ELSEVIER

Il Farmaco 57 (2002) 243–251

IL FARMACO

www.elsevier.com/locate/farmac

Invited Review

Identification and mapping of small-molecule binding sites in proteins: computational tools for structure-based drug design

Christoph Sotriffer *, Gerhard Klebe

Department of Pharmaceutical Chemistry, Philipps-University Marburg, Marbacher Weg 6, D-35032 Marburg, Germany

Received 13 December 2001

Abstract

The number of protein structures is currently increasing at an impressive rate. The growing wealth of data calls for methods to efficiently exploit structural information for medicinal and pharmaceutical purposes. Given the three-dimensional (3D) structure of a validated protein target, the identification of functionally relevant binding sites and the analysis ('mapping') of these sites with respect to molecular recognition properties are important initial tasks in structure-based drug design. To address these tasks, a variety of computational tools have been developed. Approaches to identify binding pockets include geometric analyses of protein surfaces, comparisons of protein structures, similarity searches in databases of protein cavities, and docking scans to reveal areas of high ligand complementarity. In the context of binding-site analysis, powerful data mining tools help to retrieve experimental information about related protein–ligand complexes. To identify interaction hot spots, various potential functions and knowledge-based approaches are available for mapping binding regions. The results may subsequently be used to guide virtual screenings for new ligands via pharmacophore searches or docking simulations. © 2002 Éditions scientifiques et médicales Elsevier SAS. All rights reserved.

Keywords: Cavity detection; Binding-site characterization; Protein structures; Protein-ligand complexes

1. Introduction

A drug must interact with a biological target molecule, most often a protein, to exert a physiological function. Conversely, protein function is almost invariably linked with the specific binding of substrates or endogenous ligands. Given the fundamental importance of these recognition events at the molecular level, a key role for the understanding of drug action and hence the rational design of drugs is attributed to structural information about the interacting molecules. Accordingly, three-dimensional (3D) structures of protein targets represent the preferred starting point for drug design projects [1–5].

The current initiatives of structural genomics are a clear reflection of the generally recognized importance of protein structures for biomedical research [6]. With the access to sequences of entire genomes of various organisms, the goal of these efforts is to provide a comprehensive view of the protein structure universe. High-throughput X-ray crystallography and NMR spectroscopy are expected to yield some 10^4 experimental protein structures within the next years, such that (preferably) at least one experimental structure is determined for every protein sequence family. Computational homology modeling techniques could then be applied to obtain structural models for virtually every protein in nature. Whatever the strategy and the actual success of the structural genomics efforts will be, they are likely to accelerate the growth rate of protein structural information beyond the already impressive pace at which new structures are currently deposited to the protein databank PDB [7,8]. Consequently, efficient methods are required that allow to exploit this wealth of structural data for the purpose of drug design.

Once the 3D structure of a protein is given, the strategy to follow in a design project clearly depends on the additional information that is available about the biological target of interest. In general, three cases may be distinguished:

* Corresponding author.
E-mail address: sotriffer@mailer.uni-marburg.de (C. Sotriffer).

PII: S0014-827X(02)01211-9

1.1. The binding site is unknown

Normally, at least the general function of the target protein is known, but it may not yet be understood in structural terms. This requires methods to identify the binding region that should be targeted by a drug to interfere with the protein's function. In addition, it will increasingly be the case that structural genomics provides structures of proteins whose actual biochemical function has yet to be assigned. Such situations call for methods to infer protein function from the 3D structure before an actual design process can start. Since function is normally linked to binding, the tasks of elucidating functional aspects and identifying binding pockets are interrelated to some extent.

1.2. The approximate binding region is known, but neither information about its characteristics, nor about ligand interactions is available

This situation is more commonly encountered and requires a careful analysis of the protein structure in the binding area to identify regions most favorable for interaction, indicating 'hot spots' where certain functional groups might preferably bind. This way a functional map of the binding site is generated that can guide the placement of potential ligands. Methods of this kind may also directly include the possibility to place or build-up ligands within the binding site or to generate a pharmacophore model for database searching.

1.3. The binding site is known and crystal structures with bound ligands are available

Given a well-defined binding site, methods for docking and virtual screening can be applied. Implicitly, however, these methods still rely on binding site analyses, as some sort of functional or energetical map is generally required to guide the search for suitable ligand positions.

In the following, a selection of computational methods addressing the tasks of binding site identification and analysis is briefly presented. Slight emphasis is given to knowledge-based approaches. These try to use the wealth of structural information contained in crystal structures of protein–ligand complexes and small molecules to derive rules how a ligand could interact with a protein receptor. Rules of this kind can help solving the aforementioned tasks without having to resort on frequently incomplete and simplified models. In this sense they can also help to overcome deficiencies in the current understanding of molecular recognition in biological systems.

2. Identification of potential binding sites

The specific, functional binding of small molecules is usually mediated by depressions in the protein surface. Accordingly, binding sites are generally referred to as cavities, grooves, pockets, or clefts. The phenomenon of small-molecule binding in surface depressions is ultimately a consequence of the physical principles governing molecular recognition: high affinity can only be gained by sufficiently large interaction interfaces, and specificity is more easily obtained within environments that already impose geometric constraints.

Given the preponderance of pockets or cavities as binding site locations, computational tools have been developed to detect such depressions on the protein surface in order to localize potential binding sites. These methods normally rely on purely geometric criteria, differing, however, in the algorithmic approaches being used. Three examples of fast geometric cavity identification programs are LIGSITE [9], Automatic PROtein POcket Search ('APROPOS') [10], and Putative Active Site with Spheres ('PASS') [11].

LIGSITE is based on the earlier developed POCKET program [12] and embeds the considered protein in a regularly spaced grid. Lattice intersections coinciding with a protein atom's van der Waals sphere are discarded and the remaining lattice points are scored according to their degree of burial in surface depressions. The degree of burial is determined by scanning the grid lines along the three Cartesian axes and the four cubic diagonals for areas that are enclosed by protein atoms on both sides. Adjacent lattice points of high burial are clustered to reveal contiguous cavities. For a small test set of ten protein–ligand complexes the program was reported to identify the correct location of the binding site with high precision in each case [9]. It should be noted, though, that apparently in all test cases the binding site corresponded to the largest pocket on the surface (cf. remarks further below).

Instead of using grid representations of the protein-surrounding space, APROPOS and PASS follow different approaches. The APROPOS algorithm uses a so-called alpha-shape description of the protein surface. Pockets are identified by comparing surfaces generated with different levels of resolution, i.e. an envelope surface describing the global shape of the protein and a suitably detailed surface reflecting the local structure. Based on tests with more than 300 proteins, the method was reported to locate binding sites with high reliability [10]. The program CAST is a further, more recent example of a method based on alpha-shape theory [13].

In PASS, the protein is first coated with a layer of spherical probes. These probes are then filtered to eliminate those that either clash with the protein, are not sufficiently buried, or are located too close to a more buried probe. A new layer of probes is then grown onto the scaffold of all previously identified probes and filtered as before. Growing and filtering are repeated until no probes in a new layer survive the filters. For all spheres of the final set, probe weights are computed which are proportional to the number of probe spheres in the vicinity and the extent to which they are buried. Out of the probes with the highest weights, the so-called 'active-site points' (ASP) are determined, which should represent the center of a potential binding site. For a test set of 20 apo-protein crystal structures, PASS was able to identify the location of the ligand-binding site in 12 cases as top-ranked ASP, in 16 cases as an ASP among the top three [11].

Purely geometric tools for binding site identification are generally well suited to localize all significant cavities or depressions on the protein surface. However, if for a given protein multiple cavities are found, the problem arises how to recognize which of these is a functionally relevant binding site. The question, what distinguishes a binding site from other cavities on the protein surface, is actually of quite fundamental interest [14]. While a simple general answer can hardly be given, the characterization of cavities in terms of physicochemical parameters is likely to point into the right direction for finding a solution to this problem. Purely geometric criteria, however, seem to suffice in favorable cases. Size, shape, and burial extent of protein cavities dictate the geometry of ligands that can be favorably accommodated, as good steric fit is usually a minimal requirement for high-affinity binding. In comparative analyses it has indeed been found that enzyme active sites are often characterized by a particularly large and deep cleft [13,15]. Accordingly, in many cases the active site of an enzyme can be recognized successfully by finding the largest cleft on the protein surface.

Methods going beyond a purely geometrical analysis need to score the various cavities according to some physico-chemical criteria or using an energy function. This is normally performed by scanning the surface for areas of high complementarity with respect to certain molecular fragments or entire ligands. As a matter of fact, it is also the nature of the ligand that determines which cavity is addressed as a binding site. In principle, any docking method should, therefore, be capable of identifying potential binding sites. In practice, however, standard docking methods are normally not used for this purpose, mostly because they are not efficient enough to scan entire protein surfaces in reasonable computing time. In addition, standard scoring functions are frequently not able to provide clear-cut discrimination between alternative binding locations. Nevertheless, some docking methods have been developed with the explicit intention of identifying possible ligand binding sites. An example is the approach presented by Ruppert et al. [16]. Here, the protein is coated with molecular fragments, or probes, and the position of each fragment is scored with a function parameterized on experimental binding energies to give an estimated affinity value for each probe position. The binding site is then detected by screening for regions in which high-affinity probes cluster and localizing the cluster with highest overall affinity (score). Another example is the vdW-FFT method described by Bliznyuk and Gready [17,18]. Based on van der Waals-energy terms evaluated on a regular cubic grid around the protein, fast Fourier transform techniques are used to perform a systematic scan of the entire protein surface for possible ligand orientations to identify the best geometrical matches. A set of best-matching ligand orientations is subsequently refined by molecular mechanics energy minimization, followed by evaluation of binding energies using Poisson–Boltzmann-type calculations. The top-ranked ligand orientations found in this way should elucidate the actual binding site.

All approaches mentioned so far try to identify binding sites by relying exclusively on the 3D structure of the protein under consideration. Frequently, however, information about function and binding sites of related systems is already available. Proteins of related function often share a comparable recognition pocket. With a minimum of functional information available, the binding site of a new structure may, therefore, be detected by comparison with other proteins of the same function. Conversely, a comparison of pockets on the surface of different proteins may allow to detect functional relationships. Accordingly, comparisons based on some sort of similarity with well-characterized proteins of known structure and function can provide an additional route to the identification of functionally relevant binding sites of the query protein. As with any approaches that make use of existing knowledge, the scope and success rate of these methods will further increase as the number of solved protein structures continues to grow.

In general, the essence of these methods to infer protein function from 3D structure is 'similarity' or 'homology'. Traditionally, bioinformatics assigns functional data by searching for relatives in sequence databases [19]. However, many relationships can only be detected from the 3D structure, which is more conserved during evolution than sequence similarity [20]. Various algorithms are available for comparing protein structures in 3D to recognize structurally related proteins [21]. These programs are efficient enough to perform rapid searches of entire structural databases such as the PDB. The results of mutual comparisons for all known protein structures are themselves stored in databases that provide classifications of protein structures, in part with functional annotations. An example for such a database is CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new) [22,23]. NCBI's 'Entrez Structure' is another, highly integrated database service that allows to search for neighbors in sequence and structure (http://www.ncbi.nlm.nih.gov/Structure) [24]. Yet another example is the FSSP database (http://www2.ebi.ac.uk/dali/fssp/fssp.html) [25–27], which stores the structural neighbors of all proteins in the PDB. FSSP stands for 'Fold classification based on Structure-Structure alignment of Proteins'. The algorithm used for the alignment and classification is provided as free network service through the Dali server (http://www2.ebi.ac.uk/dali). Coordinates of a new protein structure submitted to the server are compared against all others in the PDB. This can reveal functionally interesting similarities that are not detectable by comparing sequences. An example is barley endochitinase, an enzyme involved in plant defense reactions [21]. Sequence analysis and site-directed mutagenesis studies failed to identify the active site, but structural comparisons revealed a similarity with lysozyme subclasses. Importantly, location and composition of the active site and key structural residues were found to be conserved in endochitinase.

Although all these tools may provide first hints on function and binding site location, the relationship between structure and function is by no means simple and straightforward. A similar fold does not necessarily imply a similar biochemical function and proteins with different folds can also show the same function and catalytic mechanism (as for example the serine proteases trypsin and subtilisin). It is, therefore, often advisable to go beyond the comparison of protein folds or global structural motifs in order to look at local structural motifs, i.e. at the details of a protein's active site. Local structural motifs, such as the catalytic triads of enzymes, can capture the essence of the biochemical function and thus be used to assign function [28–30].

A new approach to detect functional similarity independent of sequence and fold homology goes beyond the simple search for structural motifs and uses instead physicochemical comparisons of protein cavities [31]. This is based on the rationale that protein function is often intimately connected with the recognition of ligands, which usually occurs in well-defined clefts or cavities of the protein surface. In enzymes, for example, elementary steps of the catalyzed reaction require a strictly defined spatial arrangement of the reaction partners. This in turn means that the determinants of molecular recognition need to be highly conserved in their relative orientation. The conservation of molecular recognition patterns between binding sites of functionally similar proteins should, therefore, allow to identify functional relationships and to localize binding sites by searching for similarities within protein cavities. To capture the features that are essential for molecular recognition, such a search should be based on surface-exposed physico-chemical properties.

More in detail, the method works as follows: using the aforementioned LIGSITE algorithm to detect depressions on protein surfaces, cavities are retrieved from the entire body of protein crystal structures and stored in a new database called CAVBASE. The atomic coordinates of the residues flanking the cavity are reduced to a set of generic pseudo centers, classified according to five properties: hydrogen bond donor, acceptor, mixed donor/acceptor, hydrophobic aliphatic, and aromatic contacting group. These pseudo centers are further examined for their surface exposure and assigned to the nearest lattice surface intersections (the cavity detection algorithm is based on a lattice representation of the protein-near space). The cavity shape, the set of assigned descriptors of exposed recognition properties, and the corresponding surface patches are all stored in CAVBASE. The way the information is stored allows for fast and efficient comparisons within large data sets. The implemented search algorithm tries to detect common subgraphs generated by nodes that correspond to pairs of pseudo centers of equivalent properties and similar mutual distances. Appropriate tolerances have been incorporated to consider structural variations resulting from conformational flexibility of the protein and inherent limitations of the accuracy of protein structure determination. The results of a comparison are ranked by scoring the matches in terms of the assigned surface-exposed physicochemical properties. Scanning a cavity of interest ('query cavity') against a sample of several thousand binding pockets should ideally yield as top-ranked results binding pockets that exhibit local surface similarity and share binding motifs with the query cavity. As this approach is entirely based on physicochemical properties exposed toward the surface rather than amino acid complementarity, it allows the detection of relationships independent of any sequence or fold homology. This is confirmed by examples carried out with chorismate mutases and serine proteases. Searches within a cavity database of more than 5000 entries retrieved proteins of similar function as top-ranked hits even in cases where the corresponding systems do not show any significant sequence or fold homology [31]. This new database of binding-site cavities characterized by essential recognition elements may, therefore, serve as powerful tool to detect functionally important sites in new proteins that might be targeted in drug design projects.

3. Analysis and mapping of binding sites

The purpose of structure-based drug design is to identify or construct molecules that bind with high affinity to a structurally defined binding site of a target protein. The binding site specifies structural and physicochemical constraints that must be met by any putative ligand [32]. It is hence imperative to analyze the constitution of the binding site by mapping the characteristics that are essential for ligand recognition.

As a starting point it is frequently useful to retrieve all structural information that is already available about the system under consideration. Databases with efficient user interfaces and powerful query tools are required for this purpose. With respect to protein–ligand complexes such data-mining can be performed with the help of REceptor-LIgand dataBASE (‘RELIBASE’) (http://relibase.ccdc.cam.ac.uk) [33,34]. RELIBASE has been designed as a fast and flexible tool for the retrieval, visualization, and analysis of information about protein–ligand complexes. It contains all the PDB structures together with additional data such as ligand atom- and bond types, substructures, protein sequence similarity, and crystallographic packing. As such it can be used to compare structures, binding sites, and ligand binding modes. The query system allows to perform ligand-similarity or substructure searches, searches for similar binding sites (in terms of sequence homology), and protein–ligand interaction searches to analyze interaction patterns between ligand fragments and amino acid functional groups. Search results may be superimposed for simultaneous display and the retrieved interaction geometries can be tabulated. This frequently provides useful clues about preferred interaction motifs and helps to reveal important characteristics of a binding site.

Prerequisite for tight ligand binding are specific interactions formed with protein atoms in the binding site. These are usually non-covalent in nature (e.g. ionic interactions, hydrogen bonds, van der Waals forces) and should in sum exceed unfavorable contributions such as desolvation or immobilization of translational and rotational degrees of freedom. A detailed analysis of the receptor site should, therefore, identify ‘hot spots’ of binding, i.e. those regions where most favorable non-covalent interactions are formed.

Several approaches directed toward this task are available. Most of them try to determine favorable binding locations by placing atom probes, molecular fragments, or small molecules at various points in the binding site and evaluating their interactions. Such methods have been classified as ‘fragment location’, ‘fragment placement’, and ‘site-point connection’ methods [35]. The simple display of the results (hot spots) together with the receptor structure can already be used as valuable guide for the design of new ligands. Some methods also allow to use these results directly for subsequent ligand construction or docking to the binding site [35].

A first class of methods is based on some sort of energy function to identify regions favorable for interaction with particular ligand functional groups. Frequently, methods of this kind use a discrete 3D lattice to position probe atoms or groups within the binding site. The archetypal program of this class is GRID [36]. It places probes such as methyl, hydroxyl, ammonium, or carbonyl at regularly spaced grid points within the active site. At each grid point the implemented energy function is used to calculate the interaction energy between the probe and the protein. Following this concept a functional map of the binding site is constructed which indicates the most favorable regions for placing ligand groups with similar properties to the probes. Visualization of the maps by contouring at appropriate energy levels reveals the binding-site hot spots. Since the first introduction of GRID new types of probes and energy functions have been developed to further enhance the reliability of the method [37–39].

It is worth noting, though, that in principle any scoring function could be applied to perform such grid-based mappings and hot-spot analyses. Scoring functions are normally used in the context of docking to estimate binding affinities (for a brief review of currently used scoring functions see [40]). Many docking methods, as for example AutoDock [41–43], ICM [44–46], DOCK [47,48], or ProDock [49,50], make use of grid representations to speed up the energy evaluation during the docking process. These implicit binding site maps on which the docking relies could also be analyzed explicitly, primarily by visualizing the corresponding hot spots.

Also the knowledge-based DrugScore pair-potentials have been used for grid-based hot-spot analyses [51]. To test their performance in such applications, a set of 158 crystallographically determined protein-ligand complexes was analyzed with respect to the spatial coincidence of hot spots with the experimentally observed occurrence of matching ligand atom types at these hot-spot sites. Depending on the atom-type classification, overall prediction rates between 74 and 85% were obtained. Results of this kind clearly highlight the relevance of such binding site analyses as guidelines in structure-based design.

An alternative to grid-based approaches is the multiple copy simultaneous search (MCSS) method [52]. Instead of using probe atoms on a regular grid, several copies (usually some thousand) of probe groups are randomly distributed over the binding site and then subjected to energy minimization along with molecular dynamics simulation. During these calculations the probes are invisible to each other and experience only the forces from the protein atoms. The probes can thus cluster in local minima, which allows to identify the most favorable interaction sites.

Various extension and variants of the original MCSS approach have been developed. It has, for example, been coupled with methods for automated ligand design, which makes use of the optimized functional-group positions, generated by MCSS [53–55]. The methodology has also been extended to include flexibility of the protein target, which is in contrast to standard MCSS where the protein is kept rigid [56]. Protein flexibility is also taken into account by a method to generate so-called dynamic pharmacophore models: here, multiple conformations of a protein binding site are considered for an MCSS-type mapping based on Monte Carlo sampling instead of the standard molecular dynamics procedure [57,58].

A useful approach for visual analysis of binding site characteristics is the mapping of physicochemical properties onto molecular surface representations. The property most commonly used in this context is the electrostatic potential. It is normally obtained by solving the Poisson–Boltzmann equation [59], for which a variety of programs is available (e.g. UHBD [60], DELPHI [61]). Various molecular visualization programs can then be used to color-code appropriate surface representations of the protein. One of the most popular programs in this context is GRASP (http://honiglab.cpmc.columbia.edu/grasp) [62]. Apart from visualization routines, GRASP also contains an internal Poisson–Boltzmann solver and may, therefore, by used as standalone tool for electrostatic-potential surface mapping. Parts of its functionality are provided through the online service GRASS (http://honiglab.cpmc.columbia.edu/surfserv.html) where users may search for PDB entries or submit their own coordinates to obtain surface representations of the corresponding structure with certain properties mapped onto them. Properties available for selection include not only the electrostatic potential, but also various simple hydrophobicity measures (based on atom type, residue type, or transfer free energy).

Since useful clues about the binding of nonpolar groups can be obtained from hydrophobic surface patches, more sophisticated ways for generating hydrophobicity maps have also been developed [63]. Here, the binding energy of a nonpolar probe sphere rolled over the protein surface is calculated based on the van der Waals interaction and the electrostatic desolvation energy of the protein. The results are color coded and mapped onto a molecular surface generated with GRASP. Comparative evaluation of the method using ten diverse protein–ligand complexes revealed a high predictive power with respect to binding modes of nonpolar groups. The binding energies of the nonpolar probe sphere are also used in the docking program SEED to direct the docking of nonpolar molecular fragments [64].

A further, completely different class of methods is given by rule-based or knowledge-based approaches, the essence of which is to make use of the information stored in the vast amount of experimental (crystallographic) data through the derivation of rules for preferred protein-ligand interaction patterns. This idea has been followed in the so-called composite crystal-field approach [65,66]. Here, the Cambridge Structural Database (CSD) of small molecule crystal structures [67] was statistically analyzed for intermolecular contact geometries of various functional groups, as found in the crystal packing of organic molecules. Search results for contacts between two functional groups *X* and *Y* were superimposed onto the *X* groups, producing scatterplots of the experimental distribution of *Y* around *X*. This composite picture of possible interaction geometries indicates orientational preferences and can thus be used to guide the placement of ligand functional groups in the protein binding site.

In the program LUDI, the results of this statistical analysis of nonbonded interactions have been translated into rules to calculate so-called interaction sites [68,69]. These interaction sites are discrete positions and vectors in space suitable for forming hydrogen bonds or filling hydrophobic pockets. As such they represent a functional map of the binding site. Since LUDI is actually a tool for de-novo design, it does not stop at this point, but proceeds by matching molecular fragments onto these sites. The placed fragments can be interpreted as a more sophisticated functional map, providing a more detailed information about chemical moieties that may be favorably placed in the binding site. In a subsequent step, the program enters a so-called link mode and tries to connect suitable fragments with small bridging groups (e.g. –CH_2– or –COO–) to form a contiguously connected molecule. The final structures are then scored using a fast empirical scoring function. Various recent examples have demonstrated the usefulness of LUDI in real-life drug design projects [70–72]. Besides LUDI, also the docking program FLEXX uses results of the composite crystal-field analysis to guide the placement of ligands into binding pockets [73,74].

The idea of analyzing small molecule crystal structures for intermolecular contacts has also resulted in the generation of an entire database of nonbonded interaction geometries, called ISOSTAR [75]. This database presents non-bonded interactions in terms of scatterplots, which show the distribution of contacting groups around a central group. These distributions can be transformed into density maps, which can then be displayed as contoured surfaces. The library contains more than 10 000 scatterplots based on nonbonded contacts observed in the CSD compiled from about 300 central groups surrounded by up to ≈40 types of different contact groups.

The program SUPERSTAR has been developed for identifying interaction sites in proteins based entirely on the information stored in ISOSTAR [76,77]. For this purpose, a template molecule (e.g. a protein binding site) is decomposed into structural fragments. The scatterplots, showing the distribution of a selected probe around these structural fragments, are superimposed on the corresponding portions of the template. The scatterplots are then translated into a 3D map that shows the propensity of the probe at different positions around the template molecule. This propensity reflects the probability of finding a contact group in a particular region in space. For a test set of 122 protein–ligand complexes, Superstar detects the correct atom type for solvent-inaccessible ligand atoms in 82–90% of the cases, depending on the atom-type classification. Recently, also PDB-based interaction fields have been added to SUPERSTAR [78]. In a comparative evaluation, they were found to be more suitable to identify hydrophobic interaction sites, but overall they appeared equally successful as the original CSD-based maps.

4. Concluding remarks

Whatever method is chosen for binding-site analysis and the generation of functional maps, the results usually serve the purpose of supporting interactive design work and providing suggestions for the tailored modification of ligands. In addition, they are of fundamental importance for effective virtual screening [79], both by 3D pharmacophore searches [80,81] and docking calculations [82]. Recent examples of successful structure-based design for targets as diverse as carbonic anhydrase [83], tRNA-guanine transglycosylase [70], and DNA-gyrase [71] have provided compelling evidence that the hierarchical, stepwise application of these techniques constitutes a promising and powerful strategy.

The presently available methods for binding-site identification and binding-site analysis are valuable computational tools to exploit protein structural information for the purpose of ligand design. Nevertheless, it is also well known that a considerable number of limitations exist that still preclude an easy, fast, and fully automated way from a target structure to a lead or even a drug. Careful application of the methods and a careful interpretation of the results is, therefore, mandatory, as is further research to improve the current methods. In addition, success also depends on the quality of the experimental data (structural, energetical, and biochemical alike) onto which part of these methods are built. High-throughput experimentation and data acquisition should, therefore, not be pursued at the expense of quality.

Generally, the prediction accuracy of most of the presented methods is roughly around 80% (as established by comparisons of calculated binding modes or interaction hot spots with experimental data). Different methods have different weaknesses and strengths, though, and it is frequently advisable to apply a combination of approaches to tackle a certain problem. If consistent results are obtained, this supports the conclusions; if alternative results are obtained, their plausibility may be assessed based on available experimental data and on the underlying approximations of the method applied. Moreover, they may also provoke new experiments to be performed.

While some of the current limitations are due to simplifications required to keep models and algorithms tractable within reasonable computing times, others are clear reflections of persisting problems in understanding and modeling the fundamental process of molecular recognition in biological systems. These include issues of protein flexibility, the interactions with water, the dynamic nature of the binding event and ligand induced steric and electrostatic effects, not to mention a proper consideration of the cellular environment. Due to the current inability to quantify binding with precision based on theoretical considerations and affordable calculations, knowledge-based methods may represent the preferred route for some time to come. Once the underlying data are processed, approaches of this kind are fast. More importantly, they take advantage of the growing body of experimental data and implicitly consider many effects that are not yet fully understood or manageable in a quantitative way.

Acknowledgements

Thanks are due to Ruth Brenk, Holger Gohlke, Judith Günther, and Daniel Kuhn (University of Marburg) for helpful discussions.

References

[1] I.D. Kuntz, E.C. Meng, B.K. Shiochet, Structure-based molecular design, Acc. Chem. Res. 27 (1994) 117–123.

[2] R.S. Bohacek, C. McMartin, W.C. Guida, The art and practice of structure-based drug design: a molecular modeling perspective, Med. Res. Rev. 16 (1996) 3–50.

[3] R.E. Hubbard, Can drugs be designed? Curr. Opin. Biotechnol. 8 (1997) 696–700.

[4] G. Klebe, Recent developments in structure-based drug design, J. Mol. Med. 78 (2000) 269–281.

[5] P.J. Gane, P.M. Dean, Recent advances in structure-based rational drug design, Curr. Opin. Struct. Biol. 10 (2000) 401–404.

[6] S.K. Burley, An overview of structural genomics, Nat. Struct. Biol. (Suppl.) 7 (2000) 932–934.

[7] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.

[8] H.M. Berman, T.N. Bhat, P.E. Bourne, Z. Feng, G. Gilliland, H. Weissig, J. Westbrook, The protein data bank and the challenge of structural genomics, Nat. Struct. Biol. (Suppl.) 7 (2000) 957–959.

[9] M. Hendlich, F. Rippmann, G. Barnickel, LIGSITE: automatic and efficient detection of potential small-molecule binding sites in proteins, J. Mol. Graph. Model 15 (1997) 359–363.

[10] K.P. Peters, J. Fauck, C. Frommel, The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria, J. Mol. Biol. 256 (1996) 201–213.

[11] G.P. Brady Jr, P.F. Stouten, Fast prediction and visualization of protein binding pockets with PASS, J. Comput. Aided Mol. Des. 14 (2000) 383–401.

[12] D.G. Levitt, L.J. Banaszak, POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids, J. Mol. Graph. 10 (1992) 229–234.

[13] J. Liang, H. Edelsbrunner, C. Woodward, Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design, Protein Sci. 7 (1998) 1884–1897.

[14] D. Ringe, What makes a binding site a binding site, Curr. Opin. Struct. Biol. 5 (1995) 825–829.

[15] R.A. Laskowski, N.M. Luscombe, M.B. Swindells, J.M. Thornton, Protein clefts in molecular recognition and function, Protein Sci. 5 (1996) 2438–2452.

[16] J. Ruppert, W. Welch, A.N. Jain, Automatic identification and representation of protein binding sites for molecular docking, Protein Sci. 6 (1997) 524–533.

[17] A.A. Bliznyuk, J.E. Gready, Identification and energetic ranking of possible docking sites for pterin on dihydrofolate reductase, J. Comput. Aided Mol. Des. 12 (1998) 325–333.

[18] A.A. Bliznyuk, J.E. Gready, Simple method for locating possible ligand binding sites on protein surfaces, J. Comput. Chem. 20 (1999) 983–988.

[19] M.A. Andrade, C. Sander, Bioinformatics: from genome data to biological knowledge, Curr. Opin. Biotechnol. 8 (1997) 675–683.

[20] C.A. Orengo, A.E. Todd, J.M. Thornton, From protein structure to function, Curr. Opin. Struct. Biol. 9 (1999) 374–382.

[21] L. Holm, C. Sander, Searching protein structure databases has come of age, Proteins 19 (1994) 165–173.

[22] C.A. Orengo, F.M. Pearl, J.E. Bray, A.E. Todd, A.C. Martin, L. Lo Conte, J.M. Thornton, The CATH database provides insights into protein structure/function relationships, Nucleic Acids Res. 27 (1999) 275–279.

[23] F. Pearl, A.E. Todd, J.E. Bray, A.C. Martin, A.A. Salamov, M. Suwa, M.B. Swindells, J.M. Thornton, C.A. Orengo, Using the CATH domain database to assign structures and functions to the genome sequences, Biochem. Soc. Trans. 28 (2000) 269–275.

[24] Y. Wang, K.J. Addess, L. Geer, T. Madej, A. Marchler-Bauer, D. Zimmerman, S.H. Bryant, MMDB: 3D structure data in Entrez, Nucleic Acids Res. 28 (2000) 243–245.

[25] L. Holm, C. Sander, The FSSP database: fold classification based on structure-structure alignment of proteins, Nucleic Acids Res. 24 (1996) 206–209.

[26] L. Holm, C. Sander, Mapping the protein universe, Science 273 (1996) 595–603.

[27] L. Holm, C. Sander, Dali/FSSP classification of three-dimensional protein folds, Nucleic Acids Res. 25 (1997) 231–234.

[28] J.M. Thornton, A.E. Todd, D. Milburn, N. Borkakoti, C.A. Orengo, From structure to function: approaches and limitations, Nat. Struct. Biol. (Suppl.) 7 (2000) 991–994.

[29] A.C. Wallace, N. Borkakoti, J.M. Thornton, TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites, Protein Sci. 6 (1997) 2308–2323.

[30] R.B. Russell, Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution, J. Mol. Biol. 279 (1998) 1211–1227.

[31] S. Schmitt, M. Hendlich, G. Klebe, From structure to function: a new approach to detect functional similarity among proteins independent from sequence and fold homology, Angew. Chem. Int. Ed. 40 (2001) 3141–3144.

[32] D.E. Clark, C.W. Murray, J. Li, Current issues in de novo molecular design, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, Wiley-VCH, New York, 1997, pp. 67–125.

[33] A. Bergner, J. Günther, M. Hendlich, G. Klebe, M. Verdonk, Use of Relibase for retrieving complex 3D interaction patterns including crystallographic packing effects, Biopolymers Nucleic Acids Sciences 61 (2002) in press.

[34] M. Hendlich, Databases for protein–ligand complexes, Acta Crystallogr. D Biol. Crystallogr. 54 (1998) 1178–1182.

[35] M.A. Murcko, Recent advances in ligand design methods, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, Wiley-VCH, New York, 1997, pp. 1–66.

[36] P.J. Goodford, A computational procedure for determining energetically favorable binding sites on biologically important macromolecules, J. Med. Chem. 28 (1985) 849–857.

[37] D.N. Boobbyer, P.J. Goodford, P.M. McWhinnie, R.C. Wade, New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structure, J. Med. Chem. 32 (1989) 1083–1094.

[38] R.C. Wade, P.J. Goodford, Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to form more than two hydrogen bonds, J. Med. Chem. 36 (1993) 148–156.

[39] R.C. Wade, K.J. Clark, P.J. Goodford, Further development of hydrogen bond functions for use in determining energetically favorable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds, J. Med. Chem. 36 (1993) 140–147.

[40] H. Gohlke, G. Klebe, Statistical potentials and scoring functions applied to protein–ligand binding, Curr. Opin. Struct. Biol. 11 (2001) 231–235.

[41] D.S. Goodsell, A.J. Olson, Automated docking of substrates to proteins by simulated annealing, Proteins 8 (1990) 195–202.

[42] G.M. Morris, D.S. Goodsell, R. Huey, A.J. Olson, Distributed automated docking of flexible ligands to proteins: parallel applications of AUTODOCK 2.4, J. Comput. Aided Mol. Des. 10 (1996) 293–304.

[43] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, A.J. Olson, Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function, J. Comput. Chem. 19 (1998) 1639–1662.

[44] R. Abagyan, M. Trotov, D. Kuznetsov, ICM-A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation, J. Comput. Chem. 15 (1994) 488–506.

[45] M. Totrov, R. Abagyan, Flexible protein-ligand docking by global energy optimization in internal coordinates, Proteins (Suppl.) (1997) 215–220.

[46] M. Totrov, R. Abagyan, Protein–ligand docking as an energy optimization problem, in: R.B. Raffa (Ed.), Drug-receptor thermodynamics: introduction and applications, Wiley, Chichester, 2001, pp. 603–624.

[47] E.C. Meng, B.K. Shoichet, I.D. Kuntz, Automated docking with grid-based energy evaluation, J. Comput. Chem. 13 (1992) 505.

[48] D.A. Gschwend, I.D. Kuntz, Orientational sampling and rigid-body minimization in molecular docking revisited: on-the-fly optimization and degeneracy removal, J. Comput. Aided Mol. Des. 10 (1996) 123–132.

[49] J.Y. Trosset, H.A. Scheraga, Reaching the global minimum in docking simulations: a Monte Carlo energy minimization approach using Bezier splines, Proc. Natl. Acad. Sci. USA 95 (1998) 8011–8015.

[50] J.Y. Trosset, H.A. Scheraga, PRODOCK: Software package for protein modeling and docking, J. Comput. Chem. 20 (1999) 412–427.

[51] H. Gohlke, M. Hendlich, G. Klebe, Predicting binding modes, binding affinities and 'hot spots' for protein–ligand complexes

using a knowledge-based scoring function, Persp. Drug Discov. Design 20 (2000) 115–144.

[52] A. Miranker, M. Karplus, Functionality maps of binding sites: a multiple copy simultaneous search method, Proteins 11 (1991) 29–34.

[53] A. Caflisch, Computational combinatorial ligand design: application to human alpha-thrombin, J. Comput. Aided Mol. Des. 10 (1996) 372–396.

[54] A. Caflisch, H.J. Schramm, M. Karplus, Design of dimerization inhibitors of HIV-1 aspartic proteinase: a computer-based combinatorial approach, J. Comput. Aided Mol. Des. 14 (2000) 161–179.

[55] C.M. Stultz, M. Karplus, Dynamic ligand design and combinatorial optimization: designing inhibitors to endothiapepsin, Proteins 40 (2000) 258–289.

[56] C.M. Stultz, M. Karplus, MCSS functionality maps for a flexible protein, Proteins 37 (1999) 512–529.

[57] H.A. Carlson, K.M. Masukawa, K. Rubins, F.D. Bushman, W.L. Jorgensen, R.D. Lins, J.M. Briggs, J.A. McCammon, Developing a dynamic pharmacophore model for HIV-1 integrase, J. Med. Chem. 43 (2000) 2100–2114.

[58] H.A. Carlson, K.M. Masukawa, J.A. McCammon, Method for including the dynamic fluctuations of a protein in computer-aided drug design, J. Phys. Chem. A 103 (2000) 10213–10219.

[59] B. Honig, A. Nicholls, Classical electrostatics in biology and chemistry, Science 268 (1995) 1144–1149.

[60] J.D. Madura, J.M. Briggs, R.C. Wade, M.E. Davis, B.E. Luty, A. Ilin, J. Antonsiewicz, M.K. Gilson, B. Bagheri, L.R. Scott, J.A. McCammon, Electrostatics and diffusion of molecules in solution: simulations with the University of Houston Brownian Dynamics program, Comput. Phys. Commun. 91 (1995) 57–95.

[61] A. Nicholls, B. Honig, A rapid finite difference algorithm, utilizing successive over-relaxation to solve the Poisson–Boltzmann equation, J. Comput. Chem. 12 (1991) 435–445.

[62] A. Nicholls, K.A. Sharp, B. Honig, Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons, Proteins 11 (1991) 281–296.

[63] M. Scarsi, N. Majeux, A. Caflisch, Hydrophobicity at the surface of proteins, Proteins 37 (1999) 565–575.

[64] N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt, A. Caflisch, Exhaustive docking of molecular fragments with electrostatic solvation, Proteins 37 (1999) 88–105.

[65] G. Klebe, The use of composite crystal-field environments in molecular recognition and the de novo design of protein ligands, J. Mol. Biol. 237 (1994) 212–235.

[66] P. Murray-Rust, J.P. Glusker, Directional hydrogen bonding to sp2- and sp3-hybridized oxygen atoms and its relevance to ligand-macromolecule interactions, J. Am. Chem. Soc. 106 (1984) 1018–1025.

[67] F.H. Allen, J.E. Davies, J.J. Galloy, O. Johnson, O. Kennard, C.F. Macrae, E.M. Mitchell, G.F. Mitchell, J.M. Smith, D.G. Watson, The development of version-3 and version-4 of the Cambridge Structural Database system, J. Chem. Inf. Comput. Sci. 31 (1991) 187–204.

[68] H.J. Boehm, LUDI: rule-based automatic design of new substituents for enzyme inhibitor leads, J. Comput. Aided Mol. Des. 6 (1992) 593–606.

[69] H.J. Boehm, The computer program LUDI: a new method for the de novo design of enzyme inhibitors, J. Comput. Aided Mol. Des. 6 (1992) 61–78.

[70] U. Graedler, H.D. Gerber, D.M. Goodenough-Lashua, G.A. Garcia, R. Ficner, K. Reuter, M.T. Stubbs, G. Klebe, A new target for shigellosis: rational design and crystallographic studies of inhibitors of tRNA-guanine transglycosylase, J. Mol. Biol. 306 (2001) 455–467.

[71] H.J. Boehm, M. Boehringer, D. Bur, H. Gmuender, W. Huber, W. Klaus, D. Kostrewa, H. Kuehne, T. Luebbers, N. Meunier-Keller, F. Mueller, 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 (2000) 2664–2674.

[72] H.J. Boehm, D.W. Banner, L. Weber, Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors, J. Comput. Aided Mol. Des. 13 (1999) 51–56.

[73] M. Rarey, B. Kramer, T. Lengauer, G. Klebe, A fast flexible docking method using an incremental construction algorithm, J. Mol. Biol. 261 (1996) 470–489.

[74] B. Kramer, M. Rarey, T. Lengauer, Evaluation of the FLEXX incremental construction algorithm for protein–ligand docking, Proteins 37 (1999) 228–241.

[75] I.J. Bruno, J.C. Cole, J.P. Lommerse, R.S. Rowland, R. Taylor, M.L. Verdonk, ISOSTAR: a library of information about non-bonded interactions, J. Comput. Aided Mol. Des. 11 (1997) 525–537.

[76] M.L. Verdonk, J.C. Cole, R. Taylor, SUPERSTAR: a knowledge-based approach for identifying interaction sites in proteins, J. Mol. Biol. 289 (1999) 1093–1108.

[77] M.L. Verdonk, J.C. Cole, P. Watson, V. Gillet, P. Willett, SUPERSTAR: improved knowledge-based interaction fields for protein binding sites, J. Mol. Biol. 307 (2001) 841–859.

[78] D.R. Boer, J. Kroon, J.C. Cole, B. Smith, M.L. Verdonk, SUPERSTAR: comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein–ligand interactions, J. Mol. Biol. 312 (2001) 275–287.

[79] G. Schneider, H.J. Böhm (Eds.), Virtual Screening for Bioactive Molecules, Wiley-VCH, Weinheim, 2000.

[80] T. Langer, R.D. Hoffmann, Virtual screening: an effective tool for lead structure discovery, Curr. Pharm. Des. 7 (2001) 509–527.

[81] T. Fox, E.E. Haaksma, Computer based screening of compound databases: 1. Preselection of benzamidine-based thrombin inhibitors, J. Comput. Aided Mol. Des. 14 (2000) 411–425.

[82] R. Abagyan, M. Totrov, High-throughput docking for lead generation, Curr. Opin. Chem. Biol. 5 (2001) 375–382.

[83] S. Grueneberg, B. Wendt, G. Klebe, Subnanomolar inhibitors from computer screening: a model study using human carbonic anhydrase II, Angew. Chem. Int. Ed. 40 (2001) 389–393.