feature

Contribution of structural biology to clinically validated target proteins

Masumi Mori[1], Naoko Ogawa[1], Kunihiro Tanikawa, Sanae Dodo, Sotaro Shibayama, Shigeyuki Yokoyama and Akiko Tanaka

We identified six groups of diseases expected to cause serious future health issues on the basis of a WHO report. Approved drugs for these diseases were associated with 409 target proteins; however, the percentage of selected proteins with full-length structural information deposited in the Protein Data Bank (PDB) was only 9.8%. The reason for the low percentage may be as a result of a disproportionate number of intractable proteins with multiple transmembrane regions and variable, or undefined glycosylation patterns, which impede protein preparation and crystallization, in such druggable proteins. We stress the importance of structural analysis of proteins, especially approved-drug target proteins, and the development of new methods to enable structural analyses of presently intractable proteins. In addition, we present an overview of large structural biology projects.

According to DiMasi *et al*. [1], the average cost of discovering and developing a new drug is US$ 802 million and is rising at an annual rate of 7.4% above inflation. In the 1950s, pharmaceutical companies invested 5% of sales revenue into research and development. By 1980 this had risen to 9% and by 2002, the average was 16%, with some firms spending well over 20% [2]. These spiraling costs threaten the sustainable ability of pharmaceutical companies to create innovative drugs. A recent study, based on data from the 10 biggest drug companies, suggested that the average success rate of drug candidates from a Phase I clinical study to registration is approximately 11% [3]. DiMasi reported that over one third of all terminations in drug-development projects were due to problems with efficacy (37.6%) for the period of 1987–1992 [4]. In 2000, Kola and Landis also indicated that a major cause (approximately 30%) of attrition was lack of efficacy [3]. Efficacy must be the key driving force to reduce attrition rates and consequently improve the performance of drug discovery.

During the past two decades, the pharmaceutical industry has adopted cutting-edge technologies. Among these technologies, structure-based drug design (SBDD), based on the three-dimensional structural information of target protein–ligand molecular complexes, has become a very attractive method to create high quality drug candidates [5–7]. Captopril, an angiotensin-converting enzyme (ACE) inhibitor for treating cardiovascular diseases was the first drug developed using a structure–activity relationship that was based on structural information of its inhibitor compounds and its related protein, bovine carboxypeptidase A [6,8]. After the 1980s, the determination of three-dimensional protein structures accelerated as a result of technological developments on protein sample preparation steps, high-throughput crystallization, and structure determination processes [5]. Important drugs developed through rational design, based on the target protein–drug molecule complex structures, include nelfinavir and amprenavir as human immunodeficiency virus (HIV) drugs [9,10]; zanamivir and oseltamivir, as influenza drugs [11]; the kinase inhibitors imatinib [12] and dasatinib [13], for treating chronic myelogenous leukemia, and alogliptin and others whose target protein is dipeptidyl peptidase IV [14,15].

Corresponding author: Tanaka, A. (aktanaka@riken.jp)

[1] These authors contributed equally to this work.

1359-6446/06/$ - see front matter © 2008 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2008.03.008

The combination of structural biology and SBDD will bring new success to future drug discovery.

Diseases requiring new drug development

First, we determined the priority of diseases that require new drugs. Our analysis was intended to identify those diseases with greatest unmet medical need, based upon the evaluation of disability adjusted life years (DALYs) reported in 2006 (http://www.who.int/healthinfo/statistics/bodprojections2030/en/index.html). WHO Global Burden of Disease (GBD) studies provided statistics on DALYs for more than 135 diseases and injuries, categorized by age, sex, and region [16,17]. DALYs are the sum of the future years of lifetime impacted by premature mortality or any mental or physical disability caused by a disease or injury, and thus DALYs have been used as a standard measure for estimating the societal burden of diseases and as an outcome indicator in cost-efficiency analysis [18].

Global DALYs in 2030 of communicable diseases, non-communicable diseases, and injuries are expected to be 462 953 000, 870 289 000 and 209 575 000, respectively. The data indicate the importance of the future lifetime loss by non-communicable diseases. Analysis of DALYs by income group indicated that communicable diseases have a high score (449 454 000) only in lower income countries. This indicates the crucial issue that economic factors result in communicable diseases being disproportionately serious in lower income countries. Therefore, we have focused on non-communicable diseases for the purpose of analysis in this paper.

