INTRODUCTION

The long path from genomic data to a new drug can conceptually be divided into two parts (see left side of Figure 1). The first task is to select a target protein whose molecular function is to be moderated, in many cases blocked, by a drug molecule binding to it. Given the target protein, the second task is to select a suitable drug that binds to the protein tightly, is easy to synthesise, is bio-accessible and has no adverse effects such as toxicity. The knowledge of the three-dimensional structure of a protein can be of significant help in both phases. The steric and physicochemical complementarity of the binding site of the protein and the drug molecule is an important, if not the dominating, feature of strong binding. Thus, in many cases, the knowledge of the protein structure affords well-founded hypotheses of the function of the protein. If the structure of the relevant binding site of the protein is known in detail, we can even start to employ structure-based methods in order to develop a drug binding tightly to the protein.

In this paper bioinformatics methods for prediction aspects of the protein structure are described and their use towards the goal of drug design is discussed. The possibilities and limitations of using protein structure knowledge towards the goal of developing new drugs therapies are also discussed.

NOTIONS OF PROTEIN FUNCTION

The increased accessibility of genomic data and, especially, that of large-scale expression data has opened new possibilities for the search for target proteins. This development has prompted large-scale investments into the new technology by many pharmaceutical companies. The respective screening experiments rely critically on appropriate bioinformatics support for interpreting the generated data. Specifically, methods are required to identify interesting differentially expressed genes and to predict the function and structure of putative target proteins from differential expression data generated in an appropriate screening experiment.

Protein function is a colourful notion whose meaning can range over several levels:

- a very general classification (globular, enzyme, hormone, structural, viral capsid protein, transmembrane protein, etc.);
- biochemical function (biochemical reaction, enzyme specificity, binding partners, cofactors);
- classification via broad cellular function (interaction with DNA and other proteins, cellular localisation);
- broad phenotypic function (changes observed for organisms with deleted or mutated genes);
- identification of detailed physiological function such as the localisation in a metabolic or regulatory pathway and the associated cellular role of the protein;
- identification of molecular binding partners and their mode of interaction with the protein.

The derivation of protein function from protein sequence by theoretical means is commonly performed by transferring functional information from related proteins (e.g., from other organisms). Usually the transfer is from proteins whose function has been established with experimental evidence. The establishment of the relevant protein relationship based on sequence is complicated by some subtleties of evolutionary processes. Though it is often true that organisms share related proteins with similar sequence, similar structure and the same function simply because they originate from a common ancestor and they still fulfill their role within the cellular processes, mutations occur independently after speciation events. Depending on the extent of the evolutionary changes, the recognition of homology or orthology among proteins can be difficult, but still in these cases consistent evidence for relatedness should be expected on the sequence, structure and function levels. Sometimes, the situation is complicated because of gene duplications within a species leading to paralogous copies of the same gene. These paralogous copies are subject to evolutionary changes and the evolutionary pressure on structure or function is much relaxed for all but one copy, which still serves the original purpose, such that greater deviations in sequence, structure and function occur for these copies. As still considerable, i.e., significantly more than random, sequence similarity among paralogous proteins can be observed, this messes up the situation, leading to erroneous transfer of functions to already functionally disabled or functionally completely different proteins.
Therefore, in the following, we have to distinguish between three notions. Similarity is a quantitative measure on the sequence, structure or function level. Homology is used when there is a clear established or potential (assumed, predicted) evolutionary relationship between proteins. The term orthology, in addition, indicates homologous proteins with (established or potential) the same or at least similar function. The notion of paralogy, in contrast, is used, when homologous proteins are expected to have evolved enough to expect changes in function (with or without a change in 3D structure).

For drug design, we need to know more of the function of the protein than follows from just a general classification. It would be best both to know natural binding partners and to have a detailed structural model of the binding sites of the protein.

**METHODS OF PREDICTING PROTEIN FUNCTION**

There are a number of ways to predict protein function from sequence. Most of them are based on sequence similarity. A large database of protein sequences is screened for ‘model sequences’ that exhibit a high level of similarity to the query protein sequence. Sequence alignment tools such as BLAST and PSI-BLAST are the work-horses of such analyses. If one or more model sequences are found that exhibit a sufficiently high level of similarity to the query sequence and about whose function we have some knowledge, then conclusions may be possible on the function of the query sequence. If the homology is above, say, 40 per cent and functionally important motifs are conserved then we can hypothesise that the query sequence has a function that is quite similar to that of the model sequence. As the level of similarity decreases, the conclusions on function that can be drawn from sequence similarity become less and less reliable. Classifications of proteins into families that form clusters of structurally or functionally related proteins are helpful in the prediction of protein function in these cases. There are several protein classifications available on the internet that can serve for this purpose.

