The Development and Use of Databases for Ligand-Protein Interaction Studies

Hsin, Kun-Yi

A Thesis

Submitted for the Degree of

Doctor of Philosophy

Structural Biochemistry Group

The Institute of Structural and Molecular Biology

University of Edinburgh

October 2009

Abstract

This project applies structure-activity relationship (SAR), structure-based and database mining approaches to study ligand-protein interactions. To support these studies, we have developed a relational database system called **ED**inburgh **U**niversity **L**igand **S**election **S**ystem (EDULISS 2.0) which stores the structure-data files of +5.5 million commercially available small molecules (+4.0 million are recognised as unique) and over 1,500 various calculated molecular properties (descriptors) for each compound. A user-friendly web-based interface for EDULISS 2.0 has been established and is available at http://eduliss.bch.ed.ac.uk/.

We have utilised PubChem bioassay data from an NMR based screen assay for a human FKBP12 protein (PubChem AID: 608). A prediction model using a Logistic Regression approach was constructed to relate the assay result with a series of molecular descriptors. The model reveals 38 descriptors which are found to be good predictors. These are mainly 3D-based descriptors, however, the presence of some predictive functional groups is also found to give a positive contribution to the binding interaction. The application of a neural network technique called Self Organising Maps (SOMs) succeeded in visualising the similarity of the PubChem compounds based on the 38 descriptors and clustering the 36 % of active compounds (16 out of 44) in a cluster and discriminating them from 95 % of inactive compounds.

We have developed a molecular descriptor called the Atomic Characteristic Distance (ACD) to profile the distribution of specified atom types in a compound. ACD has been implemented as a pharmacophore searching tool within EDULISS 2.0. A

structure-based screen succeeded in finding inhibitors for pyruvate kinase and the ligand-protein complexes have been successfully crystallised.

This study also discusses the interaction of metal-binding sites in metalloproteins. We developed a database system and web-based interface to store and apply geometrical information of these metal sites. The programme is called **ME**tal **S**ites in **P**roteins at **E**dinburgh **U**niver**S**ity (MESPEUS; http://eduliss.bch.ed.ac.uk/MESPEUS/). MESPEUS is an exceptionally versatile tool for the collation and abstraction of data on a wide range of structural questions. As an example we carried out a survey using this database indicating that the most common protein types which contain Mg-$O_{ATP\text{-}phosphate}$ site are transferases and the most common pattern is linkage through the β- and γ-phosphate groups.

Declaration

The work presented in this thesis is the original work of the author. This thesis has been composed by the author and has not been submitted in whole or in part for any other degree.

Hsin, Kun-Yi

Acknowledgements

I would firstly like to acknowledge my supervisors Prof Malcolm Walkinshaw and Dr. Paul Taylor for the opportunity to do my Ph.D. in Edinburgh, their encouragement and advice. Work on this thesis would not have been possible without their guidance. I am grateful to Dr Marjorie Harding for her patience in teaching and explaining the knowledge of metalloprotein. I would also like to thank the examiners Prof Simon Mackay and Prof Lindsay Sawyer for their correction and recommendation.

I would like to thank Dr. Hugh Morgan for the collaboration in the work of pyruvate kinase project. A lot of appreciation goes to the people in the 3rd floor of Swann and the colleagues in the computational group, including Dr. Nicos Angelopoulos, Dr. Simon Harding, Dr. Andrew Hinton, Dr. Liz Blackburn, Yi-Gong Sheng, Dr. Steven Shave and Wissam Mehio. Steven, Wissam, Hugh and Dr. Daphne Kan need to be especially mentioned for the helpful technique support in lung cancer cell incubation. I would also be grateful to Dr. Kok-Lian Ho, Dr. Peter Brown and Shao-Fang Wang for the excellent demonstration in activating alcohol dehydrogenase.

The friends and partners, Fanny Kong, Gum-Gum, Hsiao-Che Kuo and his wife, Dr. Nien-Jen Hu and Lu Zhou, in Edinburgh whom I shouldn't forget for not only providing their amateur entertainment and cooking amazing meal, but also helping a bit for my study and this thesis.

Finally, I would like to give a sentence to my parents, brothers and CCC that is I love you and thank you very much as I never said that in my native language due to my shyness. I cannot accomplish this degree and overcome these lonely oversea days without your endless support. Hopefully, I have made you proud. Again, I love you all.

Abbreviations

ACD	Atomic Characteristic Distance
BMU	Best Matching Unit
CDK	Chemistry Development Kit
CIF	Crystallographic Information File
CSD	Cambridge Structural Database
EDULISS	EDinburgh University Ligand Selection System
GETAWAY	GEometry, Topology, and Atom-Weights AssemblY
HAcc	Hydrogen bond acceptors
HDon	Hydrogen bond donors
HTS	High-throughput screening
IC50	Half maximal (50 %) inhibitory concentration
InChI	IUPAC International Chemical Identifier
Kd	Dissociation constant
***Lm*PYK**	Pyruvate kinase of *Leishmania mexicana*
Log P	Octanol-water partition coefficient
MCS	Maximum Common Subgraph
MESPEUS	MEtal Sites in Proteins at Edinburgh UniverSity
MLogP	Moriguchi octanol-water partition coefficient
MVC	Model-View-Controller software architecture
MW	Molecular weight
nHAcc	Number of hydrogen bond acceptors

nHDon	Number of hydrogen bond donors
NMR	Nuclear Magnetic Resonance
PCA	Principal Component Analysis
PDB	Protein Data Bank
PEP	Phosphoenolpyruvate
PK	Pyruvate kinase
PSA	Polar surface area
QSAR	Quantitative Structure-Activity Relationship
RBN	Rotatable bonds
RMSD	Root mean square deviation
SAR	Structure-Activity Relationship
SDfiles	Structure-data files
SMILES	Simplified Molecular Input Line Entry System
SOMs	Self Organising Maps
SQL	Structured Query Language
VDW	Van der Waals
VHTS	Virtual high-throughput screening
Vu	V total size index / unweighted WHIM descriptors
W3D	Wiener 3D indices
Whete	Wiener-type index from electronegativity weighted distance matrix topological descriptors
WHIM	Weighted Holistic Invariant Molecular

Table of contents

List of figures

List of tables

1. Introduction

Various biological processes involve ligand-protein interactions, and these interactions play key roles in mediating enzyme catalysis, signal transduction or other protein functions (Schreyer and Blundell 2009). Generally, ligands are defined as all non-protein and non-water molecules, such as nucleic acids, metals, small organic or inorganic ions as well as peptides of up to 20 residues (Hendlich, Bergner et al. 2003), which bind to a target protein. In drug discovery projects, the general intent is to develop a new drug based on the mode of ligand-protein interactions as the new compound (or drug candidate) should specifically bind to the target protein resulting in the disruption of the native ligand-protein interactions and inhibition of the protein's activity. For example, cyclophilins are known essential proteins for the activity of Human Immunodeficiency Virus, HIV (Thali, Bukovsky et al. 1994; Gamble, Vajdos et al. 1996; Luban 1996) and malaria parasites (Bell, Wernli et al. 1994; Silverman 2004), hence the inhibition of these proteins by targeting their ligand binding sites can be a useful strategy for anti-HIV therapy (Billich, Hammerschmid et al. 1995; Bose, Mathur et al. 2003) or anti-malarial (Nickell, Scheibel et al. 1982; Berriman and Fairlamb 1998) drug development.

The starting point of drug discovery is typically from an investigation of ligand binding sites found in a target protein and various selections of structural motifs are then carried out iteratively. The motifs are structurally or chemically similar to known active compounds and are predicted to interact with the binding sites. The process commonly ends with small molecules called hits or leads which are capable of being synthesised and chemically modified and can be used for further

experimental work or biological testing (Bajorath, Klein et al. 1999). Figure 1.1 illustrates the outline of a practical and cost-effective strategy in modern drug discovery (Agrafiotis, Lobanov et al. 2002). It shows that after the genetic engineering required to clone and express the target protein, the process enters an iterative stage to refine the selection of candidate compounds. Those approaches and resources mentioned in Figure 1.1 are divided into three main components in order and are described as:

HTS (high-throughput screening): The first screening shown in Figure 1.1 is the application of HTS (high-throughput screening) to test the drug-like compounds against the preselected target. The tested compounds are derived from a probe library. HTS is a quantitative and universal ligand-protein binding assay. As a drug discovery technique it started at the end of 1980s, up to the late 1990s and has been widely used throughout the pharmaceutical industry (Roberts 2001; Macarron 2006) due to its efficient capability in screening compounds in short order, say five weeks for a million compounds (Landro, Taylor et al. 2000).

Unlike HTS which is experimental, the other theoretical method for rapidly screening compounds is the application of virtual high-throughput screening (VHTS) which has emerged and is growing greatly with the increased computational power available in recent decades (Bajorath 2002; Lee, Choi et al. 2008). It is considered as a complementary and alternative technology to HTS (Alvarez 2004). Several well-known docking programmes are available to perform virtual screening, which primarily predict the conformation and orientation of tested liagnds within a preselected binding site (Anderson 2003; Kitchen, Decornez et al. 2004). These

programmes generally possess the features of time- and cost-saving in drug discovery projects compared with HTS as they effectively eliminate undesired compounds prior to experimental testing.

SAR (structure-activity relationship): The development of SARs (structure-activity relationships) is used in the analysis of binding results given by HTS and helps formulate a set of correlated molecular properties of tested compounds based on the reported quantitative structure-activity data. SAR was initially proposed in the mid-1960s for the study of the correlation between biological activity and chemical structure as well as for the prediction of molecular partition coefficients between octanol and water (Fujita, Iwasa et al. 1964; Hansch and Fujita 1964). The SAR analysis in this step reveals how molecular descriptors of test compounds influence the binding affinity to the target protein, i.e. the interaction mechanisms between molecular properties and binding activity (Frye 1999; Gao, Williams et al. 1999). With the SAR analysis suggesting the preferential molecular properties to be included in new compound synthesis, researchers can select the preferred compounds from a computer database which would ideally store the analogues of the drug-like compounds studied in the earlier step. The selected compounds are then synthesised for subsequent testing to refine the screening criteria or may be used directly in biological testing. Various statistical modelling, clustering or neural network techniques have been applied to aid the SAR analysis (Shi, Fan et al. 1998; Schneider 2000; Yoshida and Topliss 2000).

Other than the SAR analysis integrated with HTS, some molecular properties are seen to be widely used for compound selection as they have been shown to correlate

with bioavailability. For example, the number of rotatable bonds in a compound is a commonly used filter. Previous studies indicate that when the number of rotatable bonds is greater than 10, the rat oral bioavailability tends to be decreased (Veber, Johnson et al. 2002) as well as reducing the ligand affinity to the targeted site on average (Lipinski 2004). The molecular polar surface area, commonly called the PSA, is seen to influence the compound's activity in the central nervous system (CNS) as a value of PSA less than 60-70 $Å^2$ increases the ability of drugs to diffuse through the blood-brain barrier (Kelder, Grootenhuis et al. 1999). A simpler approach to predict the drug transport across the blood-brain barrier based on the molecular properties is that if the sum of the nitrogen and oxygen atom counts in a compound is ≤ 5, it possesses a higher probability to pass through the blood-brain barrier (Norinder and Haeberlein 2002). These molecular properties and of course the molecular descriptors involved in various rule-based screenings (detailed in Chapter 2) have been commonly considered in the early stage of screening.

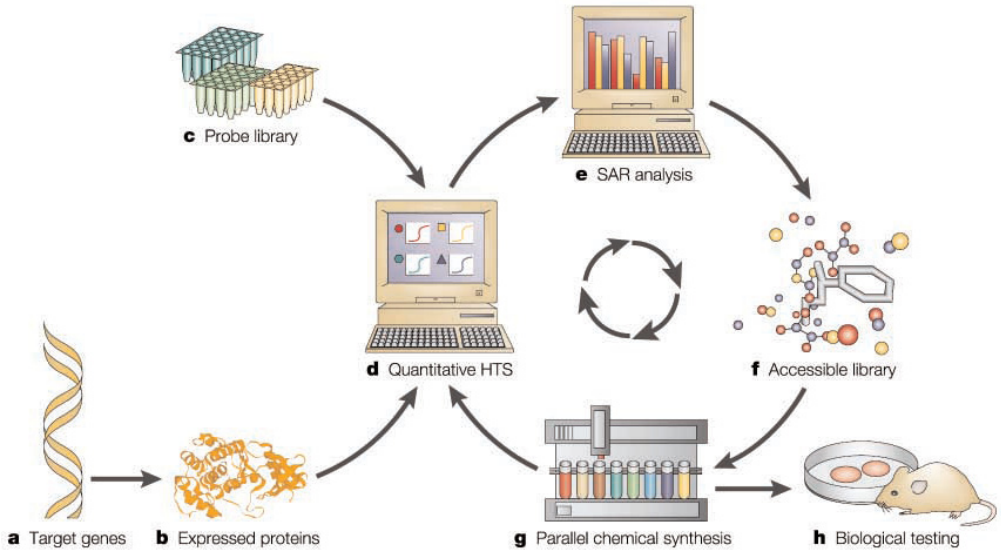
Databases (or libraries): Although the application of HTS in drug discovery has shown an impressive progression, the number of successfully marketed drugs obtained through the screening process is very low. The ratio of increased risk of failure estimated from the initial screening to the end is shown in Figure 1.2 (Oprea 2000; Hann and Oprea 2004). Approximately only one compound can reach the launched status successfully in one million tested compounds (or more) via HTS. Additionally, it on average takes 14 years to bring a compound from identified hits to an approved drug (Myers and Baker 2001).

In order to make overall discovery project more effective, researchers are working not only to optimise the performance of HTS, but also to enhance the functions of the other components shown in Figure 1.1. For example, the databases (or libraries), whether for the storage of gene information, macro- or small-molecule structure, chemical reaction or experimental activity, have been readily established and are now being rapidly enlarged (Malik, Singh et al. 2006; Bharatam, Khanna et al. 2008). Some formats to represent, store or share the above information in computer usable formats are then adopted, such as Crystallographic Information File (CIF) and Protein Data Bank (PDB) file for crystal structures (Brown and McMahon 2002; Berman, Henrick et al. 2007), FASTA for nucleotide sequences or peptide sequences (Pearson and Lipman 1988) as well as Simplified Molecular Input Line Entry System (SMILES), IUPAC International Chemical Identifier (InChI) strings and MDL Structure-Data file (SDfile) for chemical structures (Weininger 1988; Dalby, Nourse et al. 1992). These computer usable formats can also assist researchers to facilitate similarity searches by allowing them to rapidly identifing similar entries already available in databases.

In overall drug discovery, the databases (or libraries) are seen to be involved throughout the process. Therefore, the implementation, manipulation or utilisation of a bioinformatics or chemical database is becoming a crucial aspect in drug discovery. Figure 1.3 gives details of a census that show the number of papers in PubMed containing the keywords “drug discovery” and “database”. It shows that the trend for research using or regarding databases in drug discovery has obviously increased in this decade. As this thesis focuses on the database studies of ligand-protein

interaction, the following sections give a survey of some worldwide databases available for drug / ligand discovery as well as for understanding the interaction of metal atoms in proteins.

Figure 1.1. Outline of a practical strategy in modern drug discovery. (a) and (b) are to obtain the target protein. (c) is a probe library composed of drug-like compounds. (d) is to find active hits via quantitative HTS (high-throughput screening). (e) is to apply SAR (structure-activity relationship) approach based on the binding assay to formulate a set of correlated molecular properties for further compound selection. (f) is to select compounds from a computer database according to the result of previous steps. (g) is to synthesis the candidates for the further screening refinement or to be tested directly in some cellular or biological model systems (h).



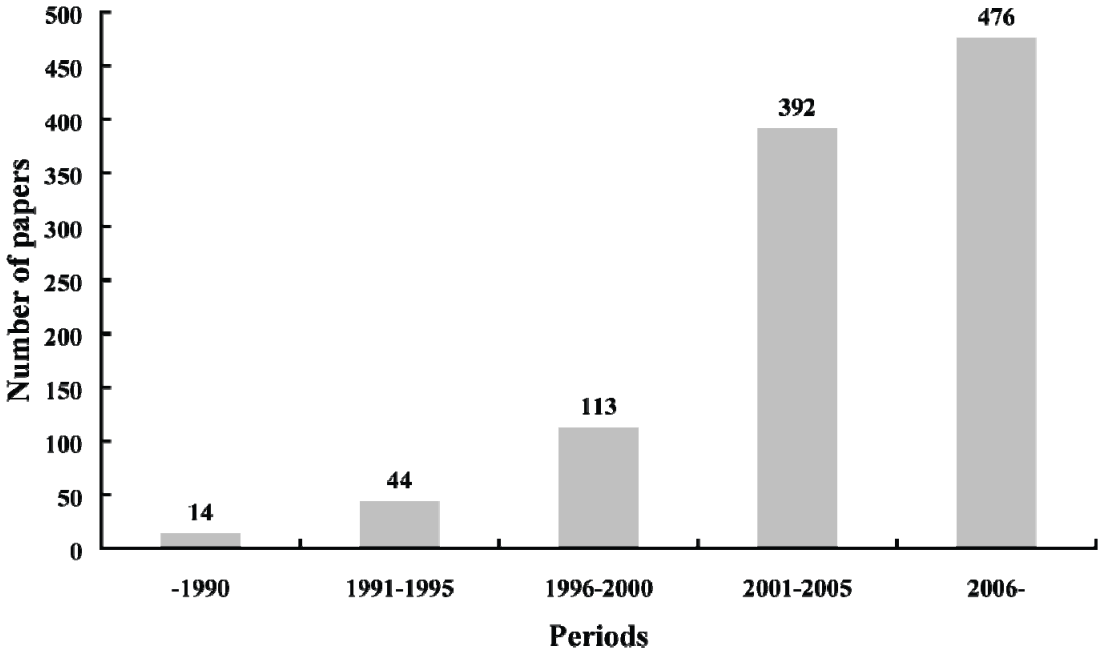
(Agrafiotis, Lobanov et al. 2002)

Figure 1.2. Cascade of a typical drug discovery. HTS hits: the compounds show activity during HTS. HTS actives: the hits are retested to confirm the binding activity and structure. Lead series: the amenable structures for further chemical synthesis. Drug candidates: compounds meet the druggable criteria, including the consideration in toxicity, pharmacokinetic properties and the activity in animal models.

HTS	Increased risk of failure	1 million
HTS hits		2000
HTS actives		1200
Lead series		50-200
Drug candidates		10
Drug		1

(Hann and Oprea 2004)

Figure 1.3. Number of papers found in PubMed from 1984 to present containing the keywords "drug discovery" and "database".



1.1 Small-molecule databases

As shown in Figure 1.1, modern drug discovery iteratively screens small molecules for their ability to bind to a preselected target and/or for their amenability to chemical synthesis. These small molecules are often presented in various formats and stored in publicly accessible databases, sometimes together with their bioassay or chemical reaction data, making the database a powerful tool in data retrieval and management (Schreiber 2000). The fundamental information stored in small-molecule databases normally includes the sketch of atoms and bonds or line notations, such as the string formats SMILES, InChI or fingerprint (Butina 1999), to represent a compound and sometimes includes the 2D or 3D coordinates of all atoms in the structure that allows researchers to see the identity of compounds by using

molecular viewers (Miller 2002) or to be used in structure-based virtual screening (Lyne 2002; Anderson 2003; Ghosh, Nie et al. 2006). On the other hand, as the lack of publicly available datasets of molecules would greatly hamper the development of chemoinformatics (Marris 2005), and so modern chemical databases should also concern their utility allowing users to retrieve the desired compounds. Table 1.1 summarises five publicly available small-molecule databases. They possess distinct features in mining facilities, documentation or data source and are useful for drug or ligand discovery, and are introduced in the following sub-sections.

Table 1.1. Overview of five publicly available small-molecule databases.

Databases	Number of Compounds	Data feature	Structure-based search	Descriptor-based search
ChemBank	+1.2 M (unique)	Bio-assay	Yes	37 items
PubChem	+19 M (unique)	Bio-assay	Yes	12 items[a]
BindingDB	+28,000	Bio-assay	Yes	Molecular weight
DrugBank	4,886	Drugs[c]	Yes	Molecular weight[b]
IBM Chemical Search Engine	+4.1 M	Patent compounds[d]	Yes	No

[a]: The function needs to integrate with structure-based search.
[b]: The function is able to integrate with molecular activity filter.
[c]: FDA-approved and Experimental drugs.
[d]: The compounds of U.S. Patents and Applications.

1.1.1 ChemBank (http://chembank.broad.harvard.edu/)

This is a small-molecule database with the data of about 2,500 high-throughput biological assays results (Seiler, George et al. 2008). There are more than 1.7

million compounds deposited in this database and 1.2 million of them are unique. This database also stores over 300 calculated molecular descriptors for each compound. Users can find the information for small molecules, assays or targets currently resident in ChemBank via its web-based interface. The interface provides not only the conventional tools allowing users to retrieve compounds, such as using substructure, similarity or descriptor search, but also offers the ability for users to search by bioassays, known function (e.g. biological process or therapeutic use) and the name of molecules or chemists (or vendors) who synthesised the compounds.

ChemBank shows an impressive performance in responding to a query and displays the result on the web pages very quickly even when the query was formulated to select all compounds extant in the database, e.g. searching for molecular weight greater than 0. The result page gives the hit compound's primary name, chemical structure, biological annotations, calculated molecular descriptor values and the substances ID, if available, referred to PubChem (http://pubchem.ncbi.nlm.nih.gov/) as well as the information of relevant HTS screening instances. The statistical data of bioassay outcomes can be presented in a histogram format to perform SAR analyses. It also provides hyperlinks to search Google for specified molecule which gives more information concerning the compound.

1.1.2 PubChem (http://pubchem.ncbi.nlm.nih.gov/)

This is a very well-known resource for efficiently identifying the biological activity information of small molecules (Fontaine, Bolton et al. 2007; Han, Wang et al. 2008; Sayers, Barrett et al. 2008), belonging to the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH; http://www.nih.gov/).

It stores more than 19 million non-redundant compounds provided by about 70 various organisations and 28 million bioassay test outcomes in 1,000 bioassays. As it represents a vast deposit, it has been widely adopted in relevant research projects such that the number of papers found in PubMed from 2005 containing the keyword "PubChem" is over 60. The web interface of the PubChem database provides a set of very complete query tools for users and has integrated its three sub-databases, i.e. PubChem Substance, PubChem Compound, and PubChem BioAssay respectively. The integration allows users to have a comprehensive overview and the query result pages also give associations between biological testing results for identical compounds. The PubChem Substance database stores the descriptions of chemical samples provided by depositors. PubChem Compound contains the unique compound structure content of PubChem Substance. PubChem BioAssay deposits the biological activity testing results from various sources. The web interface provides a set of built-in analysis facilities, including the functions "BioActivity Summary", "Structure-Activity Analysis", and "Structure Clustering" which can be found in the query result pages to perform a basic SAR analysis.

As well as the molecular structure files, the bioassay results and their descriptions are downloadable from the web interface and the experimental protocols of the bioassays are described on the page. All PubChem data is available on the PubChem FTP server and users can connect to the site freely to download the entire deposit.

1.1.3 BindingDB (http://www.bindingdb.org/)

This database holds over 28,000 compounds which have had their binding affinities in ligand-protein complexes experimentally determined (Liu, Lin et al. 2007). At

present, over 56,000 measurements of enzyme inhibition constant or isothermal titration calorimetry (ITC) experiments have been carried out against 591 preselected targets which are mainly drug-targets or candidate drug-targets whose structural data have been deposited in the PDB.

To retrieve the information from this database, users can set query criteria based on the basic features of the small compounds or the targets, such as a structure-based or similarity search, molecular weight, target sequence or their names. Alternatively, a set of activity filters, such as asking the binding affinity in a specified range of IC50 (nM) or ΔG° (kcal/mol), are available via its web-based interface. For example, the query result page shows 1,630 observations when the given question was the measured IC50 (nM) between 2 and 5 in all of the available assays. The information of various affinity measurements is then presented together with the hyperlinks connected to other databases for easy navigation, such as PDB, NCBI's MMDB (Molecular Modeling DataBase), PubChem or PubMed, in which the details of the tested lignad or target can be found.

The BindingDB website also provides a capability to perform non-structure-based virtual screening. For example, users can upload their interesting compounds to the website, a built-in piece of software will then calculate the Tanimoto similarity between each uploaded compound and each active compound found in BindingDB database based on JChem chemical fingerprints (Csizmadia 2000), and rank the uploaded compounds according to the calculated maximal similarity to any active compound. The drawback of this service is that users cannot know the maximal similarity to which active compound from the information shown on the page.

1.1.4 DrugBank (http://www.drugbank.ca/)

This database stores the information of known drugs and their target proteins (Wishart, Knox et al. 2006; Wishart, Knox et al. 2007). At present, it holds 4,886 drug chemical structures with over 2,500 non-redundant sequences of targets, including proteins or DNA. The drug collection contains more than 1,350 small molecule drugs approved by FDA (U.S. Food and Drug Administration; http://www.fda.gov/), 3,000 experimental drugs, some biotech drugs (a protein or peptide produced using living organisms such as yeast or bacteria) and nutraceuticals (nutritional supplements).

The general molecular properties of the compounds deposited in DrugBank are listed in Table 1.2. In this database, about 79 % of the compounds fit the Lipinski's rule of five (Lipinski, Lombardo et al. 1997) and overall compounds are rather diverse indicated by the wide standard deviations observed in each molecular property. For instance, two approved drugs that possess extreme values in some molecular properties are inulin (a hypoglycemic agent) and selenium sulfide (antifungals for topical use) and their chemical structures are shown in Figure 1.4 (a) and (b), respectively.

The web-based interface of DrugBank provides a set of built-in tools for searching drugs, viewing and sorting the query results as well as for downloading small molecular structure data, target sequence (protein or gene sequence in FASTA format) and extracting text or images. A unique feature of DrugBank is that it organises a very detailed report, called DrugCard, for each drug entry. A DrugCard contains more than 100 data fields listing the information of chemical,

pharmacological, pharmacogenomic and molecular biological data. The DrugCard also gives extensive hyperlinks to some useful bioinformatics, biomedical or pharmaceutical databases to show the expanded information of the specified drug. These expanded resources include KEGG, a database for understanding functions and utilities of the biological system (Kanehisa, Goto et al. 2006); ChEBI, a database for showing the chemical entities in ontology (Degtyarenko, Matos et al. 2007); RxList, a internet drug index (Hatfield, May et al. 1999); PharmGKB, a clinically oriented drug database for knowing the impact of human genetic variations on drug response (Hodge, Altman et al. 2007); PDB, PubChem, PubMed and so on.

Table 1.2. General molecular properties of the compounds deposited in DrugBank.

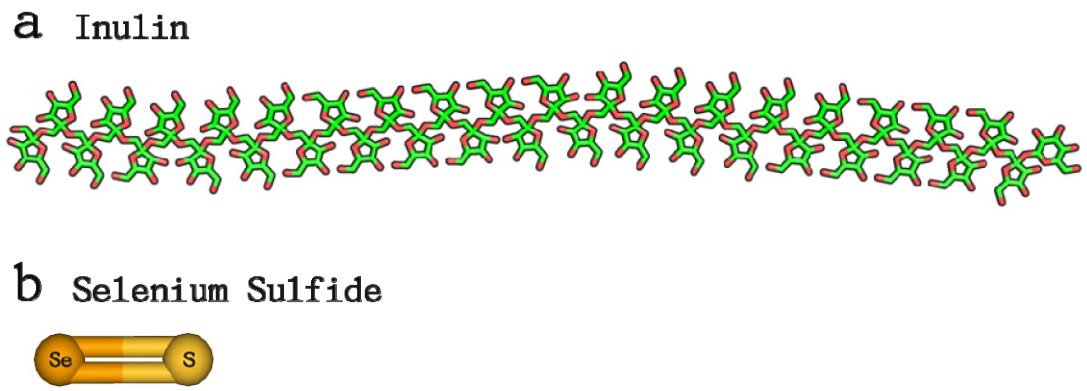
Descriptor	Max	Min	Average	Standard deviation
Molecule weight	6180.06*	46.08	392.15	346.72
Number of atoms	801*	2+	52.38	47.79
Number of bonds	838*	1+	54.17	48.96
Number of rings	38*	0	2.78	1.93
Number of HAcc	191*	0	6.90	9.13
Number of HDon	116*	0	3.11	5.96
Number of rotatable bonds	170	0	6.42	10.01
MLogP#	11.73	-68.56*	1.54	3.15

*: The compound that possesses maximum values in these observations is Inulin (CAS: 9005-80-5) and its structure is shown as Figure 1.4 (a) in 2D.

+: The compound that possesses minimum values in the number of atoms and bonds is Selenium Sulfide (CAS: 7446-34-6) and its structure is shown as Figure 1.4 (b) in 2D.

#: Calculated octanol / water partition coefficient (Moriguchi, Hirono et al. 1992).

Figure 1.4. Structures of Inulin (a) and Selenium Sulfide (b) in 2D which both are approved drugs and possess extreme values in some molecular properties observed in DrugBank.



1.1.5 IBM Chemical Search Engine (https://chemsearch.almaden.ibm.com/)

This is a chemical patent query system with a web-based interface (alpha site) allowing users to explore chemical structures from the corpus of U.S. Patents (1976 to 2005) and Patent Applications (2003 to 2005). The database contains more than 4.1 million unique compounds (Rhodes, Boyer et al. 2006). Considering the need of drug discovery, this system has identified the chemical terms for each patent entry and converted them into structures, so that it provides a patent mining facility using molecular similarity search instead of the traditional text mining, e.g. search by the name of molecules or the patent ID. Other than SMILES string, the similarity search engine is mainly based on IUPAC International Chemical Identifier (InChI) standard for the calculation of similarity.

The molecule name and patent information of hit compounds are present on the query result page. The interface also provides a molecular network function to graphically exhibit the association between the similar compounds and their patents. For example, users can select 10 hit compounds and then the network of their patents

and applicants can be displayed on a single figure. The drawback of this system is that it is still under development. The molecular search function is out of order constantly and it can only be performed by drawing a query compound or inputting a SMILES / InChi string but further adding more restricted criteria for query, such as molecular property filters.

1.2 Metalloprotein databases

Another area of ligand-protein interactions considered in this thesis is the metal-binding sites in metalloproteins. Metal atoms play a broad variety of roles in biology, such as being the enzyme cofactors or involved in catalytic active site construction and protein structure maintenance (Cox and McLendon 2000; Wittung-Stafshede 2002). The following describes some resources which provide information on metal-binding sites to aid the understanding of metalloproteins.

1.2.1 MDB (http://metallo.scripps.edu/)

MDB, the Metalloprotein Database and Browser, is a web-accessible database for metalloprotein research (Castagnetto, Hennessy et al. 2002). It collects and extracts the geometry information of more than 32,000 metal-binding sites from 8,069 structures existing at PDB. From the viewpoint of a metal-binding site, the ligands conventionally mean the residues, molecules or charged ions that interact with the metal atom (Yamashita, Wesson et al. 1990). MDB allows users to search the metal-binding sites for 37 different metal atoms which are found to be bound with proteins via the first-shell ligands (i.e. ligands bind these metals) or the second-shell ligands, (i.e. residues contact the metal-binding ligands).

The ligand types found in MDB include amino acids, nucleic acids, metal, water, anion and hetero atoms. The longest metal-ligand (first-shell) distance found in MDB is 3.5 Å. Other than specifying metal atoms, users can also select some crystallographic filters to restrict the query, such as the number of donors, structure resolution (Å), *R*-value, metal-ligand distances etc. The interface provides an interactive 3D viewer and on-line histogram plotting tool to display the geometry of hit metal sites and the statistical distribution of the metal-ligand distances, respectively. The database can be accessed using an SQL (Structured Query Language) statement directly via a given page for a more flexible query but the users would need to be familiar with the SQL language.

This resource is a useful tool for studying metal site geometries and properties. This data is crucial for the construction of new metal binding sites into a known protein scaffold or for the refinement of exiting metal sites. However, new material does not appear to have been added to this database since 2004 and the query results are not downloadable for further analysis.

1.2.2 JenaLib (http://www.imb-jena.de/IMAGE.html)

JenaLib, Jena Library of Biological Macromolecules, is an image library of biopolymer structures deposited in PDB and NDB (Nucleic Acid Database; http://ndbserver.rutgers.edu/) (Berman, Olson et al. 1992) with emphasis on visualisation and analysis (Reichert and Suhnel 2002). Two sub-databases in the JenaLib system, called Hetero Components Database and Site Database, provide the fundamental descriptions of biopolymer structures containing hetero molecules or atoms. All residues are taken into account to be the components of a site, which

have at least one atom located within a distance of 4.2 Å from a specific hetero atom or molecule.

As metal atoms are defined as hetero atoms according to the format of a PDB file, users can find the information of metallopteins from its two sub-databases. From the web-based interface of Hetero Components Database, users can simply choose a metal atom via the provided interactive periodic table of elements and the page will return the hit structures which contain the selected metal. The pre-generated static images of the hit structures are available for structural observation which emphasise the location of selected metals occurring in the structure. Alternatively, users can choose the dynamical molecule viewer (Jmol) provided to explore and manipulate the structures in 3D.

The design of JenaLib is not exactly useful for metal site investigations, as it only lists some fundamental descriptions of the biopolymer structure, such as the title in PDB file, protein class, structure resolution or relevant references, and provides their images. It does not detail the geometry of metal sites. The number of available options for a query is also limited. Users can merely specify the hetero ID or name presented in the PDB file, the metal name or search for the occurrence of any components in the environment of hetero components. Only one single option can be selected in a query.

1.2.3 PROMIS (http://metallo.scripps.edu/PROMISE/)

PROMIS, Prosthetic Groups and Metal Ions in Protein Active Sites (version 2.0), is a database focused on the relationships between protein and metal ions and prosthetic

groups (Degtyarenko, North et al. 1998) found in protein's active sites. Prosthetic groups are defined as the non-amino acid portion conjugated to a protein, such as cofactors (Nagel, Dellweg et al. 1992).

For the study of metal sites in proteins, the web-based interface of this database lists the information of known chlorophyll-containing proteins (i.e. proteins in photosynthetic organisms that contain magnesium), copper proteins, haem proteins, iron-sulphur proteins etc. These proteins have been classified by their functions, classes or the centre types of the metal in the coordination (e.g. mononuclear and binuclear centre). In order to illustrate the amino acid-metal coordination, the figures of various coordination groups are pre-generated by the MOLSCRIPT programme (Kraulis 1991). The pages also list the associated bibliography for each mentioned protein, more than 2,000 references in total.

One drawback of this resource is that it has not been updated for a long time. It was last modified 1 March 1999. Besides, its web-based interface is static without the functions for dynamic query or structural display.

1.3 Project aims

This project aims to collect the chemical structures of the vast number of commercially purchasable small molecules and to construct an accessible database with a web-based interface to maintain and utilise the data efficiently. The generation and application of the molecular properties, i.e. the descriptors, for each deposited compound have been included in this project. The approach of developing a Structure-Activity Relationship (SAR) is then used in the ligand-protein binding

study using the created molecular descriptors and the known bio-assay outcomes, in order to discover the correlation between compound activity and molecular descriptors. A structure-based mining of pharmacophore searching approach is also designed and implemented using the deposited chemical structures as material for ligand discovery. Additionally, this project involves a study dealing with the interaction of different metals with proteins. Similarly, a database system has been considered to store the geometrical information of the metal sites derived from metalloprotein structures.

1.4 Reference:

Agrafiotis, D. K., V. S. Lobanov, et al. (2002). "Combinatorial informatics in the post-genomics era." Nature Reviews Drug Discovery **1**(5): 337-346.

Alvarez, J. C. (2004). "High-throughput docking as a source of novel drug leads." Current Opinion in Chemical Biology **8**(4): 365-370.

Anderson, A. C. (2003). "The process of structure-based drug design." Chemistry & Biology **10**(9): 787-797.

Bajorath, J. (2002). "Integration of virtual and high-throughput screening." Nature Reviews Drug Discovery **1**(11): 882-894.

Bajorath, J., T. E. Klein, et al. (1999). "Computer-aided drug discovery: from target proteins to drug candidates." Pacific Symposium on Biocomputing **4**: 413-414.

Bell, A., B. Wernli, et al. (1994). "Roles of peptidyl-prolyl cis-trans isomerase and calcineurin in the mechanisms of antimalarial action of cyclosporin A, FK506, and rapamycin." Biochemical Pharmacology **48**(3): 495.

Berman, H., K. Henrick, et al. (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data." Nucleic Acids Research **35**(Database issue): D301.

Berman, H. M., W. K. Olson, et al. (1992). "The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids." Biophysical Journal **63**(3): 751-759.

Berriman, M. and A. H. Fairlamb (1998). "Detailed characterization of a cyclophilin from the human malaria parasite Plasmodium falciparum." Biochemical Journal **334**(Pt 2): 437.

Bharatam, P. V., S. Khanna, et al. (2008). "Modeling and informatics in drug design." Prelinical Development Handbook: ADME and Biopharmaceutical Propertie: 25-28.

Billich, A., F. Hammerschmid, et al. (1995). "Mode of action of SDZ NIM 811, a nonimmunosuppressive cyclosporin A analog with activity against human immunodeficiency virus (HIV) type 1: interference with HIV protein-cyclophilin A interactions." Journal of Virology **69**(4): 2451-2461.