The top five diseases with highest DALYs in 2030 are: 'psychiatric conditions' (229 403 000); 'cardiovascular diseases' (176 999 000); 'sense organ diseases' (106 120 000); 'malignant neoplasms' (105 001 000), and 'respiratory diseases' (81 094 000). Among the 'psychiatric conditions', the diseases with the highest DALYs reported are unipolar depressive disorders, and the diseases with the highest rates of increase of DALYs (between 2005 and 2030) are Alzheimer's disease and other dementias. Of the 'cardiovascular diseases', the diseases with high DALYs are ischemic heart disease and cerebrovascular disease. Among the 'sense organ diseases', the diseases with highest DALYs are cataracts and hearing loss. As an addition to the top five causes of DALYs, 'Diabetes mellitus' is noteworthy. DALYs for this disease in 2030 are expected to show a 1.63-fold increase compared with those in 2005, which is the highest increase, so we included this disease as the sixth selected disease.

Target proteins of the approved drugs

To highlight the contribution of structural biology to clinically validated target proteins, we created a dataset of the protein names and UniProt accession numbers of the target proteins for approved drugs for treating the six selected diseases (named DT dataset). As a result, a total of 409 proteins with Swiss–Prot accession numbers were selected as important target proteins from the 617 target proteins for approved drugs in the United States and Japan. To compare the results, we collected the protein names and UniProt accession numbers of all human proteins in the Swiss–Prot database (named SP dataset). Physical features of the each protein contained in the datasets were analyzed computationally and summarized in Table 1. The total percentage of proteins with full-length structural data deposited in the Protein Data Bank (PDB) as of Feb/22/2008 is only 9.8% (40/409) in the DT dataset, which was slightly higher than the 6.8% (983/14 448) in the SP dataset. Using SBDD, structural information about drug target proteins could bring about the development of more effective and improved drug molecules. Actually, the companies that have developed the approved drugs may already have structural information about the target protein–drug molecule complexes, but such information would be difficult to share with global scientists. Many scientists would welcome structural analyses of these target proteins.

To determine why many drug target proteins lack full-length structural information, we analyzed the physical characteristics of the human proteins (Table 1). In the SP dataset, the percentages of proteins with full-length structural data deposited in the PDB, in terms of the numbers of transmembrane regions (TM), were 8.8% (947/10 750) (TM = 0), 1.6% (27/1658) (TM = 1), and 0.4% (9/2040) (TM > 1). The percentage of structurally determined proteins with carbohydrate substitution was 2.8% (99/3540), and for those lacking carbohydrate substitution was 8.1% (884/10 908). These percentages indicate that the existence of multiple transmembrane regions and/or glycosylation in proteins hinders structural analysis, which is a well-known problem. In the DT dataset, the percentages of proteins with multiple transmembrane regions (TM > 1) and glycosylation were as high as 41.3% and 53.3%, respectively, while in the SP dataset, they were 14.1% and 24.5%, respectively. This means that half of the target proteins of the approved drugs for diseases with higher DALYs are proteins with multiple transmembrane regions (TM > 1) and/or glycosylation, which, as a result, hinder their full structural identification.

The DT dataset classified by the selected diseases was precisely analyzed (Table 2). The

TABLE 1

Physical characteristics of the relevant drug target proteins

	Numbers of TM			Glycosylations		Total
	0	*1*	>1	−	+	
DT dataset (409)	192 (46.9%) 34[a]	48 (11.7%) 5[a]	169 (41.3%) 1[a]	191 (46.7%) 32[a]	218 (53.3%) 8[a]	409 40[a]
SP dataset (14 448)	10,750 (74.4%) 947[a]	1,658 (11.5%) 27[a]	2,040 (14.1%) 9[a]	10,908 (75.5%) 884[a]	3,540 (24.5%) 99[a]	14,448 983[a]

To identify the UniProt accession numbers of target proteins of approved drugs for treating the selected diseases with higher DALYs, we utilized KeyMolnet® (Institute of Medicinal Molecular Design, Inc., Japan, Version 3.7, UniProt Version 8.5, Aug/22/2006 release). The chemical properties of the proteins were obtained from the UniProt database (http://au.expasy.org/sprot/), using their accession numbers. The features of the proteins were obtained from the Feature Table column in UniProt.

