTE to model side-chains and in CODA to model loops (Deane & Blundell 2001). These approaches appear to have considerable advantages in retrospective prediction or 'post-diction'—CHORAL has been used to model the protein cores of 150 members of 10 protein families, demonstrating an equivalent and sometimes superior performance compared to other modelling programs, particularly in modelling distant members of superfamilies showing low sequence identity with the templates—but still have to be tested in an objective prediction mode.

Although these comparative structural bioinformatics approaches have proved very helpful, an experimental structure will invariably be more accurate. The possibility of using high-throughput crystallography for defining structures for the majority of gene products in an organism, known as structural genomics, has recently become a reality and there are several world-wide initiatives to define three-dimensional structures of representative protein family members in several genomes (Lesley *et al.* 2002; Service 2002; Heinemann *et al.* 2003; Rupp 2003). Structures defined by these structural genomics initiatives will not replace structural bioinformatics but rather focus its applications. If the objective of defining representative structures for each family can be achieved then comparative modelling can be used to construct models for all members of each family. Structural genomics has now defined structures for about 1000 proteins after 5 years' massive investment. The small number of genes in individual genomes (from less than a thousand to thirty thousand) hides the fact that there are probably as many as 50 000 families of proteins in total even in the prokaryotes. Each species tends to have its own, peculiar families. Thus, structural bioinformatics will continue to be central to identification of new members of superfamilies, a role that is likely to make major contributions to target identification and validation.

3. TARGET VALIDATION AND THE IDENTIFICATION OF LIGAND BINDING REGIONS

Many of the structures defined by structural genomics programs are of proteins that have no known function, and which have been identified as proteins based only on their gene sequences. It is becoming increasingly important to develop computational methods that will identify sites involved in productive intermolecular interactions that might give clues about functions and binding sites for these proteins. Although sequence motif databases, such as PROSITE (Hofmann *et al.* 1999), identify specific residues likely to be involved in function, three-dimensional descriptors of functional sites have an advantage as the sites themselves are usually made from discontinuous regions of the protein sequence (Kasuya & Thornton 1999). There have been several attempts to predict functional/interaction sites computationally, for example by identifying steric strain or other types of high-energy conformations that often occur at active sites (Herzberg & Moult 1991; Heringa & Argos 1999), or through identifying clefts that can accommodate ligands (Laskowski *et al.* 1996). Almost all protein functional sites arise through mutation and selection and hence they will be the most highly conserved regions of a protein (Zvelebil *et al.* 1987; McPhalen *et al.* 1992; Irving *et al.* 2001). The most widely used method based on evolutionary conservation of sequence is 'evolutionary trace' (Lichtarge & Sowa 2002), in which residues that are conserved are highlighted on the structure.

However, restraints leading to conservation of sequence can arise from both function, and structure. In our recent approach encoded in CRESCENDO, we have tried to differentiate evolutionary restraints on protein function from those on sequence and structure (Chelliah *et al.* 2004). As discussed above, the degree of conservation of amino acid residues has been shown to be strongly dependent on the environment in which they occur in the folded protein, and substitution tables that give the likely replacements of amino acids in particular local environments have been derived

(Overington *et al*. 1990; Overington *et al*. 1992). A method to distinguish restraints placed on substitutions due to protein structure from restraints deriving from functions mediated by interactions with other molecules has been developed using these environment-specific substitution tables (ESSTs; Chelliah *et al*. 2004; figure 3). Two different sequence scoring systems, used to identify the functional sites, gave similar results. The first scoring system, termed 'conservation score' (figure 3*a*) is a modification of that of Rodionov & Blundell (1998). This score quantifies the degree of sequence conservation at an alignment position compared to the average conservation. The second score, termed the 'divergent score' (figure 3*b*), quantifies the overall difference, or divergence, between the observed and predicted substitution probabilities. The positions where ESSTs make poor predictions of the overall amino acid substitution pattern are identified using information theory. The clusters of high scoring alignment positions apparently subjected to these additional restraints in evolution correlate well with the functional sites in protein defined by experimental methods. We have also analysed conservation of local structure in homologous families of proteins and developed a term to describe structural conservation (figure 3*c*) that can be used to increase the accuracy of functional site prediction (figure 3*d*). The method relies on the clustering of residues in three-dimensional space. The method has been applied to a set of well-characterized protein families and is able to identify functional sites. The technique is fast, automatic and predicts functional sites with a high degree of accuracy.

Since the residues involved in protein interactions have strong evolutionary pressure to remain unchanged, the sites that have such evolutionary pressure would have different substitution patterns when compared to the non-interacting sites. Thus, the study of residue substitution as a function of local environments with the inclusion of functional characters should highlight the clear difference in substitution patterns between the residues near and far from the active site. In a study of enzyme families which provide a good system for studying the substitution patterns near and far from the catalytic site, new sets of ESSTs (called function-dependent environment-specific substitution tables—FD-ESSTs) that include functional restraints arising from interacting with other molecules were derived. Tests of the FD-ESSTs in the homology recognition program FUGUE showed significant improvement compared to the recognition performance obtained using the standard ESSTs and other sequence alignment programs (Chelliah *et al*. 2005). The alignment accuracies obtained by standard ESSTs and FD-ESSTs were also improved with pronounced improvements at lower percentage identities (less than 30%). The alignments near the active site improved substantially.