Bose, S., M. Mathur, et al. (2003). "Requirement for cyclophilin A for the replication of vesicular stomatitis virus New Jersey serotype". Journal of General Virology. **84:** 1687-1699.

Brown, I. D. and B. McMahon (2002). "CIF: the computer language of crystallography." Acta Crystallographica Section B: Structural Science **58**(3): 317-324.

Butina, D. (1999). "Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large

data sets." Journal of Chemical Information and Computer Sciences **39**(4): 747-750.

Castagnetto, J. M., S. W. Hennessy, et al. (2002). "MDB: the metalloprotein database and browser at the Scripps Research Institute." Nucleic Acids Research **30**(1): 379.

Cox, E. H. and G. L. McLendon (2000). "Zinc-dependent protein folding." Current Opinion in Chemical Biology **4**(2): 162-165.

Csizmadia, F. (2000). "JChem: Java applets and modules supporting chemical database handling from web browsers." Journal of Chemical Information and Computer Sciences **40**(2): 323-324.

Dalby, A., J. G. Nourse, et al. (1992). "Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited." Journal of Chemical Information and Computer Sciences **32**(3): 244-255.

Degtyarenko, K., P. Matos, et al. (2007). "ChEBI: a database and ontology for chemical entities of biological interest." Nucleic Acids Research: D344.

Degtyarenko, K. N., A. C. T. North, et al. (1998). "PROMISE: a database of information on prosthetic centres and metal ions in protein active sites." Nucleic Acids Research **26**(1): 376.

Fontaine, F., E. Bolton, et al. (2007). "Fast 3D shape screening of large chemical databases through alignment-recycling." Chemistry Central Journal **1**: 12.

Frye, S. V. (1999). "Structure-activity relationship homology (SARAH): a conceptual framework for drug discovery in the genomic era." Chemistry & Biology **6**(1): 3-7.

Fujita, T., J. Iwasa, et al. (1964). "A new substituent constant, p, derived from partition coefficients." Journal of the American Chemical Society **86**(23): 5175-5180.

Gamble, T. R., F. F. Vajdos, et al. (1996). "Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid." Cell **87**(7): 1285-1294.

Gao, H., C. Williams, et al. (1999). "Binary Quantitative Structure- Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands." Journal of Chemical Information and Computer Sciences **39**(1): 164-168.

Ghosh, S., A. Nie, et al. (2006). "Structure-based virtual screening of chemical libraries for drug discovery." Current Opinion in Chemical Biology **10**(3): 194-202.

Han, L., Y. Wang, et al. (2008). "Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem." BMC Bioinformatics **9**(1): 401.

Hann, M. M. and T. I. Oprea (2004). "Pursuing the leadlikeness concept in pharmaceutical research." Current Opinion in Chemical Biology **8**(3): 255-263.

Hansch, C. and T. Fujita (1964). "p-σ-π analysis. A method for the correlation of biological activity and chemical structure." Journal of the American Chemical Society **86**(8): 1616-1626.

Hatfield, C. L., S. K. May, et al. (1999). "Quality of consumer drug information provided by four Web sites." American Journal of Health-System Pharmacy **56**(22): 2308-2311.

Hendlich, M., A. Bergner, et al. (2003). "Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions." Journal of Molecular Biology **326**(2): 607-620.

Hodge, A. E., R. B. Altman, et al. (2007). "The PharmGKB: integration, aggregation, and annotation of pharmacogenomic data and knowledge." Clinical Pharmacology & Therapeutics **81**(1): 21-24.

Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Research **34**(Database Issue): D354.

Kelder, J., P. D. J. Grootenhuis, et al. (1999). "Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs." Pharmaceutical Research **16**(10): 1514-1519.

Kitchen, D. B., H. Decornez, et al. (2004). "Docking and scoring in virtual screening for drug discovery: methods and applications." Nature Reviews Drug Discovery **3**(11): 935-949.

Kraulis, P. J. (1991). "MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures." Journal of Applied Crystallography **24**(5): 946-950.

Landro, J. A., I. C. A. Taylor, et al. (2000). "HTS in the new millennium The role of pharmacology and flexibility." Journal of Pharmacological and Toxicological Methods **44**(1): 273-289.

Lee, H. S., J. Choi, et al. (2008). "Optimization of High Throughput Virtual Screening by Combining Shape-Matching and Docking Methods." Journal of Chemical Information and Modeling **48**(3): 489-497.

Lipinski, C. A. (2004). "Lead-and drug-like compounds: the rule-of-five revolution." Drug Discovery Today: Technologies **1**(4): 337-341.

Lipinski, C. A., F. Lombardo, et al. (1997). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." Advanced Drug Delivery Reviews **23**(1-3): 3-25.

Liu, T., Y. Lin, et al. (2007). "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities." Nucleic Acids Research **35**(Database issue): D198.

Luban, J. (1996). "Absconding with the Chaperone: Essential; Cyclophilin-Gag Interaction in HIV-1 Virions." Cell **87**(7): 1157-1160.

Lyne, P. D. (2002). "Structure-based virtual screening: an overview." Drug Discovery Today **7**(20): 1047-1055.

Ma, D., X. Hong, et al. (1996). "A cyclosporin A-sensitive small molecular weight cyclophilin of filarial parasites." Molecular & Biochemical Parasitology **79**(2): 235-241.

Macarron, R. (2006). "Critical review of the role of HTS in drug discovery." Drug Discovery Today **11**(7-8): 277-279.

Malik, A., H. Singh, et al. (2006). "Databases and QSAR for Cancer Research." Cancer Informatics **2**: 99-111.

Marris, E. (2005). "Chemistry society goes head to deal with NIH in fight over public database." Nature **435**: 718-719.

Miller, M. A. (2002). "Chemical database techniques in drug discovery." Nature Reviews Drug Discovery **1**(3): 220-227.

Moriguchi, I., S. Hirono, et al. (1992). "Simple method of calculating octanol/water partition coefficient." Chemical and Pharmaceutical Bulletin **40**(1): 127-130.

Myers, S. and A. Baker (2001). "Drug discovery-An operating model for a new era." Nature Biotechnology **19**(8): 727-730.

Nagel, B., H. Dellweg, et al. (1992). "Glossary for chemists of terms used in biotechnology." Pure and Applied Chemistry **64**: 143-168.

Nickell, S. P., L. W. Scheibel, et al. (1982). "Inhibition by cyclosporin A of rodent malaria in vivo and human malaria in vitro." Infection and Immunity **37**(3): 1093-1100.

Norinder, U. and M. Haeberlein (2002). "Computational approaches to the prediction of the blood-rain distribution." Advanced Drug Delivery Reviews **54**(3): 291-313.

Oprea, T. I. (2000). "Current trends in lead discovery: Are we looking for the appropriate properties?" Molecular Diversity **5**(4): 199-208.

Pearson, W. R. and D. J. Lipman (1988). "Improved tools for biological sequence comparison." Proceedings of the National Academy of Sciences **85**(8): 2444-2448.

Reichert, J. and J. Suhnel (2002). "The IMB jena image library of biological macromolecules: 2002 update." Nucleic Acids Research **30**(1): 253.

Rhodes, J., S. Boyer, et al. (2006). Mining patents using molecular similarity search, World Scientific Pub Co Inc P:304.

Roberts, S. A. (2001). "High-throughput screening approaches for investigating drug metabolism and pharmacokinetics." Xenobiotica **31**(8-9): 557-589.

Sayers, E. W., T. Barrett, et al. (2008). "Database resources of the National Center for Biotechnology Information." Nucleic Acids Research: D5.

Schneider, G. (2000). "Neural networks are useful tools for drug design." Neural Networks **13**(1): 15-16.

Schreiber, S. L. (2000). "Target-oriented and diversity-oriented organic synthesis in drug discovery." Science **287**(5460): 1964.

Schreyer, A. and T. Blundell (2009). "CREDO: A protein-ligand interaction database for drug discovery." Chemical Biology & Drug Design **73**(2): 157-167.

Seiler, K. P., G. A. George, et al. (2008). "ChemBank: a small-molecule screening and cheminformatics resource database." Nucleic Acids Research **36**(Database issue): D351.

Shi, L. M., Y. Fan, et al. (1998). "Mining the NCI anticancer drug discovery databases: genetic function approximation for the QSAR study of anticancer ellipticine analogues." Journal of Chemical Information and Computer Sciences **38**(2): 189-199.

Thali, M., A. Bukovsky, et al. (1994). "Functional association of cyclophilin A with HIV-1 virions." Nature **372**(6504): 363-365.

Veber, D. F., S. R. Johnson, et al. (2002). "Molecular properties that influence the oral bioavailability of drug candidates." Journal of Medicinal Chemistry **45**(12): 2615-2623.

Weininger, D. (1988). "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules." Journal of Chemical Information and Computer Sciences **28**(1): 31-36.

Wishart, D. S., C. Knox, et al. (2007). "DrugBank: a knowledgebase for drugs, drug actions and drug targets." Nucleic Acids Research.

Wishart, D. S., C. Knox, et al. (2006). "DrugBank: a comprehensive resource for in silico drug discovery and exploration." Nucleic Acids Research **34**(Database Issue): D668.

Wittung-Stafshede, P. (2002). "Role of cofactors in protein folding." Accounts of Chemical Research **35**(4): 201-208.

Yamashita, M. M., L. Wesson, et al. (1990). "Where metal ions bind in proteins." Proceedings of the National Academy of Sciences **87**(15): 5648-5652.

Yoshida, F. and J. G. Topliss (2000). "QSAR model for drug human oral bioavailability." Journal of Medicinal Chemistry **43**(13): 2575.

2. Development of small-molecule database, EDULISS 2.0

2.1 Introduction

EDinburgh **U**niversity **L**igand **S**election **S**ystem (EDULISS) is a relational database system which holds the physicochemical properties of compounds. The initial version was created by Dr. Andrew Hinton of the Institute of Structural and Molecular Biology around 2002 to 2005. Version 1.0 of the database contained 16 prominent supplier chemical catalogues, +1.5 million compounds and over 1,500 different descriptor items per compound.

Since the suppliers are constantly synthesising novel compounds and building-blocks, the number of commercially purchasable compounds has considerably increased. Practise in this area varies, however many suppliers update their catalogues regularly, adding new compounds, removing unavailable compounds or altering the compound structure and its structure-data file (SDfile). Table 2.1 shows the growth and decline between the various versions of several catalogues which display the largest variation between editions. Furthermore, in some cases the SDfiles might be altered by vendors, such as the location of hydrogen or mineral atoms and modification to the bonding information (Figure 2.1). For the reasons mentioned above, it is necessary to update EDULISS. Unfortunately, the original EDULISS (version 1.0) only stored the supplier's catalogue location details on the file system (that is the paths of files) with their index of string position of every compound entry in database rather than the whole SDfiles. This considerably increases the difficulty in updating the SDfiles and EDULISS as the start-offset string position of every compound entry will be changed even when a single entry in the SDfile has been altered. Hence, a newly

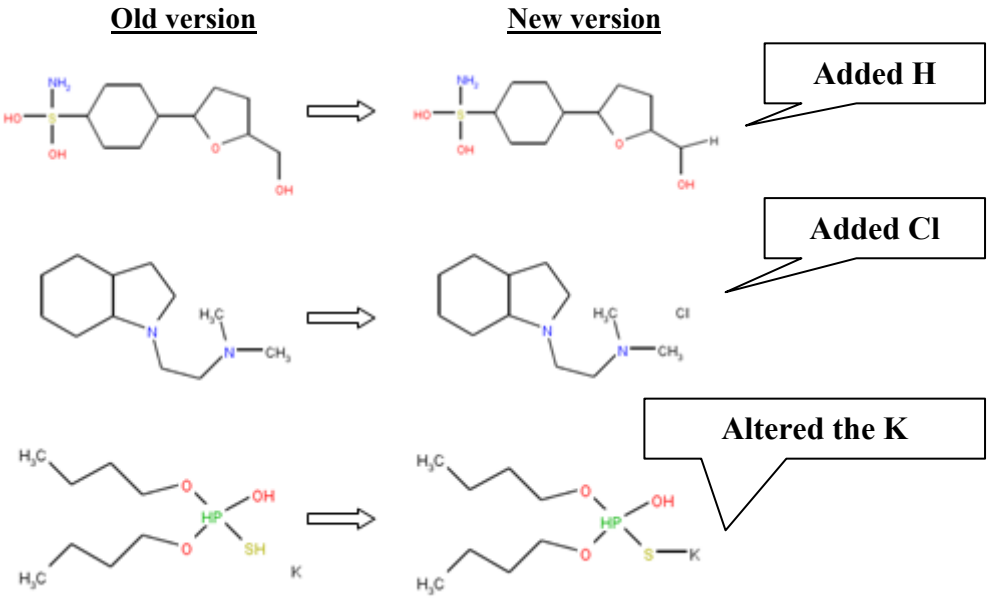
designed and updateable EDULISS database is required. Additionally, in order to provide the drug discovery community use of the resource in Edinburgh and eventually throughout the world, a web-based interface for EDULISS is also needed. The following sections present the procedure of creation and manipulation of EDULISS 2.0 with some demonstrations.

Table 2.1. Changes observed between the different versions of catalogues

Suppliers	Old Version	New Version	New Compounds	Unavailable Compounds
Asinex	96,073	130,012	55,068	21,477
IBS	140,602	360,211	252,631	33,021
ChemBridge	180,651	433,682	281,311	28,280
Iflab	100,583	147,682	80,368	33,269
Specs	232,028	196,586	1,901	37,343
Maybridge	59,676	58,855	252	1,075

Asinex: http://www.asinex.com/; IBS: http://www.ibscreen.com/; ChemBridge: http://chembridge.com/chembridge/; Iflab: http://www.lifechemicals.com/; Specs: http://www.specs.net/; Maybridge: http://www.maybridge.com/default.aspx.

Figure 2.1. Example cases of altered compounds between different versions of catalogues.

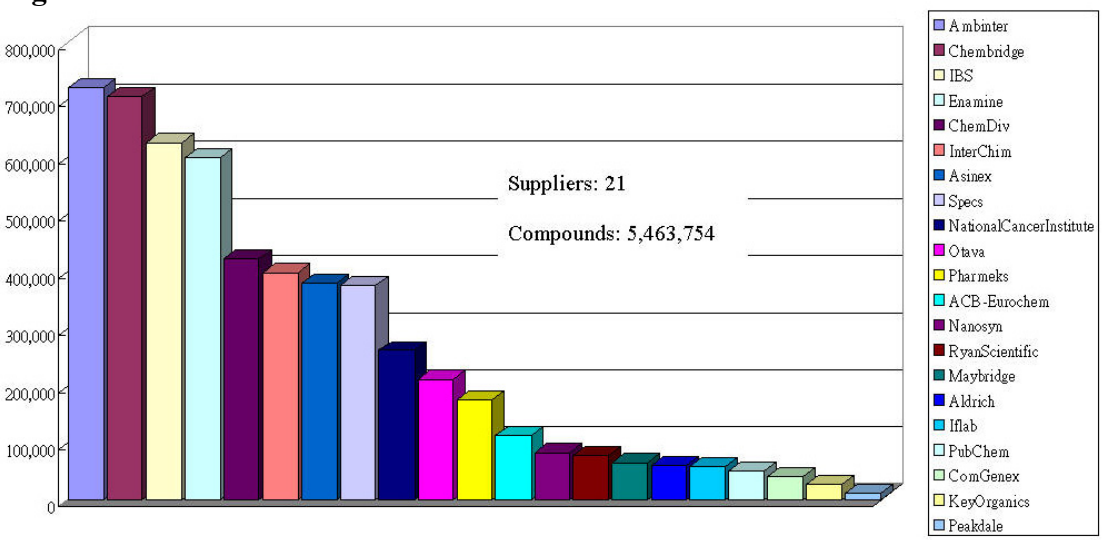


2.2 Materials and Methods

2.2.1 Collection of compound structure-data files, SDfiles

The collection of compound structure-data files for EDULISS 2.0 can be separated into two parts: 1. Collections of the new version catalogues from original (version 1.0) suppliers and additional new suppliers and 2. Collections from Zinc-Database which provides 3-dimensional, pre-selected compounds in SDfile, SMILES, and Flexi-Dock formats (Irwin & Shoichet, 2005). Compounds in the Zinc-Database must be less than 700 Da, have XLogP values of –4 to 6, and be able to be converted into 3D. Users can use simple Lipinski filters or download discreet subsets of compounds for selecting compounds that are defined as lead-like, drug-like, fragment-like and lipophilic. The total of available compounds within the Zinc-Database was 5,463,754 at the time of collection for EDULISS 2.0 (Figure 2.2).

Figure 2.2. Collection in order of found within Zinc database.



2.2.2 The treatment of SDfiles

Regardless of the source of catalogues, the compounds used for EDULISS 2.0 were entirely collected as 2-dimensional (2D) SDfile formats. Different suppliers have used differing programmes for the generation of 3-dimensional (3D) atomic coordinates. Differing programmes use alternative parameters for the calculation of energy minima, bond lengths and bond angles for the 2D to 3D conversion. Therefore, an identical 2D compound could be built with widely diverse geometry by different programmes, e.g. differing geometrical conformation and distances between each atom. For instance, the sums of interatomic distances (hydrogen included) in a benzene could be 181.774, 183.324 and 228.558 Å depending on if the 2D to 3D conversions were performed by CORINA (Sadowski and Gasteiger 1993), web-based service provided by the company of Molecular Networks GmbH (http://www.molecular-networks.com), CONCORD (discussed later) or the Chemistry Development Kit (CDK), an open source Java library for bio- and chemoinformatics (Steinbeck, Han et al. 2003), respectively. To provide consistent structure-based screening and molecular geometrical property calculation for each compound, the conversion must be treated using a single programme throughout. A well-established 3D generation software is CONCORD^{TM} (Hendrickson, Nicklaus et al. 1993) created by Pearlman. The process of generating the 3D structure of compounds in SDfile format includes two steps in which an intermediate 3D–SYBYL mol2 format is required. CONCORD will not add implicit hydrogen atoms when both input and output files are in SDfile format. The mol2 format contains implicit hydrogen that could be further transformed by CONCORD into the SDfile

format. Since EDULISS 1.0 had processed the 2D to 3D conversion by CONCORD and performed with an acceptable conversion time (0.02 seconds per compound), EDULISS 2.0 consistently utilised it for 3D generation. As EDULISS 2.0 aims to store as many compounds as possible for further various applications, the number of non-hydrogen atoms, rotatable bonds and atoms per ring had been set up as the allowed maximums of CONCORD that were 250, 400 and 24 respectively. Figure 2.3 presents the schema for the collection and conversion process of EDULISS 2.0.

After the conversion process, all of the 3D catalogues were assigned an in-house ID number to represent the suppliers uniquely, named Supplier_ID. Furthermore, the other in-house ID was given for each compound, known for obscure historical reasons, as the SPH_Number. The number format is "SPH1-XXX-XXX". When the Supplier_ID is combined with a SPH_Number, it forms a unique identification code for each compound. For example, if the Supplier_ID of the Sigma catalogue is 25 and the SPH_Number of a compound in this catalogue is SPH1-000-001, then the unique identification code is "25SPH1-000-001". This code not only plays an important role in representing a compound but is also a unique key to represent an entry throughout the EDULISS 2.0 database tables. On the other hand, catalogues always contain varied tags in the annotation block of SDfiles. For instance, the tag of compound ID given by Sigma itself is "> <CAT_NO>" compared with the tag "> <code>" in the MayBridge catalogue. These irregular tags will interfere with the function of relevant EDULISS 2.0 scripts and cause problem during further data-mining operations. Three regular tags, therefore, have been added into annotation block for each compound in SDfiles, including the tags of SPH number, the name of

the supplier and the original compound ID (given by the supplier). The name of the supplier is referred to the name of database table that stored the 2D and 3D SDfiles. The added tags provide helpful patterns for a programmer to perform regular expression searches. Figure 2.4 is a typical 3D SDfile format after treatment.

The final numbers of successfully converted compounds of each supplier are shown in Figure 2.5. The number of the compounds offered by the suppliers is significantly different from the largest library, TimTec (http://www.timtec.net/), containing 828,288 compounds to the smallest, ontochem (http://www.ontochem.de/), containing only 703.

Figure 2.3. Schema of collection and 2D to 3D converstion process of compound structure-data files (SDfiles) for EDULISS 2.0.

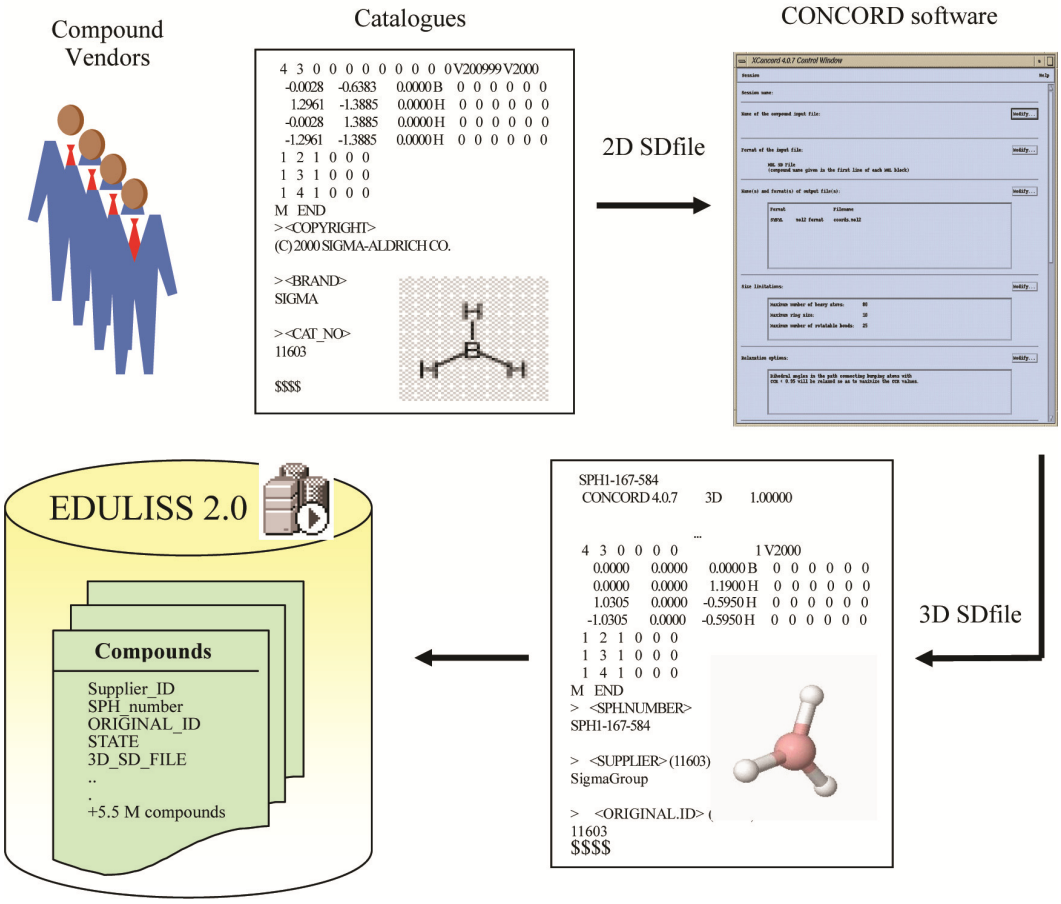


Figure 2.4. Typical compound represented as SDfile format in EDULISS 2.0.

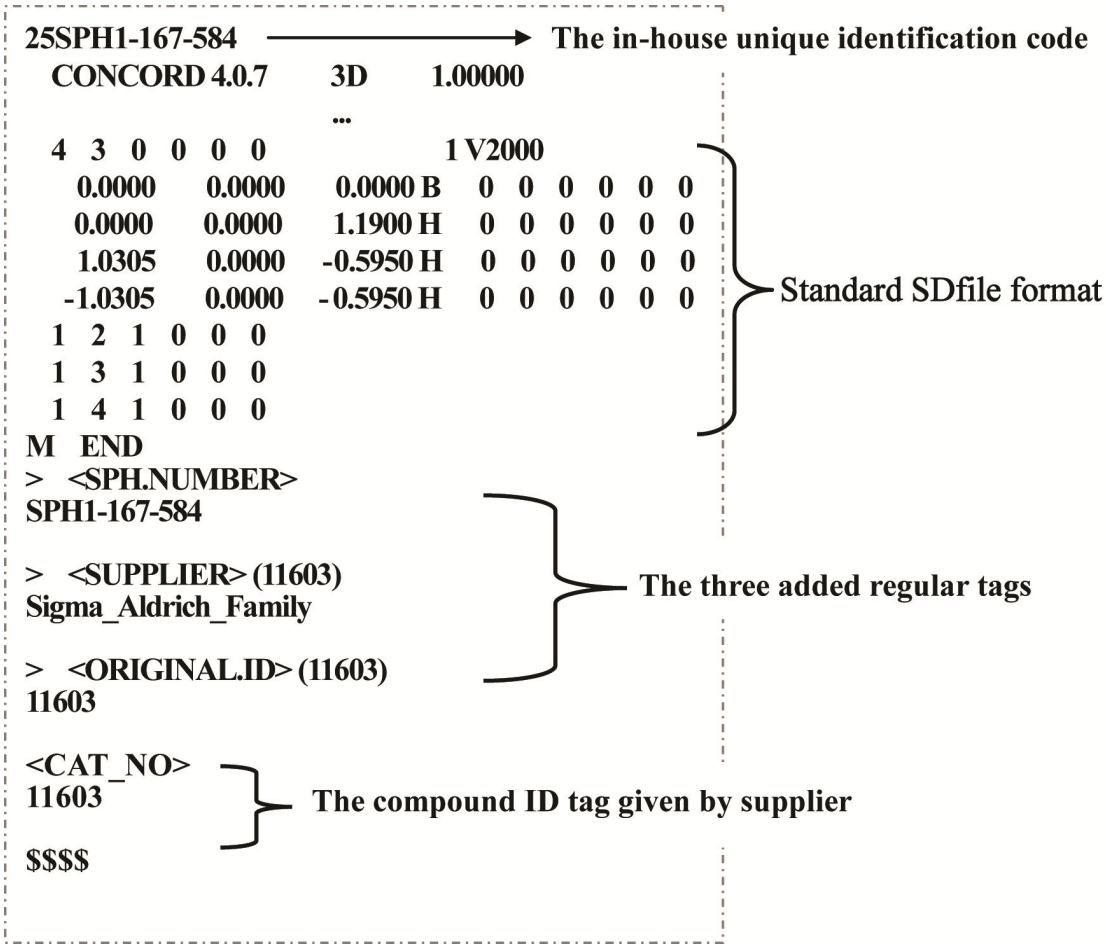
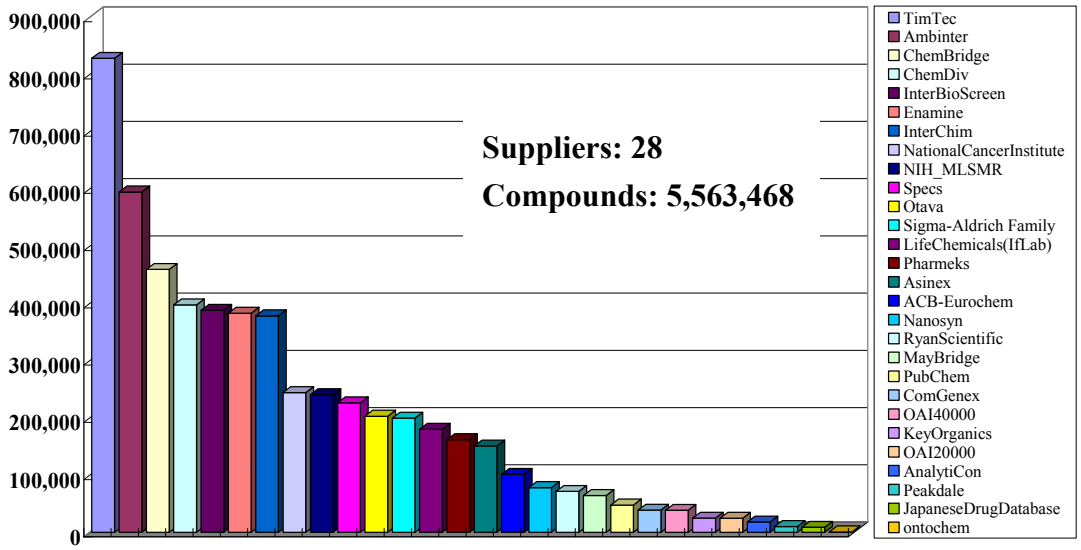


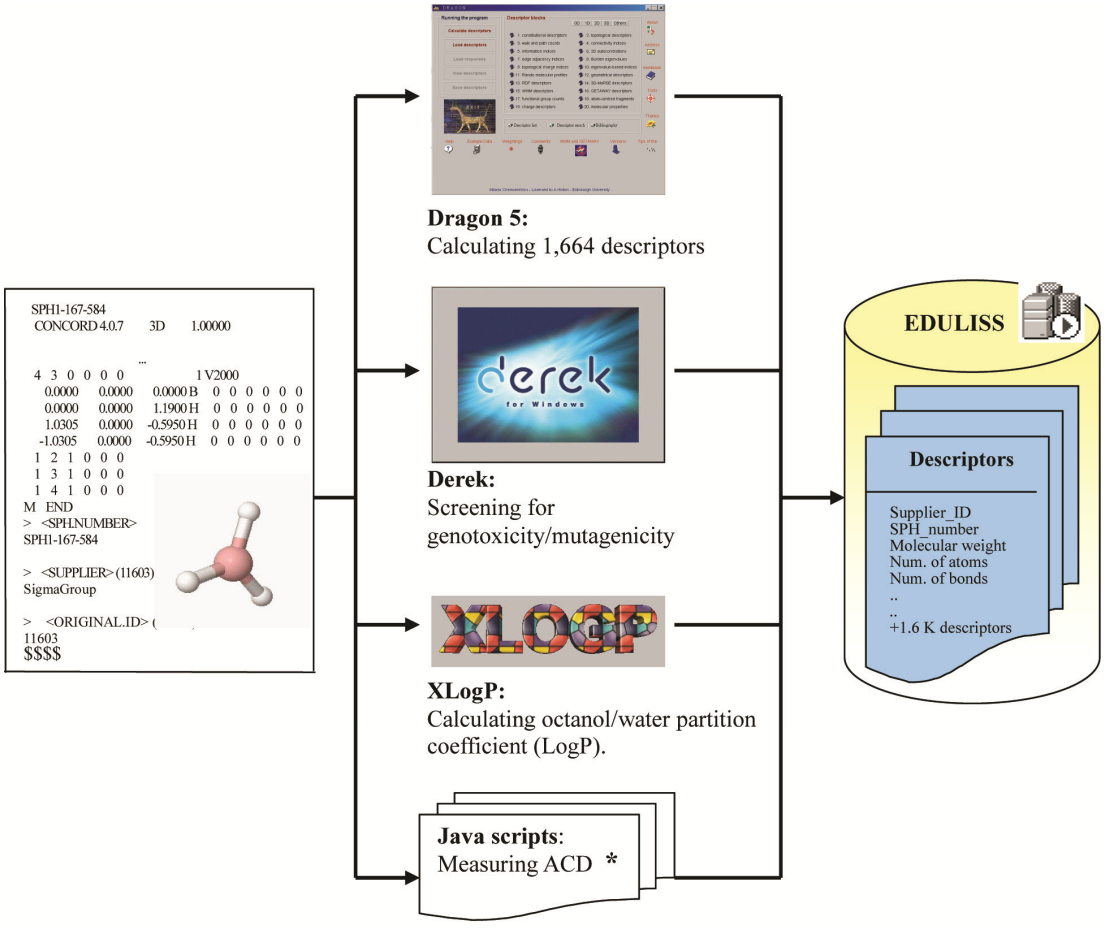
Figure 2.5. Supplier catalogues in order of size found within EDULISS 2.0



2.2.3 The extraction of molecular properties

The molecules in EDULISS 2.0 were processed by three programmes, DRAGON, DEREK and XLOGP, and series of scripts to extract their physicochemical properties. These processes are schematically shown in Figure 2.6 and detailed in the following sub-sections together with the description of the extracted properties.

Figure 2.6. Schema of generation of physicochemical properties (i.e. molecular descriptors) for each compound stored in EDULISS 2.0.



* Atomic Characteristic Distance, described in section 2.2.3.2.

2.2.3.1 Molecular descriptors

A vast majority of descriptors stored in EDULISS 2.0 were originated from the DRAGON 5.4 software, developed by Milano Chemometrics and QSAR Research Group (http://www.talete.mi.it/), which is capable of computing over 1,600 different molecular descriptors per compound. The entire descriptor list and their meanings can be found on its website. According to their characteristics, the DRAGON descriptors can be generally separated into five different classes including i) simple counts, ii) topological descriptors, iii) geometrical descriptors, iv) physiochemical descriptors and v) biological descriptors.

In class i), as the name suggests, a compound can be profiled as counts of specific features, such as counts of a series of atoms, bonds and rings, the number of functional groups and fragments, the sum of atomic properties, etc. Some of the atom- and bond-based counts also consider the specific hybridisation states and the bond order as well as taking the number of members in a ring into account. In DRAGON software these count-based descriptors are termed constitutional descriptors. The fragment-based descriptors are based on a list that contains 120 atom-centred fragments proposed by Vellarkad and colleagues (Viswanadhan, Ghose et al. 1989) and classified by the commonly occurring atomic states of carbon, hydrogen, oxygen, nitrogen, halogens, sulfur, phosphorus and selenium in organic molecules.

Class ii) and iii) are essentially shape-based descriptors mostly derived from various matrix representations of a chemical structure, usually hydrogen depleted. There are two fundamental matrices that are extended into other matrices and then used to

deduce most descriptors of these two classes. These are the adjacency matrix and the distance matrix. The samples in Figure 2.7 illustrate the two matrices and their extensions of 3-methyloctane. The adjacency matrix, commonly symbolised as **A** and denoted as **A=A(G)** where **G** is the molecular graph, is an important concept used in molecular graph theory (Lukovits 2000). It is an N×N symmetric array where N is the atom count, such that the element A_{ij} of the matrix is equal to one if atom A_i and A_j are adjacent (bonded) and zero otherwise. Two descriptor groups, namely atomic walk counts (Ruecker and Ruecker 1993) and connectivity-based descriptors are mainly based on this matrix. The adjacency can also be formed by edges (bonds) instead of atoms, known as edge-adjacency matrix (Estrada 1995; Estrada 1996; Estrada and Ramirez 1996; Estrada, Guevara et al. 1998) and denoted as **E** and **E=E(G)**. This matrix is the basis for the calculation of a series of edge-based descriptors.

The conformation of a distance matrix, usually symbolised as **D** and denoted as **D=D(G)**, is similar to an adjacency matrix but it summarises the distance information between all the atom pairs (Bonchev and Trinajstic 1977; Mihalic and Veljan 1992). The elements in a distance matrix are the topological distance D_{ij} which is the number of edges (bonds) in the shortest path between atoms. This matrix can be used to deduce the classical topological index the Wiener index (Wiener 1947b; Wiener 1947c). In the literature, the term Weiner index is used interchangeably with Wiener number, Wiener path or path number. It is calculated as half the sum of all topological distances in a molecular graph, such that the smaller is this index, the greater is the compactness of the molecule. This calculation

of this index also deduces the polarity number of a compound which is defined as the number of pairs of atoms which are separated by three bonds. The value of Wiener index of 3-methyloctane is 110 for the path number and 7 for the polarity number. There is a modified form of Wiener index which is known as Wiener 3D index and its matrix is symbolised as **G**. While the 2D distance matrix captures the constitutional features of the molecule, the 3D character is encoded in the geometric distance matrix. The Wiener 3D index simply measures the geometrical distance of atom pairs instead of the path number in the calculation (Crippen 1991; Mihalic and Veljan 1992). When the elements (D_{ij}) of a distance matrix have been converted into their reciprocals (excluding zero elements), that is $1/D_{ij}$, the resulting matrix is known as Reciprocal matrix (Ivanciuc, Balaban et al. 1993; Plavsic, Nikolic et al. 1993). There is yet another distance matrix known as the Detour matrix which counts the distance between all of the atom pairs in the longest path (Trinajstic, Nikolic et al. 1997; Rucker and Rucker 1998). Furthermore, the adjacency matrix and distance matrix can be combined for descriptor calculation. For instance, eccentric connectivity index (Sharma, Goswami et al. 1997) and eigenvalue-based descriptors (Balaban, Ciubotariu et al. 1991) accumulate the features of both matrices and are termed adjacency-cum-distance matrix and adjacency-plus-distance respectively. The matrices mentioned above can also be weighted by a set of variances, including atomic mass, van der Waals volumes, electronegativity and polarizability (Schultz, Schultz et al. 1990; Ivanciuc, Ivanciuc et al. 1998) for further computing. The calculation of the WHIM (Weighted Holistic Invariant Molecular) descriptor group (Todeschini, Lasagni et al. 1994) utilises the other conformation of the matrix where the elements are placed by the molecular coordinates (x, y, z)

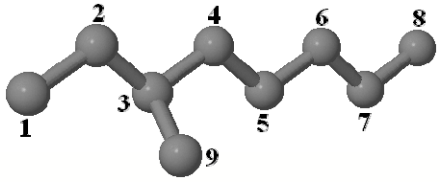
forming an N×3 matrix where N is the number of atoms. The details of this group will be described in a later chapter (section 3.2). Apart from the descriptors deduced from various matrices, some topological and geometrical descriptors are simply obtained from the measurement of the shortest path or geometrical distance between the specific atom pairs, including oxygen, nitrogen, sulfur, phosphorus and halogen.