Information about the three-dimensional structures of the proteins was cited from the PDB (http://www.rcsb.org/pdb/). We used 'pdb_seqres. txt' and 'pre-released. seq' on the ftp site (Feb/22/2008) of the database to analyze the proteins. It contained 34 501 non-redundant structures. The three-dimensional structural data in the PDB of the query proteins were searched by amino acid sequences, using the Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/blast/blast_references.html), which finds regions of local similarity between sequences. A query protein was considered as a full-length, structurally determined protein, when the sequence hit more than one sequence in the PDB, with a minimum of 95% sequence identity and a minimum matching region of 90% with the query sequence.

[a] Numbers of proteins with full-length structures deposited in the PDB. TM: transmembrane regions. DT dataset: the set of approved-drug target proteins for treating the diseases with higher DALYs. SP dataset: the set of all human proteins in the Swiss–Prot database.

TABLE 2

Physical characteristics of the drug target proteins for the six diseases

	Numbers of TM			Glycosylations		Total
	0	1	>1	−	+	
Psychiatric conditions	1 (1.3%) 1[a]	3 (4.0%) 2[a]	71 (94.7%) 0[a]	5 (6.7%) 2[a]	70 (93.3%) 1[a]	75 3[a]
Cardiovascular diseases	72 (35.8%) 7[a]	19 (9.5%) 1[a]	110 (54.7%) 1[a]	77 (38.3%) 7[a]	124 (61.7%) 2[a]	201 9[a]
Sense organ diseases	80 (72.7%) 11[a]	5 (4.5%) 1[a]	25 (22.7%) 0[a]	71 (64.5%) 10[a]	39 (35.5%) 2[a]	110 12[a]
Malignant neoplasms	97 (71.9%) 20[a]	22 (16.3%) 1[a]	16 (11.9%) 0[a]	89 (65.9%) 18[a]	46 (34.1%) 3[a]	135 21[a]
Respiratory diseases	29 (47.5%) 1[a]	0 (0.0%) 0[a]	32 (52.5%) 0[a]	27 (44.3%) 1[a]	34 (55.7%) 0[a]	61 1[a]
Diabetes mellitus	5 (16.1%) 4[a]	7 (22.6%) 0[a]	19 (61.3%) 0[a]	3 (9.7%) 2[a]	28 (90.3%) 2[a]	31 4[a]

[a] Numbers of proteins with full-length structures deposited in the PDB. TM: transmembrane regions.

diseases with higher numbers of target proteins for approved drugs were cardiovascular diseases, malignant neoplasms, and sense organ diseases. The diseases with higher numbers of more intractable targets were psychiatric conditions and diabetes mellitus. Many of these targets may be intrinsic membrane proteins on the cell surface, such as receptor proteins. Indeed, all of the known drug target proteins for treating psychiatric conditions are membrane proteins (TM = 1 or TM > 1), with the exception of acetylcholinesterase. The development of new technologies is indispensable for structural biology to contribute to the medical management of these diseases.

Structural biology efforts to analyze clinically validated target proteins

Our results demonstrate that innovative technology for the structural analysis of intractable proteins is required to accelerate the analysis of societally relevant drug target proteins. In academia, a number of consortia have been devoted to three-dimensional protein structure analyses. Many groups began their research with low-hanging fruit, that is, proteins with structures that can be analyzed in a high-throughput manner, such as small, stable proteins, as mentioned by Stevens *et al.* [19]. As a result of these initiatives, various technologies for high-throughput structure determinations have been developed [5]. The national initiatives are now in a second phase, with the Protein Structure Initiative (PSI-2) in the United States and the Target Proteins Research Program in Japan. Each initiative is now making new efforts to enable the structural analysis of proteins with more challenging and intractable features, such as intrinsic membrane proteins, proteins with sugar chains, high molecular weight, and multi-subunit proteins. Thus, structural biologists will have a strong probability of success in the development of new methods to accelerate the structural analysis of currently intractable proteins.

Structural biologists should also take immediate action toward analyzing and publishing the structures of clinically validated target proteins. Considering the conformational changes by induced-fit ligand binding, the structural analyses of the protein–ligand (drug molecule) complexes are crucial to protein science and the discovery of improved drug candidates through SBDD. When, however, structural biologists successfully analyze the structures of protein–drug molecule complexes, predict modified drug molecules with higher affinities to the protein pocket, and publish the results to share globally, there may be issues with respect to the infringement of knowledge management of pharmaceutical companies that have developed and filed the drug molecules. One strategy to avoid infringement is to obtain candidate compounds that are in the public domain and to evaluate their binding ability to the known drug pockets of target proteins, through compound screening using chemical libraries. Then their complex structures can be freely analyzed, enabling the prediction of modified ligand molecules and the search for effective modified ligand molecules. We can also publish the results without concern for the infringement issue. For protein science, the analyses of ligand-binding structures of a protein will give us clues to understand its substrate specificity and its enzymatic reaction mechanisms. Further, when our efforts of searching for effective modified ligand molecules result in finding active inhibitors with new molecular structures, the findings can be transferred to industry to develop new candidate compounds with higher affinities to the target.