4. LEAD DISCOVERY

Drugs have traditionally been identified from natural products and through *in vivo* studies of 'cause and effect'. The association of particular protein targets with disease pathways allowed a more rational approach where analogues of natural ligands could be designed. Around 10 years ago, drug discovery programs refocused their technology on the rapid assay of huge numbers of compounds. This random approach is called high-throughput screening, and aims to identify compounds with IC50s lower than 10 μM for their target proteins. The advent of this new technology required huge investment in faster systems for compound synthesis for the generation of large chemical libraries. Combinatorial chemistry was developed using both solid phase chemistry approaches, and solution phase libraries coupled with high-throughput purification platforms (Bailey 1997; Spencer 1998; Seneci & Miertus 2000; Dolle 2004). Automation of bioassays and systems for collection, storage and analysis of the very large datasets generated were also developed. However, the rate of newly registered compounds in clinical trials has not increased in proportion to the exponential increase in investment occasioned by these new robotic approaches.

The industry has, therefore, once again refocused, this time on targets and their related family members that are thought to be more tractable. Tractability of a target is based on the number of drug-like ligands for a target class, as well as knowledge of the binding sites of family members using protein structure information (Hopkins & Groom 2002). Examples of families of interest include the protein kinases and various proteinases. The classification of targets into families has allowed the design of focused compound libraries for particular families. Several approaches are now concentrating on screening very small molecules, or 'fragments' from which a lead can be designed using a knowledge, derived from biophysical assays, of how the fragment binds in the active site of the target.

In parallel, *in silico* approaches for identifying potential drug candidates have been developed. Ligand docking aims to find the optimum binding position and orientation for a compound in the active site of the proteins (Taylor *et al*. 2002). The best docking programmes correctly dock about 70–80% of ligands when tested on large sets of protein–ligand complexes (Nissink *et al*. 2002; Friesner *et al*. 2004); however, difficulties arise in trying to predict the affinities of the different compounds for the protein active site (Kitchen *et al*. 2004). Nevertheless, virtual screening has proved helpful in docking and ranking a large number of compounds so that the highest-ranking compounds can be selected for acquisition or synthesis and experimentally tested for activity against the target protein. Virtual screening provides a significant enrichment, perhaps twentyfold, of true hits in a selected subset of compounds (Boehm *et al*. 2000; Abagyan & Totrov 2001; Bajorath 2002; Lyne 2002; Shoichet *et al*. 2002; Jain 2004; Kitchen *et al*. 2004).

Both the fragment based approach and virtual screening are designed to provide more efficient sampling of chemical space by effectively decreasing the compound sample size. This optimization of the small molecule screening process has allowed experimental methods such as NMR and X-ray crystallography to contribute to drug discovery. Fragments are

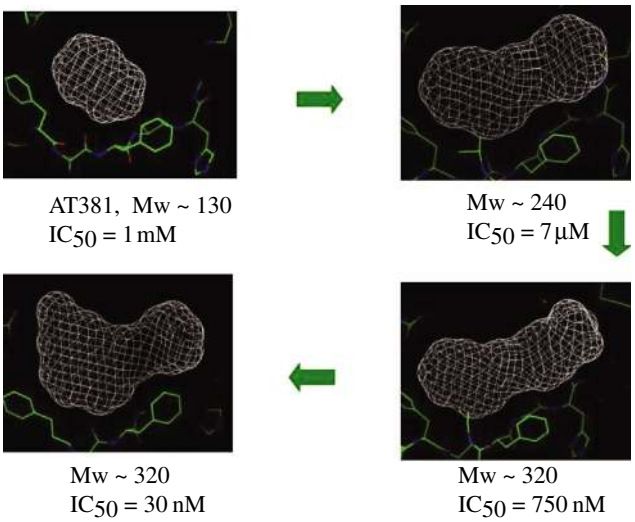

Figure 5. Pyramid hits-to-leads generation for Cdk2. This figure shows the electron density of various fragments bound to Cdk2. From around 500 compounds screened, 11 cocktails showed hits. From these fragments AT381 was selected with around 1 mM activity. Subsequent steps, represented by green arrows, are the optimization of fragment AT381 to improve potency, selectivity and ADME properties. © Astex Technology Ltd. 2005.

typically small organic molecules of between 100 and 250 Da. They will exhibit low binding affinities (approx. 100–10 mM) against target proteins and consequently cannot be identified by traditional high throughput screening. Biophysical methods must therefore be used to detect the fragments in the active site. Once a useful fragment has been identified and its binding mode defined, the fragment may provide a novel template for a larger ligand with better potency. Although the fragment hits have low affinity they often exhibit high ligand efficiency, i.e. high values for the average free energy of binding per heavy atom (Hopkins *et al.* 2004), and this property makes them attractive as start-points for optimization.

NMR spectroscopy was the first structural technique to be exploited for use in fragment screening. In 'SAR by NMR' (Shuker *et al.* 1996), perturbations to the NMR spectra of a protein are used to indicate that ligand binding is taking place and to give some indication of the location of the binding site. In the SHAPES approach, compound scaffolds derived from those most commonly found in known therapeutic agents are used and screened using NMR to detect binding (Fejzo *et al.* 1999). Recent reviews have emphasized the complementarity of NMR screening methods and crystallography in applications to inhibitor design (Muchmore & Hajduk 2003; Moore *et al.* 2004).

