Function first: a powerful approach to post-genomic drug discovery

Stephen F. Betz, Susan M. Baxter and Jacquelyn S. Fetrow

In the post-genomic era, pharmaceutical researchers must evaluate vast numbers of protein sequences and formulate novel, intelligent strategies for identifying valid targets and discovering leads against them. The identification of small molecules that selectively target proteins or protein families will be aided by knowing the function and/or the structure of the target(s). By identifying protein function first, efficiencies are gained that allow subsequent focus of resources on particular protein families of interest. This article reviews current proteomic-scale approaches to identifying function as a way of accelerating lead discovery.

Stephen F. Betz
Susan M. Baxter
Jacquelyn S. Fetrow
GeneFormatics
5830 Oberlin Drive
Suite 200
San Diego
CA 92121, USA
tel: +1 858 450 3331
fax: +1 858 450 1138
e-mail: jacquefetrow@geneformatics.com

▼ Since their inception in the early 1990s, genome-sequencing projects have been exceptionally successful, and the number of fully sequenced genomes has grown impressively. Many genome sequences are publicly available, including associated annotations and curated updates based on new experimentation [1,2]. This wealth of information has opened many new avenues of investigation. Of particular interest is the effect of genomic data on drug discovery. This review focuses on the question, 'How can we effectively use this new information to accelerate the development of new therapeutics that target gene products or their functions?'

The nucleotide sequences generated by both high-throughput sequencing and differential expression methods provide vast amounts of new information that must be evaluated. This is sometimes described as a 'deluge of targets'; however, it is more accurately described as a deluge of putative protein sequences. These proteins are not necessarily drug targets in the pharmaceutical sense today because their biochemical and biological functions are unverified and, in many cases, unknown. Indeed, some might not even be proteins that are actually expressed *in vivo*. Although the number of current drug targets derived from genomic sequences is low, genomic datasets offer the promise of a rich source of *bona fide* drug targets that are waiting to be identified.

The post-genomic era also provides significant challenges for lead discovery, that is, the identification and validation of small molecules that modulate the function of specific protein targets. The first challenge is the need for any lead discovery effort to deal with larger numbers of possible targets. Most pharmaceutical companies' discovery strategies rely on HTS of large compound libraries against a selected target protein. These methods hinge on having a reliable assay typically using cell-based reporter activity, characterized enzymatic activity or competitive binding against a known ligand. These methods have high monetary and resource costs, and limit the number of targets that can be screened simultaneously. Additionally, the number of false positive hits from HTS is often high, requiring a secondary assay to validate the true positives.

Certain screening strategies have been developed that circumvent the need for *a priori* enzymatic or ligand binding knowledge. For instance, both NMR and X-ray crystallography-based screening methods that detect binding rather than activity have been used to provide lead compounds without the need for functional knowledge [3–5]. Binding affinity has also been probed using MS techniques linked to NMR [6]. Libraries of biologically common reactants, products, inhibitors and co-factors can be screened to provide important functional information and chemical handles for further assay development and lead optimization [7]. In addition to screening technologies, the advancement of structure-guided drug design has aided the development of therapeutics in several important enzyme classes, including proteases, kinases, and

