Computational Methods Facilitate the Assignment of Protein Functions**

Gerd Folkers* and Christian D. P. Klein

A tremendously difficult task will be tackled by biochemists in the next few years: the ever-increasing amount of genomic data has to be put into a functional context. The sequencing of genomes is primarily a logistic challenge for the participating institutions. The throughput of such enterprises can be increased without limits by increasing the available resources. For the time being, the situation is different for the most important ensuing step: The elucidation of the function of the proteins that are encoded by newly identified genes. This area of research has been termed "Functional Genomics".

The DNA sequence immediately leads to the sequence of a protein. Protein sequences have been archived for many years, and thousands of sequences (for proteins whose function is often known) are publicly available.[1] Proteins with a similar sequence often fulfil identical functions; therefore, it is clear that sequence comparisons can lead to the assignment of protein functions. This approach has been pursued for some time,[2] and by continuous improvements of the basic algorithum it has become a valuable tool for geneticists and biochemists. One of the advantages of this method is that it is very fast, and it does not require any three-dimensional structural information. However, it has one fundamental problem: Proteins have evolved in a convergent manner, directed towards their present-day function. This means that proteins which are not evolutionary connected (that is, they do not have a common ancestor), and whose amino acid sequences are not similar, can have a similar or identical function.[3] For example, the enzymes trypsin and subtilisin have a very different amino acid sequence. It would not be possible to recognize their functional similarity by inspecting their primary structure since only about 10% of their sequences are identical. This is not much more than one could expect for a random pair of proteins. However, a certain section of the two enzymes is very similar: it is the region that forms the catalytic center. The amino acids that participate in the catalytic mechanism are neighbors in three-dimensional space, but not on the amino acid chain. It would not be possible to identify these two enzymes as functionally related serine proteases by a sequence-matching technique.

The conservation of the functionally important region is one of the paradigms of protein chemistry. Independent of the primary structure, the physicochemical properties of this region are similar, and in many cases even the participating amino acids are identical.

From Structure to Function

S. Schmitt, M. Hendlich, and G. Klebe have presented a new method for the identification of functional similarity that is based on the conservation of the active site of enzymes and the ligand-binding pocket of receptors.[4] The first step in their approach is to abstract the binding sites of proteins whose function and three-dimensional structure is known. A binding pocket is no longer represented by amino acids, but by the spatial arrangement of certain pseudofunctionalities such as hydrogen-bond donor or acceptor (Figure 1). This data

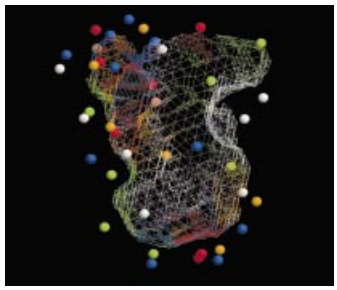

Figure 1. Surfaces and pseudo centers of a binding pocket (from ref. [4]).

reduction allows the construction of a database which is comparably easy to handle but still contains all the relevant information. The (putative) binding pockets of proteins with unknown function are treated analogously and can then be compared to the database. If a similar spatial arrangement of pseudofunctionalities in the binding pocket is found, then the function of the unknown protein can be deduced.

The concept of this method is similar to techniques that are used to identify small-molecule binding partners of proteins such as ligands or inhibitors. Programs that are frequently used in the drug-development process, such as GRID,[5] RELIBASE,[6] and others, may be regarded as "ancestors" of this new approach.

[*] Prof. Dr. G. Folkers, Dr. C. D. P. Klein
Pharmazeutische Chemie
Departement für Angewandte Biowissenschaften
ETH Zürich
Winterthurerstrasse 190, 8057 Zürich (Switzerland)
Fax: (+41) 1-635-6884
E-mail: folkers@pharma.ethz.ch

[**] We would like to thank Dr. H. Rüther, Cologne, Prof. H.-D. Höltje, Universität Düsseldorf, and Prof. Dr. P. A. Schubiger, Paul-Scherrer-Institut, Villingen (Switzerland), for critical comments on the manuscript.

  1433-7851/01/4022-4175 $ 17.50+.50/0

This is not the first attempt to use spatial features of binding pockets for the functional analysis of proteins.[7] The main advantage of the new method is the fact that it does not solely use geometric information (the shape of the pocket), but it also uses the arrangement of physicochemical properties.