Class iv) descriptors represent information related to various physicochemical traits that can be empirically measured, such as hydrophilicity index (Todeschini, Vighi et al. 1997), lipophilicity, molar refractivity (Ghose and Crippen 1987), molecular polar surface (Ertl, Rohde et al. 2000) and so forth.

The final descriptor class v) represents those descriptors that relate to the biological properties of a compound such as toxicity descriptors which include an important solubility index called Log P, a partition coefficient between octanol and water. An algorithm to calculate Log P was introduced using a regression analysis of biological activities to establish a QSAR (Hansch and Fujita 1964). The more hydrophobic a compound, the easier is its uptake into (e.g. human) fat. There are three different calculated Log P values that have been stored in EDULISS 2.0, MLogP (Moriguchi, Hirono et al. 1992) and ALogP (Viswanadhan, Ghose et al. 1989) are provided by DRAGON software as well as XLogP developed by Dr. Renxiao Wang (Wang, Fu et al. 1997; Wang, Gao et al. 2000). The XLogP method uses 80 atom-pair descriptors with associated Log P values to calculate an overall Log P value. Another important source of toxicity information in EDULISS 2.0 was obtained from the DEREK toxicity prediction software which is an expert system developed by Lhasa Ltd (Barratt, Castell et al. 2000). The predicted toxicity includes skin sensitisation,

carcinogenic, photoallergenic potential, and so on. These predicted results are outputted as a diagnosis-like report.

Figure 2.7. Examples of adjacency and distance matrix of 3-methyloctane.



Adjacency matrix, A

Atom	1	2	3	4	5	6	7	8	9
1	0	1	0	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0	0
3	0	1	0	1	0	0	0	0	1
4	0	0	1	0	1	0	0	0	0
5	0	0	0	1	0	1	0	0	0
6	0	0	0	0	1	0	1	0	0
7	0	0	0	0	0	1	0	1	0
8	0	0	0	0	0	0	1	0	0
9	0	0	1	0	0	0	0	0	0

Edge-adjacency matrix, E

Edge*	1-2	2-3	3-4	4-5	5-6	6-7	7-8	9-3
1-2	0	1	0	0	0	0	0	0
2-3	1	0	1	0	0	0	0	1
3-4	0	1	0	1	0	0	0	1
4-5	0	0	1	0	1	0	0	0
5-6	0	0	0	1	0	1	0	0
6-7	0	0	0	0	1	0	1	0
7-8	0	0	0	0	0	1	0	0
9-3	0	1	1	0	0	0	0	0

* Bond

Distance matrix, D

Atom	1	2	3	4	5	6	7	8	9
1	0	1	2	3	4	5	6	7	3
2	1	0	1	2	3	4	5	6	2
3	2	1	0	1	2	3	4	5	1
4	3	2	1	0	1	2	3	4	2
5	4	3	2	1	0	1	2	3	3
6	5	4	3	2	1	0	1	2	4
7	6	5	4	3	2	1	0	1	5
8	7	6	5	4	3	2	1	0	6
9	3	2	1	2	3	4	5	6	0

Geometrical distance matrix, G

Atom	1	2	3	4	5	6	7	8	9
1	0	1.53	2.49	3.84	4.99	6.30	7.48	8.77	2.92
2	1.53	0	1.53	2.50	3.85	4.99	6.30	7.48	2.49
3	2.49	1.53	0	1.53	2.50	3.85	4.99	6.30	1.53
4	3.84	2.50	1.53	0	1.53	2.50	3.85	4.99	2.49
5	4.99	3.85	2.50	1.53	0	1.53	2.50	3.84	2.92
6	6.30	4.99	3.85	2.50	1.53	0	1.53	2.49	4.32
7	7.48	6.30	4.99	3.85	2.50	1.53	0	1.53	5.22
8	8.77	7.48	6.30	4.99	3.84	2.49	1.53	0	6.60
9	2.92	2.49	1.53	2.49	2.92	4.32	5.22	6.60	0

2.2.3.2 A bit string to represent Atomic Characteristic Distance (ACD)

Atoms can be classified in differing ways, the atoms of hydrogen bond donors (HDon) and acceptors (HAcc) cause association of molecules by forming hydrogen bonding. Halogens (fluorine, chlorine, bromine and iodine), sulfur and phosphorus often exhibit highly reactive and commonly occur in the functional groups of organic molecules. Furthermore, the atoms of nitrogen, oxygen, sulfur and phosphorus contribute molecular polar surface area and volume (Palm, Luthman et al. 1998; Winiwarter, Bonham et al. 1998) which is useful in understanding their structure and ability to bind ligands and other molecules. Thus, these atoms have conventionally been queried for compound selection. Profiling the distribution of these atoms in a compound can be useful for the study of intermolecular interactions. Although DRAGON software can provide the information regarding the geometrical distances between these atom pairs, they are described and output as the summation of the distances in a compound. For instance, DRAGON gives a single value, 147.318 Å, which represents all of the observed distances between the three phosphorus atoms and the five nitrogens in an ATP molecule. As a result, it is difficult to understand the relative location of these atoms and this figure cannot answer a question such as "Does any compound contain phosphorus and nitrogen within 3.5 Å of each other?" In order to obtain more precisely and completely these geometrical properties, a proper strategy which should be able to describe and store this distance information of each atom pair is needed.

Atomic Characteristic Distance (ACD) in this study is designed to answer the requirement described above. It includes each distance measured from each HDon

and HAcc, halogen (fluorine, chlorine, bromine and iodine), sulfur and phosphorus to each HDon and HAcc, thus fifteen types of interatomic distance for each compound. The measurement of ACD and identification of HDon and HAcc were done by an in-house java script using CDK java tool kit and it defines HDon and HAcc as below (quoted from CDK application document):

HDon:

1. Any -OH where the formal charge of the oxygen is non-negative (that is formal charge >= 0).
2. Any -NH where the formal charge of the nitrogen is non-negative (that is formal charge >= 0).

HAcc:

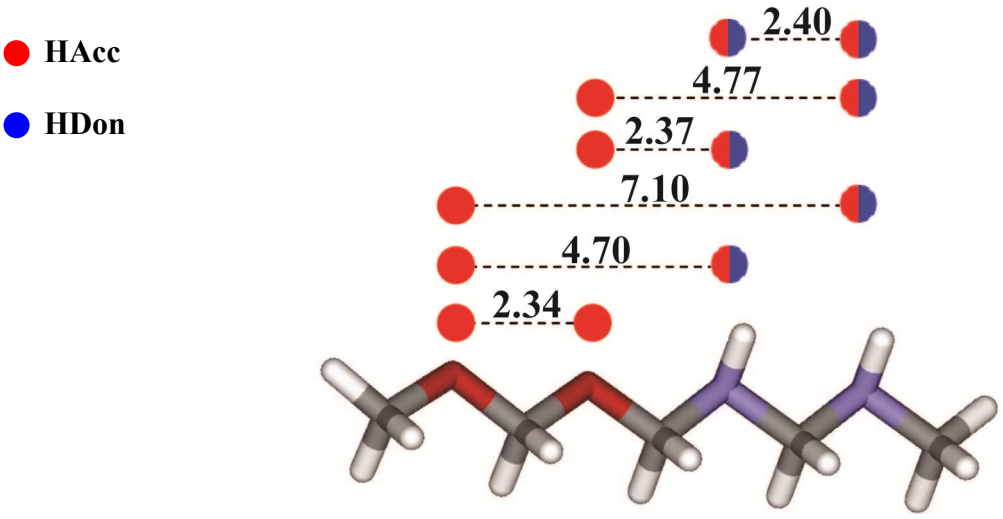
1. Any oxygen where the formal charge of the oxygen is non-positive (that is formal charge <= 0) except an aromatic ether oxygen (that is an ether oxygen adjacent to at least one aromatic carbon) or an oxygen adjacent to a nitrogen.
2. Any nitrogen where the formal charge of the nitrogen is non-positive (that is formal charge <= 0) except a nitrogen adjacent to an oxygen.

Since the number of observed distances from the entire EDULISS 2.0 compound collection (+5.5 M) is very large, recording each measurement as an individual distance in the storage is not a good solution for high-speed screening. Thus, ACD employed the bit-wise algorithm to build a bit string, a boolean array, for each compound to describe the information of distances in which a bit represents the presence (true, i.e. 1) or absence (false, i.e. 0) of a distance between specific atom pairs, thus there are fifteen bit strings for a compound to represent the fifteen types of

distance pairs. The length of a bit string is 128 long. The first bit represents a distance less than or equal to 2.50 Å (i.e. <=2.50 Å) and its next bit is 0.25 Å longer (i.e. > 2.50 and <= 2.75 Å) and so forth until the last bit representing 34.25 Å. Figure 2.8 illustrates the composition of three bit strings of the distances between HAcc and HDon pairs where the example virtual compound contains two oxygens (shown as red) and two nitrogens (shown as blue) forming four HAcc and two HDon, respectively. The bit 1, 10, 11 and 20 of HAcc - HDon string are recorded as true (i.e. 1) according to the observed distances, and other bits are recorded as false (i.e. 0). The HAcc-HAcc string has the same composition. There is only one observed distance between HDon and HDon (that is 2.40 Å), so that the overall string only contains a true value in its first bit. Since the example compound has no halogens, sulfurs and phosphorus atoms, the other twelve strings are formed by false bits entirely.

As a search process, a bit string is generated for a user-defined query distance and then compared to that of each compound. If a specific true bit in the query is not also true in the compound's bit, then the compound could not fit the query distance. A user, of course, could perform a multi-distance query in a single search. Apart from its ideal storing and manipulative facility, the bit strings also make a very fast screening for molecule searching since computers perform the necessary boolean operations very quickly.

Figure 2.8. Examples of the bit string composition of a virtual compound.



The distance scales in a bit string

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21		127	128
D	2.50	2.75	3.00	3.25	3.50	3.75	4.00	4.25	4.50	4.75	5.00	5.25	5.50	5.75	6.00	6.25	6.50	6.75	7.00	7.25	7.50		34.00	34.25

The bit string of HAcc – HDon pairs

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21		127	128
V	**1**	0	0	0	0	0	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0	**1**	0		0	0

The bit string of HAcc – HAcc pairs

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21		127	128
V	**1**	0	0	0	0	0	0	0	0	**1**	**1**	0	0	0	0	0	0	0	0	**1**	0		0	0

The bit string of HDon – HDon pairs

I	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21		127	128
V	**1**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		0	0

I: the index of bit; D: the distance belonging to a bit (Å); V: the value of bit

2.2.4 Storage of the information of molecular structures and properties

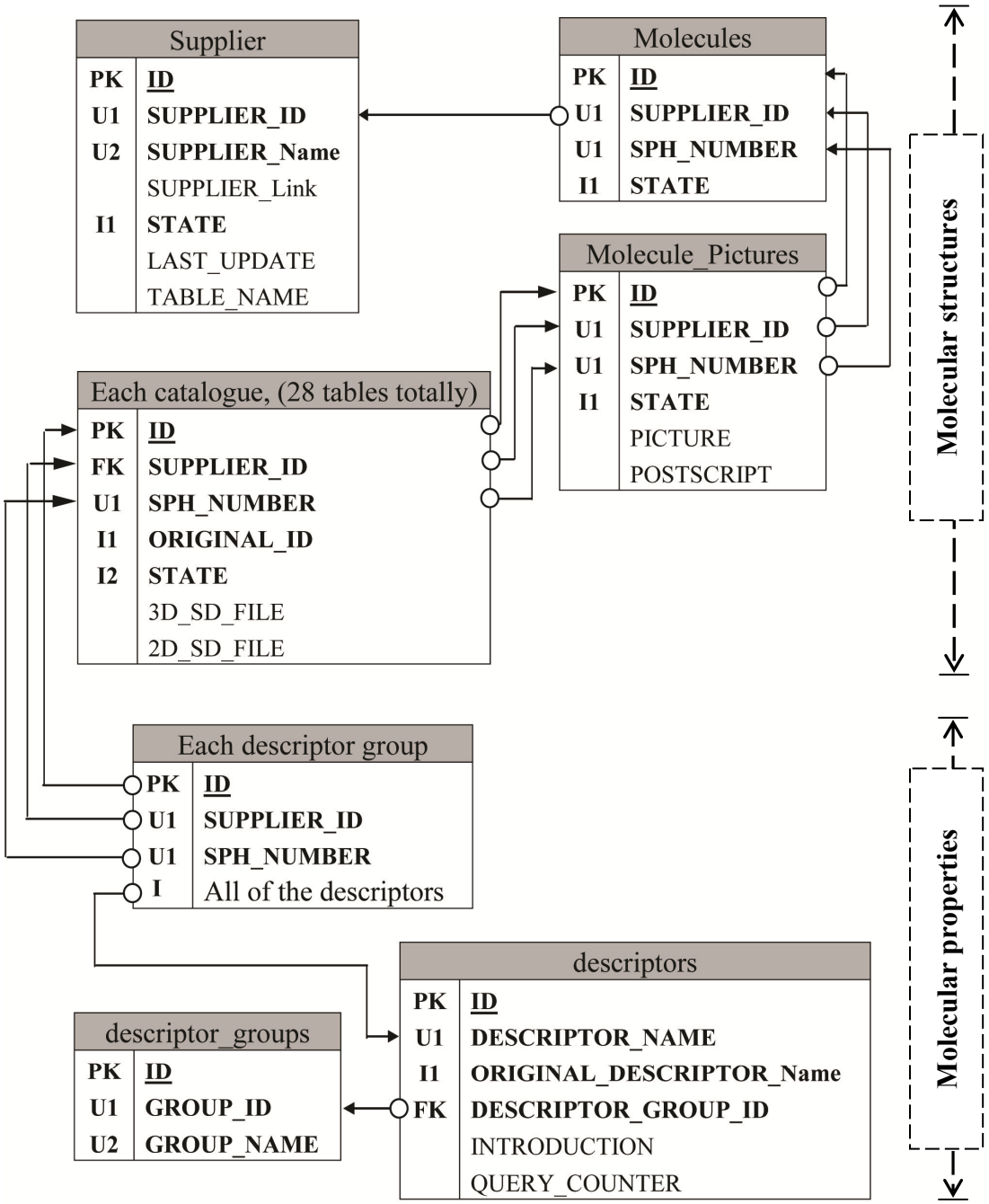
After the collection and conversion of SDfiles and the extraction of molecular properties have been done, the issue of properly storing these data must be addressed. The storage should take the data management and further application requirements into account. As mentioned in section 2.1, EDULISS 1.0 saved the SDfiles as general file system files making updating the database as difficult operation. Given the greater number of collected compounds in EDULISS 2.0, the total size of the SDfiles is over 20 gigabytes. It takes over 9 minutes to go through the files for a simple operation such as counting the total number of compounds in Sigma catalogue by a perl script. To address these problems and store the data in a way that facilitates efficient loading, incremental updating, varied querying and data subsetting, a relational database is the most reasonable choice. It is fast, efficient, compatible with various scripts and, in the case of MySQL (http://www.mysql.com/), free. This study employed MySQL version 4.1.10a and involved the design of a database schema to relationally organise data.

The major schema of EDULISS 2.0 database design is shown in Figure 2.9. The schema could be divided into two phases, the storage of molecular structures (SDfiles) and molecular properties (descriptor values), respectively. The root of molecular structure tables is the "Supplier" table which contains the fundamental information of collected suppliers, including the name, website link, ID for EDULISS 2.0, the day of last update and the table name through which their SDfile could be found. The "Molecule" table determines the entry ID and the unique identification code (described in section 2.2.2) for each compound, that is any

compound and its properties only possess a unique entry ID and identification code to represent itself throughout the database. EDULISS 2.0 also stores the compound pictures in 2D and the postscript for each. It aims to display molecule 2D pictures on web pages directly instead of taking time to invoke browser plugins, such as Java applets or server-side programmes for display of the individual structures. The actual molecular structures, both 2D and 3D SDfiles, were deposited in 28 tables for each catalogue. The names of these tables refer to the column, named TABLE_NAME, being listed in the "Supplier" table. Each of these molecular structure tables contains a boolean column, named STATE, to describe the usability of a compound, where the value is equal to zero if a compound no longer exists in the new version of catalogue and one otherwise. This design aims to keep the EDULISS 1.0 data and to ensure that the earlier studies could be continued with EDULISS 2.0. In the same way, all of the property tables refer to a root table, named "descriptor_groups". This table lists each descriptor group ID and name, and is associated with the next table, named "descriptors" where the details, such as the name, referred group ID, and short description for each descriptor can be found. This table also contains a column, named QUERY_COUNTER, to record the query frequency of a descriptor. At present, the highest frequency of queried descriptor is the set of descriptors dealing with Lipinski's rule of five. Since DRAGON software has logically separated the descriptors into twenty groups, EDULISS 2.0 database uses the same principle for the design of the molecular property tables, i.e. storing descriptor values in twenty tables.

Relational databases can perform queries very quickly given correct table design with the utilisation of indices for columns. Indices guide MySQL to directly and quickly find rows with specific column values, rather than find the relevant rows from sequential searches of the table. The EDULISS 2.0 database has deployed a number of indices in the relevant tables, especially for the columns where the descriptor values and the unique identification codes were deposited. It therefore only took 0.35 second to count the total number of compounds in Sigma catalogue, in comparison to the 9 minutes search time given in the beginning of this section.

Figure 2.9. Major schema of EDULISS 2.0 database design.



PK: Primary key; **U**: Unique key; **FK**: Foreign key; **I**: Index

2.3 The profile of molecular properties in EDULISS 2.0

The EDULISS 2.0 database stores over 5.5 million available compounds in total, containing the data from 28 different commercial and other smaller specialist compound catalogues. 2D and 3D coordinates for each molecule are stored with over 1600 constitutional, topological, geometrical, physicochemical and toxicological descriptors per compound. The descriptors can be used interactively to select the subgroups of the database and to provide profiling information.

In this database, over +3.9 million fit the Lipinski's rule of five, that is molecular weight (MW) $\leq$ 500, calculated Log P $\leq$ 5 (MLogP $\leq$ 4.15), number of hydrogen bond acceptors (nHAcc) $\leq$ 10 and number of hydrogen bond donors (nHDon) $\leq$ 5. A total of +3.4 million fit the Oprea lead-like criteria, that is MW $\leq$ 460, number of rotatable bonds (RBN) $\leq$ 10, MLogP between -4 and 4.2, nHAcc $\leq$ 9, nHDon $\leq$ 5 and number of rings (nCIC) $\leq$ 4 (Hann and Oprea 2004). The more stringent Astex Rule of three is met by 254,018 compounds, that is MW $\leq$ 300, topological polar surface area (PSA) $\leq$ 60, RBN $\leq$ 3, MLogP $\leq$ 3 (Rees, Congreve et al. 2004). The statistical profiles of some general descriptors are shown in Figure 2.10 and descriptor ranges are shown in Table 2.2. There are +3.8 million (69 %) observed molecular weights (MW) placed in the range between 300 and 500 Da and the average, 373.53 Da, fits the MW limitation in lead-like criteria. The database also contains +520,000 compounds with MWs lower than 250 Da and potentially fitting the need of fragment-based screening. A fragment is defined as a small molecule, 100-250 Da, with few functional groups (Rees, Congreve et al. 2004; Hartshorn, Murray et al. 2005).

The distribution of three calculated Log P values is shown in Figure 2.10 (b). Algorithms for predicting aqueous solubility from structure generally rely on a directly proportional relationship between Log P and solubility (Jorgensen and Duffy 2002; Delaney 2005). In EDULISS 2.0, the most observed Log P value is 4 through three types of Log P and the average of MLogP is 3.26, and hence most compounds meet the solubility requirement essential for compound solubility during screening. There are +5.4 million (99 %) compounds that contain hydrogen bond acceptor ≤ 10 (avg. is 5.38) and donor ≤ 5 (avg. is 1.27) respectively. They fit the need of conventional rule-based screenings.

The distributions of oxygen and nitrogen counts are shown in Figure 2.10 (d) and exhibit a similar tendency. Both distributions mainly range between 0 and 7 with the near averages, and contain 2.88 oxygens and 2.61 nitrogens per observed compound. Figure 2.10 also shows the distribution of sulfur and phosphorus in (d) whereas Figure 2.10 (e) displays a set of halogen atoms (fluorine, chlorine, bromine and iodine). These atoms have often been queried for compound selection. The phosphorus containing compounds are very rare in the whole collection since only some 26 thousand have been observed. In the whole database, over 2.2 million (41 %) sulfur containing compounds have been observed and they mainly have one or two sulfurs. The most frequently occurring halogen is the chlorine atom as +1.2 million compounds can be found in the EDULISS 2.0 collection. Apart from chorine, the database does not have many compounds that contain other halogen atoms. There is an interesting distribution occurring in fluorine atoms. Although the fluorine containing compounds are rare, they mainly have one or three fluorine atoms.

Referring to the distribution shown in Figure 2.10 (f), almost all compounds deposited in EDULISS 2.0 are composed of one or more rings (98 %) and each compound contains about three rings on average. The rings in a compound are mainly formed by 6-membered and then 5-membered rings.

From the perspective of protein-ligand docking research, the relevant geometric conformation of small molecules is of particular interest. Figure 2.11 profiles the distribution of Wiener 3D index (W3D) of the EDULISS 2.0 compounds, which represents the compactness of a compound, as well as the distances of the ACD atom pairs previously described in section 2.2.3.2. To highlight the contrast, it also indicates the locations of glycine (W3D is 113.079 Å), the smallest amino acid, and adenosine triphosphate (ATP and its W3D is 7,264.493 Å) molecules around the distribution curve in Figure 2.11 (a). As it exhibits a left-skewed tendency, it can be seen that there are +3.3 million compounds (62 %) with geometric compactness relatively more then ATP. Figure 2.11 (b), (c) and (d) show the distances from HAcc, HDon, S and P to HAcc and HDon respectively. The number of observed within 2.5 Å are much greater than those in longer distances, but the distributions of halogens do not display the same tendency as those which are shown in Figure 2.11 (e) and (f). At this short distance, 2.5 Å, it can be assumed that the atoms of O, N, S and P were synthesised in the same functional group or occur adjacent to each other. The overall distributions show that the distance between the atoms and HAcc and HDon are mainly within 10 Å and halogens can be longer at 15 Å, but it is rare for them to have distances greater than 17.5 Å.

Figure 2.10. Molecular property profiles in the EDULISS 2.0 database. (a) MW. (b) three types of Log P. (c) number of HAcc and HDon. (d) number of N, O, P and S. (e) number of halogen atoms. (f) number of rings, 5-membered and 6-membered rings. The vertical axes are the number of compounds.

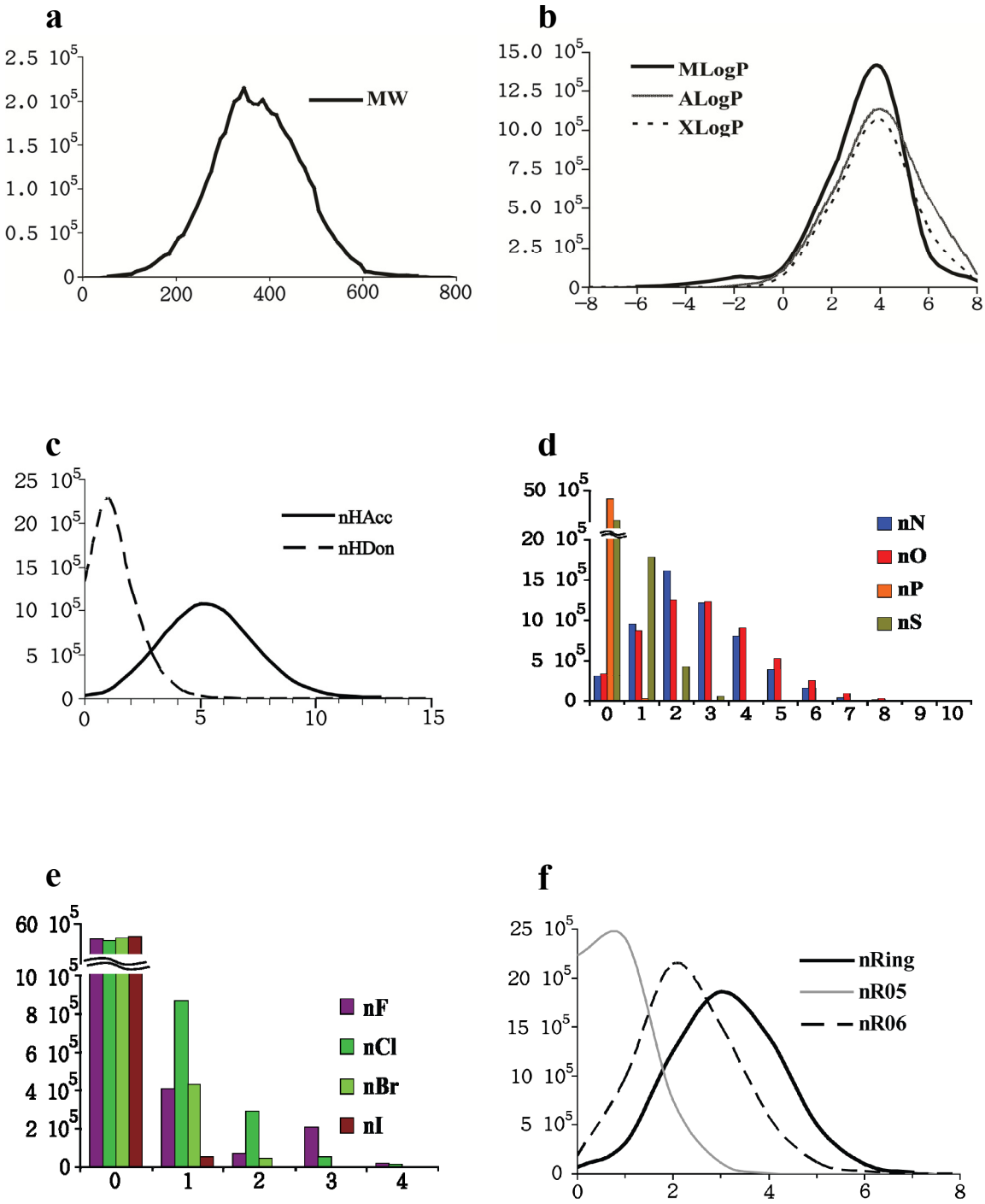


Figure 2.11. Geometric conformation profiles of compounds in the EDULISS 2.0 database. (a) Wiener 3D index. (b) The distances between hydrogen bond donor (HDon) and acceptor (HAcc) atoms. (c) to (f) The distances between S, P and halogens and HDon and HAcc atoms, respectively. The vertical axes are the number of compounds and the horizontal are the distances (Å).

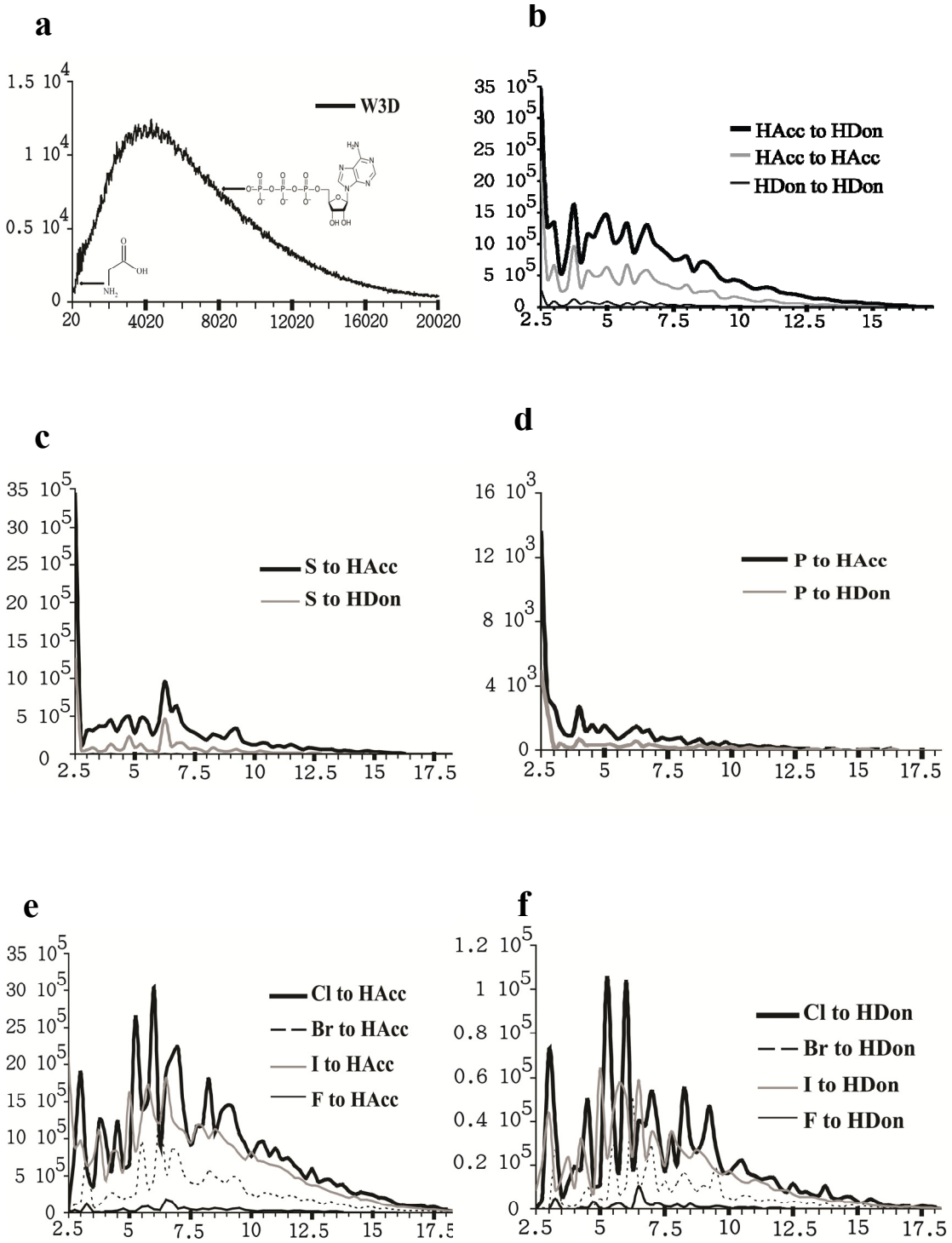


Table 2.2. Descriptor ranges for the +5.5 million compounds stored in the EDULISS 2.0 database.

Descriptor	Max	Min	Average	Standard deviation
Molecule weight	2,952.20	13.84	373.53	96.82
Number of atoms	395	2	44.90	12.46
Number of oxygen atoms	61	0	2.88	1.74
Number of nitrogen atoms	49	0	2.61	1.51
Number of HAcc	89	0	5.38	2.16
Number of HDon	62	0	1.27	1.10
Number of phosphorus atoms	9	0	0.01	0.09
Number of sulfur atoms	21	0	0.51	0.69
Number of fluorine atoms	51	0	0.26	0.85
Number of chlorine atoms	23	0	0.31	0.66
Number of bromine atoms	14	0	0.10	0.34
Number of iodine atoms	6	0	0.01	0.11
Number of bonds	399	1	47.03	13.24
Number of aromatic bonds	96	0	13.34	5.99
Number of rotatable bonds	107	0	5.30	2.77
Number of rings	24	0	3.13	1.18
Number of 5-membered rings	16	0	0.78	0.77
Number of 6-membered rings	17	0	2.32	1.08
MLogP	134.05	-34.80	3.26	3.30
Sum of atomic VDW volumes [a] ($Å^3$)	221.83	0.71	29.25	7.81
Topological polar surface area [b] ($Å^2$)	1,416.43	-11.58	74.06	42.32
Wiener 3D index [c] (Å)	2,021,122	0.92	6969.63	6190.85

a: VDW stands for van der Waals, scaled on carbon atom; **b**: Using N, O, S and P polar contributions; **c**: hydrogen included in the calculation.

2.4 The facilities of EDULISS 2.0

As described in section 2.2.4, EDULISS 2.0 has a set of relationally organised tables designed to hold the molecule's SDfiles, their properties (descriptor values) and relevant data. Researchers can access EDULISS 2.0 by logging into the database directly or by using scripts, such as Perl, Python, Java or C++. The application libraries of these programming languages have provided the essential MySQL driver which acts as an interface between the database and programmers. All information in EDULISS 2.0 can be retrieved by the above approaches. However, a basic knowledge of SQL (Structured Query Language) is required. It is a relational data language providing a consistent and English keyword-oriented set of facilities, including data definition, manipulation and control for query (Chamberlin, Astrahan et al. 1976). In order to deal with the case where the users are unfamiliar with SQL, a user-friendly web-based interface of EDULISS 2.0 has been established to provide a series of straightforward facilities to satisfy the need of data-mining for drug discovery. Details of the interface and facilities of EDULISS 2.0 will be described in following sections.

2.4.1 The construction of the web-based interface

The web-based interface of EDULISS 2.0 uses Java Servlet technology (see http://java.sun.com/products/servlet/) and JavaServer Pages (JSP, see http://java.sun.com/products/jsp/) to build the web pages. The web site utilises Apache Tomcat (see http://tomcat.apache.org/index.html) as the web server and the runtime environment for Java technologies mentioned above. To enable extensibility in functionality and for ease of maintenance, the web site was constructed using the

Model-View-Controller (MVC) software architecture. MVC separates the data model, user phase, and control logic into three individual components so that modifications to one component can be made with minimal impact to the others. Figure 2.12 is the schema of this web site. When a task is requested by a client, the servlet, (i.e. the controller), dispatches the request to an appropriate functional script, (i.e. the model), in order to carry out computing jobs. Upon the completion of the task, the process will return to the servlet and then dispatched to the relevant web pages, i.e. the views, to display the query result on the client's browser. Although the MVC makes the web site look more complicated and its development more time-consuming, it is easier to extend its functions whenever necessary. The web site developer can plug a new script in the model phase for some specific function, such as a new similarity search approach, with minimal or no amendment to other components. Furthermore, the MVC architecture provides a safer server environment because the controller phase separates the model phase from the client. The components of the model phase run on the server side and manage access to databases or files in the file system, and therefore often contain privileged information, such as the account, password and relevant folder paths, of the database and server. This separation prevents the server information from being exposed.

2.4.2 Retrieval of SDfiles and the molecular properties

The EDULISS 2.0 web site is available at http://eduliss.bch.ed.ac.uk/. Its homepage is shown in Figure 2.13. The web site provides four different data-mining approaches to select preferred compounds and their properties, including descriptor- or rule-based mining, molecule similarity search, Atomic Characteristic Distance

(ACD) mining and mining by molecule ID. Users can simply choose one of them in the home page and then perform the data-mining. Before the web site performs the data-mining, users can specify the preferred catalogues. Shown in Figure 2.14 (a) is a list for catalogue selection with supplier's link, the number of deposited compounds and the date of latest update. Although it is possible to select one or more catalogues for a query, the server will only extract a maximum of two million hits per batch a limit instituted to ensure stable service. After the catalogue has been selected, the four mining approaches can be performed and these will be described in the following sections.

2.4.2.1 Descriptor- or rule-based mining

Since EDULISS 2.0 has more than 1,600 molecular descriptors for each compound, it allows users to determine a series of descriptor items for a query. Figure 2.14 (b) lists the available options for descriptor items. From the top, the first twenty groups are the DRAGON descriptors with 1,664 items in total. When the link has been clicked, the page will display the entire descriptors belonging to the group, so that the users can choose various items and set the preferred values for the query. Shown in Figure 2.14 (c) is an example to determine the number of carboxylic acids (nRCOOH) of a compound. The page displays the general statistical profile of the item and allows the users to set a range for the values or specify the logical operator. Since it is not easy to find out the preferred descriptor from thousands of items, the web site has a function that helps users to search for the proper descriptor by its key word. Apart from the options of DRAGON descriptors, the users can also perform a rule-based query, such as Lipinski rule of five, Astex rule of three or Oprea lead-like.

If Lipinski rule of five has been chosen, the page will bring up the relevant descriptors directly and the users can amend them if necessary. The list shown in Figure 2.14 (b) also gives three convenient links to help the users set up the criteria quickly. These links include "Top 10 selected Descriptors" that brings up the ten most frequently selected descriptors for query; the link "User Defined" that allows the users to set up their preferred items, and the link "Your Last Query Condition" that records the users' last criteria for query. Since the criteria have been determined, the page will display the selected items with their values and the SQL statement. It provides a final opportunity to amend the criteria before submission. As shown in Figure 2.14 (d) for demonstration, we select the catalogues of Sigma and MayBridge with Lipinski rule of five as the criteria, and then this query selects 223,687 compounds in total. On the page, the users can download the SDfiles of hit compounds with their descriptor values. The descriptor information can be added to the annotation block of an SDfile or written in a tab-separated file individually. The tab-separated file can be imported into some software directly, such as Microsoft Excel, and it is convenient for further treatment. Alternatively, the users can choose to display the details for each compound in the form of 2D pictures or further display 3D structures graphically by Jmol, an open-source Java viewer for chemical structures in 3D (see http://jmol.sourceforge.net/). As shown in Figure 2.14 (e), the page also displays the number of identical compounds that can be found in the whole EDULISS 2.0 collection.