1359-6446/02/$ – see front matter PII: S1359-6446(02)02398-X

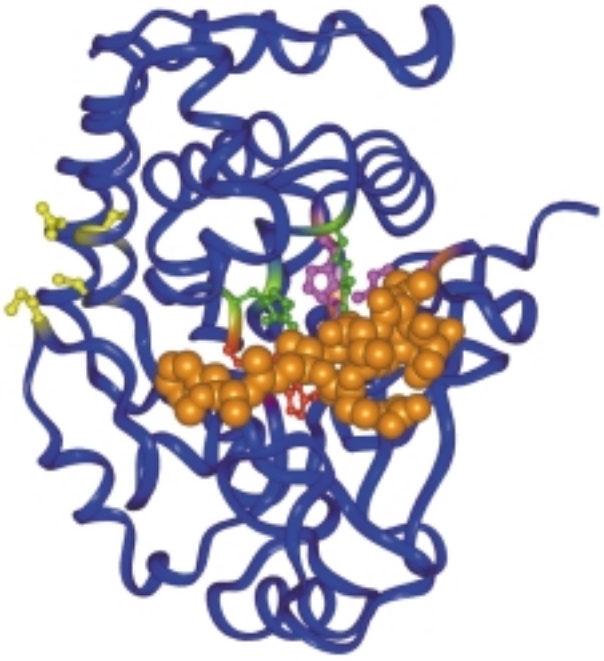

Figure 1. Proteins contain multiple features important for molecular function, including structure, activity, specificity and regulation. The crystal structure of a serine-threonine phosphatase [Protein Data Bank (PDB) code 1fjm] bound to microcystin-Lr toxin inhibitor (orange space-filling model) reveals key amino acid positions and inhibitor binding-site geometry. The molecular function of this protein is completely described by four different functional sites: two metal binding sites (pink and green side-chains); a catalytic and substrate binding site (red residues); and a redox site (yellow residues) [62] at the surface of the protein. Information about the overall phosphatase activity is informative, but additional information about the active-site geometry and a separate, putative modulating site yields knowledge that can be leveraged in lead discovery.

phosphatases among others [8–11]. Nonetheless, both HTS and structural-biology based methods typically rely on analyzing single proteins serially, limiting the scope of potential targets to be examined. To address the potentially large target set(s) generated by proteomics, new methods that leverage the best features of existing methods but that operate automatically and effectively manage resource costs and timescales must be developed.

Another challenge facing pharmaceutical research is the need for selectivity in lead identification. A full understanding of proteomic data provides a unique opportunity to compare the interactions of small molecules with many proteins encoded by a given genome (or even several genomes), particularly those that are functionally related. The opportunity to develop, from the early stages of the discovery process, truly selective compounds and libraries is now at hand. Such a capability should ultimately decrease the failure rate of compounds in development resulting from cross-reactivity-dependent toxicity, thus lowering the cost and time of the drug discovery process.

A common starting point used in the mining of genomic sequences for drug leads involves identifying the function of gene products. Information about pathway involvement, catalytic activity, protein class, or active-site chemistry, among other functional features (Fig. 1), provides knowledge that can be used in therapeutic design. With functional information in hand, decisions about assay development, screening, and validation are more straightforward and scalable. In fact, once function has been identified across a genome, enumeration of the entire functional 'family' of interest allows the potential for screening or design of small molecules that are specific for individual members within each functional family – a 'functional family' approach to drug discovery [12–14]. Parallel large-scale processes and analyses to identify function first will be key for this lead discovery approach to be successful in the post-genomic era.

Protein function assignment methods for identifying new targets and lead compounds

Biological function is defined quite widely and in cases where experimental evidence is lacking, any level of functional annotation can provide information useful to discovery efforts. Initially, function is often inferred by sequence similarity at either the DNA or amino acid level. If the sequences are 'similar enough', the pre-determined function from one sequence is automatically transferred to the other sequence, a practice that can lead to misannotation, misinterpretation and a disturbing propagation of errors [15].

Experimental methodologies have opened up several strategies for large-scale functional annotation at the cellular level. Expression monitoring [serial analysis of gene expression (SAGE) or expression array profiling] is a common approach used to identify sets of gene products involved in cellular function [16]. The interest in parallel protein analysis has led to advances in designing protein chips [17,18] and protein-interaction methods [19] for arraying protein sets for function. Notably, researchers have recently used tandem-affinity purification (TAP) to identify protein complexes in the yeast proteome and thus link, on a large-scale, protein networks or complexes that operate together to carry out cellular function [20]. The advantage to these methods is that gene product function can be identified within the complex biological environment, yielding valuable target validation data. Several disadvantages limit the use of these methods, including gene products that have low copy numbers, are transiently expressed,

or produce inactive proteins. Recent comparison has suggested that the false positive of these methods is extremely high [21].

The most informative and powerful cellular approaches combine function annotation with small-molecule lead discovery. Expression monitoring approaches have been used to tease out biological effects of drugs on particular functional classes [22,23], to tweak signaling pathways [24], and to qualify small-molecule modulators against a proteome [25]. Phage display methods have been reported to be useful in identifying cellular targets for small molecules [26,27]. New parallel approaches offer details about the chemistry of the target or the likely shape of a suitable ligand. Chemical approaches such as affinity tagging [28,29] or 'click' chemistry [30] provide functional information about active-site chemistries and geometries. Selective chemical probes that take advantage of covalent inhibition have been designed to target specific protein families with similar active-site chemistries [31,32]. These chemistry-based methods have the distinct advantage of detecting active proteins in the cell and provide a starting platform for refining small-molecule affinity and selectivity. Importantly, these approaches when integrated with biochemical experiments provide pharmaceutical researchers with small-molecule leads, even without *a priori* knowledge of protein target identification or structure.