The Molecular Libraries Screening Centers Network in the United States provides such ligand information to academic researchers. The Target Proteins Research Program in Japan is funding large chemical libraries to provide candidate ligand molecules to the program members. On the basis of these new national project frameworks, structural biologists in academia will be able to work with companies by sharing structural information about clinically validated target proteins, in order to contribute to future medical progress.

Acknowledgements

We are grateful to Dr. T. Sasazuki, International Medical Center of Japan, and Dr. T. Nagano and Dr. H. Kimura, The University of Tokyo, for their valuable advice and encouragement. This work was supported partly by the National Project on Protein Structural and Functional Analyses, the Development of Fundamental Technology for Protein Analyses and the Targeted Proteins Research Program, from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References

1 DiMasi, J.A. *et al.* (2003) The price of innovation: new estimates of drug development costs. *J. Health Econ.* 22, 151–185

2 Booth, B. and Zemmel, R. (2004) Prospects for productivity. *Nat. Rev. Drug Discov.* 3, 451–457

3 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715

4 DiMasi, J.A. (2001) Risks in new drug development: approval success rates for investigational drugs. *Clin. Pharmacol. Ther.* 69, 297–307

5 Kuhn, P. *et al.* (2002) The genesis of high-throughput structure-based drug discovery using protein crystallography. *Curr. Opin. Chem. Biol.* 6, 704–710

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

7 Bajorath, F. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1, 882–894

8 Hassall, C.H. *et al.* (1982) The design of a new group of angiotensin-converting enzyme inhibitors. *FEBS Lett.* 147, 175–179

9 Kaldor, S.W. *et al.* (1997) Viracept (nelfinavir mesylate, AG1343): A potent, orally bioavailable inhibitor of HIV-1 protease. *J. Med. Chem.* 40, 3979–3985

10 Kim, E.E. *et al.* (1995) Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J. Am. Chem. Soc.* 117, 1181–1182

11 Varghese, J.N. (1999) Development of neuraminidase inhibitors as anti-influenza virus drugs. *Drug Dev. Res.* 46, 176–196

12 Schindler, T. *et al.* (2000) Structural mechanism for STI-571 inhibition of Abelson tyrosine kinase. *Science* 289, 1938–1942

13 Tokarski, J. *et al.* (2006) The structure of Dasatinib (BMS-354825) bound to activated ABL kinase domain elucidates its inhibitory activity against imatinib-resistant ABL mutants. *Cancer Res.* 66, 5790–5797

14 Feng, J. *et al.* (2007) Discovery of alogliptin: a potent, selective, bioavailable, and efficacious inhibitor of dipeptidyl peptidase IV. *J. Med. Chem.* 50, 2297–2300

15 Lubbers, T. *et al.* (2007) 1,3-Disubstituted 4-aminopiperidines as useful tools in the optimization of the 2-aminobenzo[a]quinolizine dipeptidyl peptidase IV inhibitors. *Bioorg. Med. Chem. Lett.* 17, 2966–2970

16 Murray, C.J.L. and Lopez, A.D. (1996) Evidence-based health policy—lessons from the global burden of disease study. *Science* 274, 740–743

17 Mathers, C.D. and Loncar, D. (2006) Projections of global mortality and burden of disease from 2002 to 2030. *Plos. Med.* 3, 2011–2030

18 Rushby, J.F. and Hanson, K. (2001) Calculating and presenting disability adjusted life years (DALYs) in cost-effectiveness analysis. *Health Policy Plan.* 16, 326–331

19 Stevens, R.C. *et al.* (2001) Global efforts in structural genomics. *Science* 294, 89–92

Masumi Mori
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

Naoko Ogawa
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

Kunihiro Tanikawa
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Sanae Dodo
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

Sotaro Shibayama
Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Shigeyuki Yokoyama
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Akiko Tanaka
RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
Graduate School of Pharmaceutical Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan