ELSEVIER

Progress and problems in the exploration of therapeutic targets

Chanjuan Zheng, Lianyi Han, Chun W. Yap, Bin Xie and Yuzong Chen

Bioinformatics and Drug Design Group, Department of Pharmacy and Department of Computational Science, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543

Drugs exert their therapeutic effect by binding and regulating the activity of a particular protein or nucleic acid target. A large number of targets have been explored for drug discovery. Continuous effort has been directed at the search for new targets and more-extensive exploration of existing targets. Knowledge of these targets facilitates the understanding of molecular mechanisms of drugs and the effort required for drug discovery and target searches. Areas of progress, current focuses of research and development and the difficulties in target exploration are reviewed. The characteristics of the currently explored targets and their correlation to the level of difficulty for target exploration are analyzed. From these characteristics, simple rules can be derived for estimating the difficulty level of target exploration. The feasibility of predicting druggable proteins by using simple rules and sequence-derived physicochemical properties is also discussed.

Pharmaceutical agents generally exert their therapeutic effect by binding to and regulating the activity of a particular protein or nucleic acid target [1,2]. Modern drug discovery has been primarily based on the search of leads directed against a preselected target and the subsequent testing of the derived drug candidates [1,2]. Continuous efforts and interest have been directed at the discovery of new targets, as well as more-extensive exploration of the targets of successful drugs [1,2]. Rapid advances in genomics [3], structural genomics [4], proteomics [5], unravelling of molecular mechanisms of diseases [6], the development of experimental target identification and validation technologies [7,8] have enabled the identification of new targets, the discovery of clues to the molecular mechanism of drug actions and adverse drug reactions, and the understanding of the pharmacogenetic implications of the variation of the DNA sequence, expression profiles and post-transcriptional processing of targets [6,9].

These advances have led to the discovery of a growing number of targets [10–12]. Moreover, existing targets have continuously been explored for the development of more-effective drugs such as subtype-specific agents [13–15]. The number of reported new and existing targets has increased from ~500 targets reported in a 1996 survey [2] to 1494 distinct protein subtypes and 41 nucleic acids described in the current version of the Therapeutic Target Database (TTD, http://bidd.nus.edu.sg/group/ttd/ttd.asp) [16]. Apart from the emergence of new targets, the significant increase in the number of targets is partly caused by a combination of the increase of subtype-specific agents directed at a subtype of an existing target (that splits into two or more targets) and the accumulation of information about previously unknown or unreported targets of existing drugs or investigative agents [13–15].

Targets can be divided into two groups. One consists of successful targets that are targeted by at least one marketed drug, and the other contains research targets that are targeted only by investigational agents not approved for clinical use at present. Analysis of both groups of targets provides useful information about general trends, the current focus of research and development, areas of success and difficulties in the exploration of targets. Knowledge of target characteristics, such as protein sequence features, structural properties, proteomic profiles, pathway affiliation and roles, and tissue-distribution patterns, is also useful for molecular dissection of the mechanism of action of drugs and for predicting features to guide drug design and target discovery [7,8,17,18].

Corresponding author: Chen, Y.Z. (phacyz@nus.edu.sg)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.03.012

Success stories in the exploration of therapeutic targets

Targets of certain disease classes appear to be more successfully explored than others, which can be exhibited by the statistical profiles of the successful targets. There are 268 successful targets in TTD [16] compared with 120 successful targets in the 1996 survey [19]. The most highly explored group of successful targets is that of neoplasms, followed by infectious and parasitic diseases, nervous system and sensory organ disorders, and circulatory system diseases.

Examples of the well-established successful targets in these classes are: estrogen receptor (breast cancer) and gonadotropin-releasing hormone (prostate cancer) in the neoplasms; HIV-1 protease (AIDS) and penicillin-binding proteins (bacterial infections) in the class of infectious and parasitic diseases; acetylcholinesterase (Alzheimer's disease), catechol-O-methyl-transferase (Parkinson's disease), 5-hydroxytryptamine 1D receptor (migraine), and mu- and kappa-opioid receptors (drug dependence) in the class of nervous system and sensory organ disorders; and angiotensin-converting enzyme (hypertension, cardiac failure, arrhythmias) in the class of circulatory system diseases.

More-recent successful targets can be found in the list of FDA-approved drugs 2000–2005, together with the drugs approved during this period, examples are outlined in Table 1 (for a more extensive list see online supplementary material Table 1). There are 70 identifiable targets, most of which have been targeted by drugs marketed before 2000 [16]. Therefore, it appears that the majority of the successful targets have been continuously explored for deriving new therapeutic agents. Examples for continuously explored targets during 2000–2005 are 5-hydroxytryptamine (5HT) receptors with 11 drugs, adrenoceptors with seven drugs and cyclooxygenase (COX) with five drugs.

In total, 16 new successful targets have emerged since 1996 [20]. The largest group of these targets is in the neoplasm family, showing that cancer continues to be a major focus of new-target exploration. Examples of these targets are: receptor protein tyrosine kinase erbB2 (HER2/neu) with Herceptin® approved in 1998 for HER2-positive metastatic breast cancer; BCR/ABL tyrosine kinase with Gleevec® approved in 2001 for chronic myeloid leukaemia; hepatitis B virus (HBV) DNA polymerase with Hepsera® approved in 2002 for HBV; phosphodiesterase 5 with Viagra® approved in 1998 for erectile dysfunction; and COX-2 with Celebrex® approved in 1998 for arthritis.

Current research and development initiatives

Investigating the research targets, particularly those targeted by recently reported or patented agents, provides useful clues about how drug development has been focused on targets of specific disease classes. There are 894 research targets in TTD [16], compared with 79 research targets in the 1996 target survey [19]. The distribution pattern of these research targets with respect to disease classes, along with that of successful targets, is shown in Figure 1. As expected, the number of investigational targets is higher than the number of successful targets for every disease class. The only exception is the class of congenital anomalies, where minimal success seems to have been made in the identification of useful targets. This is partly a result of the current preference for surgical therapies as primary treatment options [21]. The lack of knowledge about the mechanism of congenital anomalies [22] also makes it difficult to identify useful targets for the diseases in this class.