**A number of these databases** (Pfam, PROSITE, PRINTS, ProDom, SWISSPROT+TREMBL) are currently being united in the InterPro database. Since protein function is basically tied to protein domains, protein domain analysis is an integral part of the methodology that leads to protein family databases. Since only 20–40 per cent of the protein sequences in a genome such as Mycoplasma genitalium, M. janaschii and M. tuberculosis have significant sequence similarity to proteins of known function, we need to be able to make conclusions on the function of proteins that exhibit no significant sequence similarity to suitable model proteins. As the similarity between query sequence and model sequence decreases below a threshold of, say, 25 per cent, safe conclusions on a common evolutionary origin of the query sequence and the model sequence can no longer be made. However, it turns out that, in many cases, the protein fold can still be reliably predicted, and in several cases even detailed structural models of protein binding sites can be generated. Thus, especially in this similarity range, protein structure prediction – again together with the identification of conserved sequence
or spatial motifs – can help to ascertain aspects of protein function. Other sources of information beside sequence similarity have been explored in order to gain insight into protein function. These methods are represented by five arrows pointing downwards in the top right part of Figure 1. The following comments on these methods apply in the order from left to right:

- Sequence alignment has long been used for ascertaining protein function. This is the standard method and we commented on it above. This approach is only reliable if there is high sequence similarity such that we can argue about orthologous proteins, since we know the function of one of the proteins.

- Recently, the Rosetta stone method has been introduced. This method uses over 20 completely sequenced genomes and analyses evolutionary correlations of two domains being fused into one protein in one species and occurring in separate proteins in another species. From these classifications the method establishes pairwise links between functionally related proteins and elicits putative protein–protein interactions.

- For the same purpose, the phylogenetic profile method analyses the co-occurrence of genes in the genomes of different organisms.

- The analysis of change of phenotype based on mutated genes (eg by knock-out experiments) yields important information on aspects of protein function.

- In the future, the analysis of genetic variations among individuals, eg single nucleotide polymorphisms (SNPs), will be helpful in ascertaining protein function beyond mere disease linkage or association (right arrow in Figure 1).

None of these methods looks at protein structures, and thus we do not discuss them in more detail here. While these methods are reported to generate significant insight into protein function on a higher level and to point to putative target proteins, in the end, drug design can be expected to necessitate structural knowledge of either the target protein or its binding partners.

METHODS FOR PREDICTING PROTEIN STRUCTURE

In the authors' view, computational methods for predicting protein structure from sequence alone are still well out of range, although, there are recent methodical advances – sometimes called mini-threading – that are based on the assembly of fragments (see eg ROSETTA). In contrast, modelling protein structures after folds that have been seen before has become quite a powerful method for protein structure prediction. Here, the query sequence is aligned (threaded) to a model sequence whose three-dimensional structure is known (the template protein). All proteins in a given protein structure database – usually, an appropriate representative set of structures are tried — and each template is ranked using heuristic scoring functions. The score reflects the likelihood that the query sequence assumes the template structure. The approach of modelling a protein structure after a known template is called homology-based modelling and the selection of a suitable template protein is often done via protein threading.

Protein threading has three major objectives: first, to provide orthogonal evidence of possible homology for distant related protein sequences; second, to detect possible homology in cases where sequence methods fail; and third, to improve structural models for the query sequence via structurally more accurate alignments.

There are several successful protein threading methods, including methods based on hidden Markov models.
dynamic programming methods based on profiles;49–51

environment compatibility (ie contact capacity potentials as used in the protein threader 123D).52

These programs are very fast. A mid-size protein sequence can be threaded against a database of about 1,500 protein structures in a few minutes on a PC or workstation. However, the underlying methods assume that the assignment of chemical properties to spatial regions in the protein is the same in the query protein and the template protein. This is not the case, in practice, especially if one compares proteins with partly different folds or different functions. Extensions of the homology-based modelling approach to proteins with very similar protein structures but different chemical make-up require the solution of algorithmically provably hard problems and thus necessitate much more computing time. There are:

- heuristic approaches based on distance-based pair potentials of mean force;53–56
- optimal or approximate combinatorial tree search techniques.57–59

Such approaches need hours to thread a protein through a database of 1,500 templates. However, they can yield more accurate alignments and models of binding pockets of proteins.