Nonetheless, optimization of a lead to a clinical development candidate involves iterative chemistry cycles to improve targeting, specificity, transport and toxicology profile. There is a need to understand how small molecules could potentially interact with many other proteins, not only the particular 'target' of therapeutic value. Although recent efforts [33–35] have shown the great power of linking protein engineering with chemical genomics to study protein function on a large-scale, engineered targets and unnatural pathway design strategies do not, in all probability, offer options for therapeutic intervention in humans. At this later stage of drug development, structural and physicochemical characterization of the target protein, its potential binding site(s), and analysis for other potential competing binding sites on other proteins are invaluable. Meaningful incorporation of this knowledge requires parallel genome-wide functional assignment in which function is determined early in the discovery process, providing thorough identification of entire families with a low false-positive rate.

Use of structure for function identification

Protein structure information is not necessarily available for all sequences that are currently pharmaceutical targets, and even in some cases such as GPCRs, structural information is severely limited [36]. Structural proteomic efforts [37] aim to provide high-throughput solutions to experimental structure determination, while mapping protein structure space, much as the genomic sequencing projects provided a high-throughput solution to genome sequencing. With representative structures covering 'fold space', computational methods can be used to predict the folds of closely related proteins, or fold family members. Homology modeling methods are usually robust enough to generate reasonable structure models when sequence similarity between an experimentally characterized sequence and an unknown sequence is >30%. Accurate computational models can provide valuable information, allowing some evaluation of the physicochemical properties of the protein's active, or functional, sites for lead discovery and development; however, many proteins are not 30% identical to proteins of known structure.

Currently, experimental structural efforts are resource intensive and are limited to proteins that can be cloned, expressed and purified in sufficient quantities to allow target structure determination. Proteins must form diffraction-quality crystals for X-ray crystallography or be well behaved in solution at sufficient concentrations for NMR spectroscopy. Many, although not all, proteins will satisfy these criteria [38]. Only after protein samples are available can structural biology methods determine high quality structures. At this time, none of these methods is fully automatable on a large-scale. The availability of sufficient public and private funding [38,39] ensures that many of the current difficulties in protein production will be tackled and eventually solved.

Even if structural determination was cheap and easy, knowing a protein's tertiary structure is not necessarily enough to determine the function of a protein or identify potential small-molecule binding sites. Protein structure determination is often viewed as one method for determining molecular function. The structure of the target protein is compared with that of a protein with pre-determined function, and if their structures are 'similar enough' a functional inference is made. Unfortunately, as in sequence genomics, this approach only works some of the time [40]. Recent studies of currently available structures have shown that functional inference can correctly predict molecular function for <50% of the proteins analyzed [41,42]. Early reports [43] from structural genomics efforts indicate that it is not straightforward to identify biochemical function or functional sites for unannotated proteins even with high-resolution structures in hand. In addition, proteins with different structures can support the same activity, complicating the problem of structure-based functional assignment (Fig. 2). These findings highlight the need for