The ratio between research and successful targets in a disease class gives a good indication about the extent of the exploration of

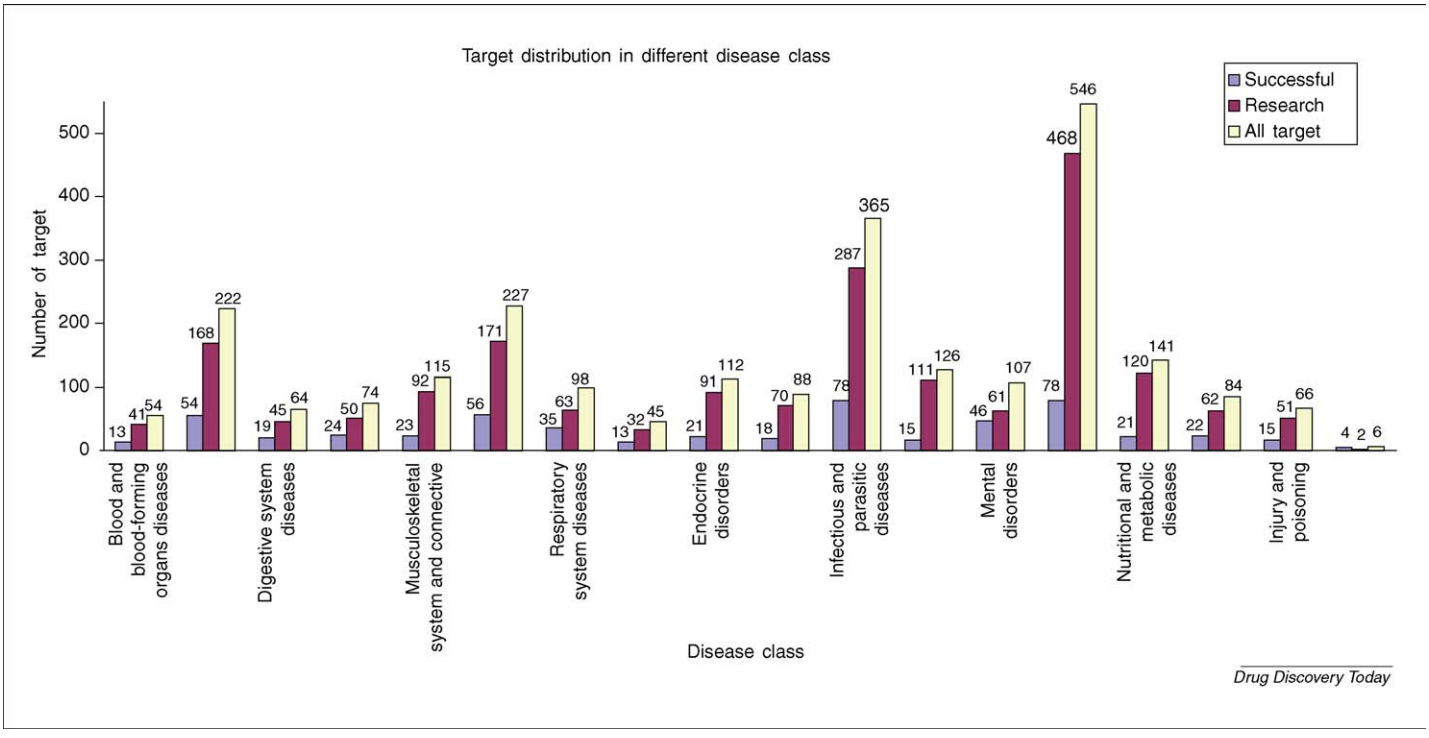

FIGURE 1

Comparison of the distribution pattern of successful targets and research targets in disease classes. The blue bars represent successful targets, the red bars represent research targets and the yellow bars represent all targets.

TABLE 1

Successful targets of FDA-approved drugs 2000–2005 (as of September 2005)