The advent of high-throughput crystallography means that hundreds or even thousands of small molecules can be screened, and binding sites for the molecules in the protein accurately defined. The approach depends on soaking crystals with single molecules or cocktails of compounds. As protein crystals contain extensive solvent-filled channels, making up around 50% of their volume, small molecules will usually diffuse rapidly into the crystals

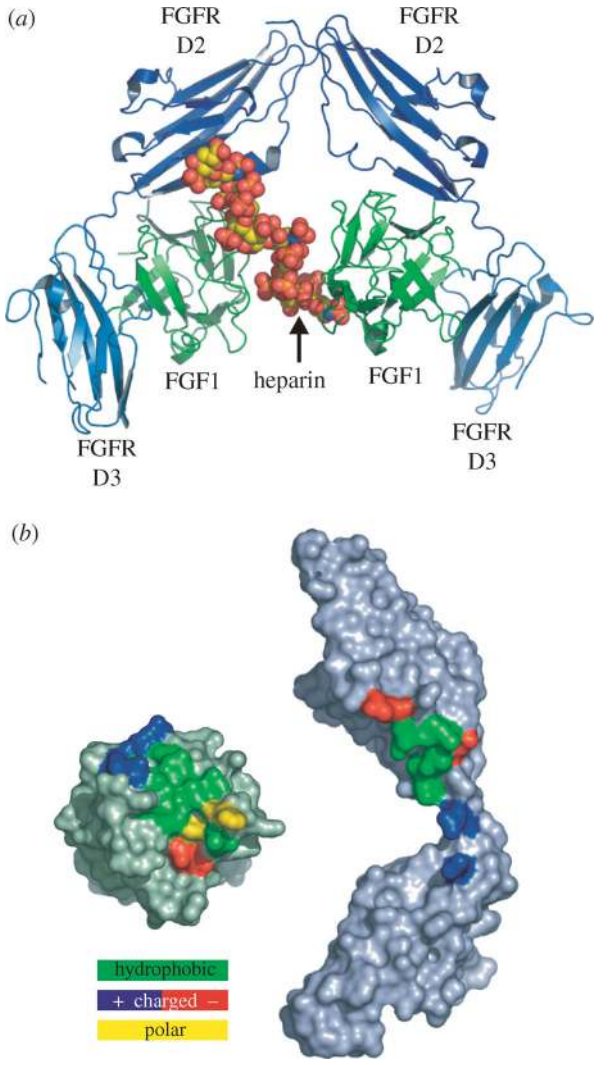

Figure 6. (*a*) The fibroblast growth factor (FGF; green) and its receptor (blue) contain globular domains that form a complex with the co-factor heparin, without significant changes to their 3D structures. (*b*) A more detailed examination of the interaction between the FGF and domain 2 (top as shown) of the receptor shows that binding sites in both proteins consist of a discontinuous epitope on surfaces that are comparatively flat for protein structures. The binding sites broadly consist of a hydrophobic centre bordered by charged and polar patches.

and interact as if they were in solution, provided that the binding site is not occluded by the crystal packing. The small molecules can then be visualized using difference Fourier techniques by collecting sets of X-ray data on each soaked crystal under identical conditions. Automatic procedures have been developed to facilitate the rapid structure solution of protein–ligand complexes by interpreting and analysing the X-ray data without the need for manual intervention. The molecules within each cocktail are fitted to the electron density in turn and then ranked according to how well they fit. This provides complete automation of the system once the initial protein crystals have been characterized and the structures solved.

There is an increasing number of examples in the literature where X-ray crystallography has been used as a tool to identify fragment 'hits' (Verlinde *et al.* 1997;

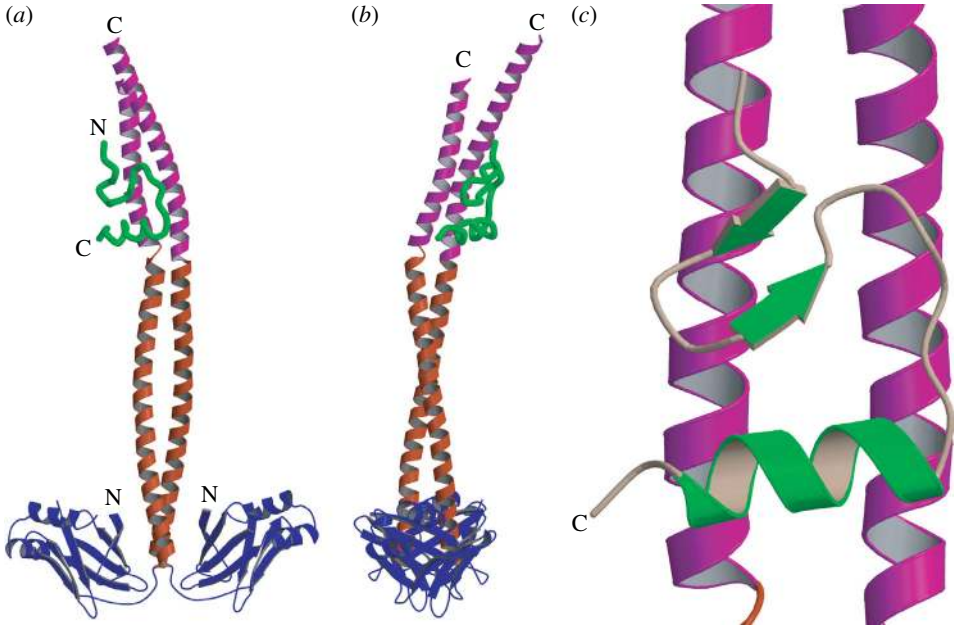

Figure 7. The human non-homologous end joining protein Xrcc4 binds to a flexible linker between tandem BRCT domains of DNA ligase IV, imposing structure on the linker through the interaction. (*a*), (*b*) show the full Xrcc4 structure with ligase linker bound (shown in green); (*c*) shows a close-up of the interaction, highlighting the structure that is imposed on a previously unstructured peptide.