It may be surprising for someone experienced in molecular modeling to learn that the method of Schmitt et al. uses only five types of pseudofunctionalities and still gives good results. Programs that are used to determine the intermolecular interaction properties of small molecules use much more sophisticated representations. Electrostatic properties, for example, are ignored (or treated indirectly) by the method of the Klebe research group—which has the advantage that there is no need to calculate such properties before a prediction is made.

The most important prerequisite for the method of Schmitt, Hendlich, and Klebe is that the three-dimensional structure of the protein must be known. It is not sufficient to know the gene or amino acid sequence of the protein. Three-dimensional structures of proteins are usually determined by X-ray crystallography, with NMR spectroscopy gaining importance during the last few years.[8] Both methods, however, are not traditionally known to be fast or easy. The huge commercial potential associated with this field means that attempts are now being made to establish high-throughput methods for the determination of protein structures.[9] Furthermore, theoretical methods are gaining importance,[10] and it is often possible to make predictions that are sufficiently accurate for all practical purposes. Some approaches are based on comparisons of primary structure, whereas others do not need any additional, empirical information.[11]

At the moment, the genomic data that is available for higher organisms appears to contain a considerable amount of errors.[12] If a faulty protein sequence is used to predict the structure and function of a protein, then the abstraction that is made by Schmitt et al. might be advantageous, since the influence of a "false" amino acid on the representation of the binding pocket and on the outcome of the prediction is reduced.

Genes, Proteins, Drugs

The main motivation for the sequencing of genomes is the hope of finding new therapeutic strategies. Drugs are seldomly targeted at the genetic material; intervention is usually made at the protein level. It does not make any sense to select a protein with unknown function as the target for drug development and therapeutic intervention. Figure 2 shows the relevance of genomic data for the development of new drugs: The DNA sequence immediately gives the protein sequence. The three-dimensional structure of the protein can be derived by theoretical methods or is determined experimentally. The next step, the assignment of a function, is critical for the overall success and quite difficult to perform. Genetic engineering techniques are frequently used to generate organisms that lack the gene in question (for example, so-called "knock-out-mice") so that the effect of this mutation on the organism can be studied. It is often difficult (or impossible in the case of humans) to generate such organisms,

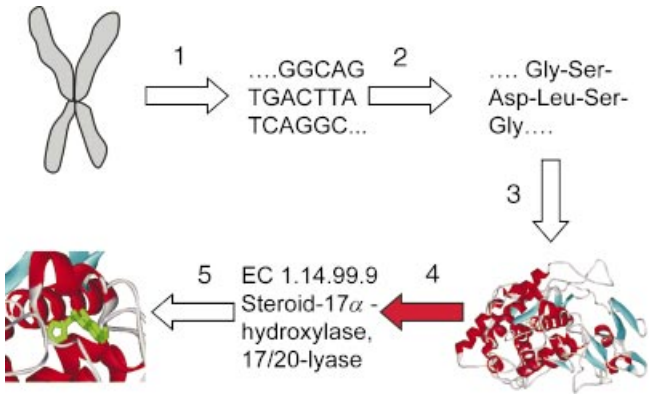

Figure 2. 1: Genome sequencing; 2: translation to primary structure; 3: deduction or determination of 3D structure; 4: functional classification; 5: design of ligands or inhibitors.

and the interpretation of the results can be problematic. Attempts are underway to automatize the experimental procedures for the generation of transgenic animals.[13] Nevertheless, valid theoretical methods can be a very important resource in functional genomics.

The functional analysis is the decisive step when it comes to establish a protein as a target for therapeutic intervention. Once the validity of a target is confirmed, inhibitors or antagonists can be identified or designed using, for example, structure-based design methods.