2.4.2.2 Molecular structure similarity search

Apart from the descriptor-based mining, EDULISS 2.0 is able to perform molecular structure similarity searches. The web site provides a molecule editor, JME Molecular Editor (see http://www.molinspiration.com/jme/), which is a Java applet for drawing directly the query structure as shown in Figure 2.15 (a). After drawing the query structure and submitting it, the web site will convert the query into a 3D format by CONCORD software and generate its SMILES string as well as calculate some common molecular descriptors using the CDK tool kit, and these are shown in Figure 2.15 (b). The SDfile and 3D structure of the query are also displayed. At present, EDULISS 2.0 is capable of performing similarity searches in terms of topological and geometrical approaches. The topological approach bases on the compound path number and polarity number, i.e. Wiener index (described in section 2.2.3.1), to compare the index of the query compound with each target, i.e. the EDULISS 2.0 collection, and then measure the similarity between them using Euclidean distance, i.e. the shorter the distance, the more the query is similar to those of the target; thus the connectivity of bonds of hit compounds should be similar to those of query structure. The users can limit the number of hit compounds, such as searching for the top 100 or more. Figure 2.15 (c) displays the search result of which the compounds are the top four and topologically similar to query structure. For comparison, the four compounds are placed on the right side in each frame and the query structure is placed on the left side iteratively. Similarly, the query structure and hit compound can be displayed in a 3D picture and their SDfiles can be downloaded.

Apart from the topological approach, the web site also provides geometrical similarity searches based on a 3D similarity measurement developed by my colleague Steven Shave (paper in progress). It evaluates the 3D similarity between two specified molecules by a method that generates some descriptors used to geometrically characterise the structure and then stores the descriptor values in a binary file for the similarity calculation (For details, see http://eduliss.bch.ed.ac.uk/eduliss/Eduliss_introduction.jsp?topic=UFSRAT). The implementation of this similarity search was written in C++ and then plugged into EDULISS 2.0 web site. Although the web site was built using a Java solution, the integration of the C++ implementation and the web site can be achieved smoothly as the web site was designed using the MVC architecture mentioned in section 2.4.1.

2.4.2.3 Mining by Atomic Characteristic Distance (ACD)

This facility aims to pull out compounds that contain the specific atoms introduced in some distances between the atoms and HAcc or HDon as described in section 2.2.3.2. On the page, the users can select a series of the atoms and determine the HAcc or HDon, and then specify the preferred distances between them to be the query criteria. Figure 2.16 shows the criteria that aim to pull out the compounds containing the four atomic characteristic distances between the specified atom pairs. If the users click the "add" button, the page will bring up an additional entry to add one more criterion to a query whenever necessary. Alternatively, the users can click the "–" button to remove the entry. The query result will be displayed and can be downloaded as described in section 2.4.2.1.

2.4.2.4 Mining by molecule ID

It is the simplest mining approach in EDULISS 2.0 for selecting a compound from the EDULISS 2.0 collection. The users can simply specify a catalogue and then input the identification code of the compound using either the in-house SPH number or the code a supplier assigns. For instance, if a user selects Sigma as the supplier and inputs the SPH number as SPH1-131-349, the web site will retrieve the referred compound and display relevant information on the page, as shown in Figure 2.17. The page not only displays the hit compound in 2D and 3D graphs, but also lists the identical compounds found in the whole database. The result in the example shows two identical compounds that can be found in the catalogues of InterBioScreen and ChemBridge. This function is practically useful for the purpose of compound purchase. For example, if this compound is out of stock at Sigma or has a long lead time to delivery, a researcher can purchase the compound from the other two suppliers instead. This also facilitates prices comparison of the compound.

Figure 2.12. Schema of the web-based interface of EDULISS 2.0.

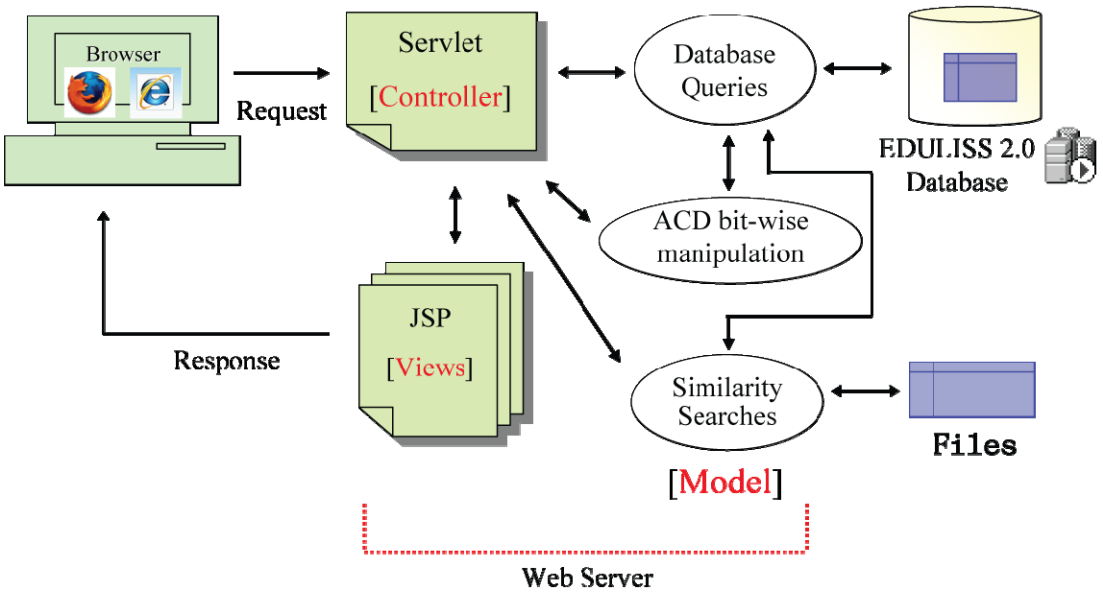


Figure 2.13. The home page of EDULISS 2.0 web-based interface.

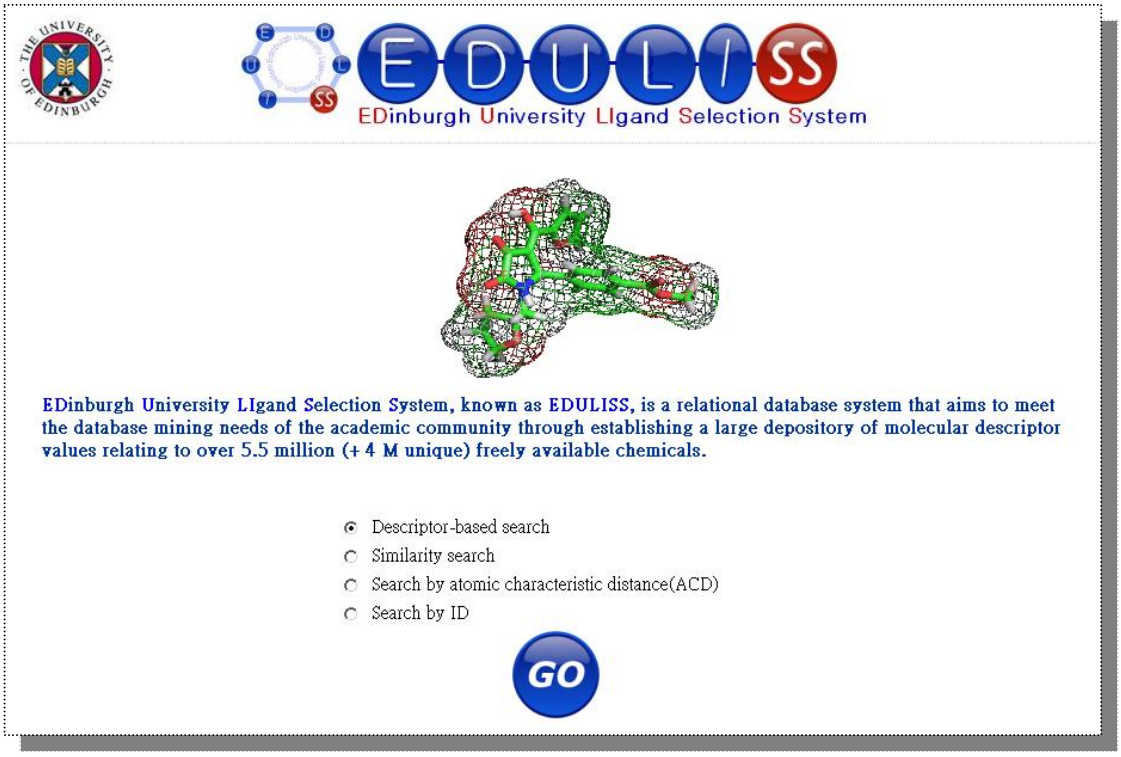


Figure 2.14. Components of pages for descriptor-based mining. (a) A list of available of suppliers. (b) A list of available of descriptors. (c) The setting of descriptor value. (d) Query results. (e) Details of hit compounds.

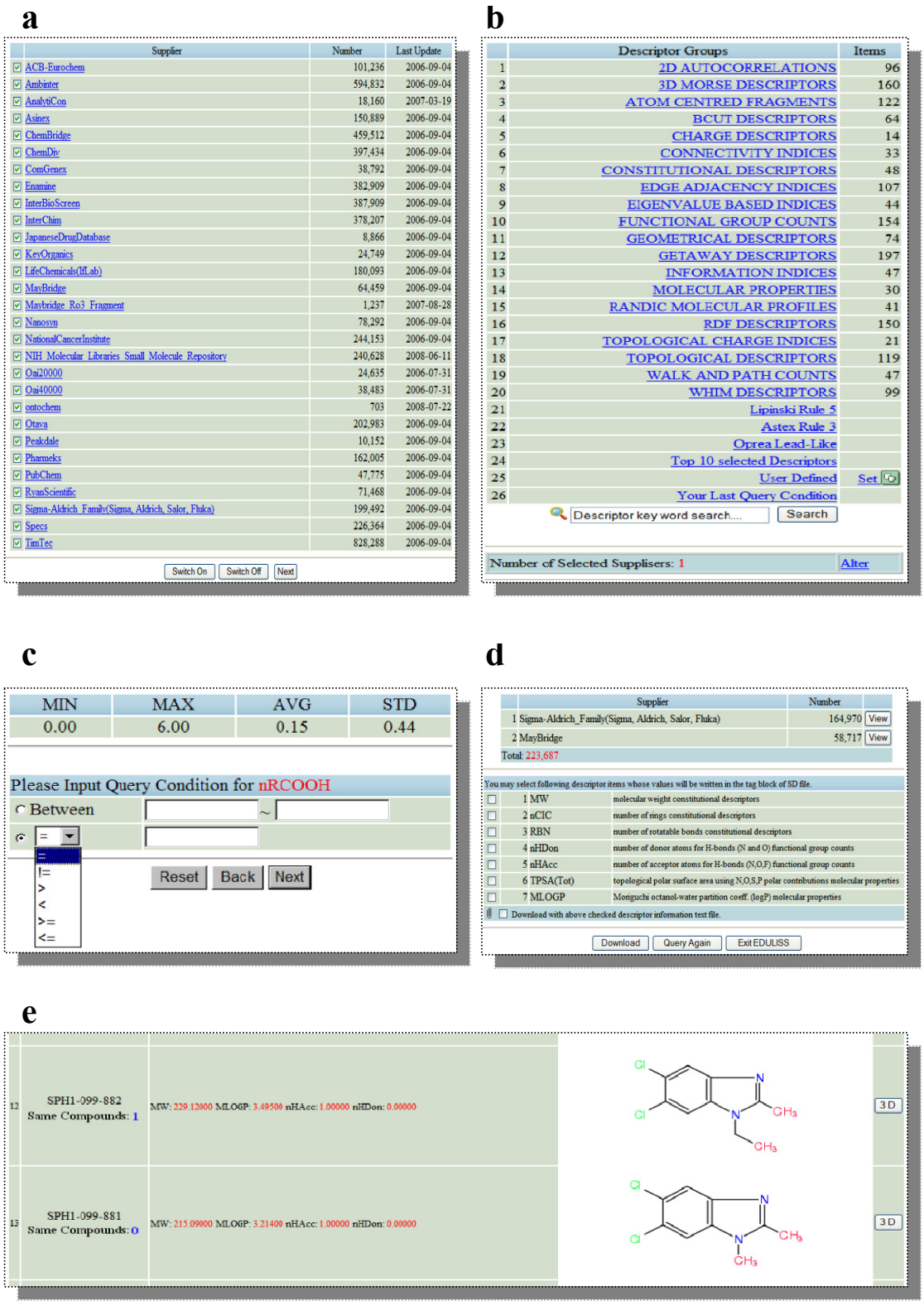
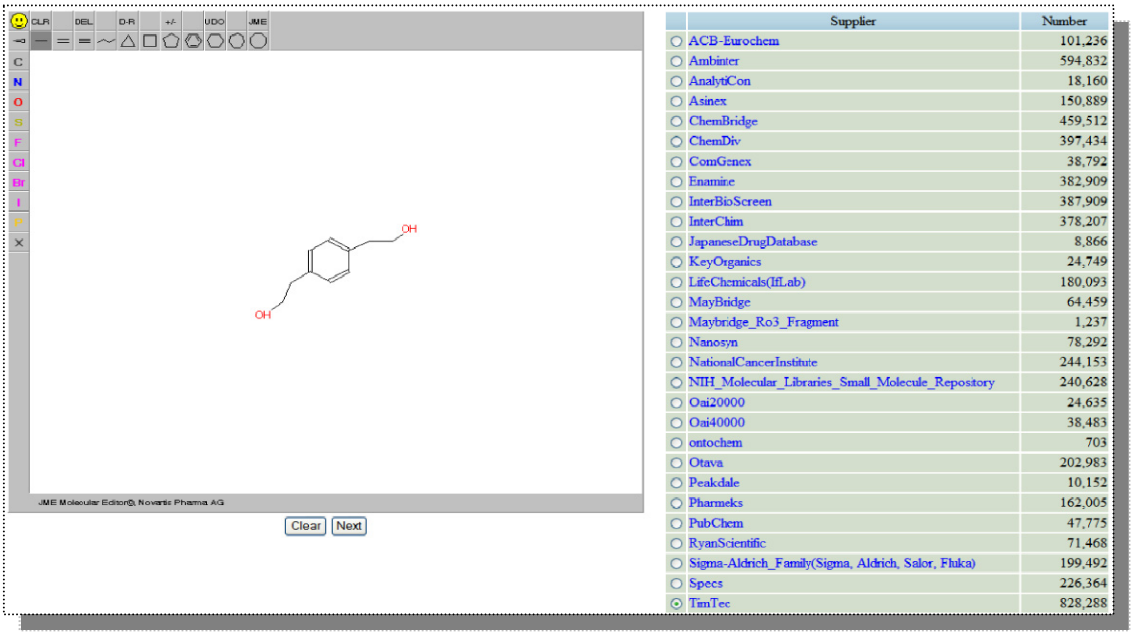
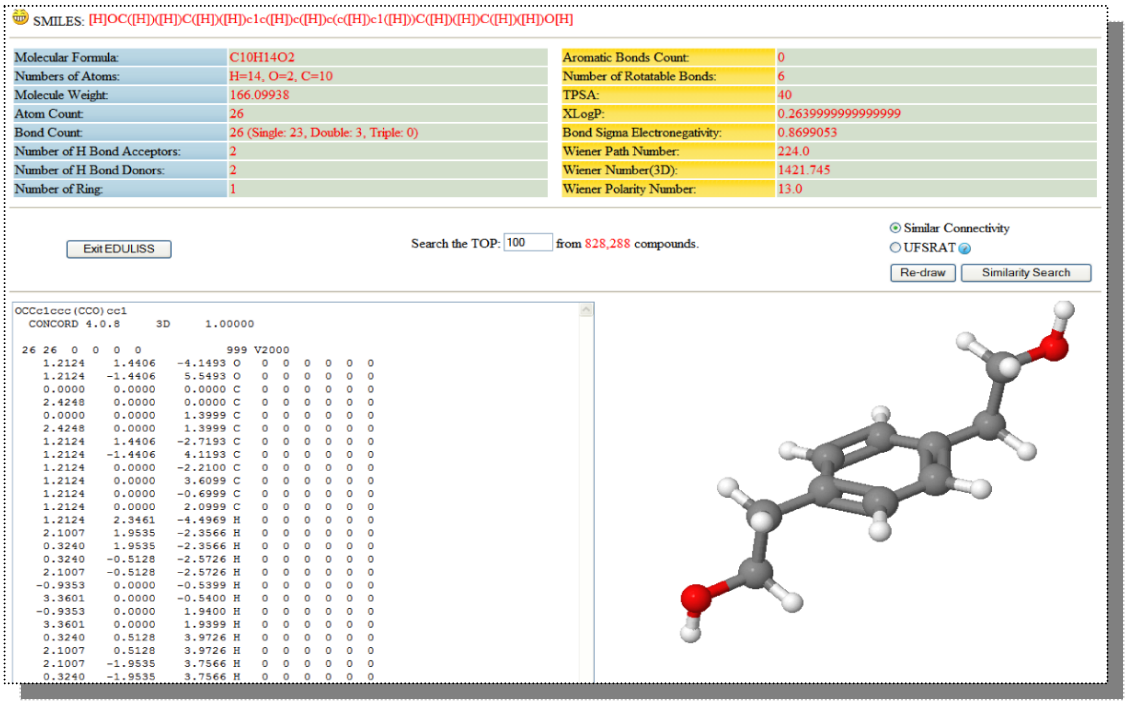


Figure 2.15, Components of pages for molecular structure similarity search. (a) The molecule editor for drawing directly the query structure. (b) Details of the query structure, including its SMILES string, general descriptors and 3D SDfile. Users can choose an approach to perform similarity search and determine the number of the hit compounds on this page. (c) The similarity search result of topological approach. The compound on the left of each frame is the query structure and those on the right are the hit compounds.

a



b



c

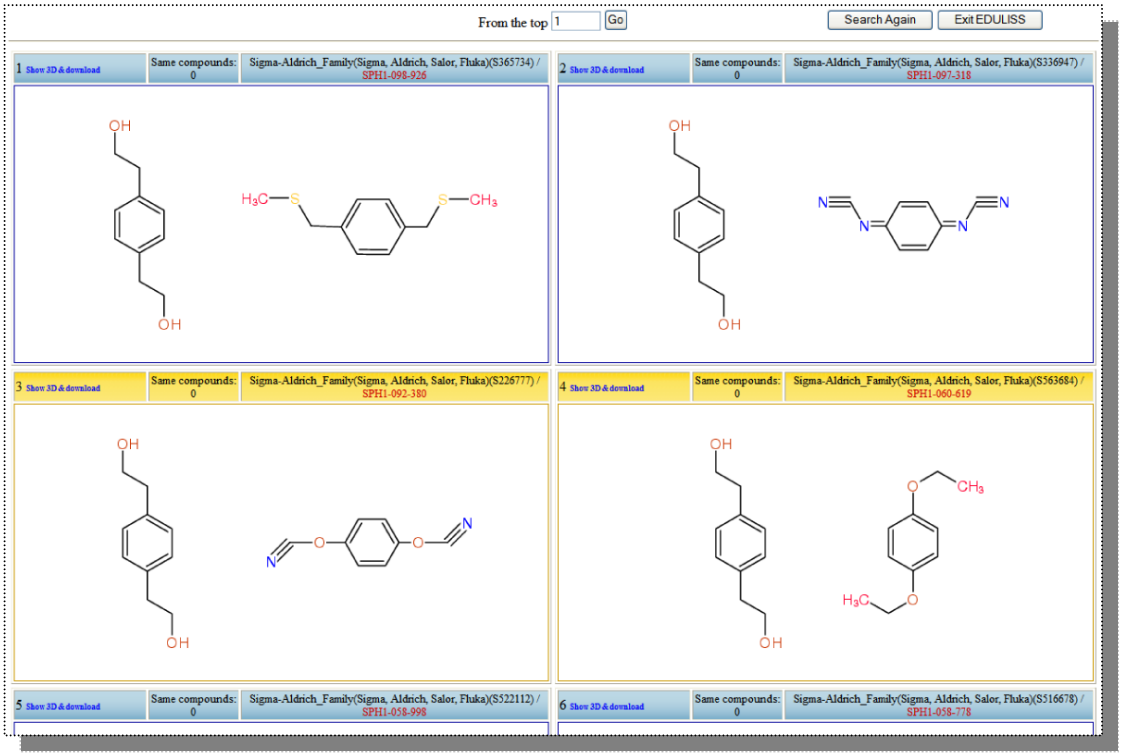


Figure 2.16. Components of pages for Atomic Characteristic Distance (ACD) mining approach.

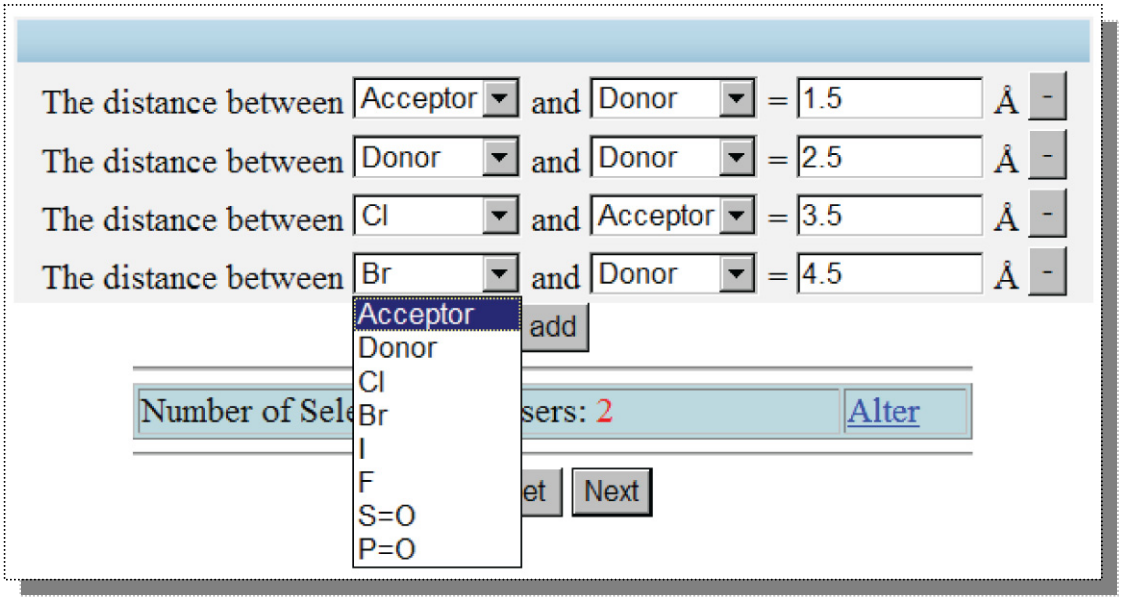
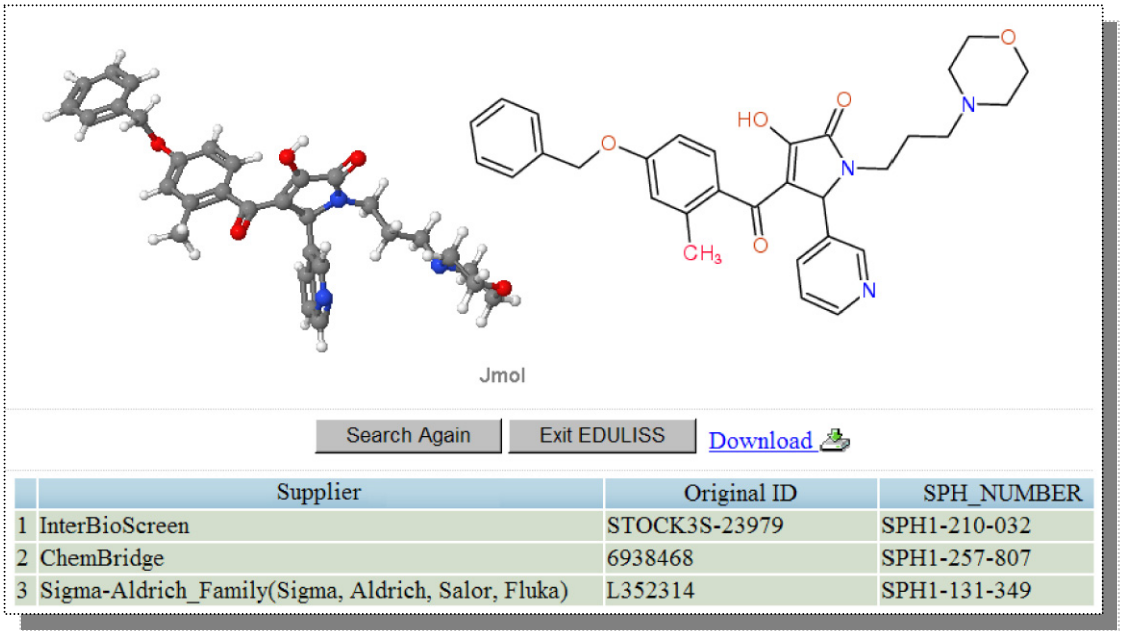


Figure 2.17. Components of pages for the retrieval by a compound identification code. This page lists the identical compounds that can be found in the entire EDULISS 2.0 collection.



	Supplier	Original ID	SPH_NUMBER
1	InterBioScreen	STOCK3S-23979	SPH1-210-032
2	ChemBridge	6938468	SPH1-257-807
3	Sigma-Aldrich_Family(Sigma, Aldrich, Salor, Fluka)	L352314	SPH1-131-349

2.5 Reference:

Balaban, A. T., D. Ciubotariu, et al. (1991). "Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors." Journal of Chemical Information and Computer Sciences **31**(4): 517-523.

Barratt, M. D., J. V. Castell, et al. (2000). "Development of an expert system rulebase for the prospective identification of photoallergens." Journal of Photochemistry & Photobiology, B: Biology **58**(1): 54-61.

Bonchev, D. and N. Trinajstic (1977). "Information theory, distance matrix, and molecular branching." The Journal of Chemical Physics **67**: 4517.

Carr, R. A. E., M. Congreve, et al. (2005). "Fragment-based lead discovery: leads by design." Drug Discovery Today **10**(14): 987-992.

Chamberlin, D. D., M. M. Astrahan, et al. (1976). "SEQUEL 2: A Unified Approach to Data Definition, Manipulation, and Control." IBM Journal of Research and Development **20**(6): 560-575.

Congreve, M., R. Carr, et al. (2003). "A 'Rule of Three'for fragment-based lead discovery?" Drug Discovery Today **8**(19): 876-877.

Crippen, G. M. (1991). "Chemical distance geometry: Current realization and future projection." Journal of Mathematical Chemistry **6**(1): 307-324.

Delaney, J. S. (2005). "Predicting aqueous solubility from structure." Drug Discovery Today **10**(4): 289-295.

Ertl, P., B. Rohde, et al. (2000). "Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties." Journal of Medicinal Chemistry **43**(20): 3714-3717.

Estrada, E. (1995). "Edge Adjacency Relationships and a Novel Topological Index Related to Molecular Volume." Journal of Chemical Information and Computer Sciences **35**(1): 31-33.

Estrada, E. (1996). "Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes." Journal of Chemical Information and Computer Sciences **36**: 846-849.

Estrada, E., N. Guevara, et al. (1998). "Extension of edge connectivity index. Relationships to line graph indices and QSPR applications." Journal of Chemical Information and Computer Sciences **38**: 428-431.

Estrada, E. and A. Ramirez (1996). "Edge Adjacency Relationships and Molecular Topographic Descriptors. Definition and QSAR Applications." Journal of Chemical Information and Computer Sciences **36**: 837.

Ghose, A. K. and G. M. Crippen (1987). "Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity

relationships. 2. Modeling dispersive and hydrophobic interactions." Journal of Chemical Information and Computer Sciences **27**(1): 21-35.

Hann, M. M. and T. I. Oprea (2004). "Pursuing the leadlikeness concept in pharmaceutical research." Current Opinion in Chemical Biology **8**(3): 255-263.

Hansch, C. and T. Fujita (1964). "p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure." Journal of the American Chemical Society **86**(8): 1616-1626.

Hartshorn, M. J., C. W. Murray, et al. (2005). "Fragment-based lead discovery using X-ray crystallography." Journal of Medicinal Chemistry **48**(2): 403-413.

Hendrickson, M. A., M. C. Nicklaus, et al. (1993). "CONCORD and CAMBRIDGE: comparison of computer generated chemical structures with x-ray crystallographic data." Journal of Chemical Information and Computer Sciences **33**(1): 155-163.

Ivanciuc, O., T.-S. Balaban, et al. (1993). "Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices." Journal of Mathematical Chemistry **12**: 309-318.

Ivanciuc, O., T. Ivanciuc, et al. (1998). "Design of topological indices. Part 10. Parameters based on electronegativity and covalent radius for the computation of molecular graph descriptors for heteroatom-containing molecules." Journal of Chemical Information and Computer Sciences **38**: 395-401.

Jorgensen, W. L. and E. M. Duffy (2002). "Prediction of drug solubility from structure." Advanced Drug Delivery Reviews **54**(3): 355-366.

Lukovits, I. (2000). "A compact form of the adjacency matrix." Journal of Chemical Information and Computer Sciences **40**(5): 1147-50.

Mihalic, Z. and D. Veljan (1992). "The Distance Martrix in Chemistry." Journal of Mathematical Chemistry **11**: 223-258.

Moriguchi, I., S. Hirono, et al. (1992). "Simple Method of Calculating Octanol/Water Partition Coefficient." Chemical & Pharmaceutical Bulletin **40**(1): 127-130.

Palm, K., K. Luthman, et al. (1998). "Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors." Journal of Medicinal Chemistry **41**(27): 5382-5392.

Plavsic, D., S. Nikolic, et al. (1993). "On the Harary index for the characterization of chemical graphs." Journal of Mathematical Chemistry **12**: 235-250.

Rees, D. C., M. Congreve, et al. (2004). "Fragment-based lead discovery." Nature Reviews Drug Discovery **3**(8): 660-672.

Rucker, G. and C. Rucker (1998). "Symmetry-Aided Computation of the Detour Matrix and the Detour Index." Journal of Chemical Information and Computer Sciences **38**: 710-714.

Ruecker, G. and C. Ruecker (1993). "Counts of all walks as atomic and molecular descriptors." Journal of Chemical Information and Computer Sciences **33**(5): 683-695.

Sadowski, J. and J. Gasteiger (1993). "From atoms and bonds to three-dimensional atomic coordinates: automatic model builders." Chemical Reviews **93**(7): 2567-2581.

Schultz, H. P., E. B. Schultz, et al. (1990). "Topological organic chemistry. 2. Graph theory, matrix determinants and eigenvalues, and topological indexes of alkanes." Journal of Chemical Information and Computer Sciences **30**(1): 27-29.

Sharma, V., R. Goswami, et al. (1997). "Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies." Journal of Chemical Information and Computer Sciences **37**: 273-282.

Steinbeck, C., Y. Han, et al. (2003). "The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics." Journal of Chemical Information and Computer Sciences **43**(2): 493-500.

Todeschini, R., M. Lasagni, et al. (1994). "New molecular descriptors for 2D and 3D structures. Theory." Journal of Chemometrics **8**: 263-272.

Todeschini, R., M. Vighi, et al. (1997). "3D-modelling and prediction by WHIM descriptors. Part 8. Toxicity and physico-chemical properties of environmental priority chemicals by 2D-TI and 3D-WHIM descriptors." SAR and QSAR in Environmental Research **7**(1-4): 173-93.

Trinajstic, N., S. Nikolic, et al. (1997). "THE DETOUR MATRIX IN CHEMISTRY." Journal of Chemical Information and Computer Sciences **37**(4): 631-638.

Vellarkad, N. V., K. G. Arup, et al. (1989). "Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics." Journal of Chemical Information and Computer Sciences **29**: 163-172.

Viswanadhan, V. N., A. K. Ghose, et al. (1989). "Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics." Journal of Chemical Information and Computer Sciences **29**(3): 163-172.

Wang, R., Y. Fu, et al. (1997). "A new atom-additive method for calculating partition coefficients." Journal of Chemical Information and Computer Sciences **37**(3): 615-621.

Wang, R., Y. Gao, et al. (2000). "Calculating partition coefficient by atom-additive method." Perspectives in Drug Discovery and Design **19**: 47-66.

Wiener, H. (1947a). "Structural Determination of Paraffin Boiling Points." Journal of the American Chemical Society **69**(1): 17-20.

Wiener, H. (1947b). "Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons." Journal of the American Chemical Society **69**(11): 2636-2638.

Wiener, H. (1947c). "Influence of Interatomic Forces on Paraffin Properties." The Journal of Chemical Physics **15**: 766.

Winiwarter, S., N. M. Bonham, et al. (1998). "Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach." Journal of Medicinal Chemistry **41**(25): 4939-4949.

3. The recognition of unique compounds

As EDULISS 2.0 holds millions of compounds, it is of great interest to be able to know the number of unique compounds throughout the collection. The following sections describe procedures to identify unique compounds in EDULISS 2.0.

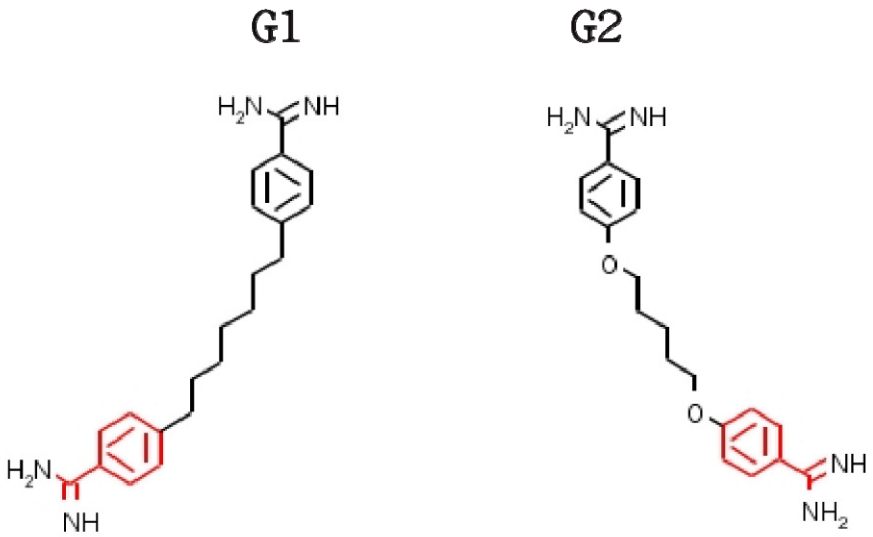
3.1 The comparison of compounds using Maximum Common Subgraph (MCS)

Conventionally, the recognition of the difference between compounds requires the pair-wise comparison of all the structures. There is a very elegant and pure 2D graph theory algorithm to perform the comparison known as Maximum Common Subgraph, MCS (Barrow and Burstall 1976; Bayada, Simpson et al. 1992; Raymond, Gardiner et al. 2002). To compare two compounds it constructs the product graphs for each of the two structures as G1 and G2 and then finds the clique which corresponds to the largest matching of the bonds and atoms between G1 and G2. A measure of the similarity between the two compound based on MCS can be given by the Tanimoto coefficient:

$$T = \frac{C}{A + B - C} \qquad 0 \le T \le 1 \qquad \textit{eq. 3.1}$$

where T is the index representing the similarity between the two structures, i.e. G1 and G2; A and B are the numbers of non-hydrogen atoms in G1 and G2 respectively; C is the number of non-hydrogen atoms in the MCS which can be found in G1 and G2. Figure 3.1 illustrates the mapping of two graphs (compounds) using MCS and the maximum common subgraphs are coloured by red. The Tanimoto coefficient is 0.2195. If G1 and G2 are entirely matched together, i.e. the coefficient is 1.00, they are considered as isomorphic.

Figure 3.1. Mapping of two graphs (compounds) using MCS. The red subgraphs are the maximal common substructure between the two compounds.

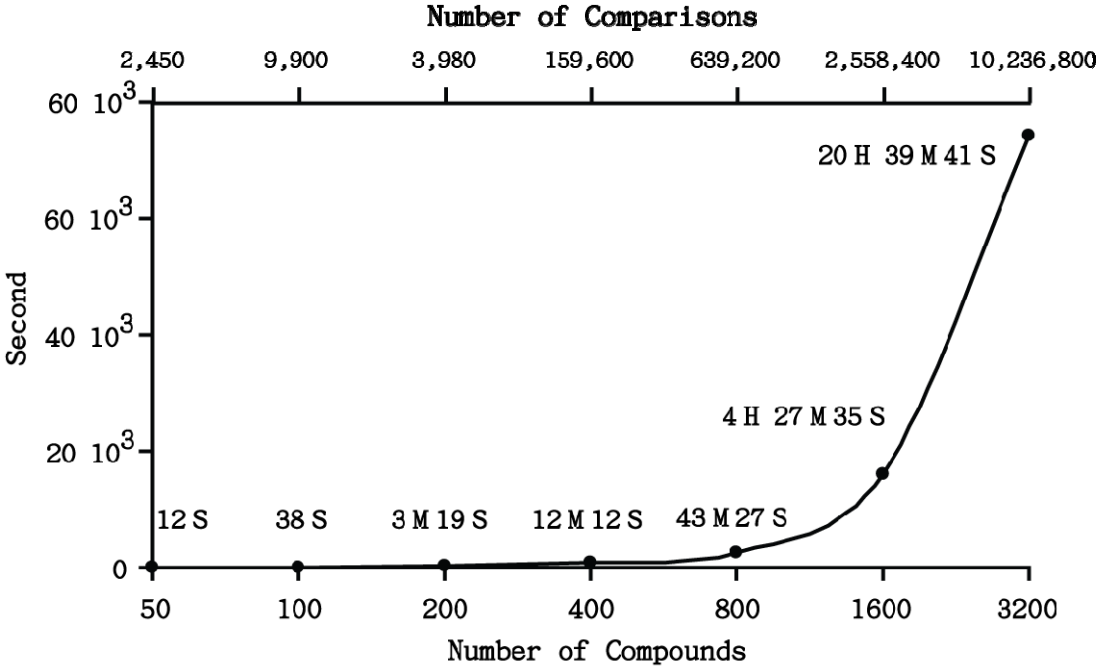


3.2 The survey of high discrimination descriptors

Although the MCS is able to identify isomorphic compounds in terms of the structure graph, the drawback of the application of MCS in finding the unique compounds from a great dataset is that the number of pair-wise comparisons can be very large as the required number is N×(N−1) where N is the number of compounds. To compare throughout the collection in EDULISS 2.0, the estimated number of the pair-wise comparison is over 3×10^{13}, as the database holds +5.5 million compounds. Figure 3.2 illustrates the requirement of the essential run time of the pair-wise comparisons for the seven datasets containing 50, 100, 200, 400, 800, 1,600 and 3,200 compounds, respectively. These comparisons were performed using an in-house programme CONCOMP written in C++ which implements the MCS theory. The tasks ran on a workstation equipped with an Intel(R) Pentium(R) D 3.00GHz CPU. As the result, the required run time grows dramatically raised in a geometric fashion from within

an hour to near a day when the dataset contains greater than 800 compounds. The rate of growth is directly proportional to the present number of pair-wise comparisons. Thus, it is impossible to go through the whole EDULISS 2.0 collection by this pure comparison.

Figure 3.2. Required run time of pair-wise comparisons for a given number of compounds. The lower X axis gives the number of compounds to be compared and the upper X axis shows the required comparisons for this number of compounds.



Inspecting the numbers of heavy atoms (i.e. C, O, N, P and S) in the examined compounds prior to performing the structure comparisons is a conventional method to improve the comparison speed. Unfortunately, there are only 5,599 compounds in EDULISS containing the numbers of these heavy atoms uniquely and so billions of comparisons are still required. If the compounds can be efficiently distinguished as unique, or characterised into clusters in which the compounds have identical specific

properties (i.e. descriptor values), the required comparisons can be considerably decreased because it is not necessary to run the inter-cluster comparisons. For example, dividing ten compounds into three clusters as 3, 3 and 4, the number of comparisons can be decreased from 90, i.e. 10×9, to 24, i.e. (3×2)+(3×2)+(4×3), times. The more the clusters, the less the comparisons are required. To implement this strategy, a survey for finding highly discrimination descriptors has been carried out. The survey uses the Sigma catalogue containing 199,267 compounds to be a training set and utilises each DRAGON descriptor, 1,664 items, to distinguish the compounds from the training set. Shown in Figure 3.3 is the result of the top 20 high discrimination descriptors in which the bars represent the numbers of unique compounds being recognised by the descriptors and the curve is the accumulation of the unique compound amount. It shows that the four descriptor groups, i.e. geometrical group (coloured by red), topological group (coloured by green), eigenvalue based indices (coloured by black) and WHIM (Weighted Holistic Invariant Molecular) descriptors (coloured by blue), exhibit the highest discriminating power among the 20 groups. The top 20 descriptors are able to recognise over 171,000 (+85 %) unique compounds in total. The most efficient item is W3D (Wiener 3D index, described in section 2.2.3.1) and is noted as point A on the curve, which is able to find over 147,000 (+73 %) unique compounds alone. From the shape of the curve, it is clear that points B and C exhibit a higher growth rate in finding unique compounds and they are Whete (topological group) and Vu (WHIM descriptors) respectively. In total, there are over 167,000 (+84 %) compounds that can be distinguished as unique since they have different values of W3D, Whete and Vu. These three descriptors can further characterise the rest of

compounds and group them into over 5,700 clusters allowing efficient use of MCS. Shown in Figure 3.4 is the distribution of the number of the rest of compounds in the clusters and the compounds in a cluster have the same values of the three descriptors. As each cluster has less than 4 compounds on average and only 20 clusters have over 50 compounds, the required pair-wise comparisons using MCS are only 310,754 in total which is less than the requirement of 800 compounds shown in Figure 3.2. The details of the Whete and Vu descriptors are described as following.

Whete: this descriptor is a Wiener-type index and is deduced from a distance matrix whose elements, i.e. the path numbers, are weighted by a series of atomic and bond parameters with the relative electronegativity values using carbon as standard. Electronegativity describes an element's ability to attract electrons and the electronegativities of main group atoms for this descriptor's calculation were estimated based on the Pauling's scale (Sanderson 1983). The fundamental of the Whete calculation is the same as the Wiener index but plus a weighting scheme for considering the presence of heteroatoms and multiple bonds. Figure 3.5 lists the atomic and bond weighting parameters used by Ivanciuc and colleagues (Ivanciuc, Ivanciuc et al. 1998), and the weighted matrix of a sample compound. In the figure, table (a) lists the atomic parameters based on relative electronegativity and listed in table (b) are the bond parameters of relative electronegativity where *i* and *j* are the identifiers of the atoms in a compound. The bond parameters are varied depending on the bond orders. During the construction of the distance matrix, the element D_{ij} of the matrix should be weighted by the atomic parameters listed in table (a) if *i* equals *j*, and weighted by the bond parameters listed in table (b) otherwise. Table (c) is the

weighted matrix of the sample compound whose structure is similar to 3-methyloctane shown in section 2.2.3.1 but it contains a double bond (C1=C2) instead of a single bond (C1−C2) and a carbon has been replaced by a heteroatom, oxygen. Thus, the elements of the matrix regarding C1 and the oxygen are weighted by the referred parameters listed in (a) and (b). As the original Wiener index only takes the path number into account for its calculation, the index values of the sample compound and 3-methyloctane both are 110 but the Whete values of these two compounds are 104.282 and 110 respectively. They can be distinguished successfully because of the consideration for the presence of the multiple bond and heteroatom in the calculation of Whete.

Vu: it is one of the descriptors in the WHIM group which was first proposed by Todeschini and colleagues (Todeschini, Lasagni et al. 1994). The indices in this group contain the information of a molecular structure in terms of its size, shape, symmetry and atom distribution and have found application in fields such as those using Quantitative Structure-Activity Relationship (QSAR; (Todeschini, Vighi et al. 1996)) or modelling physicochemical properties of organic compounds (Todeschini, Gramatica et al. 1995; Gramatica, Navas et al. 1998). They can be divided into two subgroups according to their calculation, as directional and non-directional descriptors. The procedure for the calculation of WHIM descriptors is schematically shown in Figure 3.6. The utilisation of Principal Component Analysis (PCA) is the core of WHIM's calculation. PCA is a useful and common statistical technique for finding patterns in the data of high dimension. The calculation of these descriptors is performed in the following six steps.

1. The starting point is the use of the molecular geometry, i.e. the coordinates (x, y, z), to construct an N×3 matrix where the N is the number of atoms and the three columns for x, y and z coordinates.

2. The second step is the application of a weighting scheme to characterise some particular aspects of the molecule, such as atomic mass, van der Waals, electronegativity and polarizability. A set of weighted matrices are defined and used in the later step. Similar to the weighting scheme of Whete, the weighting parameters are a series of relative values using carbon as standard and Todeschini and Gramatica have particularly tabulated the parameters in their publication (Todeschini and Gramatica 1998).

3. This step aims to centre the molecular coordinates. It makes the coordinates independent of the origin, referring other matrices a unique reference point, i.e. invariance to translation.

4. PCA is performed to diagonalise the weighted covariance matrix from the centred molecular matrix which is calculated by the decomposition process of eigenvalue and eigenvector. The PCA can also lead to three principal component axes for the next step.

5. There is a set of new atomic coordinates obtained by projecting the old axes on to the three principal component axes. The new atomic coordinates construct a score matrix which is the same dimension as the original molecular matrix, i.e. invariance to rotation.

6. Finally, for each weighting scheme a number of descriptors are calculated from each component. According to the decomposition results of eigenvalues and eigenvector along each component, the statistical parameters are able to represent the characteristics of the compound. In the directional group, when the descriptors are directly defined by the eigenvalues λ1, λ2, λ3 and represent the variances of the atoms along each component, they are related to the molecular size. If the descriptors are constituted by the eigenvalue proportions, they represent a global view of the molecular shape. When the descriptors are calculated from an information content index on the symmetry along each component, they represent the symmetry of the compound. Furthermore, the tendency of data points, i.e. the atom projections, around the origin and along the principal axes can represent the atom distribution. The descriptors in the non-directional group are directly derived from the directional group. Vu belongs to the non-directional group and obtained from an unweighted matrix. This descriptor is deduced from the eigenvalues λ1, λ2 and λ3 in three different ways shown as below:

$$T = \lambda 1 + \lambda 2 + \lambda 3$$

$$A = \lambda 1\lambda 2 + \lambda 1\lambda 3 + \lambda 2\lambda 3$$

$$\mathrm{Vu} = T + A + \lambda 1\lambda 2\lambda 3$$

The *T* and *A* are related to linear and quadratic contributions respectively and the Vu considers the *T*, *A* and the third-order term. Thus a complete expression of the molecular size can be represented by this index.

Figure 3.3. Top 20 items of the result of high discrimination descriptor survey. The bars are the numbers of the unique compounds found by referred descriptors. The curve displays the accumulation of recognised unique compounds.

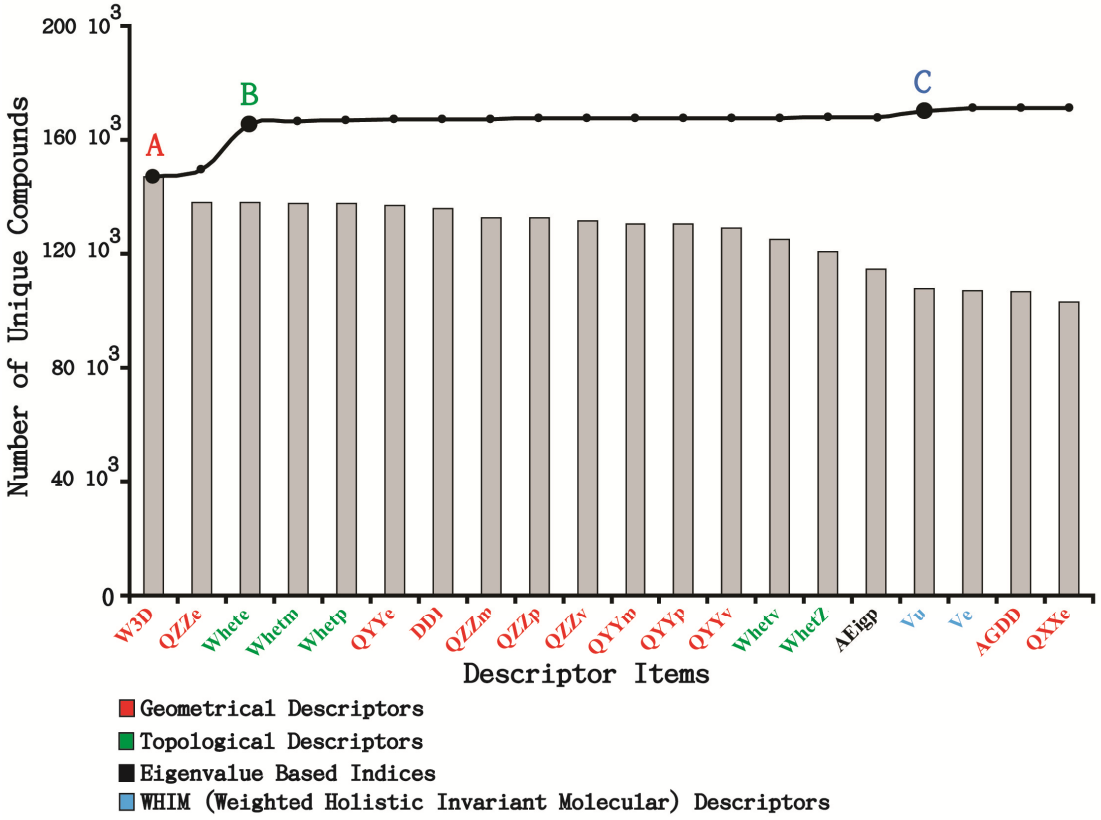


Figure 3.4. Distribution of the number of compounds in each cluster in which the compounds have the same values of the three descriptors, i.e. W3D, Whete and Vu.

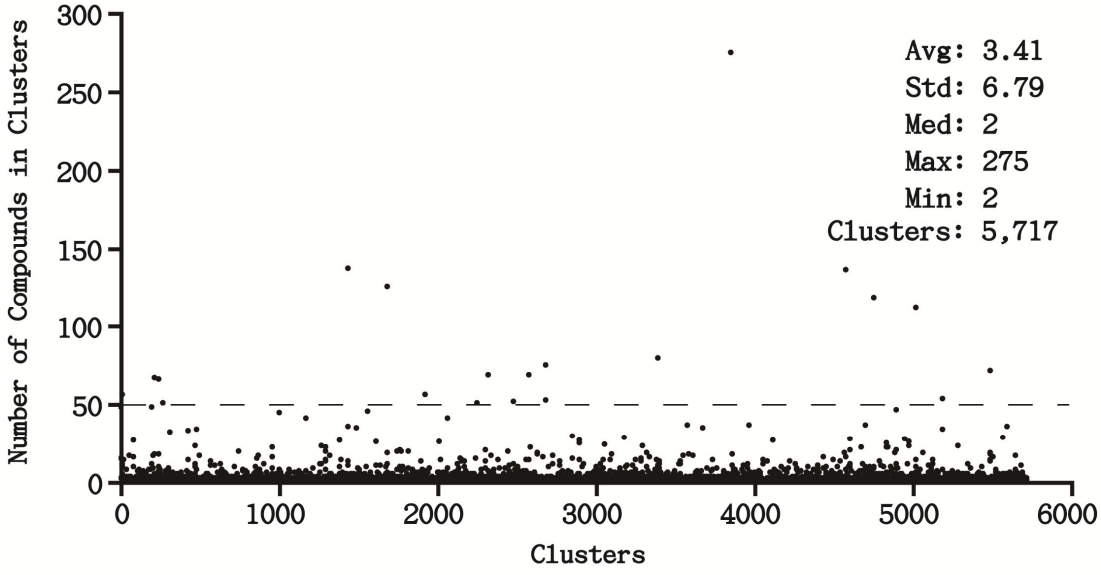


Figure 3.5. Example of the weighted distance matrix for the Whete's calculation. (a) Atomic parameters based on relative electronegativity (denoted as X). (b) Bond parameters based on relative electronegativity where *i* and *j* are the identifiers of the atoms in a compound. (c) The weighted matrix of the sample compound which contains a double bond (C1=C2) and a heteroatom, oxygen.

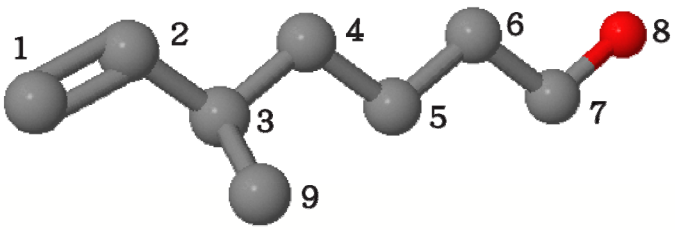
a

Atom	AwX
B	-0.175
C	0.000
N	0.130
O	0.229
F	0.308
Si	-0.067
P	0.079
S	0.190
Cl	0.277
As	-0.057
Se	0.087
Br	0.196
Te	-0.048
I	0.093

b

Atom*i*	Atom*j*	Single	Double	Triple	Aromatic
C	C	1.000	0.500	0.333	0.667
C	N	0.870	0.435	0.290	0.580
C	O	0.771	0.386		
C	F	0.692			
C	Si	1.067			
C	P	0.921			
C	S	0.810	0.405		0.540
C	Cl	0.723			
C	Se	0.913			
C	Br	0.804			
C	Te	1.048			
C	I	0.907			
N	N	0.757	0.379		0.505
N	O	0.671	0.336		0.447
O	S	0.624	0.312		

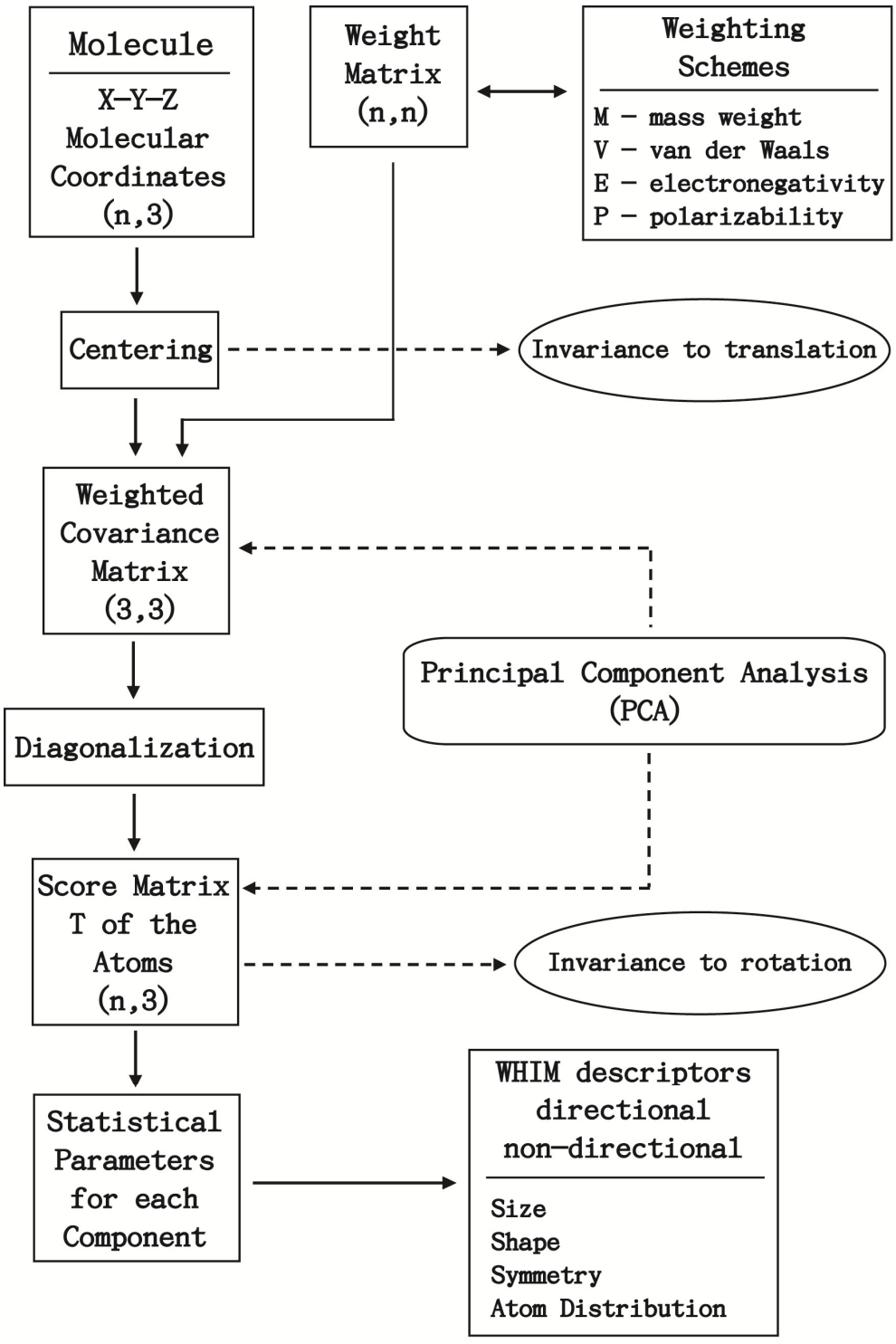
(Ivanciuc, Ivanciuc et al. 1998)



c

Atom	1	2	3	4	5	6	7	8	9
1	0	0.5	1.5	2.5	3.5	4.5	5.5	6.271	2.5
2	0.5	0	1	2	3	4	5	5.771	2
3	1.5	1	0	1	2	3	4	4.771	1
4	2.5	2	1	0	1	2	3	3.771	2
5	3.5	3	2	1	0	1	2	2.771	3
6	4.5	4	3	2	1	0	1	1.771	4
7	5.5	5	4	3	2	1	0	0.771	5
8	6.271	5.771	4.771	3.771	2.771	1.771	0.771	0.229	5.771
9	2.5	2	1	2	3	4	5	5.771	0

Figure 3.6. Flow chart of the procedure for the calculation of the WHIM descriptors.



(Todeschini and Gramatica 1998)

3.3 Result

Utilising these high discrimination descriptors, i.e. W3D, Whete and Vu, with the clustering approach in the whole EDULISS 2.0 collection, these three descriptors are able to find 3,117,625 unique compounds prior to performing MCS comparisons and characterise the rest into 845,193 clusters. The number of required pair-wise comparisons using MCS is 6,495,096. Thus the recognition of unique compounds in EDULISS 2.0 can be done very efficiently and the total number of unique compounds present is 4,011,697. Additionally, in these +4 million unique compounds with distinct molecular formula and molecular weight are only 328,902 (8.2 % out of +4 M) and 321,520 (8 % out of +4 M) respectively. In the final analysis, although inspecting the atomic types or examining the molecular weights can recognise about 8 % of unique compounds prior to performing pair-wise comparison, we did not apply these two attributes in coordination with the molecular compactness, size and atomic parameters in the recognition of unique compounds.

3.4 Reference:

Barrow, H. G. and R. M. Burstall (1976). "Subgraph Isomorphism, Matching Relational Structures and Maximal Cliques." Information Processing Letters **4**(4): 83-84.

Bayada, D. M., R. W. Simpson, et al. (1992). "An algorithm for the multiple common subgraph problem." Journal of Chemical Information and Computer Sciences **32**(6): 680-685.

Gramatica, P., N. Navas, et al. (1998). "3D-modelling and prediction by WHIM descriptors. Part 9. Chromatographic relative retention time and physico-chemical properties of polychlorinated biphenyls (PCBs)." Chemometrics and Intelligent Laboratory Systems **40**(1): 53-63.

Raymond, J. W., E. J. Gardiner, et al. (2002). "RASCAL: Calculation of Graph Similarity using Maximum Common Edge Subgraphs." The Computer Journal **45**(6): 631-644.

Sanderson, R. T. (1983). "Electronegativity and bond energy." Journal of the American Chemical Society **105**(8): 2259-2261.

Todeschini, R. and P. Gramatica (1998). "New 3 D molecular descriptors: The WHIM theory and QSAR applications." Perspectives in Drug Discovery and Design **9**: 355-380.

Todeschini, R., P. Gramatica, et al. (1995). "Weighted holistic invariant molecular descriptors. Part 2. Theory development and applications on modeling physicochemical properties of polyaromatic hydrocarbons." Chemometrics and Intelligent Laboratory Systems **27**(2): 221-229.

Todeschini, R., M. Lasagni, et al. (1994). "New molecular descriptors for 2D and 3D structures. Theory." Journal of Chemometrics **8**: 263-272

Todeschini, R., M. Vighi, et al. (1996). "Modeling and prediction by using WHIM descriptors in QSAR studies: toxicity of heterogeneous chemicals on Daphnia magna." Chemosphere **32**(8): 1527-1545.

4. The correlation between molecular descriptors

This chapter presents an investigation of the correlation between molecular descriptors, using the descriptors that have been calculated by DRAGON software. As the number of descriptors is large (1,664 items), this study mainly focuses on the descriptors which have been implicated in and frequently discussed in the studies of QSAR, and uses a statistical approach to reveal the relationship between them. The following sections describe the procedures and results of this investigation.

4.1 Materials and methods

The Pearson correlation coefficient (γ) has been applied in this study. It is the most commonly utilised measure of correlation between a pair of interval- or ratio-scaling variables (Miller 1996). As the DRAGON descriptors are mainly expressed by interval scaling or calculated using these scaling concepts, the coefficient is proper for this study. The correlation coefficient γ of variables x and y is defined as:

$$\gamma(x,y) = \frac{\sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{N}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{N}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{N}\right)}} \qquad \textit{eq. 4.1}$$

where the *N* is the number of subjects, i.e. the number of used compounds in the present study. The γ(x, y) value varies between +1 and -1. The size of the coefficient value indicates the degree of the relationship between two variables and its sign represents the direction of the relationship. A positive value means the direct correlation between x and y, and a negative γ indicates the reverse pattern, and γ of zero is obtained when there is no linear relationship between x and y.

The calculation time for the entire database would be unacceptably long and therefore a test set was taken from the EDULISS collection by random selection of 500,000 entries (i.e. compounds), each entry having the values of 1,664 DRAGON descriptos. The statistical analysis was carried out using the SAS 9.1 statistical package (SAS 2004). Table 4.1 is the profile of some general descriptors of the test set. Comparing this table to the molecular property profiles of the EDULISS database shown in Figure 2.10 and Table 2.2, the general distribution of this test set is consistent with the whole EDULISS collection. Thus, the test set is qualified to represent the molecular properties in the database for this study.

Table 4.1. Descriptor ranges for the test set of 500,000 compounds*.

Descriptor	Max	Min	Average	Standard deviation
Molecule weight	2,834	41.06	371.45	94.80
Number of atoms	395	4	44.58	12.11
Number of bonds	399	3	46.72	12.88
Number of rings	24	0	3.17	1.13
Number of oxygen atoms	61	0	3.08	1.63
Number of nitrogen atoms	31	0	2.71	1.38
Number of HAcc	88	0	5.34	2.11
Number of HDon	54	0	1.67	0.96
MLogP	78.12	-19.25	3.68	2.72
Wiener 3D index (Å)	1,293,163	9.81	6,781	6,591

*: HAcc: hydrogen bond acceptors; HDon: hydrogen bond donors.

4.2 Results and discussions

4.2.1 The correlations of overall descriptors

Figure 4.1 shows a colour scale image of the correlation matrix for each descriptor pair where blue and red represent the direct ($0<\gamma\leq1$) and reverse ($0>\gamma\geq-1$) correlation, respectively. Each pixel represents a correlation coefficient and the main diagonal the coefficients of the descriptors correlated to themselves, i.e. $\gamma(x, x) = 1$. Other pairs of coefficients are counted only once, diagonal elements are excluded (i.e. used $\gamma(x, y)$ and excluded $\gamma(y, x)$). Thus, a total of $(1664^2)/2 = 1{,}384{,}448$ values are considered. Obviously, the dimension of the matrix is too high to detail the analysis of the relationship between each pair of descriptors. However, the matrix is able to indicate that some descriptors are highly correlated to others.

Figure 4.1. Colour scale image of the correlation matrix for each pair of descriptors.

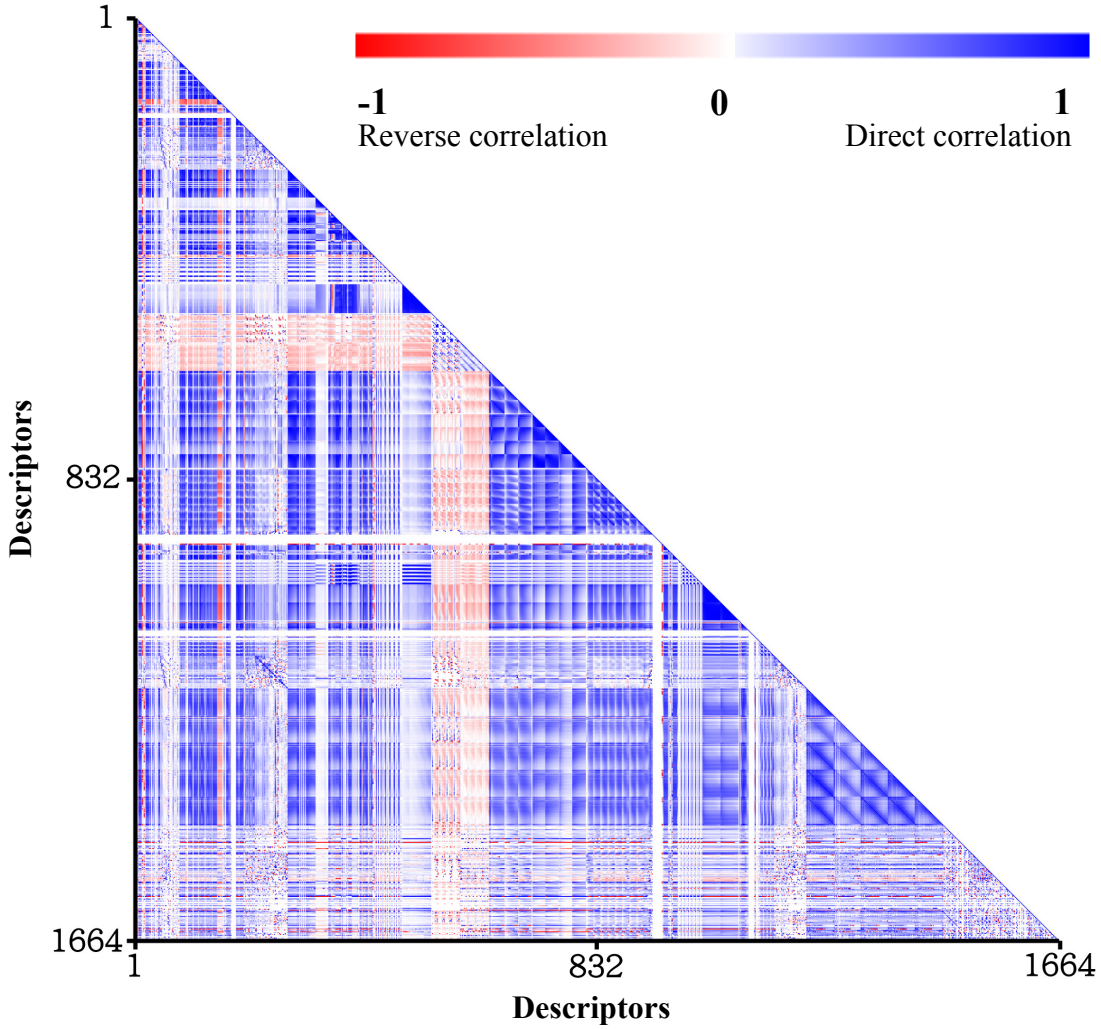


Figure 4.2 shows a histogram of the distribution of correlation coefficients. Bars are filled with the same colour scale as used in Figure 4.1. The observations of diagonal elements and $\gamma(x, x)$ are excluded, i.e. $(1664^2-1664)/2 = 1{,}383{,}616$ values are used. In the cases of direct correlation (i.e. blue side) the mean and median of coefficients are 0.19 and 0.07, and in reverse correlation (i.e. red side) the values are -0.15 and -0.09 respectively. The overall distribution shows that most descriptors have little correlation to others. However, there are still over 20,000 pairs of descriptors that present strong linear relationships, i.e. $\gamma(x, y) \geq |0.9|$. They mainly tend to occur in

the same descriptor classes. As DRAGON software has logically separated the descriptors into twenty groups, the high correlation tendency, for example, can be easily seen by Figure 4.3 which shows the intercorrelations of two descriptor groups including Burden eigenvalues (64 items) and Randic molecular profiles (41 items). Burden eigenvalue descriptors are obtained from the positive and negative eigenvalues of a molecular adjacency matrix, weighting the diagonal elements by a series of atomic parameters (Burden 1989; Burden 1997). Randic molecular profiles are derived from the geometry matrix of a compound, defined as the average row sum of its entries raised at the k^{th} power and normalise the averages by the factorial k! in turn (Randic 1995). The DRAGON software computes Randic molecular profiles for k = 1 to 20. As a result, the descriptors belonging to the same group present high correlations between themselves, as their calculations are often based on the same or similar principle and weighted by certain parameters iteratively. The group of Burden eigenvalues is an instance which uses atomic masses, van der Waals volumes, electronegativities and polarizabilities to weight elements of the adjacency matrix. Consequently, Figure 4.3 (a) shows a clear correlation gradation.

Figure 4.2. Histogram of correlation coefficients. Bars are filled with the same colour scale as used in Figure 4.1. The number of used observations is 1,383,616 (i.e. 1664^2-1664)/2) as the diagonal elements and γ(x, x) are excluded. The vertical axis represents the number of descriptor pairs. The horizontal axis is the scales of correlation coefficients.

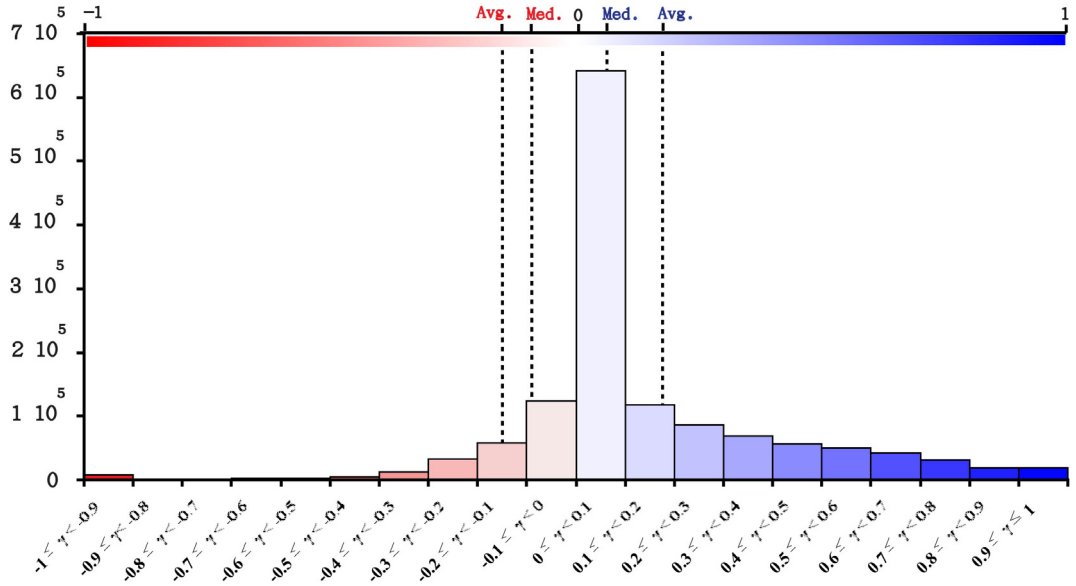
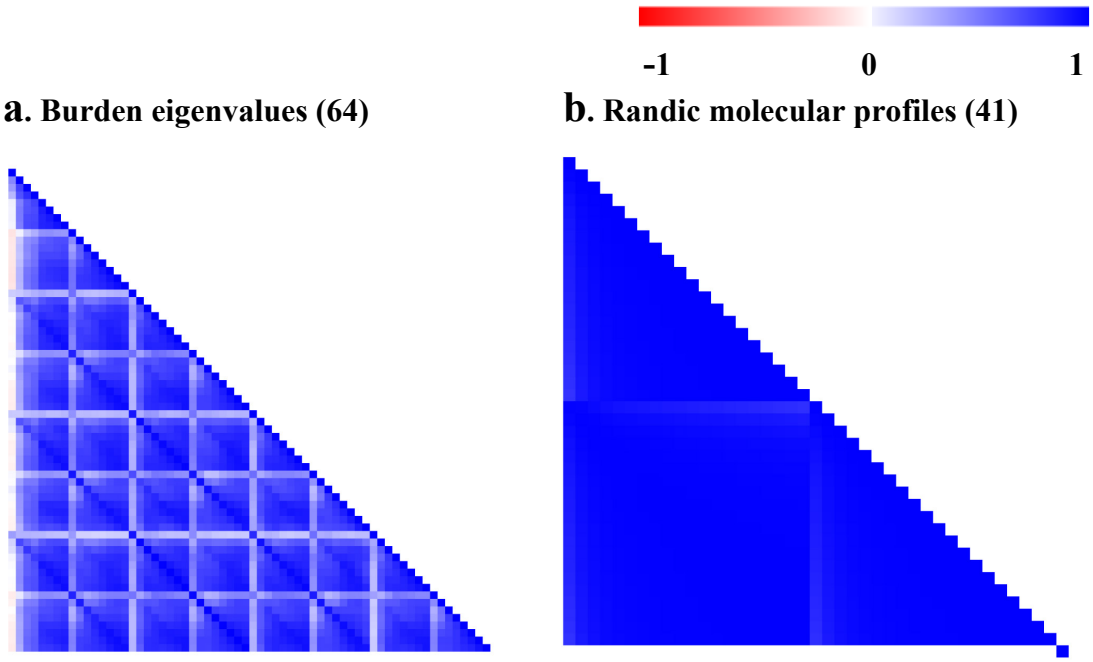


Figure 4.3. Colour scale image of the correlation matrix of two groups of descriptors.