The process of protein threading selects a suitable template protein for a protein query sequence and computes an alignment of the backbone of the two proteins that is the starting point for generating a structural model for the query protein based on the structure of the template protein. What is left is to place the side chains of the query protein and to model the loops of the query protein that are not modelled by the template structure. These two tasks are performed by homology-based modelling tools such as:

- Modeller60–64 and ModBase;65
- Swiss-Model;66,67
- or commercial versions included in Quanta (MSI) or Sybyl (Tripos, Inc.).

For protein side-chain modelling there are two contrasting approaches based on knowledge deduced from structural databases and methods such as energy minimisation and molecular dynamics,68 respectively. Methods based on side-chain rotamer libraries that have been created via the analysis of the protein structure database are usually employed to get a first model. Energy minimisation or molecular dynamics69 is often used to refine the model. Such methods have been in use for crystallography/nuclear magnetic resonance (NMR) for many years and are available in several program packages and tools (Charmm,70 GROMOS/GROMACS71,72 and many others73,74). In general these methods are quite computer-intensive and can only be exercised on one or a few proteins.

Generally, the backbone alignment is an input to homology-based modelling tools and the quality of the derived models is highly sensitive to the accuracy of the provided alignments.

Loops are modelled by a related host of methods. Loops that involve more than about five residues are still hard to model.75–78

The evaluation of the accuracy of assigning a protein fold (general protein architecture) to a query sequence is commonly based on generally accepted fold classifications such as SCOP79 or CATH.80 The quality of backbone alignments is much harder to rate, and no generally accepted scheme is available, as of today.80–84 Rating the quality of protein structure models is generally based on the root mean square (rms) deviation of the model and the actual structure on a selected set of residues. The problem here is that the model must be superposed with the actual structure. There are several tools that perform this...
task – DALI/FSSP,85-86 SSAP,87 VAST,88 PROSUP89 or SARF90 – and they can yield different results. Thus, there is no accepted gold standard for protein structure superposition. However, for the purpose of rating the structures of target proteins, the available superposition methods are sufficient.

PERFORMANCE OF PROTEIN STRUCTURE PREDICTION METHODS

There are strong efforts to render the quality of protein structure prediction methods more transparent and easier to evaluate. The centre of these efforts is the bi-annual CASP experiment, which rates protein structure prediction methods on blind predictions and aims at developing standardised and generally agreed upon assessment procedures both for fold identification and the evaluation of alignment accuracy as well as homology models. A blind prediction is a prediction of the three-dimensional structure for a protein sequence at a time, at which the actual structure of the protein is not known (yet). After the structure has been resolved, the prediction is compared with the actual structure. There have been three issues of the CASP experiment,91 the fourth one follows this year. The CASP experiment has been a significant help in providing a more solid basis for assessing the power of different protein structure prediction methods.

For fold recognition, detectable progress has been observed from CASP1 to CASP2. In CASP3, similar performance as in CASP2 was achieved on more difficult targets. There appears to be a certain limit of current fold recognition methods, which is still well below the limit of detectable structural similarity (via structural comparisons). In addition, in CASP3 several groups produced reasonable models of up to 60 residues for ab initio target fragments.

In CASP3 from 43 protein targets, 15 could be classified as comparative homology modelling targets, ie related folds and accompanying alignments could be derived beyond doubt. For more than half of the 21 more difficult cases reasonable models could be predicted by at least one of the participating prediction teams. In addition, the CAFASP subsection of the assessment has demonstrated that 10 out of 19 folds could be solved via completely automatic application of the best threading methods without any manual intervention.

Methods for refining rough structural models towards the true native structure of the query protein are also not straightforward. This is an active area of research.92 A combination of protein threading followed by homology-based modelling cannot create genuinely novel protein structures. But it turns out to be quite sensitive in creating structure models based on known folds. Models that have been reasonably accurate (eg down to 1.4Å for some 60 amino acids of the active site of herpes virus thymidine kinase93) have been reported in blind studies of proteins with a sequence identity to the template protein of as low as 10 per cent. Correct folds can be assigned in many cases, even if the query sequence and the suitable template exhibit a very low level of sequence similarity (down to 5 per cent, ie far below the level of random sequence similarity of 17–18 per cent in optimal alignments).