The database that has been generated by the research group of Klebe could be used for another purpose which is not immediately evident—it might enable us to predict unwanted side-effects. The ideal drug molecule binds exclusively and with high affinity to its target protein. Undesired effects are usually caused by it binding to other proteins. Until now, the only way of assessing such unwanted binding processes is to perform biological tests, which tend to be costly and difficult. Therefore, side-effects are often identified at a later development stage, when considerable investments have been made. The method of the Klebe research group will make it possible to perform a "virtual" search for undesired interaction partners of new drug molecules by searching the database for enzymes or receptors that are prone to bind the candidate molecule. This strategy would be the inversion of a frequently used technique for the search for new ligands, the so-called "virtual screening" (Figure 3 A). One would not look for small

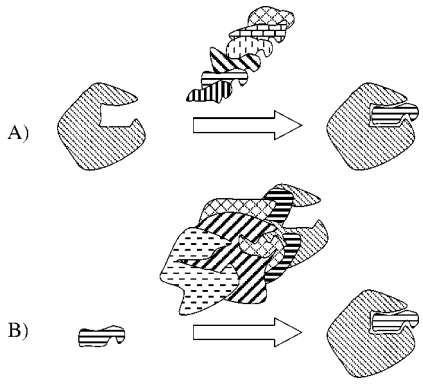

Figure 3. Virtual search for a ligand (A) and a protein (B).

  1433-7851/01/4022-4176 $ 17.50+.50/0

ligands, but for their macromolecular binding partners (Figure 3B). This type of "inverse" virtual screening is not limited to the search for side-effects: For example, one could search for the target protein of biologically active molecules with unknown mechanisms of action.

[1] SWISS-PROT: A. Bairoch, B. Boeckmann, *Nucleic Acids Res.* **1994**, *22*, 3578–3580; http://www.expasy.ch/sprot/, **2001**; BLAST: S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipton, *J. Mol. Biol.* **1990**, *215*, 403–410; http://www.ncbi.nlm.nih.gov/BLAST/, **2001**.

[2] S. B. Needleman, C. D. Wunsch, *J. Mol. Biol.* **1970**, *48*, 443–453; W. R. Pearson, D. J. Lipman, *Methods Enzymol.* **1990**, *183*, 63–98.

[3] A. Fersht, *Structure and Mechanism in Protein Science*, 2nd ed., W. H. Freeman, New York, **1999**.

[4] S. Schmitt, M. Hendlich, G. Klebe, *Angew. Chem.* **2001**, *113*, 3237–3241; *Angew. Chem. Int. Ed.* **2001**, *40*, 3141–3144.

[5] P. J. Goodford, *J. Med. Chem.* **1985**, *28*, 849–857.

[6] M. Hendlich, *Acta Crystallogr. Sect. D* **1998**, *54*, 1178–1182; http://relibase.ebi.ac.uk/, **2001**.

[7] D. Fischer, R. Norel, H. Wolfson, R. Nussinov, *Proteins Struct. Funct. Genet.* **1993**, *16*, 278–292; S. L. Moodie, J. B. Mitchell, J. M. Thornton, *J. Mol. Biol.* **1996**, *263*, 486–500; N. Kobayashi, N. Go, *Eur. Biophys. J.* **1997**, *26*, 135–144; M. Rosen, S. L. Liang, H. Wolfson, R. Nussinov, *J. Mol. Biol.* **1998**, *11*, 263–277; M. Stahl, C. Taroni, G. Schneider, *Protein Eng.* **2000**, *13*, 83–88.

[8] A. Saegusa, *Nature* **1998**, *392*, 219.

[9] Protein Structure Factory (Berlin), http://userpage.chemie.fu-berlin.de/~psf/, **2001**.

[10] P. Bork, D. Eisenberg, *Curr. Opin. Struct. Biol.* **1998**, *8*, 331–332; J. Skolnick, J. S. Fetrow, *Trends Biotechnol.* **2000**, *18*, 34–39; D. Frishman, H. W. Mewes, *Prog. Biophys. Mol. Biol.* **1999**, *72*, 1–17.

[11] D. J. Osguthorpe, *Curr. Opin. Struct. Biol.* **2000**, *10*, 146–152.

[12] See, for example, S. Karlin, A. Bergman, A. J. Gentles, *Nature* **2001**, *411*, 259–260.

[13] B. P. Zambrowicz, G. A. Friedrich, E. C. Buxton, S. L. Lilleberg, C. Person, A. T. Sands, *Nature* **1998**, *392*, 608–611.

[14] H. J. Böhm, G. Schneider, *Virtual Screening for Bioactive Molecules*, Wiley-VCH, Weinheim, **2000**.

  1433-7851/01/4022-4177 $ 17.50+.50/0