4.2.2 The correlations of some fundamental descriptors

Figure 4.4 is a matrix excerpted from Figure 4.1 to show the correlations between 41 fundamental descriptors and their introductions are printed in gray. The descriptors include the groups of simple counts, some atomic properties (van der Waals volumes, electronegativities, polarizabilities), shape-based descriptors, physicochemical traits and biological properties (MLogP and ALogP). The details and calculations of W, W3D, Wiener-type (i.e. Whetm, Whetv, Whete and Whetp) indices and WHIM (i.e. Vu, Vm, Vv, Ve, Vp and Vs) descriptors have been described in previous chapter. The Ss descriptor, Kier-Hall electrotopological states or E-states, was first proposed by Hall and colleagues (Hall, Mohney et al. 1991) and were developed from graph theory as an index of the molecular skeletal group and combines both the electronic character and the 2D feature of each skeletal atom in a molecule. It has been tested and applied in the field of QSAR (Butina 2004).

As expected, molecular weight is directly correlated to the number of non-H atoms ($\gamma = 0.95$) and bonds ($\gamma = 0.84$) as well as with the general atomic properties such as Sv ($\gamma = 0.91$), Se ($\gamma = 0.85$), Sp ($\gamma = 0.90$) and Ss ($\gamma = 0.88$). The "size" of a molecule does not absolutely determine its molecular weight as the coefficient of γ(MW, Vu) is 0.37 but the "compactness" of a molecule displays relatively higher correlations to MW as γ(MW, W) = 0.69 and γ(MW, W3D) = 0.65. Previous studies indicated that the number of rotatable bonds increases with molecular weight (Ajay, Walters et al. 1998) whereas it shows a median correlation in this study as γ(MW, RBN) = 0.55. The number of hydrogens gives higher contributions to the sums of Sv, Se and Sp than the number of heavy atoms.

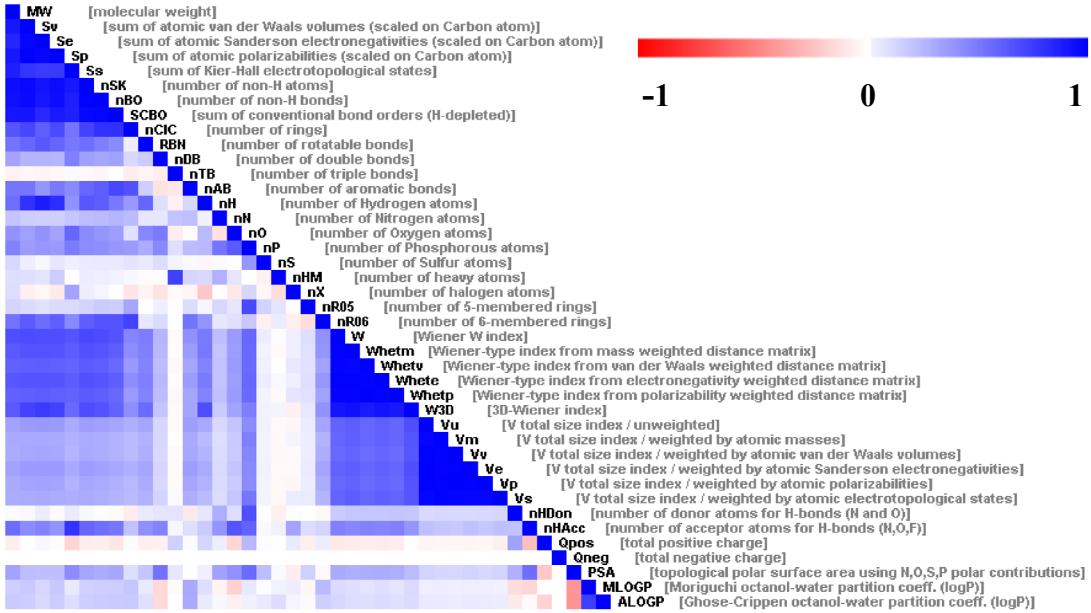
Comparing the values of γ(nCIC, nR06) and γ(nCIC, nR05), the correlation coefficient between the number of rings and 6-membered rings is two-fold greater than the coefficient of 5-membered ring pair as 0.74 vs. 0.35. It is consistent with the molecular profile shown in Figure 2.10 (f), as in the EDULISS collection, the rings in a compound are mainly 6-membered and then 5-membered.

In the case of inter-correlations, the descriptors of Wiener-type indices appear to be in a statistically significant relationship between each other though they have been weighted by various parameters respectively. The descriptors of the WHIM group have a similar tendency. On the other hand, as the descriptors of Wiener-type indices and WHIM group are deduced from 2D and 3D matrices respectively, the pair correlations between them are 0.62 on average. It tends to agree with previous research which suggested that the geometrical descriptors as a class are not very different from the topological indices (Braun, Kerber et al. 2004).

The five descriptors that include calculated values of octanol/water partition coefficients (i.e. MLogP and ALogP described in previous chapter), PSA, nHAcc and nHDon, have shown to be highly interesting for drug discovery and development as these parameters are easy to understand and are known to be related to the molecular transport properties, particularly intestinal absorption and blood-brain barrier penetration (Blake 2000; Clark and Pickett 2000). The PSA mentioned here is a simple topology based method and is calculated from the summation of the surface contributions of 43 polar fragments noting the atom types and their bonding pattern, including max bond order, sum of bond order, number of neighbours, hydrogen count, formal charge, aromatic bonds, number of 3-membered rings,

numbers of single-bonds, double-bonds and triple-bonds (Ertl, Rohde et al. 2000). The descriptors regarding the molecular weight, "size" and "compactness" show poor correlations to these five descriptors. The correlation coefficients are from γ(Vu, ALogP) = 0.09 to γ(MW, nHAcc) = 0.58. The number of hydrogen bond donor (nHDon) and acceptor (nHAcc) do not simply depend on the number of oxygen and nitrogen as their definitions take the adjacency of atoms and atomic formal charge into account. Similarly, the total positive charge (Qpos) exhibits a reverse correlation to nHAcc and direct to nHDon. As the nHDon exhibits a lower correlation to the polar surface area than nHAcc (γ(PSA, nHDon) = 0.31 and γ(PSA, nHAcc) = 0.57) and the Qpos is also reversely correlated to PSA, it could be presumed that a fragment with positive charge gives lower contribution to form the PSA. In the cases of MLogP and ALogP, they display the reverse correlation to PSA as the Log P coefficients use negative values to express the higher solubility of a compound.

Figure 4.4. Colour scale image of the correlation matrix for each pair of 41 descriptors.



4.2.3 The correlations between five selected descriptors and 151 molecular functional groups

Table 4.2 summarises the statistical profile of absolute values of correlation coefficients between the five selected descriptors, including nHAcc, nHDon, PSA, MLogP and ALogP, and the counting of 151 chemical functional groups. The 151 items and their 2D pictures with introductions can be found on the DRAGON website (http://www.talete.mi.it/help/dragon_help/index.html?ListTopolDesc). There are only three items which show linear correlation with the five descriptors, including number of aromatic dithioesters (nArCSSR), number of thiophosphates (nPO4) and number of phosphanes (nPR3).

Table 4.2. Statistical profile of absolute values of correlation coefficients between the five selected descriptors, including nHAcc, nHDon, PSA, MLogP and ALogP, and the counting of 151 chemical functional groups.

correlated pairs*	Max	Min	Average	Standard deviation
$\|\gamma$(nHAcc, 151 functional groups)$\|$	0.91	0	0.14	0.17
$\|\gamma$(nHDon, 151 functional groups)$\|$	0.80	0	0.12	0.16
$\|\gamma$(PSA, 151 functional groups)$\|$	0.96	0	0.12	0.17
$\|\gamma$(MLogP, 151 functional groups)$\|$	0.78	0	0.07	0.12
$\|\gamma$(ALogP, 151 functional groups)$\|$	0.75	0	0.06	0.12

*: each entry in this column contains 151 observations, e.g. nHAcc against 151 functional groups.

nHAcc: number of hydrogen bond acceptors; nHDon: number of hydrogen bond donors; PSA: topological polar surface area using N, O, S and P polar contributions.

The higher correlations occur in the pairs of following:

nHAcc vs. nArCSSR (γ = 0.91)

nHAcc vs. nPO4 (γ = 0.83)

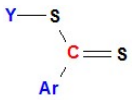
nHDon vs. nPO4 (γ = 0.80)

PSA vs. nArCSSR (γ = 0.96)

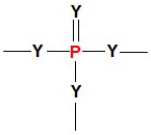
PSA vs. nPO4 (γ = 0.85)

MLogP vs. nPR3 (γ = 0.78)

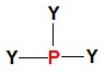
ALogP vs. nPR3 (γ = 0.75).



nArCSSR:
Y = aromatic or aliphatic group linked through C



nPO4:
Y = O or S



nPR3:
Y = H, C, halogens

The nArCSSR and nPO4 groups make higher contributions to hydrogen bonding and polar surface, and the nPR3 group tends to affect the calculated Log P values of compounds.

It is generally thought that the variety of molecular toxicity mainly depends on the nature and location of functional groups (Turabekova and Rasulev 2005). In particular, the calculated Log P value has been used as a principal descriptor of lipophilicty or hydrophobicity as well as related to the biological properties of a compound such as toxicity because the experimental measurement of Log P values is expensive, time consuming and labour intensive. At present, the most common and accepted approach to calculate Log P values is the fragmental or additive method (Thompson, Hattotuwagama et al. 2006). MLogP and ALogP are examples of this type as their Log P values are obtained by summing the contributions of each

particular fragment (functional groups). However, apart from the items mentioned above, none of other functional groups is statistically correlated to the calculated Log P values and the other four descriptors according to the results. It could be presumed that the relationship between functional groups and these five descriptors does not only simply involve the count but also other factors, such as their location in a compound.

4.3 Reference:

Ajay, A., W. P. Walters, et al. (1998). "Can we learn to distinguish between" drug-like" and" nondrug-like" molecules?" Journal of Medicinal Chemistry **41**(18): 3314.

Blake, J. F. (2000). "Chemoinformatics predicting the physicochemical properties of rug-like molecules." Current Opinion in Biotechnology **11**(1): 104-107.

Braun, J., A. Kerber, et al. (2005). "Similarity of molecular descriptors: The equivalence of Zagreb indices and walk counts". Communications in Mathematical and in Computer Chemistry. **54:** 163-176.

Burden, F. R. (1989). "Molecular identification number for substructure searches." Journal of Chemical Information and Computer Sciences **29**(3): 225-227.

Burden, F. R. (1997). "A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix." Quantitative Structure-Activity Relationships **16**(4): 309-314.

Butina, D. (2004). "Performance of Kier-Hall E-state Descriptors in Quantitative. Structure Activity Relationship (QSAR) Studies of. Multifunctional Molecules." Molecules **9**(12): 1004-1009.

Clark, D. E. and S. D. Pickett (2000). "Computational methods for the prediction of 'drug-likeness'." Drug Discovery Today **5**(2): 49-58.

Ertl, P., B. Rohde, et al. (2000). "Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties." Journal of Medicinal Chemistry **43**(20): 3714-3717.

Hall, L. H., B. Mohney, et al. (1991). "The electrotopological state: structure information at the atomic level for molecular graphs." Journal of Chemical Information and Computer Sciences **31**(1): 76-82.

Miller, S. (1996). Experimental Design and Statistics. London and New York, Routledge.

Randic, M. (1995). "Molecular Shape Profiles." Journal of Chemical Information and Computer Sciences **35**(3): 373-382.

SAS (2004). SAS Institute Inc., Cary, NC, USA.

Thompson, S. J., C. K. Hattotuwagama, et al. (2006). "On the hydrophobicity of peptides: Comparing empirical predictions of peptide log P values." Bioinformation **7**(1): 237-241.

Turabekova, M. A. and B. F. Rasulev (2005). "QSAR Analysis of the Structure Toxicity Relationship of Aconitum and Delphinium Diterpene Alkaloids." Chemistry of Natural Compounds **41**(2): 213-219.

5. The application of Structure-Activity Relationship (SAR) in ligand-protein binding study

This chapter presents a study on modelled prediction of ligand-protein binding by a statistical method which is based on Structure-Activity Relationship (SAR) modelling. The construction of the model involves comparison of known experimental ligand-protein binding outcomes against a series of molecular descriptors. The model aims to reveal those molecular descriptors which are correlated with ligand activity, and to predict the probability of ligand-protein binding.

PubChem bioassay data, which is an NMR based screening assay for a human FKBP12 protein (PubChem AID: 608), has been used in this study. The depositor source of this bioassay was provided by the Burnham Center for Chemical Genomics (BCCG; http://sdccg.burnham.org/metadot/index.pl) which was formerly the San Diego Center for Chemical Genomics.

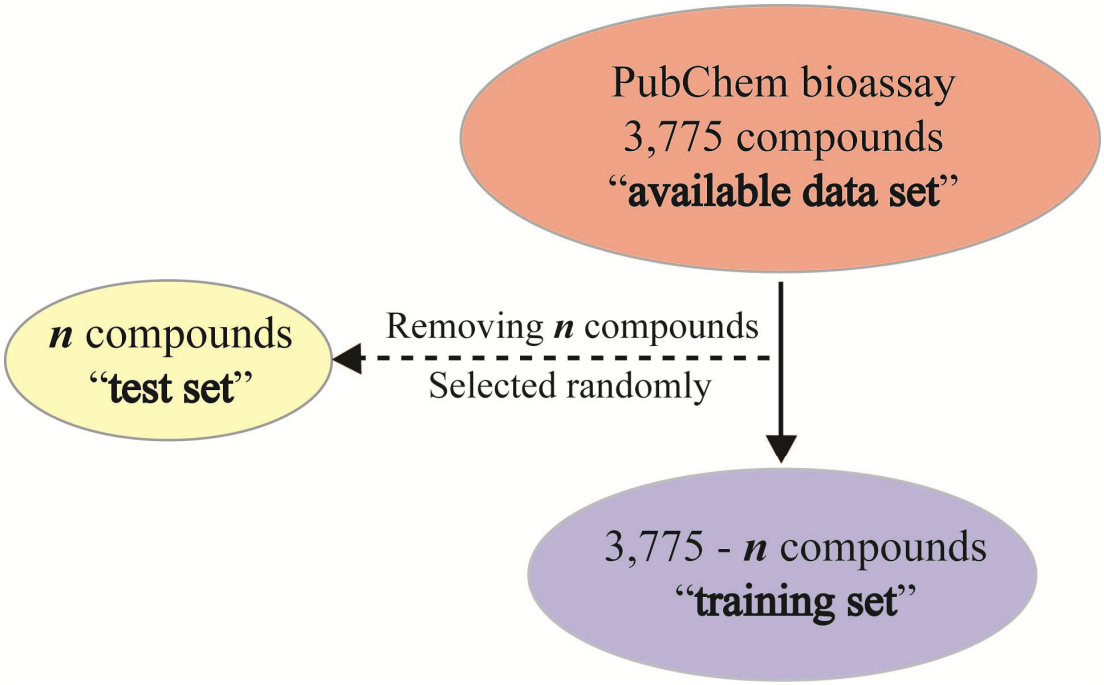
FKBP12 is critical for life and occurs in high concentrations in all cells. Thus, it can be expected to control fundamental functions of cell biology. An *in vivo* experiment shows that the mutant mice died during the embryonic stage when FKBP12 had been deleted (Shou, Aghdasi et al. 1998). Furthermore, FKBP12 has been identified as a receptor to the immunosuppressant drug FK506 (Liu, Farmer et al. 1991; Aghdasi, Ye et al. 2001). This protein catalyses the *cis-trans* transition of the peptide bond at proline residues, which is a rate-limiting step in protein folding (Kay 1996; Iida, Furutani et al. 1998).

In the PubChem bioassay, the human FKBP12 was expressed in an *Escherichia coli* strain. There were 3,775 compounds assembled for the NMR based screening. The compounds were selected from four different sources, including a set of building blocks or scaffold compounds from MayBridge (http://www.maybridge.com/), Life Chemicals (http://www.lifechemicals.com/), ChemBridge (http://www.chembridge.com/) and a collection of 602 natural products from NIH Molecular Libraries Screening Centers Network (MLSCN; http://mli.nih.gov/mli/mlscn/index.php). Despite the presence of many natural products, the average overall molecular diversity of other sets is greater than 86 % measured by the UNITY module of SYBYL software (UNITY Tripos Inc.) (Stebbins, Zhang et al. 2007). The compound activity is classified by the binding affinity (estimated dissociation constant, Kd) of the compound against the target. Compound Kd values less than 500 μM are treated as active, otherwise they are classified as inactive. The detail and protocol of this NMR based screening assay can be found on PubChem website (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=608). The final result of this screening was that only 44 compounds (structurally shown in Appendix 1) were identified as active and 3,731 as inactive.

There are two statistical approaches applied in the present study for different purposes: Logistic Regression Model and Self Organising Maps (SOMs). The PubChem data set is split into a training set and a test set and they are defined in Figure 5.1. The compounds assembled for the PubChem bioassay are called the "available data set" (shown as red in Figure 5.1); the compounds randomly removed from the available data set for model validation are called the "test set" (shown as yellow); the remaining compounds for regression model building are called the

"training set" (shown as blue). These terms will be used consistently in the subsequent sections. The molecular properties (i.e. descriptor values) of the compounds involved in these analyses were calculated using the DRAGON programme. The two statistical approaches are described in the following sections 5.1 and 5.2.

Figure 5.1. Terms of data sets for modelling exercises. The compounds assembled for the PubChem bioassay are called the "available data set" (shown as red); the randomly removed compounds for model validation are called the "test set" (shown as yellow); the remaining compounds for regression model building are called the training set (shown as blue).

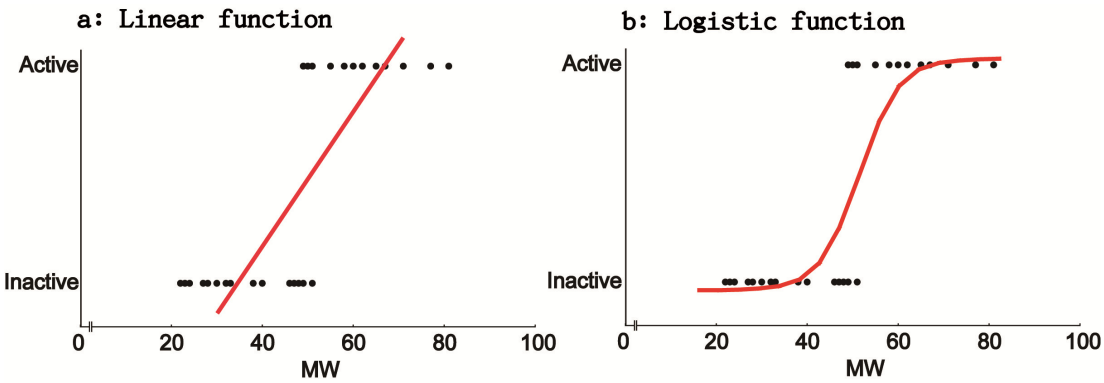


5.1 Logistic Regression Model

Logistic regression characterises the relationship between a response variable and a series of predictor variables. As the PubChem bioassay has examined and defined the compounds as active or inactive, i.e. a binary format, the Logistic Regression Method is adequate for modelling the probability of the compound's bioactivity. In

this binary response model, the response, Y, of the experimental outcome means one of two possible values which are denoted as 1 and 0 conventionally. For example, $Y = 1$ if a compound is active, whereas $Y = 0$ if a compound is inactive. Assuming X' is a vector of independent variables, i.e. the molecular descriptors, then $P = (Y = 1 \mid X')$ is the probability of response $Y = 1$ to be predicted. Unlike general linear functions that would give a bad fit (Figure 5.2 (a)), the Logistic Regression Model provides an S-shape function so that the fit can be largely improved (Figure 5.2 (b)) when the Y axis represents a binary outcome, such as Yes/No or Active/Inactive, and the X axis represents an independent variable, such as molecule weight. This regression model also ensures that whatever evaluation of the response one obtains, it always gives a value to represent the calculated probability between 0 and 1 that can be easily converted into a binary response using a proper threshold value such as 0.5.

Figure 5.2. Schematic plots of general linear and logistic functions. The Y axes represent a binary response, such as the activity of the compounds. The X axes represent an independent variable, such as the molecule weight.



The logistic regression is modelled in the form:

$$P = (Y = 1 \mid X') = \frac{e^{f(x)}}{1 + e^{f(x)}} \qquad \text{eq. 5.1}$$

$$f(x) = \beta_0 + \beta_1 \chi_1 + \beta_2 \chi_2 + \beta_3 \chi_3 + ...\beta_n \chi_n \qquad \textit{eq. 5.2}$$

where the letter *P* is the event probability. The letters β_0 to β_n are the regression coefficients and χ_1 to χ_n are the molecular descriptors. During the model building process, the most predictive molecular descriptors are first identified from the quality of a single descriptor logistic model according to the p-values which were estimated for each of the descriptors. The p-value indicates the significance level for the Wald Chi-Square Test which is used to examine whether the independent variable (i.e. the descriptor) significantly affects the dependent variable (i.e. compound activity) or not (Peng, Lee et al. 2002). Then, the model omitted the descriptors whose p-value was greater than 0.05 from further analysis. The remaining potentially predictive descriptors were used in the procedure of stepwise selection. The selection strategy starts from the inclusion of the most significant predictive descriptor in a regression model and sequentially applies the next one to obtain the most appropriate combination of descriptors and the best model (instanced as Figure 5.5). The selection aims to improve the prediction of the model and to reduce the number of predictors in a model.

To check the predictive ability of the model we can apply it to data unused in the model building, using a separate hold-out test sample or leave-one-out cross-

validation standard method to validate this technique (Hawkins, Basak et al. 2003). As the leave-one-out cross-validation is much more computationally intensive involving repeating the full modelling analysis a total of sample number plus one times, i.e. n+1, this study adopts the hold-out sample treatment (shown as Figure 5.1). The statistical analysis was carried out using the SAS 9.1 statistical package (SAS 2004) and the sampling designs for modelling exercise are described in sections 5.3.1 and 5.3.3.

5.2 Self Organising Maps (SOMs)

The SOM is a data visualisation and clustering technique developed by Teuvo Kohonen (Kohonen 1997). As it is difficult to visualise high dimensional data, SOM's strategy is to reduce data dimensions and to produce a map usually in 1 or 2 dimensions to plot the similarities of the observations through the application of self-organising neural networks. The technique of SOMs has been applied in many different fields, such as its utilisation in revealing distinct gene expression patterns, the application in various microarray experiments, the classification of cancers and studies of social science (Roussinov and Chen 1998; Khan, Wei et al. 2001; Nguyen and Rocke 2002; Wang, Delabie et al. 2002).

Figure 5.3 is an example of using SOM to cluster randomly coloured squares to illustrate the idea of a SOM. The main purpose of SOM is to project high dimensional data on a lower dimensional map and reveal the similarity of the data. In order to easily explain its schema, all samples (i.e. the squares shown in Figure 5.3 (a)) have been presented in colour. Thus, the three dimensional data for each sample is represented by the variables of red, blue and green. For instance, the green square

with a black border shown in Figure 5.3 (a) is composed of the three RGB values (R: 0.01, G: 0.79 and B: 0.23). After processing using the SOM algorithm, the multidimensional data can be projected onto a lower order surface to facilitate visualisation, for example, a two dimensional image map similar to the case of Figure 5.3 (c). The generation of the SOM in the present study was carried out by an in-house perl script and the process to construct the examples shown in Figure 5.3 is described by the following steps.

1. The data of each sample is normalised into one value range if the attributes of the data are diverse, such as the molecular weight and the number of atoms of a compound. For demonstration, the 100 samples shown in Figure 5.3 (a) were randomly represented in colour and this normalisation step can be omitted as the RGB values are arranged equivalently between 0 and 1.

2. An initialised map is constructed as shown in Figure 5.3 (b). The squares of this map were generated randomly. The dimensions and the value range of each square's data must match these in each sample shown in Figure 5.3 (a). Therefore, the data of squares is composed of three RGB values each of which is between 0 and 1. The number of map's squares should be greater than the number of samples but the greater the squares count is, the more the time-consuming it is to compute the map. Figure 5.3 (b) shows a 100×100 map, i.e. 10,000 squares on the map. The location (i.e. the x and y positions of the square on the map) and the RGB values of a square can be called a weight vector, and each weight vector possesses a unique location on the map. Sometimes, the weights are referred to as neurons because SOMs are a type of neural network.

3. Following the completion of the normalisation of samples and the initialisation of the map, the main algorithm of SOMs can be applied. A sample is chosen randomly from Figure 5.3 (a) and is examined to calculate which square throughout the map, i.e. Figure 5.3 (b), is most similar to the selected sample. The weight of the square with the shortest distance to the weight of the sample is the winner. The winner square is commonly called as the Best Matching Unit (BMU). The distance (i.e. similarity) between the selected sample and squares on the map can be obtained using the method of Euclidean distance:

$$Dist = \sqrt{\sum_{i=0}^{n} (p_i - q_i)^2} \qquad \textit{eq. 5.3}$$

where p_i and q_i are values at the ith member of the sample data. n is the number of data dimensions of the sample. In this case, the n represents 3, i.e. the RGB values.

4. This step is to scale the neighbours of the winner square (i.e. the BMU). This scaling process actually contains two parts including determining the neighbours and adjusting the neighbour's weights. In the field of neural networks, this step is also called a learning process.

In order to determine the neighbours of BMU, the radius of the BMU's neighbourhood is calculated. The radius starts large and commonly equals the width or height of the map, e.g. 100 in this case, but it is reduced during each time-step. Any squares identified within the radius are considered as the BMU's neighbours

and their weights are adjusted to be more like the selected sample. The closer the distance between an identified square and the BMU, the more its weight is altered.

The determination of the rate to diminish the radius and to adjust the square's weight can be obtained using Gaussian function. Figure 5.4 plots the Gaussian function to illustrate the case of diminishing radius. As one moves out from the peak, the radius decreases. For the instance of the diminishing radius, the function can be denoted as following:

$$\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\lambda}\right) \qquad t = 0,1,2,3,4.....n \qquad \textit{eq. 5.4}$$

where the letter sigma, σ_0, is the initial radius at the time-step, t_0, and the letter lambda, λ, is a time constant that can be deduced from the number of iterations, n, divided by log (radius of the map). As shown by the Gaussian function (illustrated in Figure 5.4), the number of neighbours diminishes over time. The "square's weight" is adjusted by another calculation function as follows:

$$L(t) = L_0 \exp\left(-\frac{t}{n}\right) \qquad t = 0,1,2,3,4.....n \qquad \textit{eq. 5.5}$$

where the L_0 is the initial rate for the adjustment, and the letters t and n are the time-step and number of iterations.

The relevant parameters used in the case of Figure 5.3 are that the initial radius, σ_0, is 100; the number of iterations, n, is 1000; and the initial rate for the adjustment, L_0, is 0.1.

5. Repeat step 3 for n iterations.

6. Once the iterations have been finished, a final SOM map can be constructed as illustrated in Figure 5.3 (c) and then the samples shown in Figure 5.3 (a) can be located on the map using the same approach for finding the BMU. For example, the green sample in Figure 5.3 (a) can be placed on the square of the final map with black border, and so on for other samples. Once all of the samples have been positioned on the final map, the similar samples can be grouped together and visualised easier.

Figure 5.3. Schema of a Self Organising Map. For demonstration, the 100 samples (i.e. each square) in (a) were randomly presented in colour which is composed of three variables, i.e. red, blue and green. For example, the green square with a black border is composed of RGB values as R: 0.01; G: 0.79; B: 0.23. The RGB values are arranged equivalently between 0 and 1. (b) is an initialised map composed of 10,000 squares. The dimensions and the value ranges of each square's data must match these in each sample, i.e. three dimensional data (RGB) and data values are between 0 and 1. (c) is the final map developed from the initial map (b) after processing the SOM algorithm. The samples can be then located on the final map according to similarity, i.e. to find the most similar colour. The green square in (a) can be placed on the position with black border, and so on for other samples. The similar samples, therefore, can be grouped together and visualised easier.

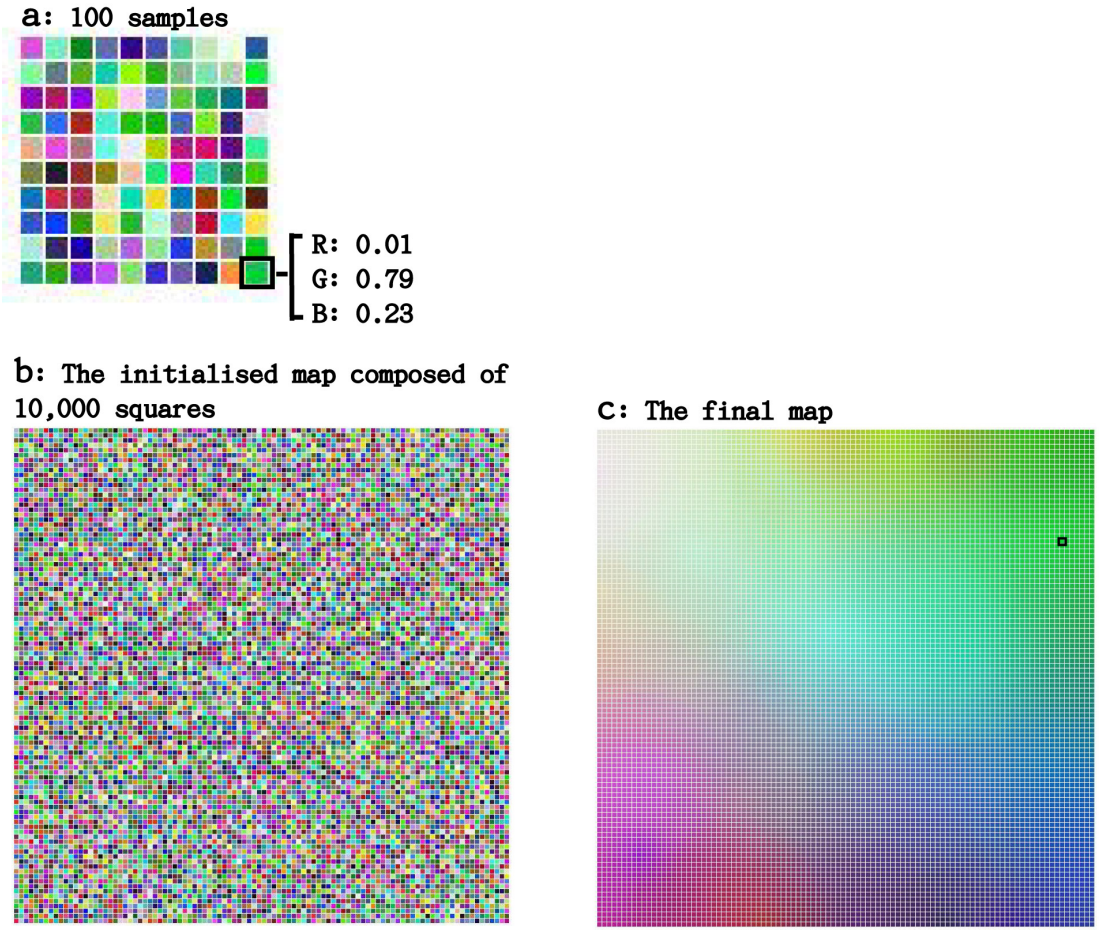
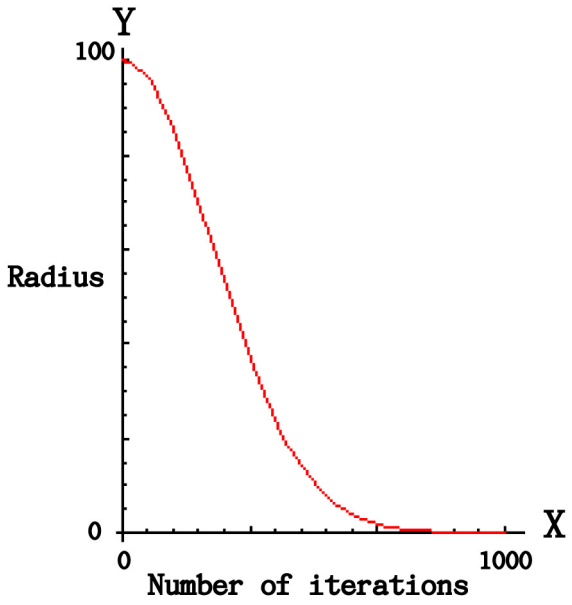


Figure 5.4. Plot of Gaussian function for the instance of the diminishing radius to determine the neighbours of BMU during each time-step (described in section 5.2 step 4). The X axis is the number of iterations and Y is the radius to scale the neighbours of the winner square (i.e. the BMU). As the function shown, the radius diminishes over time.



5.3 Results and discussion

Of the 3,768 compounds (out of 3,775) used in the present study, 44 were active (1.16 %) and the rest inactive. The number of utilised descriptor items for each compound is 1,532. Table 5.1 is the profile of some general descriptors of the test set. A comparison between this table with the molecular property profiles of the EDULISS database (i.e. Table 2.2) shows that the compounds are on average smaller in size but higher in solubility.

Table 5.1. Descriptor ranges of the compounds for the FKBP12 bioassay.

Descriptor	Max	Min	Average	Standard deviation
Molecular weight	1701.27	88.18	250.02	86.52
Number of atoms	177	12	31.52	13.34
Number of bonds	187	12	32.82	14.02
Number of rings	11	0	2.30	1.01
Number of oxygen atoms	46	0	2.43	2.24
Number of nitrogen atoms	7	0	2.11	1.43
Number of HAcc	46	0	4.43	2.06
Number of HDon	25	0	1.42	1.45
MLogP	9.49	-9.79	1.59	1.26
Wiener 3D index (Å)	185654.60	196.42	3237.22	8034.07

5.3.1 Building the logistic model

50 compounds, 1 active and 49 inactive, were removed from the available data set randomly to be the test set and then the remainder used as the training set to build the logistic model (illustrated as Figure 5.1). The process of model building and selection of predictive descriptors are summarised in Figure 5.5. An initial result is that 38 descriptors are found to be good predictors with statistical significance at a p-value level of 0.05, and these are listed in Table 5.2. These descriptors with p-values below 0.05 and the best correlations with biological activity measured by NMR were selected to provide the best model. These predictive descriptors are mainly calculated from three-dimensional representation of a molecule (21 out of 38), and are the classes of geometrical descriptors, including RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors and GETAWAY descriptors. Only 5 items are

derived from two-dimensional, compared with the 12 one-dimensional descriptors (i.e. functional groups and atom-centred fragments shown in Table 5.2). The result seems to indicate that the 3D-based descriptors have higher potential effectiveness in such an SAR modelling study.

3D MoRSE, a geometrical-based descriptor group with the full name "3D Molecule Representation of Structures based on Electron diffraction", is derived from infrared spectra simulation using a scattering function suggested by Soltzberg and Wilkins (Soltzberg and Wilkins 1977). These descriptors are fundamentally composed of 32 equidistant values between 0 and 31 $Å^{-1}$. Each value can consequently be weighted by a series of atomic properties for fitting in a biological study. Previous work shows that the fundamental principle of 3D MoRSE descriptors has been successfully applied to a 2D to 3D conversion programme known as CORINA and neural networks, as well as to distinguish compounds binding to the target receptors (Gasteiger, Sadowski et al. 1996; Schuur, Selzer et al. 1996; Schuur and Gasteiger 1997).

GETAWAY, which stands for GEometry, Topology, and Atom-Weights AssemblY, is a set molecular descriptors deduced from a leverage matrix called molecular influence matrix (MIM). It is a symmetric A×A matrix where A represents the number of atoms (Consonni, Todeschini et al. 2002). The elements of the matrix are calculated from the spatial coordinates of the molecule atoms and then weighted by different atomic properties, such as atomic mass, polarisability, van der Waals volume and electronegativity. As the GETAWAY descriptors contain the relevant

information of molecular geometry, they have good prediction capability in physicochemical property modelling (Consonni, Todeschini et al. 2002).

The model fitting exercise also reveals the presence of some molecular functional groups and fragments that are remarkably associated with the prediction of compound activity. The 2D chemical structures of the predictive functional groups are shown in the legend of Table 5.2. Although the first appearance of this built model tends to be more related to the 3D-descriptors, these 1D-descriptors should not be neglected as these molecular sub-fragments can be considered as "structure making" factors (Fedorowicz, Zheng et al. 2004). Furthermore, the number of nRCONR2, nRSR and nThiazoles shows a statistically significant difference between active and inactive compounds measured by the analysis of the ANOVA procedure. The averages of these functional group counts in the active and inactive compounds are 0.42 vs. 0.05 (nRCONR2), 0.44 vs. 0.11 (nRSR) and 0.12 vs. 0.04 (nThiazoles) respectively. The result of the modelling and ANOVA analysis shows that the presence of these predictive functional groups tends to give a positive contribution to the binding interaction between the target and compounds.