Disease class	Disease category	Number of targets	Target	Number of drugs	Drug
Circulatory system diseases	Circulation disorders	2	Endothelin receptor	1	Tracleer (bosentan)
			Type-1 angiotensin II receptor	3	Benicar; Diovan; Teveten HCT (eprosartan mesylate–hydrochlorothiazide)
	Heart disorders	2	HMG-CoA reductase	1	Caduet (amlodipine–atorvastatin)
			β-1 adrenergic receptor	1	Betapace AF tablet (Sotalol)
Infectious and parasitic diseases	Bacterial infections	1	DNA topoisomerase II	1	Avelox I.V. (moxifloxacin hydrochloride)
	Fungal infections	1	1,3-β-glucan synthase	1	Cancidas
	Parasitic infections	1	Dihydrofolate reductase	1	Malarone (atovaquone; proguanil hydrochloride) tablet
	Viral infections	4	HBV polymerase	1	Baraclude (entecavir)
			DNA polymerase	1	Hepsera (adefovir dipivoxil)
			HIV-1 protease	4	Aptivus (tipranavir); Kaletra capsules and oral solution; Lexiva (fosamprenavir calcium); Reyataz (atazanavir sulfate)
			HIV-1 reverse transcriptase	3	Sustiva; Trizivir (abacavir sulfate; lamivudine; zidovudine AZT) tablet; Viread
Neoplasms	Breast cancers	2	Cytochrome P450 19	1	Femara (letrozole) tablets
			Estrogen receptor	1	Faslodex (fulvestrant)
	Gastrointestinal cancers	3	Epidermal growth factor receptor	1	Erbitux (cetuximab)
			Tyrosine protein kinase	1	Gleevec (imatinib mesylate)
			Vascular endothelial growth factor	1	Avastin (bevacizumab)
	Leukaemia	3	Tyrosine protein kinase	1	Gleevec (imatinib mesylate)
			Ribonucleotide reductase	1	Clolar (clofarabine)
			DNA polymerase	1	Clolar (clofarabine)
	Lung cancers	2	HER1–EGFR tyrosine kinase	1	Tarceva (erlotinib, OSI 774)
			Epidermal growth factor receptor	1	Iressa (gefitinib)
	Lymphoma	1	B-lymphocyte antigen CD20	1	Bexxar
	Mesothelioma	3	Dihydrofolate reductase	1	Alimta (pemetrexed for injection)
			Glycinamide ribonucleotide formyltransferase	1	Alimta (pemetrexed for injection)
			Thymidylate synthase	1	Alimta (pemetrexed for injection)
	Reproductive organ cancers	1	Gonadotropin-releasing hormone	3	Eligard (leuprolide acetate); Plenaxis (abarelix for injectable suspension); Viadur (leuprolide acetate implant)
Nervous system and sensory organ diseases	Eye disorders	5	Vascular endothelial growth factor	1	Macugen (pegaptanib)
			β-1 adrenergic receptor	1	Betaxon (levobetaxolol)
			DNA topoisomerase II	1	Quixin (levofloxacin)
			Prostaglandin F2-α receptor	1	Travatan (travoprost ophthalmic solution)
			Topoisomerase IV	1	Quixin (levofloxacin)
	Headache	2	5-hydroxytryptamine 1B receptor	4	Axert (almotriptan malate) tablets; Frova (frovatriptan succinate); Relpax (eletriptan hydrobromide); Zomig-ZMT (zolmitriptan)
			5-hydroxytryptamine 1D receptor	4	Axert (almotriptan malate) tablets; Frova (frovatriptan succinate); Relpax (eletriptan hydrobromide); Zomig-ZMT (zolmitriptan)
	Insomnia	1	Melatonin MT1 and MT2 receptor	1	Rozerem (ramelteon)
	Neuronal disorders	1	DNA topoisomerase II	1	Novantrone (mitoxantrone hydrochloride)
	Parkinson's disease	4	D(1B) dopamine receptor	1	Apokyn (apomorphine hydrochloride)
			D(2) dopamine receptor	1	Apokyn (apomorphine hydrochloride)
			D(3) dopamine receptor	1	Apokyn (apomorphine hydrochloride)
			D(4) dopamine receptor	1	Apokyn (apomorphine hydrochloride)

TABLE 2

Some of the research targets explored for the new investigational agents described in the US patents approved between 2000–2005 (as of September 2005)

Target	Target group	Number of US patents 2000–2005	Year of first reported compound investigation	Targeted diseases
Matrix metalloproteinase(MMP)-12	MMPs	1	2004	Ulcerative colitis; Crohn's disease; lung damage by cigarette smoke; atherosclerosis; gastro–intestinal ulcers; emphysema
Tumour necrosis factor-α converting enzyme	ADAMs	29	1998	Inflammation; cancers
Cathepsin K	Cysteine proteases	25	1996	Autoimmune diseases; cartilage degradation; osteoporosis; pulmonary disorders
Phosphodiesterase (PDE) 4	PDEs	51	1995	Inflammation; obstructive diseases; asthma
Cathepsin S	Cysteine proteases	16	1994	Autoimmune diseases; osteoporosis; arthritis; muscular dystrophy; inflammation
α v β 5 integrin receptor	Integrin receptors	18	1993	Cancers; osteoporosis; arteriosclerosis; restenosis; ophthalmic disorders
MMP-3	MMPs	10	1992	Multiple sclerosis; heart failure; cancer; inflammation; pain; arthritis; osteoporosis; autoimmune disorders
α v β 5 integrin receptor	Integrin receptors	26	1991	Cancers; osteoporosis; arteriosclerosis; restenosis; ophthalmic disorders
Farnesyl protein transferase (FPT)	FPTases	26	1990	Cancers; restenosis; psoriasis; endometriosis; atherosclerosis; viral infections
MMP-2	MMPs	14	1988	Cancers; rheumatoid arthritis; osteoarthritis; multiple sclerosis
Cathepsin L	Cysteine proteases	15	1976	Autoimmune diseases; myocardial infarct; stroke; inflammation; muscular dystrophies

additional targets for diseases in that class. The classes with the largest ratios are neoplasms (468:78), infectious and parasitic diseases (287:78), nervous system and sensory organ disorders (171:56) and circulatory system diseases (168:54). These numbers indicate that intensive efforts are directed at diseases such as cancer (468 targets) [23,24], cardiovascular diseases (120 targets) [25,26], inflammation (113 targets) [27], diabetes (65 targets) [28], arthritis (64 targets) [29], obesity (57 targets) [30,31] and Alzheimer's disease (44 targets) [32,33]. The highest number of research targets and the highest number of successful targets are found in the same disease class, the neoplasms, indicating a continuous research and development focus in this important field.

A more detailed picture of the current trend of target exploration can be obtained from the research targets described in recently approved US patents for investigational agents. Table 2 shows some of the research targets described in the US patents (for a more extensive list see supplementary material Table 2). Research targets covered by a large number of patents are: matrix metalloproteinases, 46 patents for cancers, tissue ulceration, abnormal wound healing, periodontal disease, bone disease, diabetes, arthritis, atherosclerosis and inflammation; and phosphodiesterase 4, 51 patents for inflammation, asthma, prostate diseases and osteoporosis.

Among the 395 identifiable targets described in the US patents approved in 2000–2005, only 62 (15.6%) are subtype-specific targets that have been explored for the development of subtype-specific drugs. During the same period, of the 70 targets of FDA-approved drugs only 11 (15.7%) are subtype-specific targets. The percentage of subtype-specific targets in the US patents is essentially the same as that of the FDA-approved drugs, which seems to indicate the persistently high level of difficulty for finding subtype-specific therapeutic agents.

Several new targets have emerged from the US patents approved during this time. Most of these targets are intended for the treatment of high-impact diseases needing effective or more treatment options, reflecting the intensive effort from the pharmaceutical industry and research community for finding innovative methods in the treatment of such diseases. Examples of these newly emerged targets are 88 kDa glycoprotein growth factor for cancer (US patent 6,670,183), anandamide amidase for pain (US patent 6,579,900), gamma-secretase for Alzheimer's disease (US patent 6,448,229), and orexin receptor 1 for obesity (US patent 6,677,354).

Difficulties in target exploration

Although a large number of research targets, as well as successful targets, are being intensively explored, drug development productivity has been short of expectations [34]. It typically takes 12 years after the design effort is initiated to develop a marketable drug against a target [34], and drug failures have frequently been attributed to sloppy early target validation [35]. These difficulties are likely to be associated with the molecular and cellular characteristics of a drug target. Examination of difficulties in target exploration reported in the literature provides useful information about factors that affect target exploration, subsequently helping to identify characteristics of a target that contribute to these difficulties.

The difficulties of target exploration reported in the literature are mainly concerned with drug-binding specificity, efficacy, toxicity and pharmacokinetic properties. For instance, corticotrophin-releasing factor (CRF) receptor subtypes have been implicated in

behavioural and endocrine responses to stress and emotion, including fear, anxiety and aggression. Receptor subtypes, such as CRF1, have been explored for the treatment of these diseases [36] and the first antagonists were reported in 1986 [37]. Selective non-peptide antagonists have shown anxiolytic and antidepressant effects under certain conditions. However, the chemical space defined by known antagonists is fairly narrow, making it difficult for finding safe and effective candidates [36]. One compound is in Phase II trials for the treatment of major depressive disorder [36].

By contrast, cysteinyl leukotriene receptor 1 is a successful target, reported to have been explored initially in 1986 [37]. It has been targeted for the treatment of asthma and the first FDA-approved drug, accolate, appeared in 1996 [38]. A comparison of the characteristics of these two targets shows that CRF1 has 51 homologues outside the target-protein family, compared with five homologues for cysteinyl leukotriene receptor 1. Homologous proteins are searched from human protein entries in the SWISS-PROT database using PSI-BLAST (with the default similarity criterion, E-value <0.001) [39].

There are seven other research targets, reported to have been first explored in 1986 and intensively studied in subsequent years without producing a marketed drug. These are apolipoprotein(a), cyclophilin, plasminogen activator inhibitor-1, integrin α-2, thromboxane A2 receptor, proto-oncogene protein c-Fos, and interleukin-4. These targets either have high numbers of homologues outside the target family (>6) or are affiliated with multiple tissues (>3) or pathways (>3). The number of affiliated tissues for each target is obtained from SWISS-PROT [40], with the description that each target is primarily distributed in these tissues, and those of affiliated pathways are obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [41]. Statistical analysis of the characteristics of all of the long-term research targets (described in a later section) suggests that high numbers of homologues outside the target family and affiliation with multiple tissues and pathways are the major factors leading to the low success rate for the exploration of these targets.

The derived target characteristics depend on the choice of parameters of bioinformatics tools and the quality of data sources. In estimating the number of similarities of proteins of each target, a stricter PSI-BLAST cut-off, E-value = 0.001, was used. This value has been reported to give reliable predictions of homologous relationships [42] and it can be used to find 16% more structural relationships in the SCOP database than when using a standard sequence similarity with a 40% sequence-identity threshold [43]. The majority of protein pairs that share 40–50% (or higher) sequence-identity differ by <1 Å RMS deviation [44,45], and a larger structural deviation probably alters drug-binding properties. Therefore, the adopted E-value seems to be reasonable for selecting similarity proteins relevant to the binding of a common set of drugs. None-the-less, small percentages of protein pairs of higher sequence-identity have been found to differ by larger RMS deviations [45] and some protein pairs of lower sequence-identity might also have high structural similarity, which probably affects the accuracy of our analysis to some extent.

In estimating the number of affiliated tissues of each target, relevant data from SWISS-PROT were used. We were able to find the published literature for 92% of these data and random checks of these publications confirm the quality of the data. We have also used the level-4 tissue-distribution data from another database, TissueDistributionDBs (http://genome.dkfz-heidelberg.de/menu/tissue_db/index.html), to derive the tissue-distribution pattern of the same set of 158 targets. A target is assumed to be primarily distributed in a tissue if no less than 8% of the total protein contents are distributed in that tissue. Approximately 28%, 24%, 19%, 10%, 6%, 6%, 5% and 1% of these targets were found to be affiliated to 1–8 tissues, respectively, this is similar to those derived from SWISS-PROT data, even though the definition and content of these databases are different. Therefore, our estimated tissue-distribution profiles are stable but the exact percentages can differ to some degree.

Infectious disease targets are primarily microbial or viral proteins that are crucially involved in the infection or growth of the invading species, differing from other targets (human proteins or nucleic acids) involved in the modulation of human physiology. Although the success rate for the exploration of these anti-infective targets is also determined by the general characteristics of each target, additional factors also significantly influence drug development efforts and outcomes. For instance, with the first reported investigation in 1975, there are 12 infectious disease research targets, many of which have not been followed up because of the scope and the type of populations and regions affected, profitability considerations and/or target prioritization for those species with multiple targets [46]. Drug development against anti-infective targets can also be hindered by the rapid appearance of drug-resistance mutations [47], the disappearance of infectious strains [48] and the evolution of virulence strains into non-virulence ones [49,50].

Several targets have been used for hormone replacement therapies. Examples of these targets are erythropoietin receptor for anemia [51], estrogen receptor for postmenopausal women [52], insulin receptor for diabetes mellitus [53], as well as growth hormone secretagogue receptor type 1, lutropin-choriogonadotropic hormone receptor and steroid hormone receptor, estrogen-related receptor-1 (ERR1), for growth hormone deficiency [54]. In general, these targets can be divided into three classes. The first includes targets with a protein or peptide as a hormone substrate (e.g. erythropoietin receptor, insulin receptor), the second includes targets with a steroid as a hormone substrate (e.g. estrogen receptor and steroid hormone receptor ERR1) and the third includes targets with a tyrosine-derived chemical as a hormone substrate. A problem for exploring the targets of the first class is the difficulty in finding a suitable mutant of the hormone with sufficiently strong and specific binding to the intended target [55]. Targets of the second and third class often contain a higher number of homologues and are distributed in multiple tissues, which are the primary reasons for the difficulty in finding safe and highly potent analogues of the hormones [56,57].

Target characteristics and their correlation to the difficulty level of drug development

Pharmacodynamic, toxicological and pharmacokinetic properties primarily arise from interactions of a drug with its target and other molecular entities. Therefore, certain common characteristics are expected for good targets [19]. Good targets should play crucial roles in disease processes, be structurally novel for drug specificity, have few homologues having similar binding sites in humans, should not be heavily involved in other key pathways and should

TABLE 3

Profiles of some of the innovative targets of drugs approved by the FDA since 1994

Target	Year of first reported compound investigation	Year of first FDA approval	Target exploration time (years)	Number of human similarity proteins outside target-protein family	Number of human similarity proteins in target-protein family	Number of tissues target is primarily distributed	Number of pathways target is distributed	Predicted target difficulty level	First FDA-approved drug
Maltase-glucoamylase, intestinal	1967	1995	28	1	12	3	2	D	Precose
Mineralocorticoid receptor	1975	2002	27	31	101	Many	?	D	Eplerenone
Prostaglandin G/H synthase 2	1975	1998	23	33	13	4	1	D	Celebrex
Acetyl-CoA carboxylase 2	1975	1994	19	30	0	3	5	D	Glucophage
Inosine-5′-monophosphate dehydrogenase 2	1979	1995	16	4	10	4	1	D	CellCept
Phosphodiesterase 5	1984	1998	14	3	74	5	1	D	Viagra
Myeloid cell surface antigen CD33	1987	2000	13	2	21	2	1	E	Mylotarg
Type-1 angiotensin II receptor	1984	1995	11	8	388	4	2	D	Cozaar
Cysteinyl leukotriene receptor 1	1986	1996	10	5	386	2	2	E	Accolate
Receptor protein tyrosine kinase erbB-2	1988	1998	10	18	482	1	4	D	Herceptin
FK-binding protein 12	1989	1999	10	0	30	2	?	E	Rapamune
P2Y purinoceptor 12	1989	1997	8	3	280	2	?	E	Plavix

There are two target difficulty levels, E represents an 'easy' target that has a shorter expected target-exploration time and D represents a 'difficult' target that has a longer expected target-exploration time.

be selectively or minimally expressed for drug efficacy. Target molecules possess certain structural and physicochemical features for binding drug-like molecules. These characteristics probably define the sequence, structural, proteomic, pathway affiliation (functional role) and physiological profiles of therapeutic targets – and subsequently the level of difficulty for finding viable drugs against them.

Table 3 gives the target exploration (TE) time and the characteristics of some of the innovative targets of the drugs approved by the FDA since 1994, these targets had no marketed drugs before this date [20]. TE time is the number of years needed for developing a marketable drug after a target is first tested for compound design, which can be crudely estimated from the year of the first reported compound investigation and the first FDA approval. Target characteristics include the number of homologues, the number of tissues in which the target is primarily distributed and the number of pathways in which the target is involved.

The TE time seems to be statistically correlated to target characteristics. Targets with a smaller number of homologues outside the target-protein family, involved in fewer numbers of pathways and distributed in fewer numbers of tissues, tend to have a statistically shorter TE time than those with a larger number of these entities. If the difficulty level of target exploration is related to the number of these entities for a target, the number of successful targets with smaller numbers of these entities is expected to be less than those with higher numbers of these entities, which is indeed the case for the 132–190 successful targets with such information (Table 4).

One can define 'easy' targets to be those with a short TE time and the 'difficult' targets to be those with a long TE time. A simple rule for assigning easy and difficult targets can be derived based on two assumptions: first, at least 50% of the successful targets with smaller numbers of homologous proteins outside target family, involved in fewer numbers of pathways and distributed in a lower number of tissues, are considered to be easy targets; second, at least 85% of the research targets of noninfectious diseases first explored on and before 1986 are considered to be difficult targets.

By statistically analyzing the characteristics of these easy and difficult targets, the following rules are found: a target is an easy target if it has no more than five human homologues outside the target-protein family; it should be involved in no more than two pathways and primarily be expressed in no more than two tissues. These criteria ensure that the structural architecture and expression profile of easy targets can accommodate target-specific drugs that minimally interact with other pathways, tissues and functionally important (but structurally similar) binding sites.

By using this simple rule, the innovative targets in Table 3 are categorized as easy and difficult targets. The TE time of easy targets is generally no more than ten years and the TE time of the difficult targets is generally >14 years. These results suggest that it is indeed possible to use certain target characteristics as a way of estimating the level of difficulty for target exploration.

Feasibility for predicting druggable proteins by using simple rules and sequence-derived properties

Druggable proteins are proteins whose activity can be regulated by drug-like molecules, and those playing key roles in a disease process can be potentially explored as therapeutic targets [19]. With rapid advances in genomics and other aspects of disease mechanisms, there has been an increasing interest in developing methods for predicting druggable proteins from genomic and other available biological data. It is expected that the same features used for generating simple rules for classifying the difficulty level

TABLE 4

Statistics of characteristics of successful targets

Category	Human similarity proteins outside target family		Human similarity proteins in target family		Target pathways participating		Tissue distribution		Subcellular location		Biochemical class	
Number of targets in statistics	190		190		132		158		153		268	
Item	Number of similarity proteins	% of targets with this number of similarity proteins	Number of similarity proteins	% of targets with this number of similarity proteins	Number of pathways	% of targets in this number of pathways	Number of tissues	% of targets primarily distributed in this number of tissues	Location	% of targets primarily distributed in location	Class	% of targets in class
Statistical data	0–5	51%	0–5	33%	1	49%	1	28%	Membrane	60%	Enzymes	50%
	6–10	19%	6–30	23%	2	27%	2	25%	Cytoplasm	16%	Receptors	23%
	11–20	10%	31–100	18%	3	11%	3	15%	Nucleus	10%	Channels and transporters	12%
	21–40	11%	>100	26%	4	2%	4	8%	Extra-cellular and secreted	8%	Nuclear receptors	6%
	41–80	5%			5	3%	5	3%	Mitochondrion	3%	Factors and regulators	3%
	>80	4%			6	2%	6	1%	Endoplasmic reticulum	2%	Structural proteins	2%
					7		7	2%	Peroxisome	1%	Nucleic acids	2%
					8	3%	8	4%			Other binding proteins and antigens	2%
					9	1%	9	1%				
					≥10	2%	≥10	13%				

of target exploration can be used for developing rules and methods for facilitating the prediction of druggable proteins from sequence-derived information, which is particularly useful because sequence information is usually more readily available.

The following simple rules can be introduced for a quick search of druggable proteins based on their characteristics. These rules are derived from the analysis of all of the known targets and are based on the assumption that it is applicable to a majority of the known targets (>75%). A druggable protein belongs to one of a limited number of target-representing protein domain families (out of a total of 7973 families, currently known successful and research targets are represented by 92 and 412 families, respectively). Sequence variation between the drug-binding domain of the protein and the drug-binding domains of other members of the same family must allow a sufficient degree of differential binding to a 'rule-of-five' molecule in this common binding site. Preferably, the protein has <6 homologues outside its family (51% of the successful targets with identifiable drug-binding domain have <6 human similarity proteins). Moreover, the protein is preferably involved in no more than two pathways (76% of the successful targets with pathway information are associated with no more than two pathways). For diseases that tend to concentrate on specific organs or tissues, protein is preferably expressed in no more than five tissues in human (53% of the successful targets with tissue distribution information are primarily distributed in no more than two tissues). Proteins with a higher number of homologues, or association with a higher number of human pathways or distribution in a higher number of tissues, are not necessarily non-druggable. It generally increases the chance of unwanted interferences and, thus, the level of difficulty for finding viable drugs against these proteins.

Although having certain common characteristics, therapeutic targets do not necessarily bear sequence similarity to each other or to other disease-related proteins, they are from a diverse range of different families and structural folds. Therefore, a straightforward sequence similarity search against known target classes [58] and disease proteins [59] might not always be useful for identifying novel targets. It is desirable to develop druggable protein predictive tools by using methods that are not based on sequence similarity.

One strategy is to use an artificial intelligence (AI) method, such as support vector machine (SVM) [60], to develop an AI system for separating druggable and non-druggable proteins, illustrated in Figure 2. A protein sequence is translated into sequence-derived physicochemical properties that include amino acid composition, hydrophobicity, normalized van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility [60]. These properties are mathematically represented by a feature vector – feature vectors of known, druggable proteins and those of non-druggable proteins are separated by a hyperplane in a mathematically-defined hyperspace [61]. A new protein can be predicted as druggable or non-druggable based on the location of its feature vector with respect to this hyperplane.

Our preliminary study suggests that 69.8% of the known targets and 99.3% of the non-target proteins can be predicted by SVM as druggable and non-druggable, respectively. A further search of the human genome identified 1102 druggable proteins that include 153 G-protein-coupled receptors (GPCRs), 65 other receptors, 333 enzymes and 56 channels. These numbers are consistent with the estimated numbers of druggable proteins and therapeutic targets

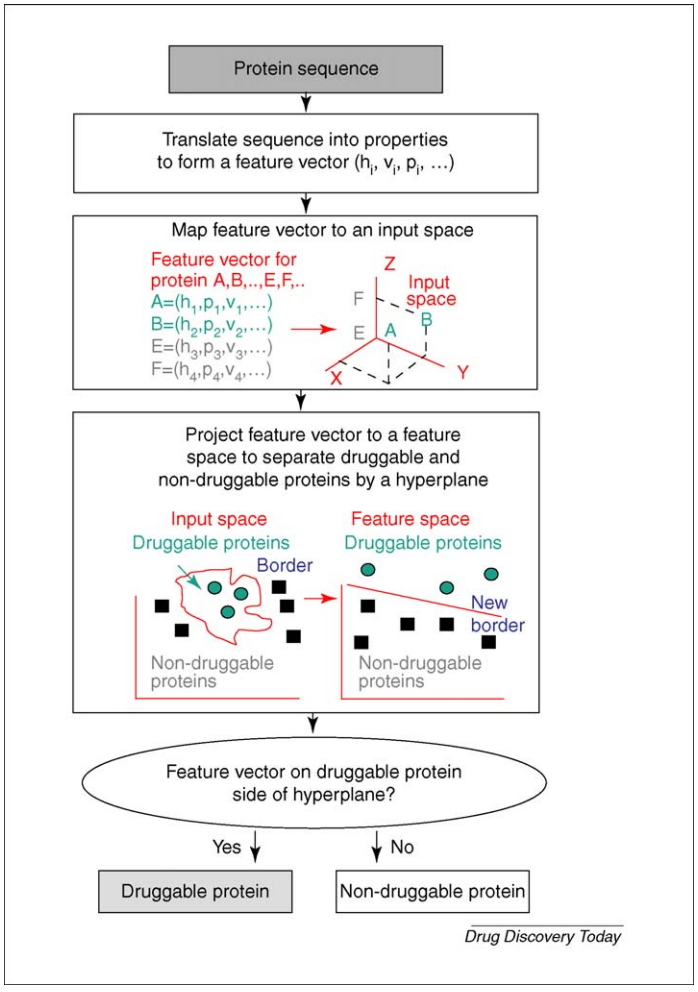

FIGURE 2

Schematic diagram illustrating the process of druggable protein prediction from the protein sequence (using a statistical learning method called Support Vector Machines). A and B are feature vectors of druggable proteins; E and F are feature vectors of non-druggable proteins; green circles are druggable proteins; black-filled squares are non-druggable proteins; feature vector (h_i, v_i, p_i, . . .) represents hydrophobicity, volume and polarizability.

in the human genome [19,62]. These seem to suggest that an AI method such as SVM has some capacity for facilitating the identification of druggable proteins from their sequences, and further investigation and development is warranted.

Concluding remarks

Strong drug design efforts and resources appear to have been continuously devoted to the design of therapeutic agents targeted at successful targets for high-impact diseases that need more treatment options or more-effective drugs. The majority of the recently approved drugs and patented investigative agents have been designed on the basis of these targets. Thus, the search of novel agents directed at these targets might be considered as a 'good bet' and, thus, one of the focuses by the pharmaceutical industry.

There is also an indication of strong interest and increasing efforts in the exploration of new targets, particularly for high-impact diseases that lack effective treatment options. These include cancer, bacterial and viral infections, cardiovascular diseases and obesity. By taking advantage of the progress in genome sequencing and in the more extensive understanding of disease mechanisms, exploration of new targets has become increasingly subtype-specific and, for some diseases, pathogen-species-specific. It is expected that more subtype-specific and pathogen-species-specific targets will be explored.

Rapid progress in genomics [3], structural genomics [4] and proteomics [5] is revolutionizing the process of target identification and drug development. In addition to the experimental methods [7,8], computational methods have recently been introduced for facilitating target identification and validation. These computational methods explore comparative sequence analysis [63] and ligand–protein inverse docking [64] for recognition of target-like and druggable proteins by analysis of their sequence or structural features. It is also feasible to use rule-based methods and statistical learning methods, such as SVM for predicting druggable proteins from sequence-derived properties and for estimating the level of difficulty of drug development. These methods could potentially be developed into useful tools for facilitating the identification of novel targets from the human and pathogen genomes. This progress, combined with advances in molecular understanding of disease processes [6], has provided opportunities for the discovery of new targets that enable the development of new therapies and personalized medicine.

Acknowledgements

We thank Cui Juan, Zhang Hailei, Li Hu, Lin Honghuang, Tang Zhiqun and Ung Choong Yong for their contributions in literature searching. This work was supported in part by grants from Singapore ARF R-151–000–031–112, Shanghai Commission for Science and Technology (04QMX1450) and the 973 National Key Basic Research Program of China (2004CB720103).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.drudis.2006.03.012.

References

1 Ohlstein, E.H. *et al.* (2000) Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol.* 40, 177–191
2 Drews, J. (1997) Strategic choices facing the pharmaceutical industry: a case for innovation. *Drug Discov. Today* 2, 72–78
3 Debouck, C. and Metcalf, B. (2000) The impact of genomics on drug discovery. *Annu. Rev. Pharmacol. Toxicol.* 40, 193–207
4 Sali, A. (1998) 100 000 protein structures for the biologist. *Nat. Struct. Biol.* 5, 1029–1032
5 Dove, A. (1999) Proteomics: translating genomics into products? *Nat. Biotechnol.* 17, 233–236
6 Macdonald, I.A. (2000) Obesity: are we any closer to identifying causes and effective treatments? *Trends Pharmacol. Sci.* 21, 334–336
7 Walke, D.W. *et al.* (2001) *In vivo* drug target discovery: identifying the best targets from the genome. *Curr. Opin. Biotechnol.* 12, 626–631
8 Ilag, L.L. *et al.* (2002) Emerging high-throughput drug target validation technologies. *Drug Discov. Today* 7 (Suppl. 18), 136–142