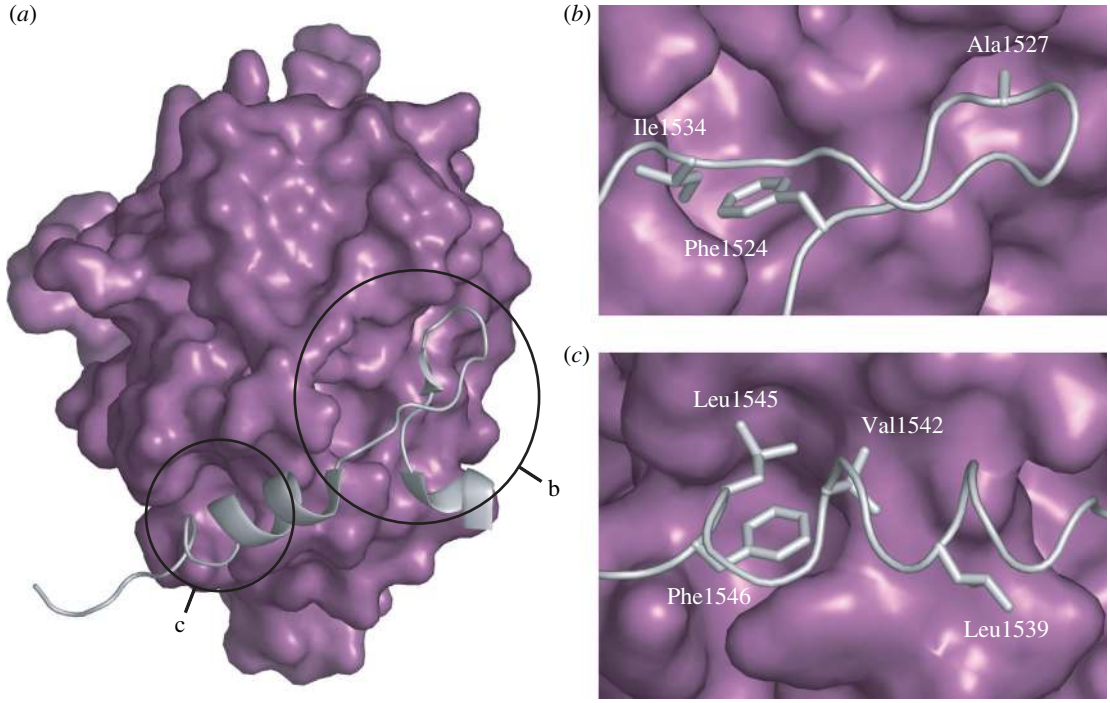

Figure 8. (*a*) Human recombinase Rad51 binds BRC repeats of BRCA2 in an interaction that is essential for function in recombination. Although this is usually essential for normal DNA repair, there is an advantage in disrupting recombination during radiotherapy and chemotherapy, which function through the introduction of DNA damage in cancerous cells. While Rad51 independently forms a stable globular structure, only upon interacting with Rad51 does the BRC peptide fold into a defined three dimension structure. Closer examination of the interaction (*b*, *c*) shows discrete regions of interaction that may be useful drug targets in disrupting the interaction and thus blocking recombination.

Nienaber *et al*. 2000; Lesuisse *et al*. 2002). Structural GenomiX (SGX) has an integrated technology platform *FAST* (http://www.stromix.com), for lead identification using high-throughput protein structure determination. Plexxikon have a process called *Scaffold-based drug discovery* for the design of useful templates that uses X-ray analysis of protein–ligand co-crystals (Hirth & Milburn 2004).

Astex Technology has developed an approach called Pyramid, where fragment libraries are screened in cocktails using X-ray crystallography. Automated molecular fragment matching and fitting in electron density is then achieved by a software procedure called AutoSolve which also ranks the candidate fragments in a cocktail (Blundell *et al*. 2002; Carr & Jhoti 2002). Figure 4 shows a schematic representation of the steps

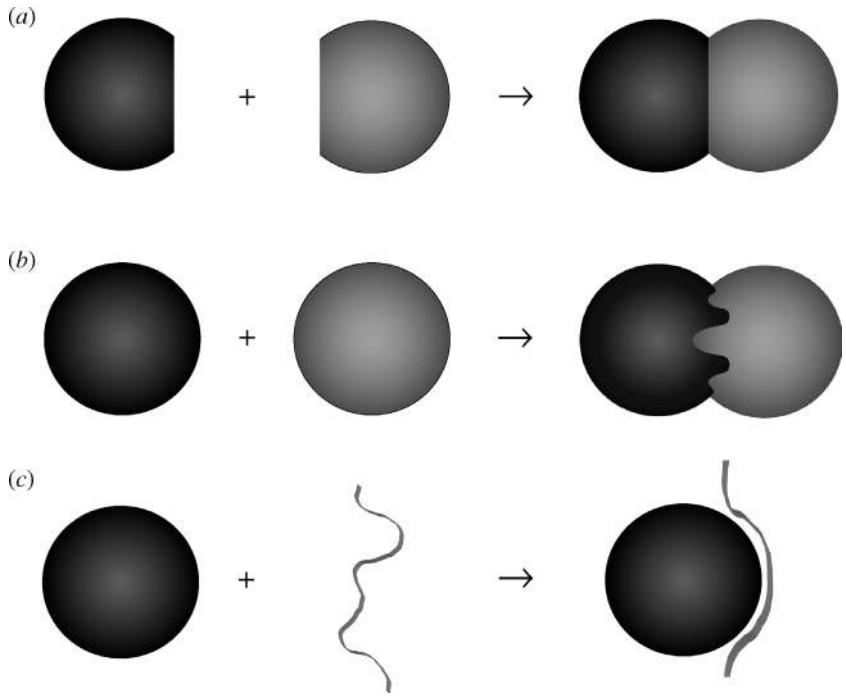

Figure 9. Protein–protein interactions can be described by three models. (*a*) Proteins of preformed globular structure that interact with no change to their structure through a discontinuous epitope. (*b*) Proteins of preformed globular structure that adapt upon interaction to form a complex of novel conformation. (*c*) A natively unstructured protein that folds upon interaction with another partner.

in the Astex Pyramid procedure. Fragment hits derived from Pyramid are subsequently optimized with carefully designed iterations in order to maintain good ligand efficiency. This process has been carried out against a number of protein targets (Gill *et al.* 2005; Hartshorn *et al.* 2005) and figure 5 shows an example of fragment binding and subsequent fragment optimization in the protein kinase Cdk2. Cdk2 is a target in the oncology disease area and molecules from this programme are now progressing towards the clinic.

The area of fragment-based lead discovery has recently been thoroughly reviewed, and many more examples of the approach are described in these articles (Erlanson *et al.* 2004; Rees *et al.* 2004).

5. NEW CHALLENGES FOR DRUG DISCOVERY IN MULTIDOMAIN AND MULTIPROTEIN PROTEIN TARGETS

One of the great internal contradictions of drug discovery in practice is that while most regulatory proteins in man, the obvious targets for new drugs, are complex proteins that are often multidomain and very usually components of multiprotein systems, most of the focus in the pharmaceutical industry is on the active sites of monomeric proteins. Is this really sustainable?

Many proteins in the higher eukaryotes are large and contain multiple domains. A typical example is the DNA protein kinase (DNA-PK), a key molecule in non-homologous end joining, which signals the assembly of the multiprotein system involved in the repair of double strand breaks (Smider *et al.* 1994; Taccioli *et al.* 1994; Blunt *et al.* 1995; Kirchgessner *et al.* 1995; Peterson *et al.* 1995). This protein is composed of a large catalytic subunit and a regulating heterodimer Ku70 and Ku80. The catalytic subunit (DNA-PKcs) can be isolated to a high degree of purity from human placenta or HeLa cells by modifying published protocols of Chan *et al.* (1996) and Gell & Jackson (1999). It is not easy to express as a recombinant protein due to its large size of 4127 amino acids. For drug discovery a divide and rule approach is required and for this boundaries of domains need to be estimated.

A program, DOMINANT (Brewerton 2004), has been written to deconvolute protein structures into their constituent domains in order that domains and domain boundaries can be classified. Given a new protein structure, DOMINANT checks the existing domain database with a structure comparison procedure called SEA to identify any recurrent domains, and then uses a procedure to identify domains from the spatial separation of secondary structures to deconvolute the remaining structure. An analysis of structurally defined domain linkers (defined using DOMINANT) and the sequence defined domain linkers (from the PFAM database) has been carried out and parameters were derived in order to produce a knowledge based method to predict the likelihood of sequences to be linkers between globular domains. Methods such as this and in particular the combination of many methods can provide clues as to where domain boundaries might be in huge multidomain proteins such as DNA-PKcs. The N-terminal region of DNA-PKcs has in fact, been shown to be made up of a large array of tandemly repeated structural motifs (Brewerton *et al.* 2004). It is thought that prediction of these repeat features in proteins will be increasingly important as they are responsible for a high number of the protein–protein interactions that take place in the

cell. A procedure, FUGUEREP, for prediction of structural repeats has been produced and used to examine DNA-PKcs and the related phosphatidyl inositol 3 kinase like kinases (Brewerton 2004; Brewerton *et al*. 2004).

Another major challenge for drug discovery arises from the very many multiprotein systems that really need to be targeted. Many of these have large surfaces of the order of 2000 Å^2, for example those involved in receptor recognition and signal transduction (see for example, Pellegrini *et al*. 2000; figure 6). This is especially true of complexes that are assembled from preformed globular domains. Not only is it difficult to bind a small molecule to the large, relatively flat surfaces of such proteins involved in protein interactions, but it is also difficult to disrupt the interaction entirely even if one did. It remains to be seen whether the emerging lead discovery approaches discussed here will prove suitable for these systems. However, recent analyses of multiprotein systems involved in cell regulation and signalling have identified a large number in which one component involves a flexible or unstructured region of the polypeptide chain (figures 7 and 8). Examples are the Xrcc4 dimer in complex with DNA ligase IV (figure 7), in which the linker region between two BRCT domains appears to organize when the complex is assembled (Sibanda *et al*. 2001). A further example (figure 8) involves the complex of the human recombinase, Rad51, and the product of the breast cancer associated gene, BRCA2 (Pellegrini *et al*. 2002), which is not only revealing in terms of the nature of the interactions and the molecular origins of cancers associated with mutations in this region of BRCA2, but also offers an encouraging and perhaps more druggable site of interaction that could be used to target agents that would be helpful during chemo- or radio-therapy. We suggest that proteins forming interactions with a ligand that comprises a continuous region of flexible peptide may be more druggable targets than where complexes are formed from preformed globular protein structures (figure 9).

6. CONCLUSIONS

Knowledge of the three-dimensional structures of protein targets is now playing a major role in all stages of drug discovery. Its place in lead optimization is well established with large teams of structural biologists recruited into all major pharmaceutical companies. The success of the method is evident from drugs now in use and new ones reaching the market. It is clear that in many companies structure-guided approaches have become central to developing good drug candidates.

But structural biology and bioinformatics show that many key targets for drug discovery are multidomain and multiprotein complexes. Such systems pose significant challenges not only for characterization using structural techniques but also because the inter-protein surfaces (figure 9) are usually comparatively flat and poor in distinguishing features, making the design of small molecule antagonists a formidable task. These challenges underline the importance of new approaches and the key roles of both academia and industry in advancing this process.

REFERENCES

Abagyan, R. & Totrov, M. 2001 High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **5**, 375–382. (doi:10.1016/S1367-5931(00)00217-9)

Bailey, N. *et al*. 1997 Solution-phase combinatorial chemistry in lead discovery. *Chimia* **51**, 832–837.

Bajorath, J. 2002 Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* **1**, 882–894. (doi:10.1038/nrd941)

Bates, P. A. & Sternberg, M. J. 1999 Model building by comparison at CASP3: using expert knowledge and computer automation. *Proteins* Suppl. 3, 47–54. (doi:10.1002/(SICI)1097-0134(1999)37:3+<47::AID-PROT7>3.0.CO;2-F)

Beddell, C. R., Goodford, P. J., Norrington, F. E., Wilkinson, S. & Wootton, R. 1976 Compounds designed to fit a site of known structure in human haemoglobin. *Br. J. Pharmacol.* **57**, 201–209.

Blundell, T. L. 1996 Structure-based drug design. *Nature* **384**(6604 Suppl.), 23–26.

Blundell, T., Sibanda, B. L. & Pearl, L. 1983 Three-dimensional structure, specificity and catalytic mechanism of renin. *Nature* **304**, 273–275. (doi:10.1038/304273a0)

Blundell, T. L., Dodson, G., Hodgkin, D. & Mercola, D. 1972 Insulin: the structure in the crystal and its reflection in chemistry and biology. *Adv. Protein Chem.* **26**, 279–402.

Blundell, T. L. *et al*. 1988 Knowledge-based protein modelling and design; 18th Sir Hans Krebs lecture. *Eur. J. Biochem.* **172**, 513–520. (doi:10.1111/j.1432-1033.1988.tb13917.x)

Blundell, T. L., Abell, C., Cleasby, A., Hartshorn, M. J., Tickle, I. J., Parasini, E. & Jhoti, H. 2002 High-throughput X-ray crystallography for drug discovery. In *Drug design: special publication* (ed. D. R. Flower), pp. 53–59. Cambridge: Royal Society of Chemistry.

Blunt, T. *et al*. 1995 Defective DNA-dependent protein kinase activity is linked to V(D)J recombination and DNA repair defects associated with the murine scid mutation. *Cell* **80**, 813–823. (doi:10.1016/0092-8674(95)90360-7)

Boehm, H. J. *et al*. 2000 Novel inhibitors of DNA gyrase: 3D structure based biased needle screening, hit validation by biophysical methods, and 3D guided optimization. A promising alternative to random screening. *J. Med. Chem.* **43**, 2664–2674. (doi:10.1021/jm000017s)

Bowie, J. U., Luthy, R. & Eisenberg, D. 1991 A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**, 164–170.

Brewerton, S. C. 2004 Structural annotation of protein sequences: tools for construct design. Ph.D. thesis, University of Cambridge.

Brewerton, S. C., Dore, A. S., Drake, A. C., Leuther, K. K. & Blundell, T. L. 2004 Structural analysis of DNA-PKcs: modelling of the repeat units and insights into the detailed molecular architecture. *J. Struct. Biol.* **145**, 295–306. (doi:10.1016/j.jsb.2003.11.024)

Campbell, S. F. 2000 Science, art and drug discovery: a personal perspective. *Clin. Sci.* **99**, 255–260.

Carr, R. & Jhoti, H. 2002 Structure-based screening of low-affinity compounds. *Drug Discov. Today* **7**, 522–527. (doi:10.1016/S1359-6446(02)02245-6)

Chan, D. W., Mody, C. H., Ting, N. S. & Lees-Miller, S. P. 1996 Purification and characterization of the double-stranded DNA-activated protein kinase, DNA-PK, from human placenta. *Biochem. Cell Biol.* **74**, 67–73.

Chelliah, V., Chen, L., Blundell, T. L. & Lovell, S. C. 2004 Distinguishing structural and functional restraints in evolution in order to identify interaction sites. *J. Mol. Biol.* **342**, 1487–1504. (doi:10.1016/j.jmb.2004.08.022)

Chelliah, V., Blundell, T. L. & Mitzuguchi, K. 2005 Functional restraints on the patterns of amino acid

substitutions: application to sequence–structure homology recognition. *Proteins: Struct. Funct. Bioinform* **61**, 722–731.

Deane, C. M. & Blundell, T. L. 2001 CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* **10**, 599–612. (doi:10.1110/ps.37601)

Dolle, R. E. 2004 Comprehensive survey of combinatorial library synthesis: 2003. *J. Comb. Chem.* **6**, 623–679. (doi:10.1021/cc0499082)

Erlanson, D. A., McDowell, R. S. & O'Brien, T. 2004 Fragment-based drug discovery. *J. Med. Chem.* **47**, 3463–3482. (doi:10.1021/jm040031v)

Fejzo, J., Lepre, C. A., Peng, J. W., Bemis, G. W., Ajay, Murcko, M. A. & Moore, J. M. 1999 The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.* **6**, 755–769. (doi:10.1016/S1074-5521(00)80022-8)

Friesner, R. A. *et al.* 2004 Glide: a new approach for rapid, accurate docking and scoring 1. Method and assessment of docking accuracy. *J. Med. Chem.* **47**, 1739–1749. (doi:10.1021/jm0306430)

Gell, D. & Jackson, S. P. 1999 Mapping of protein–protein interactions within the DNA-dependent protein kinase complex. *Nucleic Acids Res.* **27**, 3494–3502. (doi:10.1093/nar/27.17.3494)

Gill, A. L. *et al.* 2005 Identification of novel p38alpha MAP kinase inhibitors using fragment-based lead generation. *J. Med. Chem.* **48**, 414–426. (doi:10.1021/jm049575n)

Goodford, P. J., St-Louis, J. & Wootton, R. 1980 The interaction of human haemoglobin with allosteric effectors as a model for drug–receptor interactions. *Br. J. Pharmacol.* **68**, 741–748.

Hardy, L. W. & Malikayil, A. 2003 The impact of structure-guided drug design on clinical agents. *Curr. Drug Discov.*, 15–20.

Hartshorn, M. J., Murray, C. W., Cleasby, A., Frederickson, M., Tickle, I. J. & Jhoti, H. 2005 Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.* **48**, 403–413. (doi:10.1021/jm0495778)

Heinemann, U., Bussow, K., Mueller, U. & Umbach, P. 2003 Facilities and methods for the high-throughput crystal structural analysis of human proteins. *Acc. Chem. Res.* **36**, 157–163. (doi:10.1021/ar010129t)

Heringa, J. & Argos, P. 1999 Strain in protein structures as viewed through nonrotameric side chains. II. Effects upon ligand binding. *Proteins* **37**, 44–55. (doi:10.1002/(SICI)1097-0134(19991001)37:1<44::AID-PROT5>3.0.CO;2-F)

Herzberg, O. & Moult, J. 1991 Analysis of the steric strain in the polypeptide backbone of protein molecules. *Proteins* **11**, 223–229. (doi:10.1002/prot.340110307)

Hirth, K. P. & Milburn, M. V. 2004 Methods for the design of molecular scaffolds and ligands. US Patent application publication no. US2004/0171062.

Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. 1999 The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**, 215–219. (doi:10.1093/nar/27.1.215)

Hopkins, A. L. & Groom, C. R. 2002 The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730. (doi:10.1038/nrd892)

Hopkins, A. L., Groom, C. R. & Alex, A. 2004 Ligand efficiency: a useful metric for lead selection. *Drug Discov. Today* **9**, 430–431. (doi:10.1016/S1359-6446(04)03069-7)

Irving, J. A., Whisstock, J. C. & Lesk, A. M. 2001 Protein structural alignments and functional genomics. *Proteins* **42**, 378–382. (doi:10.1002/1097-0134(20010215)42:3<378::AID-PROT70>3.0.CO;2-3)

Jain, A. N. 2004 Virtual screening in lead discovery and optimization. *Curr. Opin. Drug Discov. Dev.* **7**, 396–403.

Jones, D. T. 1999 GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815. (doi:10.1006/jmbi.1999.2583)

Jones, D. T., Taylor, W. R. & Thornton, J. M. 1992 A new approach to protein fold recognition. *Nature* **358**, 86–89. (doi:10.1038/358086a0)

Kasuya, A. & Thornton, J. M. 1999 Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.* **286**, 1673–1691. (doi:10.1006/jmbi.1999.2581)

Kirchgessner, C. U., Patil, C. K., Evans, J. W., Cuomo, C. A., Fried, L. M., Carter, T., Oettinger, M. A. & Brown, J. M. 1995 DNA-dependent kinase (p350) as a candidate gene for the murine SCID defect. *Science* **267**, 1178–1183.

Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. 2004 Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949. (doi:10.1038/nrd1549)

Lapatto, R. *et al.* 1989 X-ray analysis of HIV-1 proteinase at 2.7 A resolution confirms structural homology among retroviral enzymes. *Nature* **342**, 299–302. (doi:10.1038/342299a0)

Laskowski, R. A., Luscombe, N. M., Swindells, M. B. & Thornton, J. M. 1996 Protein clefts in molecular recognition and function. *Protein Sci.* **5**, 2438–2452.

Lesley, S. A. *et al.* 2002 Structural genomics of the *Thermotoga maritima* proteome implemented in a high-throughput structure determination pipeline. *Proc. Natl Acad. Sci. USA* **99**, 11 664–11 669. (doi:10.1073/pnas.142413399)

Lesuisse, D. *et al.* 2002 SAR and X-ray. A new approach combining fragment-based screening and rational drug design: application to the discovery of nanomolar inhibitors of Src SH2. *J. Med. Chem.* **45**, 2379–2387. (doi:10.1021/jm010927p)

Lichtarge, O. & Sowa, M. E. 2002 Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.* **12**, 21–27. (doi:10.1016/S0959-440X(02)00284-1)

Lindahl, E. & Elofsson, A. 2000 Identification of related proteins on family, superfamily and fold level. *J. Mol. Biol.* **295**, 613–625. (doi:10.1006/jmbi.1999.3377)

Lombardino, J. G. & Lowe III, J. A. 2004 The role of the medicinal chemist in drug discovery—then and now. *Nat. Rev. Drug Discov.* **3**, 853–862. (doi:10.1038/nrd1523)

Lyne, P. D. 2002 Structure-based virtual screening: an overview. *Drug Discov. Today* **7**, 1047–1055. (doi:10.1016/S1359-6446(02)02483-2)

McPhalen, C. A., Vincent, M. G., Picot, D., Jansonius, J. N., Lesk, A. M. & Chothia, C. 1992 Domain closure in mitochondrial aspartate aminotransferase. *J. Mol. Biol.* **227**, 197–213. (doi:10.1016/0022-2836(92)90691-C)

Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V., Marshall, G. R., Clawson, L., Selk, L., Kent, S. B. & Wlodawer, A. 1989 Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. *Science* **246**, 1149–1152.

Moore, J., Abdul-Manan, N., Fejzo, J., Jacobs, M., Lepre, C., Peng, J. & Xie, X. 2004 Leveraging structural approaches: applications of NMR-based screening and X-ray crystallography for inhibitor design. *J. Synchrotron Radiat.* **11**, 97–100. (doi:10.1107/S0909049503023975)

Muchmore, S. W. & Hajduk, P. J. 2003 Crystallography, NMR and virtual screening: integrated tools for drug discovery. *Curr. Opin. Drug Discov. Dev.* **6**, 544–549.