The logistic model gives a scaled value between 0 and 1 presenting the predicted probability to be active for each compound, i.e. higher values for active compounds and lower values for inactive ones. The statistical profiles of the predicted values of bioassay outcomes are shown in Tables 5.3. The result of the training set (Table 5.3 (a)) shows that the model is indeed able to discriminate between the active and inactive compounds on average, as the mean of the predicted probabilities of the inactive group is particularly low. However, the model has incorrect predictions for

some observations. There are 7 active compounds whose predicted probabilities are below 0.3 while 15 inactive compounds have predicted values larger than 0.3. Applying the model to calculate the probabilities of test set (Table 5.3 (b)) shows a good but not yet totally reliable prediction. Several stricter validation tests have been carried out and will be discussed in the subsequent section.

Figure 5.5. Summary of the Logistic Regression Model building process. The model is built using a training set and is to predict the compound activity. (a) is the procedure of stepwise selection to select predictive molecular descriptors. The effect-column lists the selected independent variables, i.e. descriptors. The most significant predictive descriptor, i.e. X1334, first enters into the model and other descriptors are added sequentially to the model with a goal of obtaining the most appropriate combination of descriptors and the best model. The model omits the descriptors whose p-value was greater than 0.05 (marked by red frames) during the selection. (b) is the finally estimated intercept and regression coefficients of each selected descriptor for the probability calculation. The selected descriptors are detailed in Table 5.2.

a: Procedure of stepwise selection

	Effect		Number	Chi-Square test
Step	Entered	Removed	In	p value
1	X1344		1	<.0001
2	X1385		2	<.0001
3	X1519		3	<.0001
4	X1112		4	<.0001
18	X1392		18	0.0040
19	X1148		19	0.0019
20	X310		20	0.0140
21		X1448	19	0.0636
22	X143		20	0.0029
23		X1519	19	0.0502
24		X1525	18	0.0578
25	X899		19	0.0041
26	X1374		20	0.0074
27	X690		21	0.0043
28		X1392	20	0.3415
29	X1406		21	0.0118
30	X1214		22	0.0203
31	X984		23	0.0068
32	X1111		24	0.0130
33		X1112	23	0.3499
34	X1100		24	0.0057
35	X1434		25	0.0071
36	X1500		26	0.0041
37		X1386	25	0.0518
38	X1164		26	0.0062
39	X905		27	0.0181
40	X1340		28	0.0052
41	X492		29	0.0154
42	X1407		30	0.0213
43		X1100	29	0.1252
44	X1189		30	0.0173
45	X1032		31	0.0101
46	X1401		32	0.0074
47	X484		33	0.0121
48	X990		34	0.0221
49	X716		35	0.0205
50		X1345	34	0.0586
51		X1401	33	0.0777
52	X1467		34	0.0026
53	X1062		35	0.0031
54	X1029		36	<.0001
55	X1401		37	0.0206
56	X823		38	0.0363
57	X748		39	0.0353
58		X748	38	0.0570

b: The estimated regression coefficients

Parameter	Estimate
Intercept	60.3327
X143	−12.8285
X310	22.9408
X452	−9.6759
X484	−13.8989
X492	−20.8115
X690	−0.2945
X703	13.4753
X716	−0.4316
X823	0.2954
X899	4.3417
X904	−14.1103
X905	6.3172
X918	16.2634
X956	38.5858
X984	5.5468
X990	8.9653
X1029	−22.4764
X1032	−29.5143
X1062	25.9749
X1111	72.2312
X1142	−15.7963
X1148	−19.3798
X1164	−65.4268
X1189	9.0834
X1214	−40.9966
X1274	−179.7
X1340	9.6476
X1344	8.3399
X1352	10.4514
X1374	24.9555
X1375	5.0979
X1385	8.5028
X1401	3.6894
X1406	8.7350
X1407	7.0677
X1434	9.2309
X1467	2.3328
X1500	9.1036

Table 5.2. Potentially predictive descriptors selected by the logistic modelling exercise, i.e. the 38 best estimated descriptors shown in Figure 5.5 (b). The descriptions for each descriptor are quoted from DRAGON web site (http://www.talete.mi.it/). (continued on next two pages)

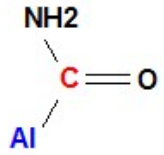
Items	Descriptions	Classes
T(O..Br)	sum of topological distances between O..Br	topological descriptors
ATS7p	Broto-Moreau autocorrelation of a topological structure - lag 7 / weighted by atomic polarizabilities	2D autocorrelations
ESpm01d	Spectral moment 01 from edge adj. matrix weighted by dipole moments	edge adjacency indices
BEHm3	highest eigenvalue n. 3 of Burden matrix / weighted by atomic masses	Burden eigenvalues
BELm3	lowest eigenvalue n. 3 of Burden matrix / weighted by atomic masses	Burden eigenvalues
G(N..N)	sum of geometrical distances between N..N	geometrical descriptors
G(O..Br)	sum of geometrical distances between O..Br	geometrical descriptors
RDF025u	Radial Distribution Function - 2.5 / unweighted	RDF descriptors
RDF110e	Radial Distribution Function - 11.0 / weighted by atomic Sanderson electronegativities	RDF descriptors
Mor05m	3D-MoRSE - signal 05 / weighted by atomic masses	3D-MoRSE descriptors
Mor10m	3D-MoRSE - signal 10 / weighted by atomic masses	3D-MoRSE descriptors
Mor11m	3D-MoRSE - signal 11 / weighted by atomic masses	3D-MoRSE descriptors
Mor24m	3D-MoRSE - signal 24 / weighted by atomic masses	3D-MoRSE descriptors
Mor30v	3D-MoRSE - signal 30 / weighted by atomic van der Waals volumes	3D-MoRSE descriptors

Mor26e	3D-MoRSE - signal 26 / weighted by atomic Sanderson electronegativities	3D-MoRSE descriptors
Mor32e	3D-MoRSE - signal 32 / weighted by atomic Sanderson electronegativities	3D-MoRSE descriptors
G2u	2st component symmetry directional WHIM index / unweighted	WHIM descriptors
E2u	2nd component accessibility directional WHIM index / unweighted	WHIM descriptors
G2e	2st component symmetry directional WHIM index / weighted by atomic Sanderson electronegativities	WHIM descriptors
Dm	D total accessibility index / weighted by atomic masses	WHIM descriptors
HATS6u	leverage-weighted autocorrelation of lag 6 / unweighted	GETAWAY descriptors
H2m	H autocorrelation of lag 2 / weighted by atomic masses	GETAWAY descriptors
HATS8m	leverage-weighted autocorrelation of lag 8 / weighted by atomic masses	GETAWAY descriptors
H3e	H autocorrelation of lag 3 / weighted by atomic Sanderson electronegativities	GETAWAY descriptors
H8p	H autocorrelation of lag 8 / weighted by atomic polarizabilities	GETAWAY descriptors
R1v+	R maximal autocorrelation of lag 1 / weighted by atomic van der Waals volumes	GETAWAY descriptors
nRCONH2	number of primary amides (aliphatic)	functional group counts
nRCONR2	number of tertiary amides (aliphatic)	functional group counts
nArCO	number of ketones (aromatic)	functional group counts
nArNO2	number of nitro groups (aromatic)	functional group counts
nN(CO)2	number of imides (-thio)	functional group counts

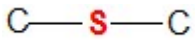
nRSR	number of sulfides	functional group counts
nPyrazoles	number of Pyrazoles	functional group counts
nIsoxazoles	number of Isoxazoles	functional group counts
nThiazoles	number of Thiazoles	functional group counts
C-018	=CHX*	atom-centred fragments
H-053	H attached to C0(sp3) with 2X attached to next C	atom-centred fragments
S-108	R=S	atom-centred fragments

*: X represents any electronegative atom (O, N, S, P, Se, halogens).

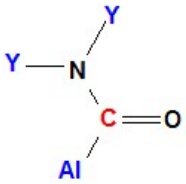
The 2D representations of the functional groups:



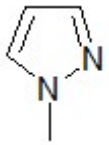
nRCONH2:
Al = H or aliphatic group linked through C



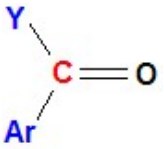
nRSR



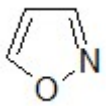
nRCONR2:
Al = H or aliphatic group linked through C
Y = Ar or Al (not H, not C = O)



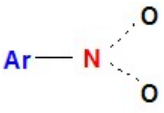
nPyrazole



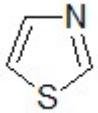
nArCO:
Ar = aromatic group linked through any atom
Y = Al or Ar



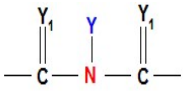
nIsoxazoles



nArNO2:
Ar = aromatic group linked through N



nThiazoles



nN(CO)2:
Y = H or C
Y1 = O or S

Table 5.3. Statistical profile of predicted probabilities of training set (a) and test set (b).

a: training set*

	Active compounds	Inactive compounds
Mean	0.697	0.004
STD	0.309	0.033
Min	0	0
Max	1	0.824
Observations	43	3,675

*: There are 50 compounds removed randomly from the PubChem available data set before model building. The removed compounds include 1 active and 49 inactive observations respectively.

b: test set*

	Active compounds	Inactive compounds
Mean	-	0
STD	-	0
Min	-	0
Max	0.99	0
Observations	1	49

*: The 50 removed compounds.

5.3.2 SOM analysis

A fundamental theorem suggests that similar structures should exhibit similar properties (Ruecker and Ruecker 1993). It follows that the active compounds should display some similar patterns which are able to discriminate themselves from the inactive compounds. Yet, it remains problematic how to reveal or describe the similarity of compounds based on types of molecular features. The logistic modelling exercise described in section 5.3.1 has selected 38 predictive descriptors

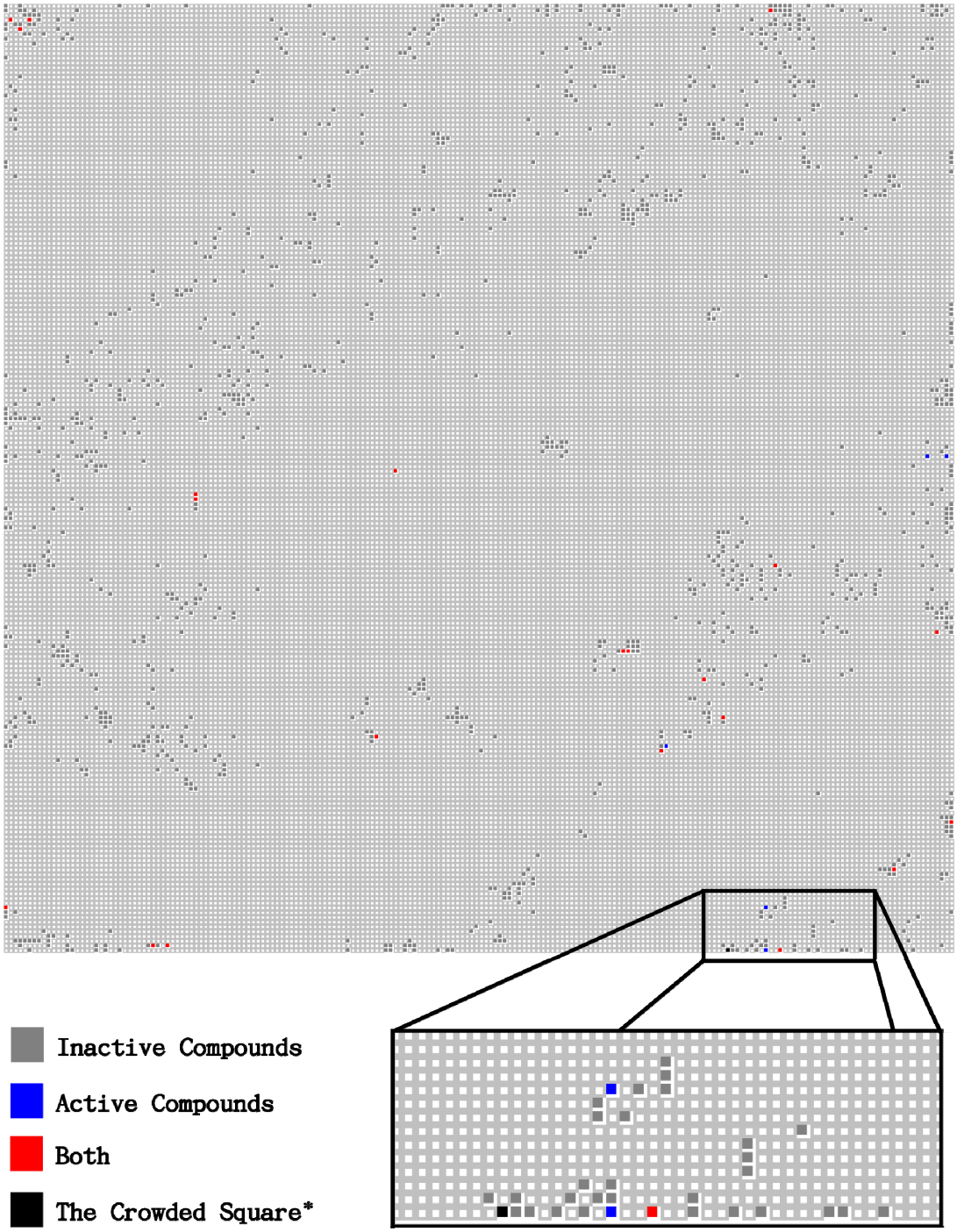
which are of greatest utility in discriminating active and inactive compounds. As mentioned above, the SOM technique is able to project high dimensional data onto a low dimension map which can be used to visualise the sample similarity. Figure 5.6 illustrates an SOM analysis summarising the similarity of the PubChem FKBP12 available data set based on the 38 predictive descriptors. The analysis groups the compounds which have similar patterns into identical positions or clusters. In order to present the clusters for easy observation, the map size has been enlarged to 200×200, i.e. 40,000 lattices for 3,768 observations in total. The blue and gray squares represent the locations of active and inactive compounds respectively. In some cases, both classes located in the same position are then shown as red squares (Figure 5.6). An initial inspection of the map shows it to be not as discriminative as expected. The squares of active and inactive compounds are mostly located in the identical positions or clusters. However, a crowded square, shown as black, is located in the amplified cluster of Figure 5.6. The crowded square contains 48 compounds of which 13 are active (30 % out of 44) whereas 35 (1 % out of 3,724) are inactive. In the amplified cluster, there are 16 active and 172 inactive compounds in total. It indicates that the analysis of SOM based on the 38 predictive descriptors succeeds in aggregating the 36 % active compounds in this cluster and discriminate them from 95 % inactive compounds, and successfully visualises the result.

It is important to note that most of the 16 active compounds show strong binding affinities: 10 out of the 16 active compounds are found in the top twenty binding results when ranked by the estimated Kd values (μM), including the two strongest binding observations, i.e. ID: 16725062 and 16725059. The 2D chemical structures

of these 16 active compounds are shown in Figure 5.7. Two common patterns occurring between these compounds include the amide (N–C=O) and sulfide (C–S–C) groups represented in blue and pink respectively. The amide group is the fundamental fragment of keto amide carbonyl. Previous studies indicate that the amide carbonyl of FK506's keto amido group forms an important part of the binding to FKBPs. Figure 5.8 shows the chemical structure of FK506 and its keto amido group is highlighted in red. The amide carbonyl portion forms a hydrogen-bonding interaction with FKBP and the ketone carbonyl oxygen engages with three aromatic residue side chains of FKBP forming a carbonyl binding pocket (Holt, Konialian-Beck et al. 1994). For the compounds that mimic FK506, the presence of this binding portion is commonly reported and considered as an important synthesis objective (Dubowchik, Vrudhula et al. 2001). Figure 5.9 demonstrates an instance of FKBP docking study in which the hit compound shows a strongest binding affinity (Kd: 0.2 μM) as it contains an amide group and the ketone carbonyl oxygen, coloured in red, forming an essential hydrogen-binding interaction with the side chain Tyr 26 (PDB ID: 1J4R). In addition, the sulfur atom also fits into the hydrophobic pocket around the edge of the binding site (Stebbins, Zhang et al. 2007). On the other hand, the items of the predictive descriptors selected by the above modelling exercise are in line with the previous studies. The functional groups listed in Table 5.2 partially fit the structural fundamentals of these two crucial binding portions and contribute to the higher binding interaction. In order to estimate if the interaction exists between these two portions, Table 5.4 shows a survey using MCS (Maximum Common Subgraph, described in chapter 3) to find the compounds that contain the groups of two sub-structures **a** (N–C=O–C–**X**–C–Y) or **b** (N–C=O–C–**S**–

C–Y) respectively. The letter **X** stands for any atom of N, O or P, and the **Y** represents an aliphatic or aromatic ring. 47 (**a**) and 39 (**b**) compounds are found in the whole PubChem available data set containing these sub-structures. The **X** in group **a**, i.e. N–C=O–C–**X**–C–Y, is actually only ever an oxygen atom. Although the number of observed compounds in group **b** is fewer than those in group **a**, 38.5 % of the observed compounds in group **b** are shown as active whereas only 4.3 % in group **a**. The compounds divide into two sets, i.e. the groups of amides (N–C=O) and sulfides (C–S–C), with both adjacent together showing a higher probability to bind with FKBP12.

Figure 5.6. Clustering of FKBP12 bioassay compounds, 3,768 (44 active and 3,724 inactive) in total, by SOM analysis. The map size is 200×200. The number of data dimensions for each compound is 38 which are the predictive descriptors selected by the logistic modelling exercise.



*: There are 48 compounds placed on the crowded (black) square. 13 of them are active and 35 of them are inactive.

Figure 5.7. The 16 active compounds placed in the amplified cluster shown in Figure 5.6. The amide (N–C=O) and sulfide (C–S–C) groups are represented in blue and pink respectively. The labels represent the compound's ID, estimated Kd value (μM) and the rank in the overall bioassay.

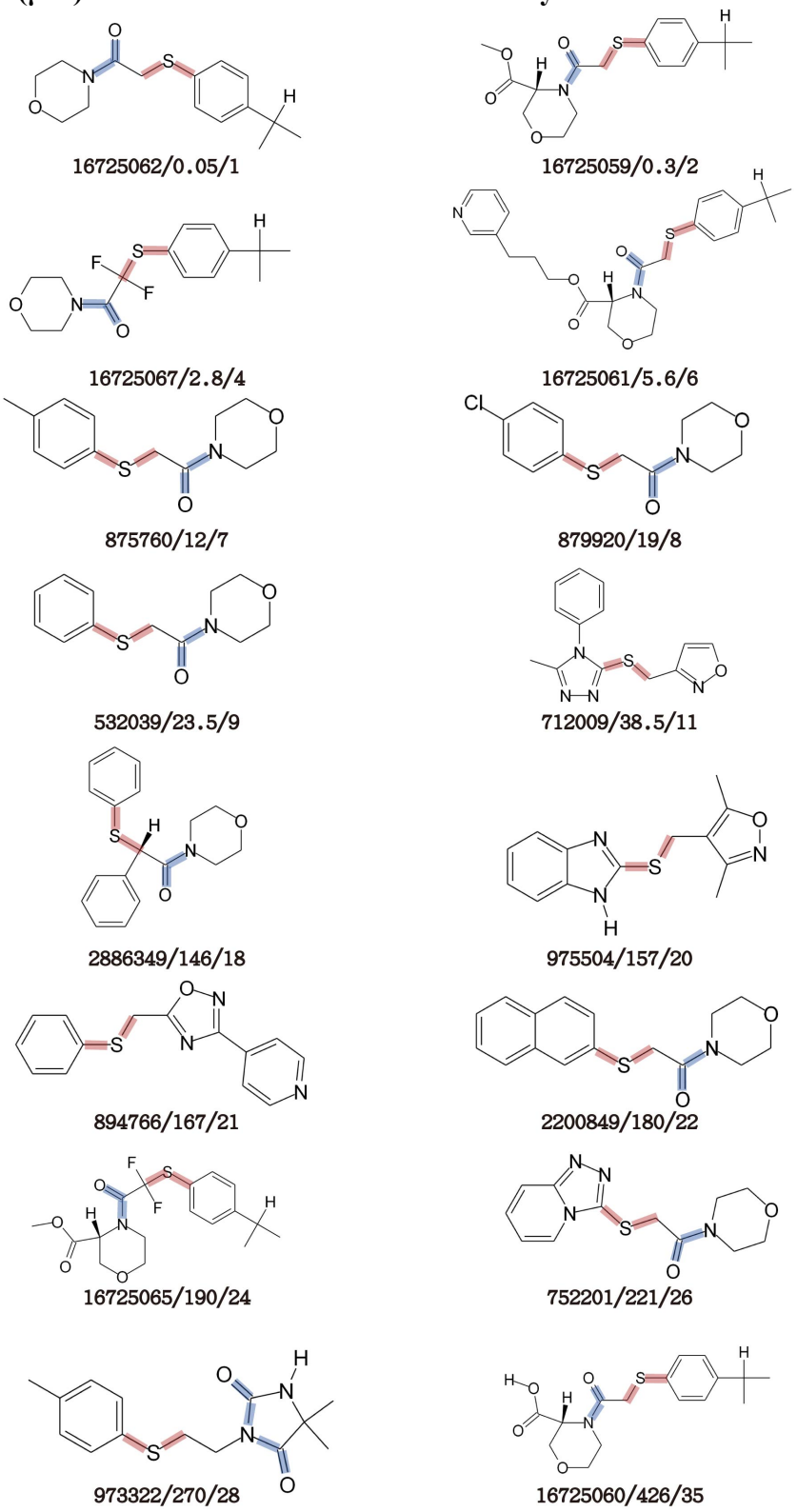


Figure 5.8. Chemical structure of FK506 with the keto amido group coloured in red.

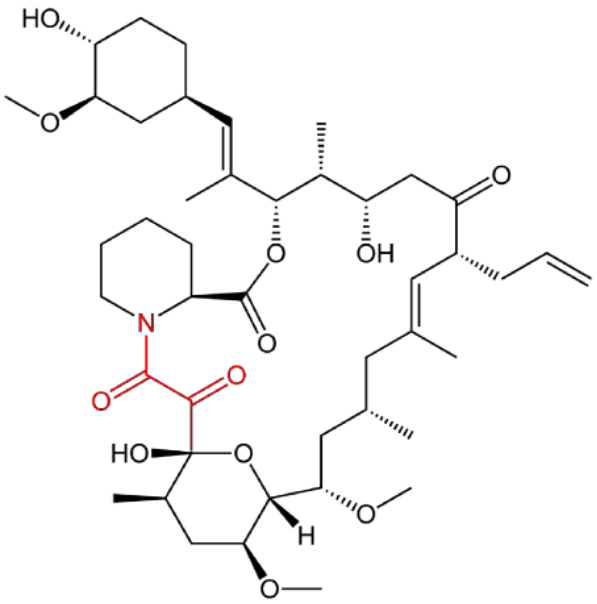
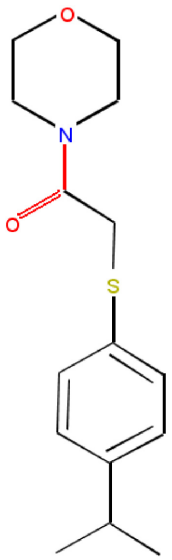
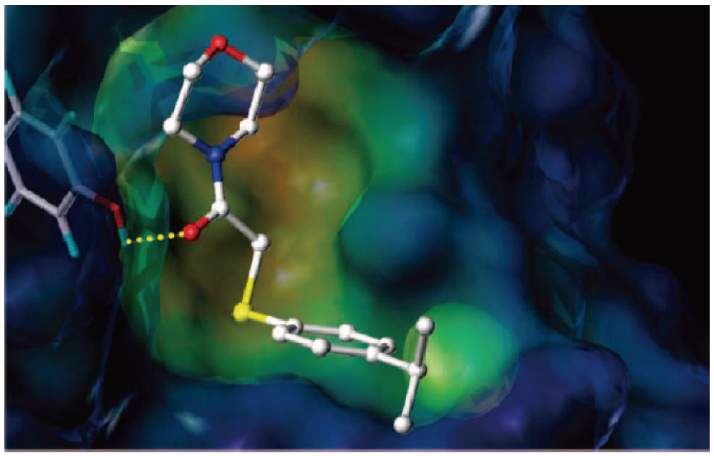


Figure 5.9. Molecular docking studies. (a) A hit compound has its amide group coloured in red; (b) A docked structure of FKBP12 (PDB ID: 1J4R) and the ligand is the hit compound. The hydrogen-bonding interaction between the side chain Tyr 26 and the ketone carbonyl oxygen of the amide group is highlighted by a dashed line.

a



b

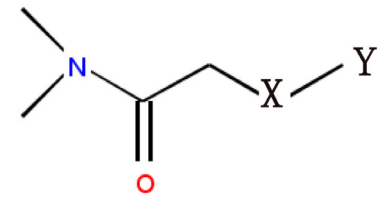


(quoted from Stebbins, Zhang et al. 2007)

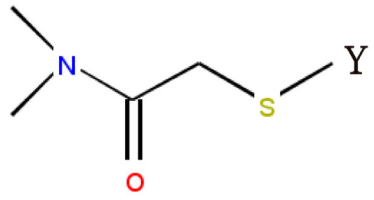
Table 5.4. Numbers of compounds containing the sub-structures of interest.

	N–C=O–C–**X**–C–Y[a]	N–C=O–C–**S**–C–Y[b]
Total of observed compounds	47	39
Active compounds	2	15
Percentage (%)	4.3	38.5

a



b



X: The atoms of N, O or P. None of atom N and P has been found within the 47 observed compounds.

Y: Aliphatic or aromatic ring.

5.3.3 Stricter model validation

In order to verify that the fitted model can be generalised to the same type of new data, the above logistic analysis has been repeated while removing a set of compounds prior to the modelling exercise. Although the built model (described in section 5.3.1) gives a good accuracy of prediction as shown in Table 5.3, the number of test set, i.e. 1 active and 49 inactive compounds, is not strict enough for model validation. This section presents the results of twenty more restrictive test sets. In each test, there are 4 active and 372 inactive compounds removed randomly from the PubChem available data set, i.e. removing about 10 % from each of the two classes. The remaining data, which is known as the training set, is then utilised for the modelling exercise. The fitted model based on the training set is called the "new model" and the model built by the entire PubChem available data set without prior removal of any compound is called the "old model". Accordingly, the old model

possesses a relative better fitting quality as it is built by the complete data set. The twenty new models are compared individually with the old model. As the results show, the tests of the new models show good predictions for the inactive compounds. 7,440 observations (372×20) in total give predicted probabilities which are extremely low from 0 to 0.15 where only three of them are greater than 0.1. However, unlike the results of the inactive compounds, the predictions for the active compounds are barely satisfactory, and are shown in Figure 5.10. The predicted values obtained from the new models tend to underestimate the probabilities of the active compounds. Only 17.5 % cases (14 out of 80 observations) in red frames in Figure 5.10 exhibit the good predictions in the tests of No. 3, 4, 5, 6, 7, 8, 11, 13, 15, 16, 17 and 20. The 2D chemical structures of these well-predicted compounds are shown in Figure 5.11. Predictably, these compounds commonly contain the groups of amides and sulfides.

5.4 Summary

The application of the Logistic Regression Model succeeds in revealing the potentially important molecular descriptors and essential sub-structures involved in the binding interaction between compounds and target. Although the built model only predicts the activity of compounds accurately for fewer cases in the stricter validation tests, it is still useful in exhibiting the features of these well-predicted compounds. These compounds can be the query template in molecular similarity searches for further application. Furthermore, the revealed descriptors can also represent the molecular similarity whilst being applied in the clustering applications such as SOMs in this study.

Figure 5.10. Predicted probabilities of 80 active compounds in the twenty stricter hold-out tests. In each hold-out test, there are 4 active and 372 inactive compounds removed randomly from PubChem available data set (i.e. removing about 10 %), and the remaining data (the training set) is then utilised for the modelling exercise, so finally twenty individual fitting models are built. The twenty fitting models using the training set are called the "new models" and the model built by the entire PubChem available data set is called the "old model". Each new model is then competed individually with the old model. The X axes list the 4 previously removed compounds (test set) and Y axes show their predicted probabilities using the new and old models, i.e. each active compound having two predicted probability values. The cases in red frames exhibit good predictions of new models. (continued on next pages)

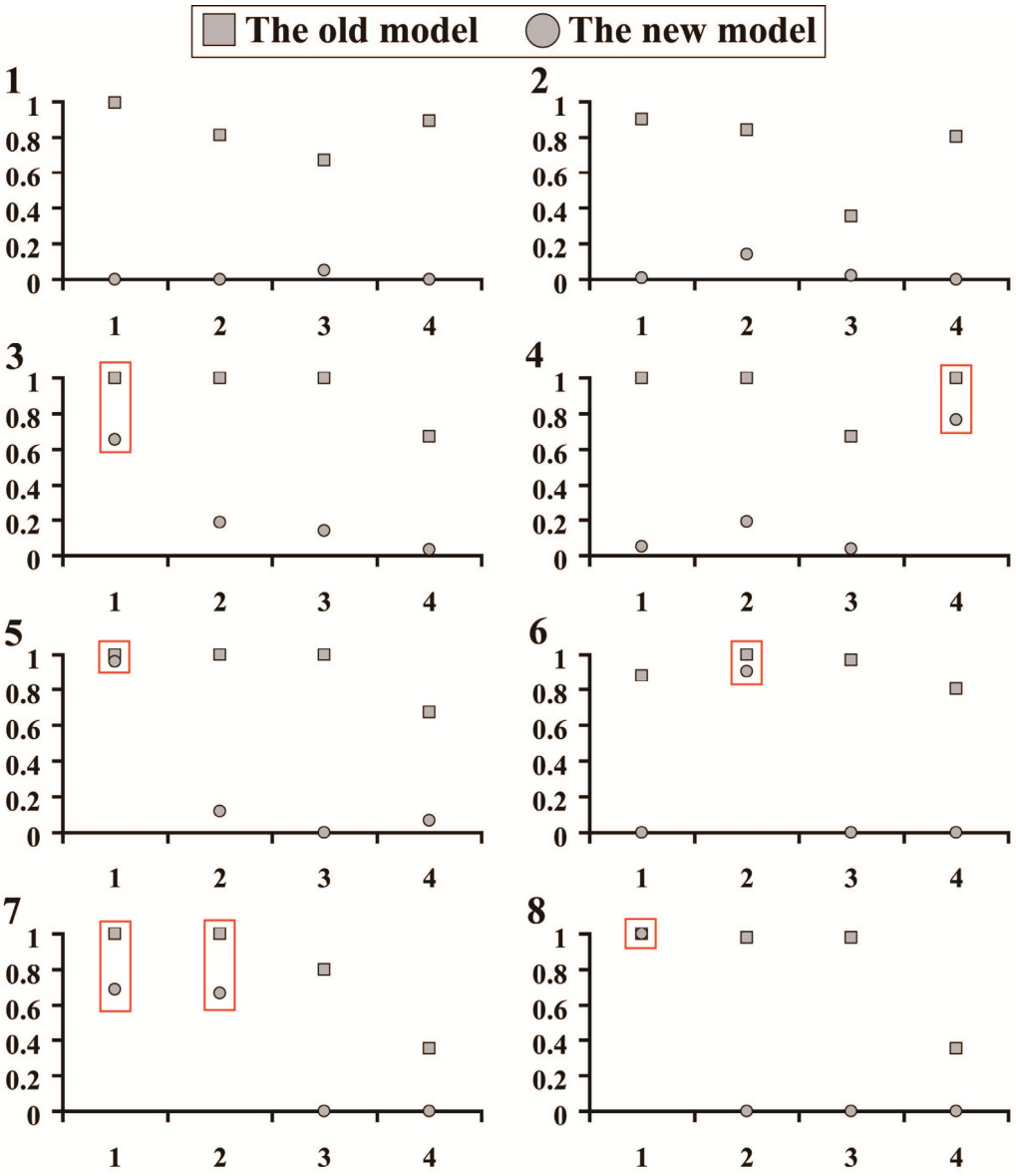


Figure 5.10. Continued.

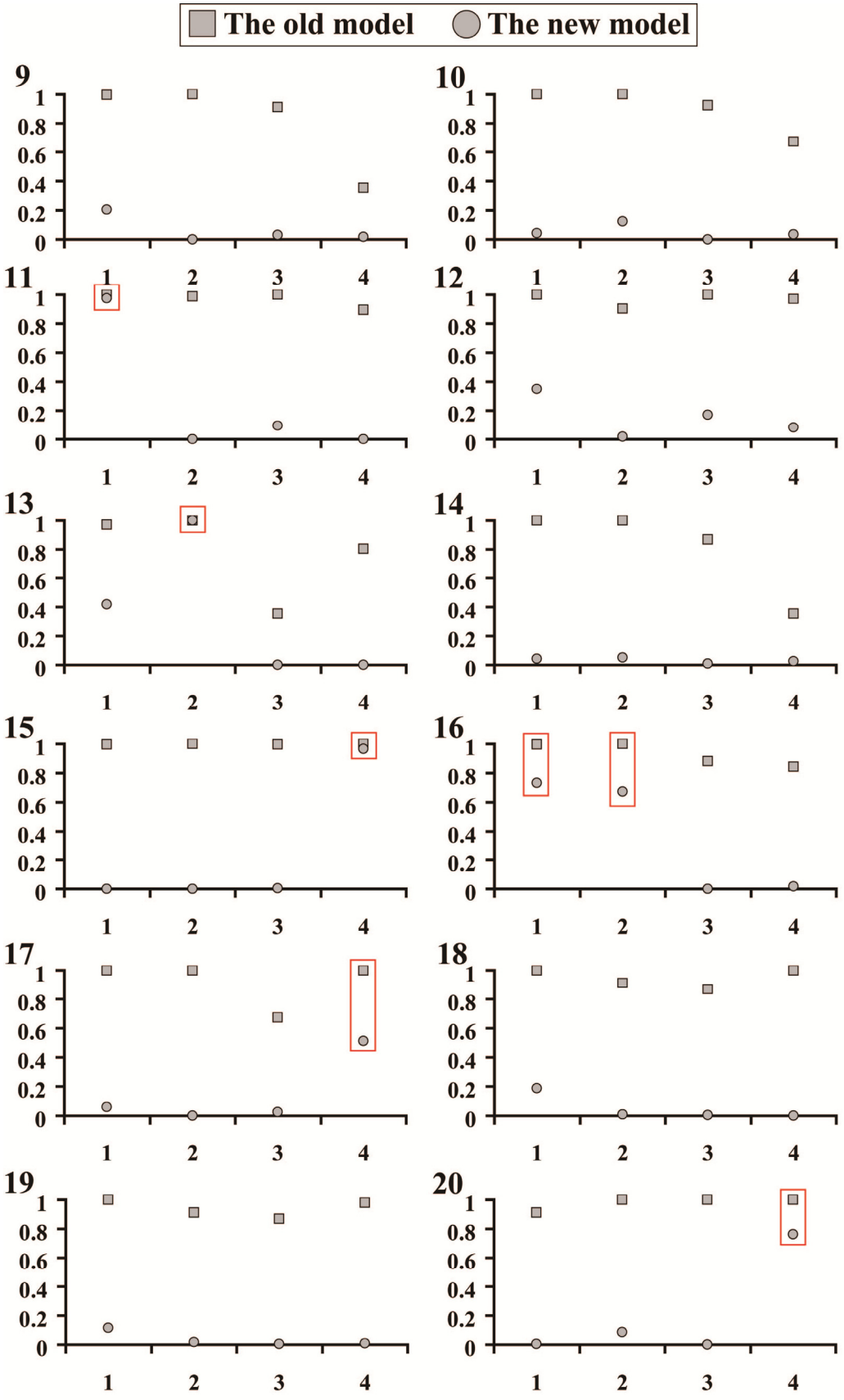
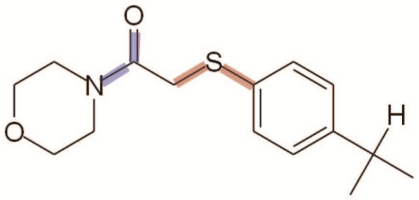
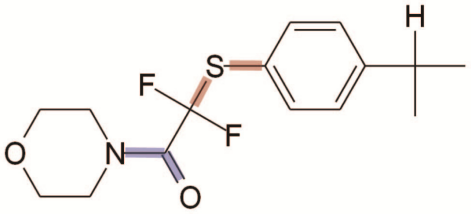


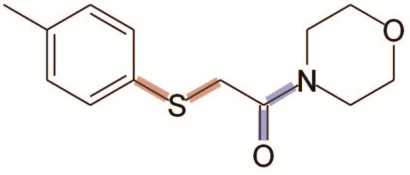
Figure 5.11. Chemical structures of the well-predicted compounds in the twenty stricter hold-out tests. The amide (N–C=O) and sulfide (C–S–C) groups are represented in blue and pink respectively. The labels represent the compound's ID, estimated Kd value (μM) and the rank in the overall bioassay.