9 Nicholls, H. (2003) Improving drug response with pharmacogenomics. *Drug Discov. Today* 8, 281–282
10 Chiesi, M. *et al.* (2001) Pharmacotherapy of obesity: targets and perspectives. *Trends Pharmacol. Sci.* 22, 247–254
11 Matter, A. (2001) Tumor angiogenesis as a therapeutic target. *Drug Discov. Today* 6, 1005–1024
12 Greenfeder, S. and Anthes, J.C. (2002) New asthma targets: recent clinical and preclinical advances. *Curr. Opin. Chem. Biol.* 6, 526–533
13 Vane, J.R. *et al.* (1998) Cyclooxygenases 1 and 2. *Annu. Rev. Pharmacol. Toxicol.* 38, 97–120
14 Torphy, T.J. and Page, C. (2000) Phosphodiesterases: the journey towards therapeutics. *Trends Pharmacol. Sci.* 21, 157–159
15 Kennedy, B.P. and Ramachandran, C. (2000) Protein tyrosine phosphatase-1B in diabetes. *Biochem. Pharmacol.* 60, 877–883
16 Chen, X. *et al.* (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res.* 30, 412–415
17 van de Waterbeemd, H. and Gifford, E. (2003) ADMET *in silico* modelling: towards prediction paradise? *Nat. Rev. Drug Discov.* 2, 192–204
18 Kennedy, T. (1997) Managing the drug discovery/development interface. *Drug Discov. Today* 2, 436–444
19 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730
20 Zambrowicz, B.P. and Sands, A.T. (2003) Knockouts model the 100 best-selling drugs–will they model the next 100? *Nat. Rev. Drug Discov.* 2, 38–51
21 Lin, P.C. *et al.* (2002) Female genital anomalies affecting reproduction. *Fertil. Steril.* 78, 899–915
22 Kobayashi, H. and Stringer, M.D. (2003) Biliary atresia. *Semin. Neonatol.* 8, 383–391
23 Buolamwini, J.K. (1999) Novel anticancer drug discovery. *Curr. Opin. Chem. Biol.* 3, 500–509
24 Elsayed, Y.A. and Sausville, E.A. (2001) Selected novel anticancer treatments targeting cell signaling proteins. *Oncologist* 6, 517–537
25 Persidis, A. (1999) Cardiovascular disease drug discovery. *Nat. Biotechnol.* 17, 930–931
26 Bicknell, K.A. *et al.* (2003) Targeting the cell cycle machinery for the treatment of cardiovascular disease. *J. Pharm. Pharmacol.* 55, 571–591
27 Lewis, A.J. and Manning, A.M. (1999) New targets for anti-inflammatory drugs. *Curr. Opin. Chem. Biol.* 3, 489–494
28 Wagman, A.S. and Nuss, J.M. (2001) Current therapies and emerging targets for the treatment of diabetes. *Curr. Pharm. Des.* 7, 417–450
29 Blake, S.M. and Swift, B.A. (2004) What next for rheumatoid arthritis therapy? *Curr. Opin. Pharmacol.* 4, 276–280
30 Bray, G.A. and Tartaglia, L.A. (2000) Medicinal strategies in the treatment of obesity. *Nature* 404, 672–677
31 Clapham, J.C. *et al.* (2001) Anti-obesity drugs: a critical review of current therapies and future opportunities. *Pharmacol. Ther.* 89, 81–121
32 Windisch, M. *et al.* (2002) Current drugs and future hopes in the treatment of Alzheimer's disease. *J. Neural Transm. Suppl.* 62, 149–164
33 Irizarry, M.C. and Hyman, B.T. (2001) Alzheimer disease therapeutics. *J. Neuropathol. Exp. Neurol.* 60, 923–928
34 Drews, J. (2003) Stategic trends in the drug industry. *Drug Discov. Today* 8, 411–420
35 Smith, C. (2003) Drug target validation: hitting the target. *Nature* 422, 341, 343, 345 *passim*.
36 Kehne, J. and De Lombaert, S. (2002) Non-peptidic CRF1 receptor antagonists for the treatment of anxiety, depression and stress disorders. *Curr Drug Targets CNS Neurol Disord* 1, 467–493
37 Young, R.N. *et al.* (1986) L-649,923, sodium (beta S*, gamma R*-4-(3-(4-acetyl-3-hydroxy-2-propylphenoxy)propylthio)-gamma-hydroxy -beta-methylbenzenebutanoate. A selective, orally active leukotriene D4 receptor antagonist. *Adv. Prostaglandin Thromboxane Leukot. Res.* 16, 37–45
38 Dunn, C.J. and Goa, K.L. (2001) Zafirlukast: an update of its pharmacology and therapeutic efficacy in asthma. *Drugs* 61, 285–315
39 Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402
40 Bairoch, A. *et al.* (2004) Swiss-Prot: juggling between evolution and stability. *Brief. Bioinform.* 5, 39–55
41 Kanehisa, M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280
42 George, R.A. and Heringa, J. (2002) Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins* 48, 672–681
43 Gerstein, M. (1998) Measurement of the effectiveness of transitive sequence comparison through a third 'intermediate' sequence. *Bioinformatics* 14, 707–714
44 Koehl, P. and Levitt, M. (2002) Sequence variations within protein families are linearly related to structural variations. *J. Mol. Biol.* 323, 551–562
45 Wood, T.C. and Pearson, W.R. (1999) Evolution of protein sequences and structures. *J. Mol. Biol.* 291, 977–995
46 Sternberg, J.M. (2004) Human African trypanosomiasis: clinical presentation and immune response. *Parasite Immunol.* 26, 469–476
47 Fikkert, V. *et al.* (2004) Multiple mutations in human immunodeficiency virus-1 integrase confer resistance to the clinical trial drug S-1360. *AIDS* 18, 2019–2028
48 Lee, N. *et al.* (2004) Effects of early corticosteroid treatment on plasma SARS-associated Coronavirus RNA concentrations in adult patients. *J. Clin. Virol.* 31, 304–309
49 Beard, C.W. and Mason, P.W. (2000) Genetic determinants of altered virulence of Taiwanese foot-and-mouth disease virus. *J. Virol.* 74, 987–991
50 Song, H. *et al.* (2005) Molecular determinants of infectious pancreatic necrosis virus virulence and cell culture adaptation. *J. Virol.* 79, 10289–10299
51 Lacombe, C. and Mayeux, P. (1998) Biology of erythropoietin. *Haematologica* 83, 724–732
52 Siddiqui, N.I. *et al.* (2005) Evaluation of hormone replacement therapy. *Mymensingh Med. J.* 14, 212–218
53 Zick, Y. (2004) Molecular basis of insulin action. *Novartis Found. Symp.* 262, 36–50
54 Petersenn, S. (2002) Structure and regulation of the growth hormone secretagogue receptor. *Minerva Endocrinol.* 27, 243–256
55 Huber, W. (2005) A new strategy for improved secondary screening and lead optimization using high-resolution SPR characterization of compound-target interactions. *J. Mol. Recognit.* 18, 273–281
56 Bursi, R. and Groen, M.B. (2000) Application of (quantitative) structure-activity relationships to progestagens: from serendipity to structure-based design. *Eur. J. Med. Chem.* 35, 787–796
57 Gooren, L.J. and Nguyen, N.T. (1999) One and the same androgen for all? Towards designer androgens *Asian J. Androl.* 1, 21–28
58 Sanseau, P. (2001) Impact of human genome sequencing for *in silico* target discovery. *Drug Discov. Today* 6, 316–323
59 Desany, B. and Zhang, Z. (2004) Bioinformatics and cancer target discovery. *Drug Discov. Today* 9, 795–802
60 Han, L.Y. *et al.* (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* 32, 6437–6444
61 Burges, C. (1998) A tutorial on Support Vector Machine for pattern recognition. *Data Min. Knowl. Disc.* 2, 121–167
62 Wise, A. *et al.* (2002) Target validation of G-protein coupled receptors. *Drug Discov. Today* 7, 235–246
63 Duckworth, D.M. and Sanseau, P. (2002) *In silico* identification of novel therapeutic targets. *Drug Discov. Today* 7 (Suppl. 11), 64–69
64 Chen, Y.Z. and Zhi, D.G. (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* 43, 217–226