STRUCTURAL GENOMICS

The goal of structural genomics projects is to solve experimental structures of all major classes of protein folds systematically independent of some functional interest in the proteins.94,95 The aim is to chart the protein structure space efficiently; functional annotations and/or assignment are made afterwards. This affords a thoroughly thought-out strategy of mixing experimental protein structure determination, eg via X-ray, with computer-based protein structure prediction. The experiments have to yield novel protein structures. The proteins to be resolved experimentally are again
selected by computer. The computer part deduces the remaining structures based on homology-based modelling and protein threading. One goal of the overall structural genomics endeavour is to have an experimentally resolved protein structure within a certain structural distance to any possible protein sequence, which allows for computing reliable models for all protein sequences.

Once a map of the protein structure space is available, this knowledge should provide additional insights on what the function of the protein in the cell is and with what other partners it might interact. Such information should add to information gained from high-throughput screening and biological assays. So far, glimpses of what will be possible could be obtained by analysing complete genomes or large sets of proteins from expression experiments with the structural knowledge available today, ie more or less complete representative sets and a quite coarse coverage of structure space.63,96,97

METHODS FOR PREDICTING PROTEIN FUNCTION FROM PROTEIN STRUCTURE

Aspects of protein structure that are useful for drug design studies typically have to involve three-dimensional structure. Predicting the secondary structure of the protein is not sufficient. Even the similarity of the three-dimensional structures of two proteins cannot be taken as an indication for a similar function of these proteins. The reason is that protein structure is conserved much more than protein function. Indeed, protein folds such as the TIM barrel (triose-phosphate isomerase) are quite ubiquitous and can be considered as general scaffolds that lend molecular stability to the protein and are not directly tied to its function. In contrast, the molecular function of the protein is tied to local structural characteristics pertaining to binding pockets on the protein surface. These characteristics are imprinted onto the protein structure by specific patterns of amino acid side chains that make up the binding pocket. The conservation of these amino acids is what makes two proteins have the same function. Since nature varies sequence quite flexibly, this level of conservation is only maintained among orthologous proteins that exhibit a high level of sequence similarity.

Thus, if the template protein from which we predict protein structure is not orthologous to the query protein, other methods of function prediction have to come to bear. It is quite natural to consider conservation patterns in the protein sequence here, such as exhibited in databases containing functional sequence motifs such as PROSITE. An alternative that has been investigated more recently is to analyse conservation in 3D space.98 Experience shows that such 'structural' motifs provide more information than motifs derived purely from sequence, even if the sequence motifs are distributed over several regions (BLOCKS+, PRINTS). An alternative that has been investigated more recently is to analyse conservation in 3D space.98 Experience shows that such 'structural' motifs provide more information than motifs derived purely from sequence, even if the sequence motifs are distributed over several regions (BLOCKS+, PRINTS). An alternative that has been investigated more recently is to analyse conservation in 3D space.98 Experience shows that such 'structural' motifs provide more information than motifs derived purely from sequence, even if the sequence motifs are distributed over several regions (BLOCKS+, PRINTS).
perform a wide range of functions.101 This limits our potential of deducing function from structure. But knowledge on which folds support a given function and which functions are based on a given fold can still help in predicting function from structure. In addition, local structural templates such as FFFs indicative for a particular function can identify similar sites and the associated function despite a globally different fold. Such 3D patterns can also discriminate among globally similar folds with respect to containing particular conserved 3D functional motifs in order to classify them into different functional categories. Though it is not easy to derive functions from resolved protein structures, the availability of structural information improves the chances compared with relying on sequence methods alone.

METHODS FOR DEVELOPING DRUGS BASED ON PROTEIN STRUCTURE

The object of drug design is to find or develop a, mostly small, drug molecule that tightly binds to the target protein, moderating (often blocking) its function or competing with natural substrates of the protein. Such a drug can be best found on the basis of knowledge of the protein structure. If the spatial shape of the site of the protein is known, to which the drug is supposed to bind, then docking methods can be applied to select suitable lead compounds that have the potential of being refined to drugs. The speed of a docking method determines whether the method can be employed for screening compound databases in the search for drug leads. A docking method that takes a minute per instance can be used to screen up to thousands of compounds on a PC or hundreds of thousands of drugs on a suitable parallel computer. Docking methods that take the better part of an hour cannot be suitably employed for such large-scale screening purposes. In order to screen really large drug databases with several hundred thousand compounds docking methods that can handle single protein/drug pairs within seconds are needed.

The high conformational flexibility of small molecules as well as the subtle structural changes in the protein binding pocket upon docking (induced fit) are major complications in docking. Furthermore, docking necessitates careful analysis of the binding energy. The energy model is cast into the so-called scoring function that rates the protein–ligand complex energetically. Challenges in the energy model include the handling of entropic contributions, and solvation effects, and the computation of long-range forces in fast docking methods. The state of the art in docking can be summarised as follows (see also Table 1).