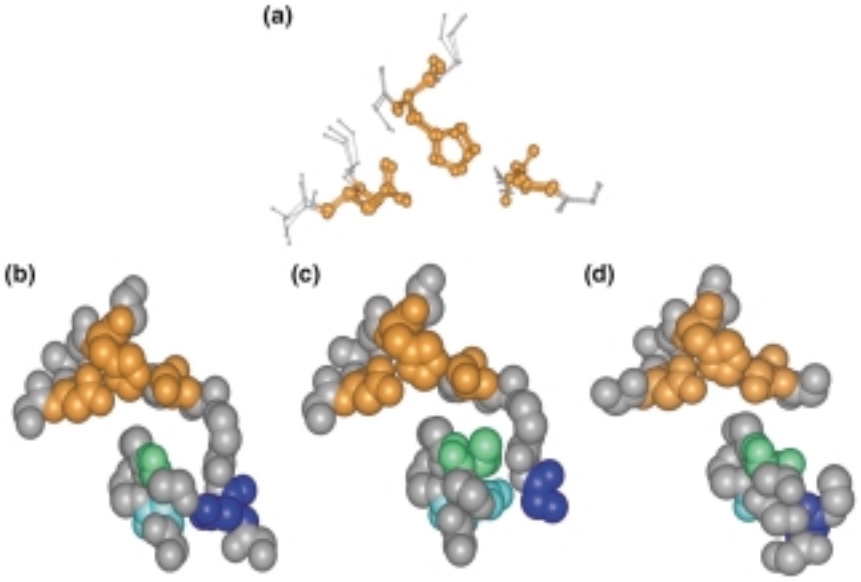

Figure 2. Functional sites contain amino acids important for both activity and specificity. The active sites of three serine proteases (trypsin clan) are shown here **(a–d)**. Trypsin [Protein Data Bank (PDB) code 1trnA] is specific for basic residues at the P1 position; elastase (PDB code 1b0fA) is specific for non-polar hydrophobic residues at the P1 position; and glutamic acid specific protease (PDB code 1hpgA) is specific for acidic residues at the P1 position. Panel **(a)** shows a superposition of the catalytic triads, demonstrating the conserved geometry underlying the chemistry common to all three proteins. The functional sites, encompassing the catalytic triad (orange sidechains) and the P1 pocket (positions key for specificity are colored green, teal and blue) for trypsin **(b)**, elastase **(c)** and glutamic acid-specific protease **(d)** are illustrated. Differences in the binding pocket geometry and chemistry lead to the selectivity of these proteins for substrates. The information that is distinct in each site could be exploited in post-genomic drug discovery.

biochemical or functional information from sources that are complementary to protein structure.

The function-first approach to structural and chemo-proteomics

To address the lack of structure and function information early in the lead discovery process, we have developed a 'function-first' approach, using unique capabilities in function and structure identification. The process starts with the identification of a set of protein sequences in the human proteome that exhibit a particular functional site that carries out catalysis, binds substrate or, perhaps, has been previously identified as an inhibitor binding site.

Instead of using sequence or global structural similarity to identify protein families, we have chosen to classify proteins based on their functional sites. For example, the kinase and serine protease protein families have well-characterized representatives that are therapeutic targets [44,45]. Despite their common family designation, protein families can include members with greatly different global folds and even different active-site structures (Fig. 2). The three-dimensional (3D) structure and the chemical characteristics of these functional sites can be analyzed to identify similarities and differences, recognizing that they are key to specific drug design, with the added advantage of being able to identify multiple functional sites within a single domain (Fig. 1). Protein structures of family members are solved by a novel combination of protein folding algorithms [46–48] coupled with sparse NMR constraints [49]. A combination of virtual screening tools and complementary experimental methods identifies small molecules that bind generally to several family members or specifically to one family member. This unique chemo-proteomic approach provides small-molecule leads on a large-scale, based on both structural and chemical information about target active sites. This approach provides crucial time and resource savings compared with large-scale parallel structure determination methods [50]. The next sections of this review outline key technologies and approaches to structural proteomics and lead discovery based on this function-first approach.

Rapid and approximate structure analysis

The first step in the function-first approach is the generation of approximate models for many protein sequences, using protein folding algorithms developed by Skolnick and co-workers [46–48]. Although these algorithms routinely perform well in CASP competitions [51,52], they suffer many of the same pitfalls of all protein folding algorithms. Because energy potentials are not all-encompassing and scoring functions can still be improved, today's protein folding algorithms typically produce only approximate models. Although not of the atomic-level detail provided by X-ray structures, these models are sufficient for automated proteomic functional analysis.

Functional site identification

We have developed a technique to identify biochemical function using Fuzzy Functional Forms™ (FFFs): motifs that describe the chemistry and geometry of functional sites in proteins [53–55]. FFFs are based on known, experimentally determined structures available in the Protein

Data Bank (PDB; http://www.rcsb.org/pdb). First, a set of key functional residues is identified in one or more functionally related protein structures. Residues are selected on the basis of the specific function, structure, or chemistry in which they participate. Next, a set of geometric constraints is defined that relates the relative location of the selected critical residues in 3D space. A key criterion for each FFF is the variability or flexibility allowed for each geometric constraint. Some geometric constraints are tightly defined, whereas others can be more varied.

The next step: a highly focused, functional family approach to structure-based drug discovery

The utility of the 'function-first' approach in drug discovery is readily seen when applied not simply to individual proteins, but when classes of targets (i.e. functional families) are examined. The goal of identifying focused families of proteins can be accomplished by taking full advantage of two pieces of information: (1) the detailed knowledge of the chemistry and functional site structure that is identified by the FFFs, and (2) using computational models to yield valuable biological and functional site information necessary for drug discovery. In this process, FFFs are used to identify all proteins in the human genome that display a given functional site [54,55].