Nienaber, V. L., Richardson, P. L., Klighofer, V., Bouska, J. J., Giranda, V. L. & Greer, J. 2000 Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* **18**, 1105–1108. (doi:10.1038/80319)

Nissink, J. W., Murray, C., Hartshorn, M., Verdonk, M. L., Cole, J. C. & Taylor, R. 2002 A new test set for validating predictions of protein–ligand interaction. *Proteins* **49**, 457–471. (doi:10.1002/prot.10232)

Overington, J., Johnson, M. S., Sali, A. & Blundell, T. L. 1990 Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. Biol. Sci.* **241**, 132–145.

Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. 1992 Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226.

Pearl, L. H. & Taylor, W. R. 1987 A structural model for the retroviral proteases. *Nature* **329**, 351–354. (doi:10.1038/329351a0)

Pellegrini, L., Burke, D. F., von Delft, F., Mulloy, B. & Blundell, T. L. 2000 Crystal structure of fibroblast growth factor receptor ectodomain bound to ligand and heparin. *Nature* **407**, 1029–1034. (doi:10.1038/35039551)

Pellegrini, L., Yu, D. S., Lo, T., Anand, S., Lee, M., Blundell, T. L. & Venkitaraman, A. R. 2002 Insights into DNA recombination from the structure of a RAD51-BRCA2 complex. *Nature* **420**, 287–293. (doi:10.1038/nature01230)

Peterson, S. R., Kurimasa, A., Oshimura, M., Dynan, W. S., Bradbury, E. M. & Chen, D. J. 1995 Loss of the catalytic subunit of the DNA-dependent protein kinase in DNA double-strand-break-repair mutant mammalian cells. *Proc. Natl Acad. Sci. USA* **92**, 3171–3174.

Rees, D. C., Congreve, M., Murray, C. W. & Carr, R. 2004 Fragment-based lead discovery. *Nat. Rev. Drug Discov.* **3**, 660–672. (doi:10.1038/nrd1467)

Rice, D. W. & Eisenberg, D. 1997 A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**, 1026–1038. (doi:10.1006/jmbi.1997.0924)

Rodionov, M. A. & Blundell, T. L. 1998 Sequence and structure conservation in a protein core. *Proteins—Struct. Funct. Genet.* **33**, 358–366. (doi:10.1002/(SICI)1097-0134(19981115)33:3<358::AID-PROT5>3.0.CO;2-0)

Rupp, B. 2003 High-throughput crystallography at an affordable cost: the TB structural genomics consortium crystallization facility. *Acc. Chem. Res.* **36**, 173–181. (doi:10.1021/ar020021t)

Sali, A. & Blundell, T. L. 1993 Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815. (doi:10.1006/jmbi.1993.1626)

Seneci, P. & Miertus, S. 2000 Combinatorial chemistry and high-throughput screening in drug discovery: different strategies and formats. *Mol. Divers.* **5**, 75–89. (doi:10.1023/A:1013824317218)

Service, R. F. 2002 Structural genomics. Tapping DNA for structures produces a trickle. *Science* **298**, 948–950. (doi:10.1126/science.298.5595.948)

Shi, J., Blundell, T. L. & Mizuguchi, K. 2001 FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* **310**, 243–257. (doi:10.1006/jmbi.2001.4762)

Shoichet, B. K., McGovern, S. L., Wei, B. & Irwin, J. J. 2002 Lead discovery using molecular docking. *Curr. Opin. Chem. Biol.* **6**, 439–446. (doi:10.1016/S1367-5931(02)00339-3)

Shuker, S. B., Hajduk, P. J., Meadows, R. P. & Fesik, S. W. 1996 Discovering high-affinity ligands for proteins: SAR by NMR. *Science* **274**, 1531–1534. (doi:10.1126/science.274.5292.1531)

Sibanda, B. L., Critchlow, S. E., Begun, J., Pei, X. Y., Jackson, S. P., Blundell, T. L. & Pellegrini, L. 2001 Crystal structure of an Xrcc4-DNA ligase IV complex. *Nat. Struct. Biol.* **8**, 1015–1019. (doi:10.1038/nsb725)

Smider, V., Rathmell, W. K., Lieber, M. & Chu, G. 1994 Restoration of X-ray resistance and V(D)J recombination in mutant cells by Ku cDNA. *Science* **266**, 288–291.

Spencer, R. W. 1998 High-throughput screening of historic collections: observations on file size, biological targets, and file diversity. *Biotechnol. Bioeng.* **61**, 61–67. (doi:10.1002/(SICI)1097-0290(199824)61:1<61::AID-BIT11>3.0.CO;2-C)

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. 1987 Knowledge based modelling of homologous proteins. Part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* **1**, 377–384.

Taccioli, G. E. *et al*. 1994 Ku80: product of the XRCC5 gene and its role in DNA repair and V(D)J recombination. *Science* **265**, 1442–1445.

Taylor, R. D., Jewsbury, P. J. & Essex, J. W. 2002 A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **16**, 151–166. (doi:10.1023/A:1020155510718)

Varghese, J. N. 1999 Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Dev. Res.* **46**, 176–196. (doi:10.1002/(SICI)1098-2299(199903/04)46:3/4<176::AID-DDR4>3.0.CO;2-6)

Venclovas, C., Zemla, A., Fidelis, K. & Moult, J. 2003 Assessment of progress over the CASP experiments. *Proteins* **53**(Suppl. 6), 585–595. (doi:10.1002/prot.10530)

Verlinde, C. L. M. J., Kim, H., Bernstein, B. E., Mande, S. C. & Hol, W. G. J. 1997 Antitrypanosomiasis drug development based on structures of glycolytic enzymes. In *Structure-based drug design*, pp. 365–394. New York: Marcel Dekker Inc.

Whittle, P. J. & Blundell, T. L. 1994 Protein structure-based drug design. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 349–375. (doi:10.1146/annurev.bb.23.060194.002025)

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. 1987 Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961. (doi:10.1016/0022-2836(87)90501-8)