Today, tools that are able to dock a molecule to a protein within seconds are still based on rigid-body docking (both the protein and ligand conformational flexibility is omitted).

Recently, fast docking tools have been adapted to screening combinatorial drug...
libraries (see, eg, Rarey and Lengauer\textsuperscript{103}). Such libraries provide a carefully selected set of molecular building blocks together with a small set of chemical reactions that link the modules. In this way, a combinatorial library can theoretically provide a diversity of up to billions of molecules from a small set of reactants. The accuracy of docking predictions lies within 50–80 per cent ‘correct’ predictions depending on the evaluation measure and the method. That means that docking methods are far from perfectly accurate. Nevertheless, they are very useful in pharmaceutical practice. The major benefit of docking is that a large drug library can be ranked with respect to the potential that its molecules have for being a useful lead compound for the target protein in question. The quality of a method in this context can be measured by an enrichment factor. Roughly, this is the ratio between the number of active compounds (drugs that bind tightly to the protein) in a top fraction (say the top 1 per cent) of the ranked drug database divided by the same figure in the randomly arranged drug database. State-of-the-art docking methods in the middle regime (minutes per molecular pair), eg FlexX\textsuperscript{104} achieve enrichment factors of up to about 15. Fast methods (seconds per pair), eg FeatureTrees\textsuperscript{105} achieve similar enrichment factors, but deliver molecules similar to known binding ligands and do not detect as diverse a range of binding molecules.

Even if the structure of the protein binding site is not known, computer-based methods can be used to select promising lead compounds. Such methods compare the structure of a molecule with that of a ligand that is known to bind to the protein, for instance, its natural substrate. Alternatives to docking for lead finding include high-throughput screening (HTS). This laboratory method allows for testing the binding affinity of up to more than several thousand compounds to the same target protein in a day. In comparison this method has the advantage that it does not have to deal with insufficiently powerful computer models, at the expense of high laboratory cost and the absence of structural knowledge on ‘why’ a compound binds to the protein.

**CONCLUSION**

In summary, the field is still in an early stage of development. \textit{Ab initio} protein structure prediction continues to be a grand challenge for which no comprehensive solution is in sight. The quality of fold prediction based on homology rises and tools has reached the stage where one can generate confident predictions for soluble proteins that in a substantial fraction (about half) of the cases provide significant threading hits in the structure database. Protein threading and homology-based prediction become especially helpful in an environment where the methods can be used in concert with experimental techniques for structure and function determination. Here, the prediction methods can exercise their strengths, which lie in being used interactively by experts and making suggestions that can be followed up by succeeding experimentation, rather than being required to provide proven fact. The process of going from structure to function is far from being automated. In a scenario that combines structure prediction methods with experimentation, the step from structure to function can be performed in a customised manner.

Protein structure prediction by homology is definitely not yet a turn-key technology. But we can expect it to enter the ‘production’ stage through the activities in structural genomics. Still the field of protein structure prediction is very busy, generating the tools and processes for raising the number of confident structure predictions and the accompanying estimates of significance. Problems for applying these results in drug design are not only that the models may not be sufficiently accurate but also that the structures of many interesting...
target proteins will not be accessible by homology-based modelling, at all, for some time to come. This includes the therapeutically particularly interesting class of membrane proteins, for which essentially no structures have been resolved.

Docking is used frequently in structure-based drug design. To the authors’ knowledge, the first drug developed with structure-based techniques was the HIV protease inhibitor Dorzolamide. In the past few years structural considerations have begun to pervade the design of new drugs. A point in case is that of the neuraminidase inhibitors for HIV. Such studies mostly involve experimentally resolved protein structures. However, even models can serve to guide drug development. Based on the experimentally resolved structure of the membrane protein bacteriorhodopsin, several groups are attempting to model binding sites of G-protein coupled receptors that are believed to be structurally similar. Nevertheless, the authors are not aware of any instance where the whole process line from the protein sequence to the lead structure has been exercised in an integrated manner and with significant help of computer predictions. The field has not reached this level of maturity yet. While structural aspects – even as predicted by the computer – can be expected to invade the search for target proteins and the development of new drugs, experimental data, where they are accessible, will always be highly welcome and often indispensable in this process.

Acknowledgements

We thank Matthias Rarey for helpful comments on this paper and Gerhard Banuczek and Gerhard Klebe for information on the state of drugs developed by structure-based techniques.

References


13. http://www.ebi.ac.uk/intrepr/


Protein structure prediction methods for drug design