Importantly, although functional sites are generally well-conserved within functional families, it is subtle differences near the functional sites (amino acid changes and/or their spatial arrangement) that lead to distinct biological activities and specificities (Fig. 2). These differences and similarities proximal to the functional site lead to a method for family sub-classification based on active-site structures and chemistries, rather than more typical global sequence comparisons that underlie evolutionary analyses. This information is key for designing small molecules that selectively inhibit a single family member while reducing cross-reactivity to other proteins in the same functional family. Importantly, this approach allows the enumeration and classification of an entire functionally related group of proteins. For example, differential expression experiments can identify a protein that might be involved in a disease state [16–18], but without a sense of the nearby active-site landscape, it is difficult to judge how 'druggable' that putative target might be.

These methods identify the set of amino acids at a given functional site, but the approach does not provide detailed structural geometry, and there will be occasions when more detailed structural information is desired. If done efficiently and quickly, this can provide key information for small molecule or library design. We have chosen to use computationally derived models to accelerate experimental structure determination. Other groups have also recognized the advantages of using sparse [56] or reduced datasets [57–60] to accelerate NMR structure determination. We have designed an integrated procedure that uses computational models (generated by comparative modeling [48] or *ab initio* methods [46]) as the starting point for structure calculations.

Benefits of the function-first approach to drug discovery

Function-first complements cell based information

The advantages of a function-first approach to structural proteomics complement the information from cellular functional proteomics methods. The parallel identification of protein function, at the cellular and molecular level, allows efficient decision-making for target validation and lead discovery. In particular, identification of biochemical function or molecular structure on a large-scale, without huge investments in time and resources, provides crucial information for the interpretation of microarray data and cellular level functional assays.

Rapid experimental analysis of structure

Using an integrated approach that rarely requires atomic level structure determination means that resource usage can be optimized. Other protein structures in the same functional family can be modeled computationally and/or refined using NMR data. This provides an approach that brings speed and efficiency to structure and function analysis. A key advantage of this approach is that protein structure models created using state-of-the-art folding algorithms can be screened automatically. Only models recognized by an FFF or that score well by modeling algorithms are considered further. Automation has allowed scientists at GeneFormatics to compute models for >25,000 human protein sequences in a matter of weeks, and make thousands of structure–function assignments. It is the combination of active-site structure and functional information that provides a more thorough and reliable assignment than sequence-based methods [53,61].

Identification of alternative drug-binding sites in known protein targets

In the human proteome, many protein sequences are expected to carry out multiple functions at the molecular level. A key advantage of using FFFs for functional site identification is the ability to analyze biochemical functional complexity. For example, one protein identified as a phosphatase by the available public annotation might actually present a catalytic site, a metal binding site, and a regulatory site, as illustrated for one protein in Fig. 1. The

single word 'phosphatase' does not adequately express this functional complexity that is so easily represented by the FFFs. Many of these alternative sites are potentially druggable sites, providing the opportunity to more specifically modulate the biological activity of a particular member of a functional family. Visibility into the multifunctional nature of a particular protein target at an early stage provides valuable information that can lead to more effective small-molecule intervention strategies.

Key structural and chemical information about the functional site

Knowledge of the chemistry and structure of the functional site is essential for exploiting the chemical experience gained over the past decades of structure-based drug design. Recognition of both the similarities and differences among a set of potential targets allows the possibility of screening or designing small molecules that are specific for individual members within each functional family – a 'functional family' approach to drug discovery.

Conclusion

Genomic information, mined effectively to gain crucial knowledge, can lead to compounds that are developed in parallel and that are more specific for their intended protein targets. By identifying and characterizing protein structure, function and active-site information early in the discovery process, pharmaceutical researchers can study multiple targets using integrated biological, structural and chemical methods simultaneously to more rapidly, and cost effectively, discover specific compounds that should lead to the first generation of drugs in the post-genomic era.

References

1 Kumar, A. *et al.* (2002) An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* 20, 58–63
2 Ouzounis, C.A. and Karp, P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol.* 3 comment2001.1-comment2001.6
3 Shuker, S.B. *et al.* (1996) Discovering high-affinity ligands for proteins: SAR by NMR. *Science* 274, 1531–1534
4 Fejzo, J. *et al.* (1999) The SHAPES strategy: an NMR-based approach for lead generation in drug discovery. *Chem. Biol.* 6, 755–769
5 Nienaber, V.L. *et al.* (2000) Discovering novel ligands for macromolecules using X-ray crystallographic screening. *Nat. Biotechnol.* 18, 1105–1108
6 Moy, F.J. *et al.* (2001) MS/NMR: a structure-based approach for discovering protein ligands and for drug design by coupling size exclusion chromatography, mass spectrometry, and nuclear magnetic resonance spectroscopy. *Anal. Chem.* 73, 571–581
7 Hajduk, P.J. *et al.* (2002) A strategy for high-throughout assay development using leads derived from NMR-based screening. *J. Biomol. Screening* (in press)
8 Klebe, G. (2000) Recent developments in structure-based drug design. *J. Mol. Med.* 78, 269–281
9 Maignan, S. and Mikol, V. (2001) The use of 3D structural data in the design of specific factor Xa inhibitors. *Curr. Top. Med. Chem.* 1, 161–174
10 Iversen, L.F. *et al.* (2001) Steric hindrance as a basis for structure-based design of selective inhibitors of protein-tyrosine phosphatases. *Biochemistry* 40, 14812–14820
11 Scapin, G. (2002) Structural biology in drug design: selective protein kinase inhibitors. *Drug Discov. Today* 7, 601–611
12 Caron, P.R. *et al.* (2001) Chemogenomic approaches to drug discovery. *Curr. Opin. Chem. Biol.* 5, 464–470
13 Wess, G. (2002) How to escape the bottleneck of medicinal chemistry. *Drug Discov. Today* 7, 533–535
14 Murcko, M. and Caron, P. (2002) Transforming the genome to drug discovery. *Drug Discov. Today* 7, 583–584
15 Baxter, S.M. and Fetrow, J.S. (2001) Sequence- and structure-based protein function prediction from genomic information. *Curr. Opin. Drug Discov. Devel.* 4, 291–295
16 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature* 405, 827–836
17 Wilson, D.S. and Nock, S. (2001) Functional protein microarrays. *Curr. Opin. Chem. Biol.* 6, 81–85
18 Zhu, H. *et al.* (2001) Global analysis of protein activities using proteome chips. *Science* 293, 2101–2105
19 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
20 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
21 von Mering, C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* 417, 399–403
22 Hardwick, J.S. *et al.* (1999) Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins. *Proc. Natl. Acad. Sci. U. S. A.* 96, 14866–14870
23 Bernstein, B.E. *et al.* (2000) Genomewide studies of histone deacetylase function in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13708–13713
24 Lewis, T.S. *et al.* (2000) Identification of novel MAP kinase pathway signaling targets by functional proteomics and mass spectrometry. *Mol. Cell* 6, 1343–1354
25 Fleming, J.A. *et al.* (2002) Complementary whole-genome technologies reveal the cellular response to proteasome inhibition by PS-341. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1461–1466
26 Sche, P.P. *et al.* (2001) Corrigendum to: 'Display cloning: functional identification of natural product receptors using cDNA-phage display'. *Chem. Biol.* 8, 399–400
27 Cochrane, D. *et al.* (2000) Identification of natural ligands for SH2 domains from a phage display cDNA library. *J. Mol. Biol.* 297, 89–97
28 Christensen, D.J. *et al.* (2001) Phage display for target-based antibacterial drug discovery. *Drug Discov. Today* 6, 721–727
29 Han, D.K. *et al.* (2001) Quantitative profiling of differentiation-induced microsomal proteins using isotope-coded affinity tags and mass spectrometry. *Nat. Biotechnol.* 19, 946–951
30 Kolb, H.C. *et al.* (2001) Click chemistry: diverse chemical function from a few good reactions. *Angew. Chem. Int. Ed. Engl.* 40, 2004–2021
31 Bogyo, M. *et al.* (2000) Selective targeting of lysosomal cysteine proteases with radiolabeled electrophilic substrate analogs. *Chem. Biol.* 7, 27–38
32 Cravatt, B.F. and Sorensen, E.J. (2000) Chemical strategies for the global analysis of protein function. *Curr. Opin. Chem. Biol.* 4, 663–668
33 Bishop, A. *et al.* (2000) Unnatural ligands for engineered proteins: new tools for chemical genetics. *Annu. Rev. Biophys. Biomol. Struct.* 29, 577–606
34 Carroll, A.S. *et al.* (2001) Chemical inhibition of the Pho85 cyclin-dependent kinase reveals a role in the environmental stress response. *Proc. Natl. Acad. Sci. U. S. A.* 98, 12578–12583
35 Bishop, A.C. *et al.* (2000) A chemical switch for inhibitor-sensitive alleles of any protein kinase. *Nature* 407, 395–401
36 Palczewski, K. *et al.* (2000) Crystal structure of rhodopsin: A G-protein-coupled receptor. *Science* 289, 739–745
37 Smith, T. (2000) A new era. *Nat. Struct. Biol.* 7, 927
38 Norin, M. and Sundstrom, M. (2002) Structural proteomics: developments in structure-to-function predictions. *Trends Biotechnol.* 20, 79–84

39 Montelione, G.T. (2001) Structural genomics: an approach to the protein folding problem. *Proc. Natl. Acad. Sci. U. S. A.* 98, 13488–13489

40 Hegyi, H. and Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* 11, 1632–1640

41 Koppensteiner, W.A. *et al.* (2000) Characterization of novel proteins based on known protein structures. *J. Mol. Biol.* 296, 1139–1152

42 Wilson, C.A. *et al.* (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* 297, 233–249

43 Yee, A. *et al.* (2002) An NMR approach to structural proteomics. *Proc. Natl. Acad. Sci. U. S. A.* 99, 1825–1830

44 Sawyers, C.L. (2002) Rational therapeutic intervention in cancer: kinases as drug targets. *Curr. Opin. Genet. Dev.* 12, 111–115

45 Katz, B.A. *et al.* (2001) Engineering inhibitors highly selective for the S1 sites of Ser190 trypsin-like serine protease drug targets. *Chem. Biol.* 8, 1107–1121

46 Kihara, D. *et al.* (2001) TOUCHSTONE: an *ab initio* protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. U. S. A.* 98, 10125–10130

47 Skolnick, J. and Kihara, D. (2001) Defrosting the frozen approximation: PROSPECTOR – a new approach to threading. *Proteins* 42, 319–331

48 Kolinski, A. *et al.* (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. *Proteins* 44, 133–149

49 Skolnick, J. *et al.* (1997) MONSSTER: a method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* 265, 217–241

50 Abola, E. *et al.* (2000) Automation of X-ray crystallography. *Nat. Struct. Biol.* 7 (Suppl), 973–977

51 Venclovas, C. *et al* (2001) Comparison of performance in successive CASP experiments. *Proteins* 45 (Suppl 5), 163–170

52 Zemla, A. *et al.* (2001) Processing and evaluation of predictions in CASP4. *Proteins* 45 (Suppl 5), 13–21

53 Fetrow, J.S. and Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. *J. Mol. Biol.* 281, 949–968

54 Fetrow, J.S. *et al.* (2001) Genomic-scale comparison of sequence- and structure-based methods of function prediction: does structure provide additional insight? *Protein Sci.* 10, 1005–1014

55 Di Gennaro, J.A. *et al.* (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.* 134, 232–245

56 Bowers, P.M. *et al.* (2000) *De novo* protein structure determination using sparse NMR data. *J. Biomol. NMR* 18, 311–318

57 Zweckstetter, M. and Bax, A. (2001) Single-step determination of protein substructures using dipolar couplings: aid to structural genomics. *J. Am. Chem. Soc.* 123, 9490–9491

58 Tian, F. *et al.* (2001) A dipolar coupling based strategy for simultaneous resonance assignment and structure determination of protein backbones. *J. Am. Chem. Soc.* 123, 11791–11796

59 Prestegard, J.H. *et al.* (2001) Nuclear magnetic resonance in the era of structural genomics. *Biochemistry* 40, 8677–8685

60 Kraulis, P.J. (1994) Protein three-dimensional structure determination and sequence-specific assignment of 13C and 15N-separated NOE data. A novel real-space *ab initio* approach. *J. Mol. Biol.* 243, 696–718

61 Fetrow, J.S. *et al.* (1998) Functional analysis of the *Escherichia coli* genome using the sequence- to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.* 282, 703–711

62 Fetrow, J.S. *et al.* (1999) Structure-based functional motif identifies a potential disulfide oxidoreductase active site in the serine/threonine protein phosphatase-1 subfamily. *FASEB J.* 13, 1866–1874