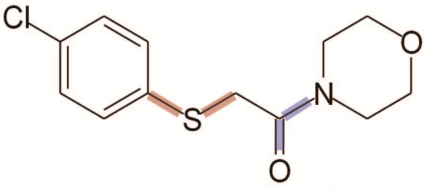
16725062/0.05/1



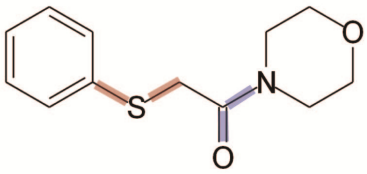
16725067/2.8/4



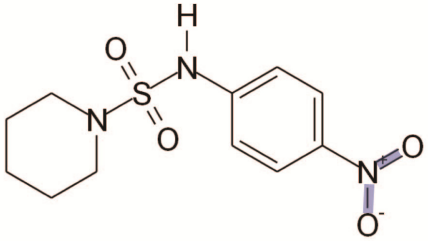
875760/12/7



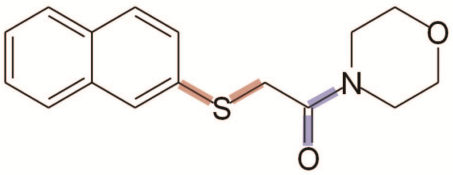
879920/19/8



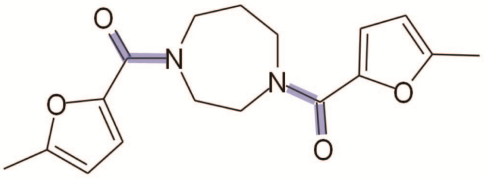
532039/23.5/12



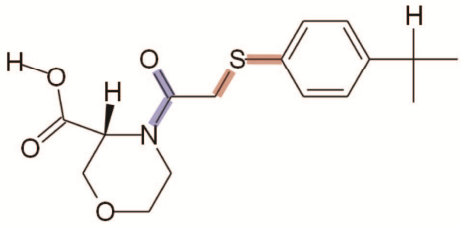
2926730/60/13



2200849/180/22



753179/204/25



16725060/426/35

5.5 Reference:

Aghdasi, B., K. Ye, et al. (2001). "FKBP12, the 12-kDa FK506-binding protein, is a physiologic regulator of the cell cycle." Proceedings of the National Academy of Sciences **98**(5): 2425.

Consonni, V., R. Todeschini, et al. (2002). "Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors." Journal of Chemical Information and Computer Sciences **42**(3): 682-692.

Consonni, V., R. Todeschini, et al. (2002). "Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors. 2. Application of the novel 3D molecular descriptors to QSAR/QSPR studies." Journal of Chemical Information and Computer Sciences **42**(3): 693-705.

Dubowchik, G. M., V. M. Vrudhula, et al. (2001). "2-Aryl-2, 2-difluoroacetamide FKBP12 Ligands: Synthesis and X-ray Structural Studies." Organic Letters **3**: 3987-3990.

Fedorowicz, A., L. Zheng, et al. (2004). "QSAR Study of Skin Sensitization Using Local Lymph Node Assay Data." International Journal of Molecular Sciences **5**: 56-66.

Gasteiger, J., J. Sadowski, et al. (1996). "Chemical information in 3D space." Journal of Chemical Information and Computer Sciences **36**(5): 1030-1037.

Hawkins, D. M., S. C. Basak, et al. (2003). "Assessing model fit by cross-validation." Journal of Chemical Information and Computer Sciences **43**(2): 579-586.

Holt, D. A., A. L. Konialian-Beck, et al. (1994). "Structure-activity studies of synthetic FKBP ligands as peptidyl-prolyl isomerase inhibitors." Bioorganic & Medicinal Chemistry Letters(Print) **4**(2): 315-320.

Iida, T., M. Furutani, et al. (1998). "FKBP-type peptidyl-prolyl cis-trans isomerase from a sulfur-dependent hyperthermophilic archaeon, Thermococcus sp. KS-1." Gene **222**(2): 249-255.

Kay, J. E. (1996). "Structure-function relationships in the FK506-binding protein (FKBP) family of peptidylprolyl cis-trans isomerases." The Biochemical Journal. **314**(2): 361-385.

Khan, J., J. S. Wei, et al. (2001). "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." Nature Medicine **7**: 673-679.

Kohonen, T. (1997). Self-Organizing Maps. New York, Springer-Verlag.

Liu, J., J. D. Farmer, et al. (1991). "Calcineurin is a common target of cyclophilin-cyclosporin A and FKBP-FK506 complexes." Cell **66**(4): 807-815.

Nguyen, D. V. and D. M. Rocke (2002). "Tumor classification by partial least squares using microarray gene expression data". Oxford University Press **18:** 39-50.

Peng, C. Y. J., K. L. Lee, et al. (2002). "An Introduction to Logistic Regression Analysis and Reporting." Journal of Educational Research **96**(1): 3-14.

Roussinov, D. and H. Chen (1998). "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation.". Communication Cognition and Artificial Intelligence, Spring.

SAS (2004). SAS Institute Inc., Cary, NC, USA.

Schuur, J. and J. Gasteiger (1997). "Infrared spectra simulation of substituted benzene derivatives on the basis of a 3D structure representation." Analytical Chemistry **69**(13): 2398-2405.

Schuur, J. H., P. Selzer, et al. (1996). "The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity." Journal of Chemical Information and Computer Sciences **36**(2): 334-344.

Shou, W., B. Aghdasi, et al. (1998). "Cardiac defects and altered ryanodine receptor function in mice lacking FKBP12." Nature **391**(6666): 489-92.

Soltzberg, L. J. and C. L. Wilkins (1977). "Molecular transforms: a potential tool for structure-activity studies." Journal of the American Chemical Society **99**(2): 439-443.

Stebbins, J. L., Z. Zhang, et al. (2007). "Nuclear magnetic resonance fragment-based identification of novel FKBP12 inhibitors." Journal of Medicinal Chemistry **50**(26): 6607-6617.

UNITY Tripos Inc. Chemical Information Software. 1699 S. Hanley Road, St. Louis, MO 63144, USA, Tripos Inc**:** UNITY.

Wang, J., J. Delabie, et al. (2002). "Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study." BMC Bioinformatics **3**(1): 36.

6. The application of Atomic Characteristic Distance (ACD) and EDULISS 2.0 for ligand discovery

This chapter presents the results of ligand discovery on the known 3D structure of pyruvate kinase using the approach of small-molecule database mining. The application of ACD (Atomic Characteristic Distance, described in chapter 2) and the facilities of EDULISS 2.0 have been used in the mining processes. The steps in this ligand discovery range from a set of simple database queries via the interface of EDULISS to a series of complicated small-molecule screenings mainly using ACD. The entire data mining process can be divided into two stages which are described as stage 1 and stage 2 respectively in this chapter. The candidate compounds are then tested for inhibition of enzyme activity, and some of the protein-ligand complex structures are also crystallised. The crystallographic experiments and bioassays mentioned in this chapter were performed by my colleague Dr. Hugh Morgan.

6.1 Materials and methods

6.1.1 Overview of target protein

The target protein used in this study is a pyruvate kinase, PK (EC 2.7.1.40), which is vital in the energy producing step of glycolysis. PK catalyses the transfer of a phosphoryl group from phosphoenolpyruvate (PEP) to ADP to yield pyruvate and ATP (Mesecar and Nowak 1997) shown as below:

$$H^+ + PEP + ADP \xrightarrow{M(I),\ M(II)} Pyruvate + ATP$$

Other than the presence of the reaction substrates, i.e. PEP and ADP, the catalysis also requires the monovalent and divalent cations, i.e. M(I) and M(II), prior to any

activity being detected (Yu, Lee et al. 2003). Both monovalent and divalent cations, including K^+, and Mg^{2+} or Mn^{2+}, are involved in the reaction. K^+ is an essential activator of PKs, whose presence causes 2-6 fold higher affinities of the relevant substrates than in the absence of K^+. The presence of K^+ induces the correct geometrical arrangement of the active site residues, allowing either PEP or ADP to bind in the pocket independently, whereas in the absence of K^+ the two substrates have an ordered binding mechanism as ADP cannot bind to the pocket before PEP (Oria-Hernandez, Cabrera et al. 2005). By contrast with K^+, the divalent cations, mainly Mg^{2+} and in some cases Mn^{2+}, cause global and local conformational changes in the pocket and directly interact with the transferred phosphoryl group to yield ATP (Matte, Tari et al. 1998).

The architectures of PKs have been isolated and characterised from a number of species. In vertebrates it has been found to exist as four isoforms, including the L-form found in the liver, the R-form in erythrocytes, the M1 form in skeletal muscle and the M2 form in the foetus (Lovell, Mullick et al. 1998). Most of PKs exist as a homotetramer and each subunit encompasses four domains. Apart from the N-terminal domain, the commonly discussed domains are labeled as A, B and C. Figure 6.1 (a) is the ribbon structure diagram of an *E. coli* PK subunit and Figure 6.1 (b) schematically shows domains A, B and C in green, blue and red, respectively. The active site located between domains B and A is the docking pocket for the relevant substrates, including PEP and ADP, as well as the required monovalent and divalent cation. The fructose-1,6-bisphosphate (FBP) binding site is located in domain C. As FBP is the allosteric effector of several species PKs, this binding site is also commonly called the effector site (Mattevi, Valentini et al. 1995). A proposed

hypothesis of the mechanism of PK's allosteric regulation is that the movement of the three domains takes place within each subunit when it is activated by the docking of PEP and FBP (Mattevi, Bolognesi et al. 1996). When the movement occurs, two types of conformational changes can be observed, including rotation of the A, B and C domains within each subunit and every subunit within the tetramer. Figure 6.2 illustrates the tetramer conformation change due to the domain movements. When domain B shows an open conformation (Figure 6.2 (a)), the enzyme is in the inactive state, called the T-state, whereas it is in the active state when domain B is closed (Figure 6.2 (b)), called the R-state. During the transition of T-state to R-state, the domain and subunit orientations cause a decrease in height and an increase in width of the tetramer.

A major difference of enzyme control property between the PKs of mammalian host and trypanosomatid protozoans is that the allosteric regulation of mammalian PK is in response to F-1,6-BP, however protozoans's PK is regulated by F-2,6-BP (Valentini, Chiarelli et al. 2000). An early study indicated that F-1,6-BP shows a similar effect to F-2,6-BP on trypanosoma brucei but it needs approximately 4000-fold higher concentrations (Schaftingen, Opperdoes et al. 1985). Apart from directly blocking the active site to inhibit the function of protozoans' PK, this unique diversity in the allosteric regulation provides a possible strategy for the effector site ligand design to interfere with PK's allosteric regulation and thereby to produce activity against related diseases caused by protozoans such as Leishmania.

Figure 6.1. Illustration of the domains, catalytic site and FBP binding site of an *E. coli* PK subunit (monomer). (a) Ribbon structure diagram. (b) Schematic representation of the domain A, B and C shown in green, blue and red, respectively. Both Figure 6.1 and 6.2 are schematically quoted from Mattevi, Valentini et al. 1995 and Mattevi, Bolognesi et al. 1996.

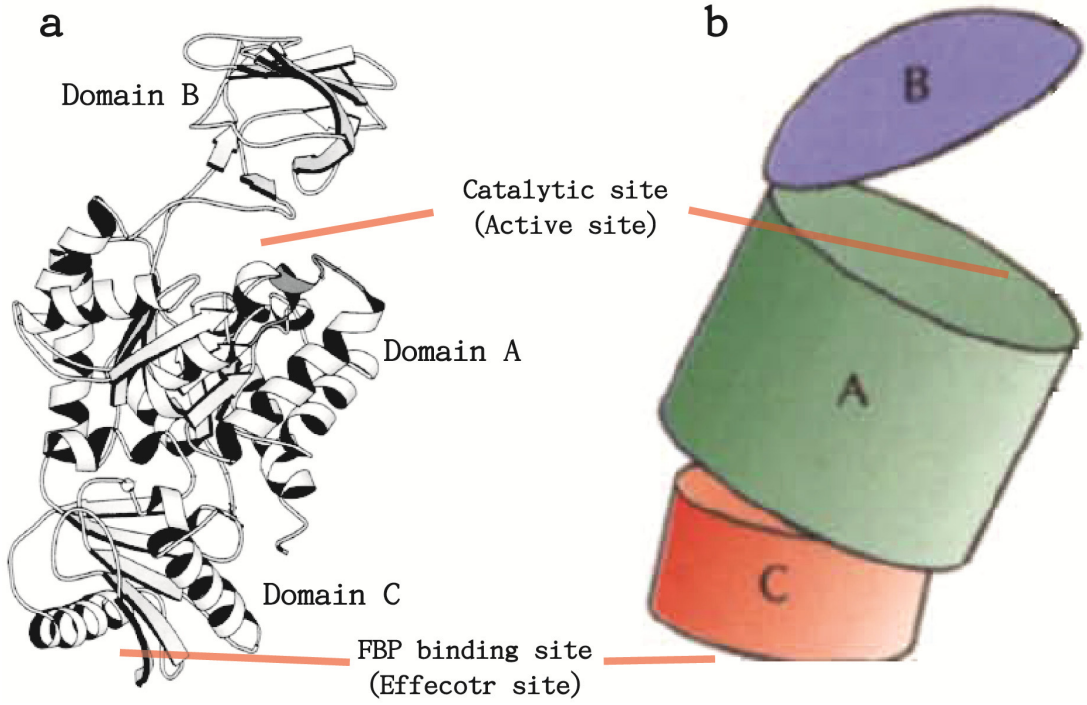
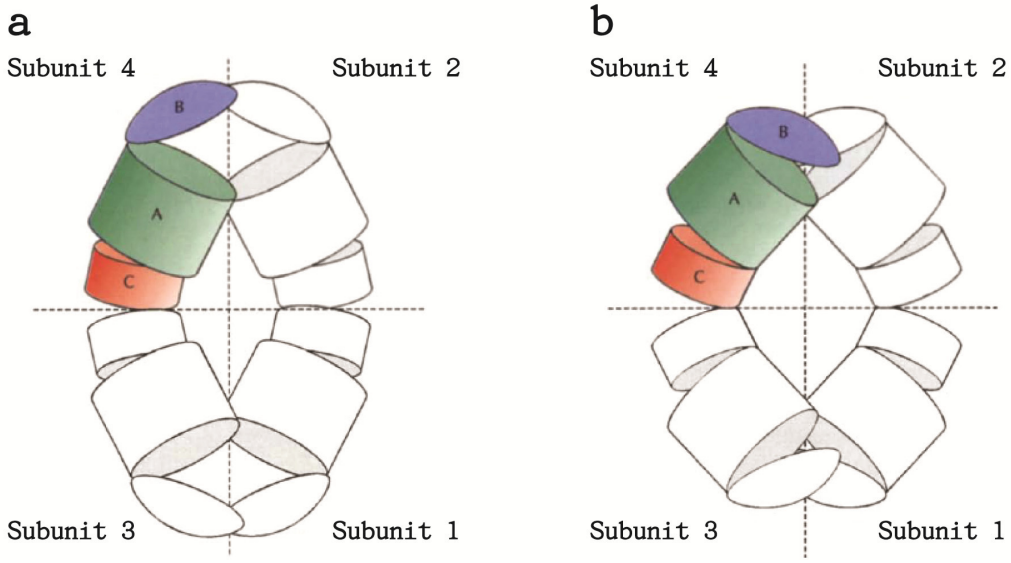


Figure 6.2. Illustration of the tetramer conformation change due to the movements of domains. (a) The T-state / inactive: domain B is opened. (b) The R-state / active: domain B is closed. From T- to R-state, the domain and subunit orientations cause a decrease in height and an increase in width of the tetramer.



6.1.2 Applied structures of target protein

For the purpose of ligand design and database mining, a known 3D protein-complex structure with a reliable resolution is needed. There is a bioassay result published on the PubChem website (PubChem AID: 361) which utilises quantitative high-throughput screening (Inglese, Auld et al. 2006) to investigate the influence of 51,415 compounds on the effectiveness of the enzyme PK from *Bacillus stearothermophilus*. According to the result, there are 602 of the tested compounds identified as either PK activators or inhibitors, but none of the 3D protein-ligand complex structures has been reported.

An X-ray structure of sulphate-bound PK of *Leishmania mexicana* (*Lm*PYK; PDB ID: 3E0V) has been published (Tulloch, Morgan et al. 2008), which is in an active-like conformation. In the crystalline state, there are three and two sulphate ions in the active and effector sites of each subunit, illustrated as Figure 6.3. Comparing this to the ATP-bound PK structure 1A49 obtained from rabbit (49.4 % sequence identity with 3E0V examined by PDBSum of EMBL-EBI; http://www.ebi.ac.uk/pdbsum/), the two sulphate ions (out of three) in the active site, i.e. γS and βS, occupy the positions close to the γ-phosphate and β-phosphate of ATP, respectively (Figure 6.3 (b)). The βS sulphate ion forms three hydrogen bonds with H54, K335 (domain A) and R175 (domain B), and these interactions construct a bridge between domain A and B. The two effector site sulphate ions, S(i) and S(ii), mimic the phosphates of F-2,6-BP (Figure 6.3 (c)). Sulphate S(i) is bound in the same position as the 6-phosphate of F-1,6-BP in the PK structures of 1A3W (yeast; 49.5 % identity), 1LIU (human erythrocyte; 48.3 % identity) and 1T5A (tumour; 49.8 % identity). The other

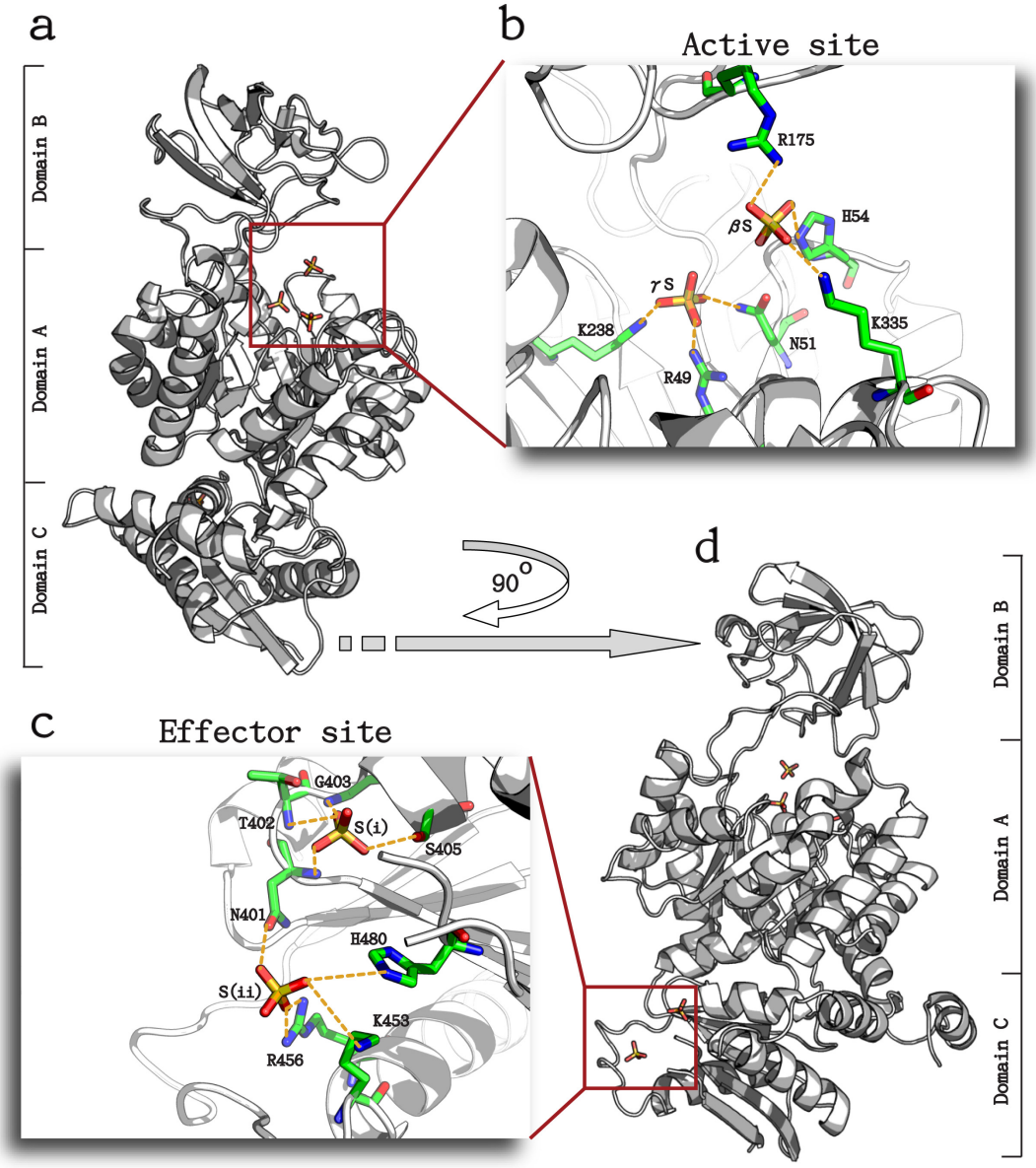
sulphate ion S(ii) is located in a position close to the 1-phosphate of F-1,6-BP found in similar structures.

We chose the sulphate-bound PK structure, i.e. 3E0V, to be the target for the ligand design in stage 1. Conversely, since we have successfully crystallised a protein-ligand complex at a resolution of 2.3 Å, whose sequence is exactly the same as *Lm*PYK 1PKL (Rigden, Phillips et al. 1999), we decide to use this structure in stage 2 to be the target. As the structure was crystallised at a neutral pH 7.2 in its active conformation, it is a more favourable target for ligand design.

6.1.3 Applied data mining approaches

The applied data mining approaches are mainly ACD (Atomic Characteristic Distance) detailed in chapter 2 and the facilities of EDULISS 2.0 which have also been developed for database mining. In order to explain the methods used more clearly, the details of ligand design will be described in the following sections.

Figure 6.3. Subunit structure (monomer) of sulphate-bound *Lm*PYK (PDB ID: 3E0V) at a resolution of 3.3 Å. (a) and (d) are orthogonal views. The active site can be seen in (b) and the effector site is shown in (c). All sulphate ions and the interacted residues are shown as sticks. The potential hydrogen bonds are shown as yellow dashed lines. The three active site sulphate ions, i.e. γS, βS and δS, are shown in (b), in which the γS and βS ions occupy the positions close to the γ-phosphate and β-phosphate of ATP comparing with the PK structure of 1A49 (rabbit), respectively. The two effector site sulphate ions, shown in (c), mimic the phosphates of F-2,6-BP. S(i) is bound in the same position as the 6-phosphate of F-1,6-BP in the PK structures of 1A3W (yeast), 1LIU (human) and 1T5A (tumour). S(ii) locates in a position close to the 1-phosphate of F-1,6-BP comparing with above mentioned structures (Tulloch, Morgan et al. 2008).



6.2 Results and discussion

This section also includes the descriptions of ligand designs and data mining processes used in stage 1 and 2.

6.2.1 Stage 1: To hit compounds that contain two or more sulphate groups with sulphur atoms that have geometries consistent with the distances between the sulphate ions in the sulphate-bound PK structure

The positions of the five sulphate ions found in the active and effector sites of the sulphate-bound PK structure, i.e. *Lm*PYK 3E0V, provide a starting point for the ligand design. Table 6.1 lists the geometrical distances between these ion pairs. We aim to select the compounds possessing two or more sulphate groups and the sulphur atoms are at the distances fitting the distances between sulphate ions found in *Lm*PYK 3E0V active or effector sites.

6.2.1.1 Design of screening criteria

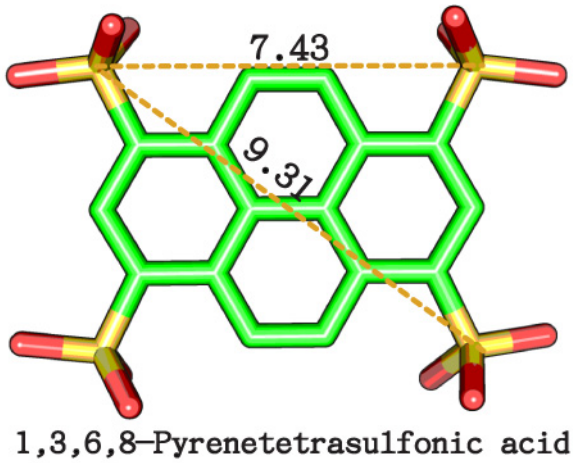
As EDULISS has a series of molecular descriptors regarding the functional groups and geometrical distances between atoms, the first screening can be accomplished via the interface of EDULISS. According to the geometrical distances between sulphate ion pairs listed in Table 6.1, we gave a tolerated range for the screening criterion as 6 to 9.5 Å. Considering the flexibility of the compounds, the preferred selection is to further choose the candidates which do not possess backbone flexibility. From the mining results, we selected 1,3,6,8-Pyrenetetrasulfonic acid for co-crystallising the ligand-protein complexes. The chemical structure is shown as Figure 6.4 together with the geometrical distances between the sulphur atoms.

Table 6.1. Distances between the five sulphate ions in the active and effector sites*. The five sulphate ions are shown in Figure 6.3 (b) and (c).

	Active site			Effector site
Subunit No.	δS - γS	γS - βS	βS - δS	S(i) - S(ii)
A	7.99	6.54	8.95	8.61
B	8.53	6.92	8.80	8.74
C	8.27	6.51	8.90	9.29
D	8.51	7.25	8.31	9.67
Range	7.99 – 8.53	6.51 – 7.25	8.31 – 8.95	8.61 – 9.67
Average	8.3	6.8	8.7	9.1

*: All of the distances are in angstrom (Å).

Figure 6.4. Pyrene-like compounds whose sulphate groups are at the distances fitting the criterion, i.e. 6 to 9.5 Å, to mimic the positions of sulphate ions found in *Lm*PYK 3E0V active and effect sites.

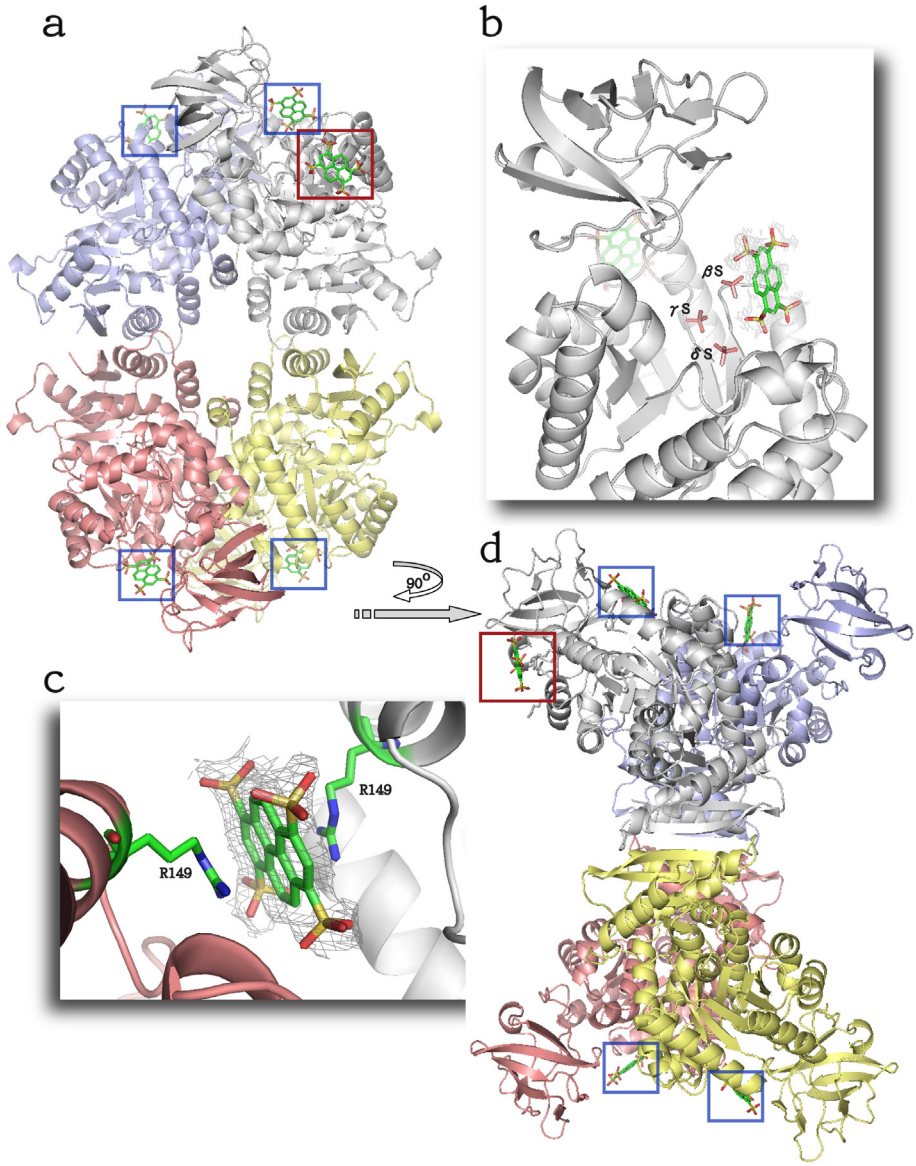


6.2.1.2 Screening and crystallisation result

The 1,3,6,8-Pyrenetetrasulfonic acid bound PK complex has been crystallised successfully at the resolution 2.1 Å and its structure model is shown as Figure 6.5. The orthogonal views of the tetramer conformation are shown as Figure 6.5 (a) and (d) where the subunit A, B, C and D are coloured by red, gray, blue and yellow, respectively. Four pyrene-like compounds, i.e. 1,3,6,8-Pyrenetetrasulfonic acid, can be found in the active sites of each subunit (marked by blue frames) and another odd one is only bound the subunit B (marked by red frames). Figure 6.5 (b) illustrates the active site of subunit B where the three pink ions, labelled as γS, βS and δS, are of the *Lm*PYK 3E0V active site sulphate ions placed by overlapping the domain A and C of the subunit B (aligning the residues 1-88 and 183-497; RMSD fit: 0.362). The pyrene-like compounds have docked on the active sites but did not exactly fit the required position, i.e. the positions of the γS, βS and δS sulphate ions. However, there is an odd pyrene-like compound bound on the edge of subunit B between *Lm*PYK tetramers within the crystal lattice. The compound interacts with R194 via π-π interaction illustrated as Figure 6.5 (c) and shown with its electron density where the red structure belongs to the neighbouring symmetry related molecule. At this binding position, the compound potentially enhances the packing arrangement in the crystal unit cells thus it promotes crystallisation and provides a better structure resolution than *Lm*PYK 3E0V (2.1 v.s. 3.3 Å). The better resolution provides more precise observation for the residues, metals or waters found in the active and effector site for further ligand design and screening. Furthermore, the pyrene-like compounds can only bind the edge of subunit B but not in the active sites at low concentration. As this is a useful property in crystallisation, this compound has been

used in lower concentration for some consequent crystallographic experiments in order to improve the structure resolution.

Figure 6.5. Structure model of the pyrene-like compound bound *Lm*PYK at the resolution 2.1 Å. (a) and (d) illustrate the tetramer conformation in orthogonal views (subunit A: red, B: gray, C: blue and D: yellow). Four pyrene-like compounds, i.e. 1,3,6,8-Pyrenetetrasulfonic acid, can be found in the active sites of each subunit (marked by blue frames) and the other one is only bound subunit B (marked by red frames). (b) shows a pyrene-like compound binding the active site of subunit B with its electron density. The three pink ions, labelled as γS, βS and δS, are the *Lm*PYK 3E0V active site sulphate ions placed by overlapping the domain A and C (aligning the residues 1-88 and 183-497; RMSD: 0.362). (c) shows the odd pyrene-like compound interacting with R194 (subunit B) via π-π interaction. The red structure belongs to the neighbouring symmetry related molecule.

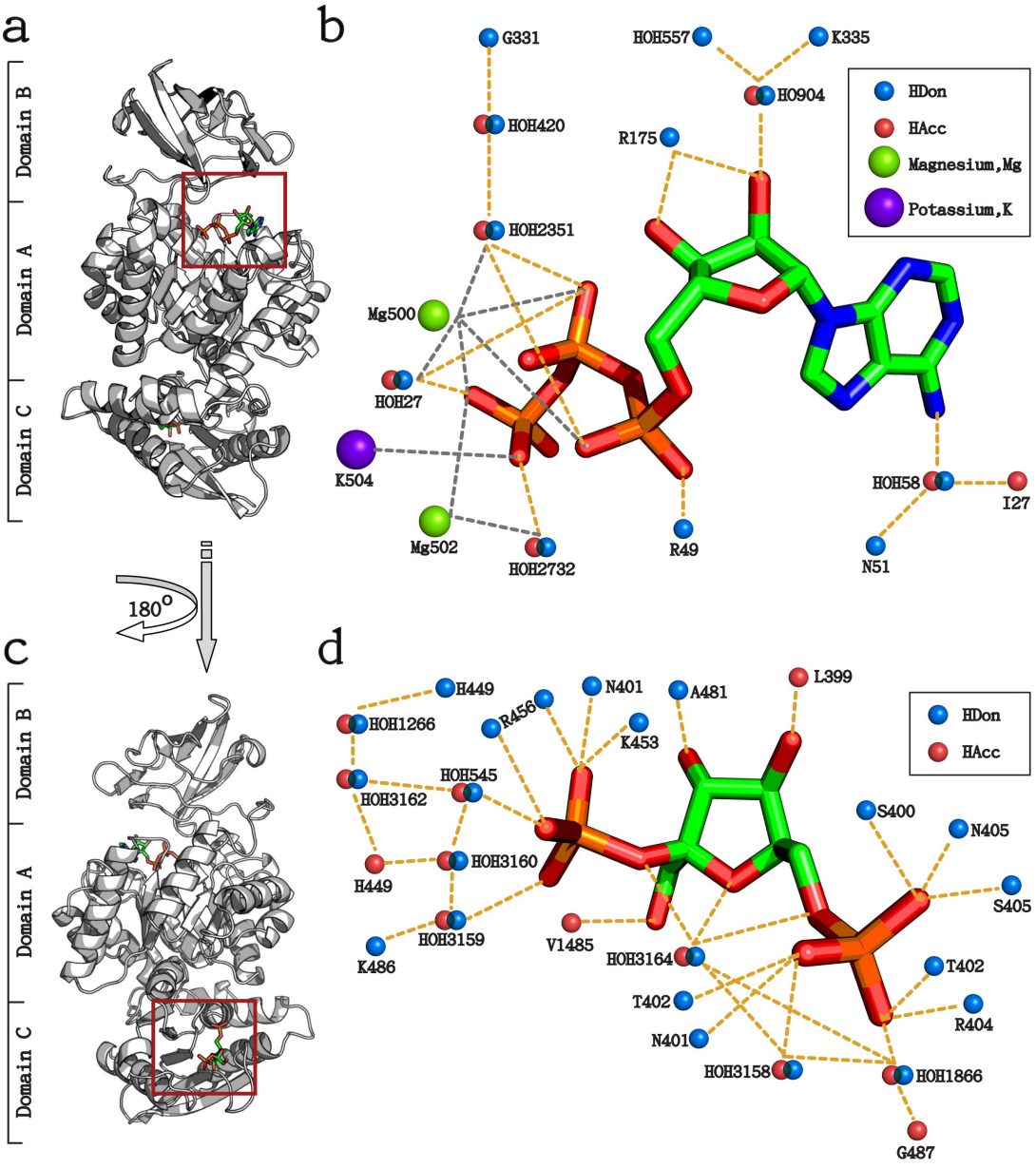


6.2.2 Stage 2: Application of Atomic Characteristic Distance (ACD) to more complex screening requirements

In stage 1, the screening simply considered the distances between identical atoms and the crystal model shows that the selected candidates appeared to bind the active sites weakly. In this stage, the screening takes both the interatomic distances between various atom types and the hydrogen bond interactions observed in the binding sites into account. Also, we use an *Lm*PYK/ATP/FBP complex structure as the target instead of the sulphate-bound PK structure used in stage 1.

As described in 6.1.2 section, the sequence of the used target protein is entirely the same as *Lm*PYK 1PKL but it is in the R-state / active conformation. There are ATP and oxalate molecules in the active site and a F-2,6-BP molecule in the effector site respectively, illustrated as Figure 6.6 (a) and (c). The potential interactions of the ATP and F-2,6-BP molecules in the two sites are schematically shown in Figure 6.6 (b) and (d). The interactions are shown as yellow dashed lines, which are defined as at a potential hydrogen bond length (≤ about 3.2 Å). Several water molecules are involved in the interactions between these two molecules and residues.

Figure 6.6. Subunit structure of the target which is an ATP- and F-2,6-BP-bound complex in R-state / active conformation, whose sequence is the same as *Lm*PYK 1PKL (Rigden, Phillips et al. 1999). (a) and (c) are reversed views. The active site ATP and effector site F-2,6-BP molecules are marked by red frames. (b) and (d) schematically diagrams the potential H-bond interactions of ATP and F-2,6-BP. The interactions are shown as yellow dashed lines, which are defined as at a potential hydrogen bond length (≤ about 3.2 Å). The hydrogen bond donors (HDon) and acceptors (HAcc) are coloured by blue and red, and labelled by their residue names and IDs, respectively. The potential ionic interactions are shown as gray dashed lines.

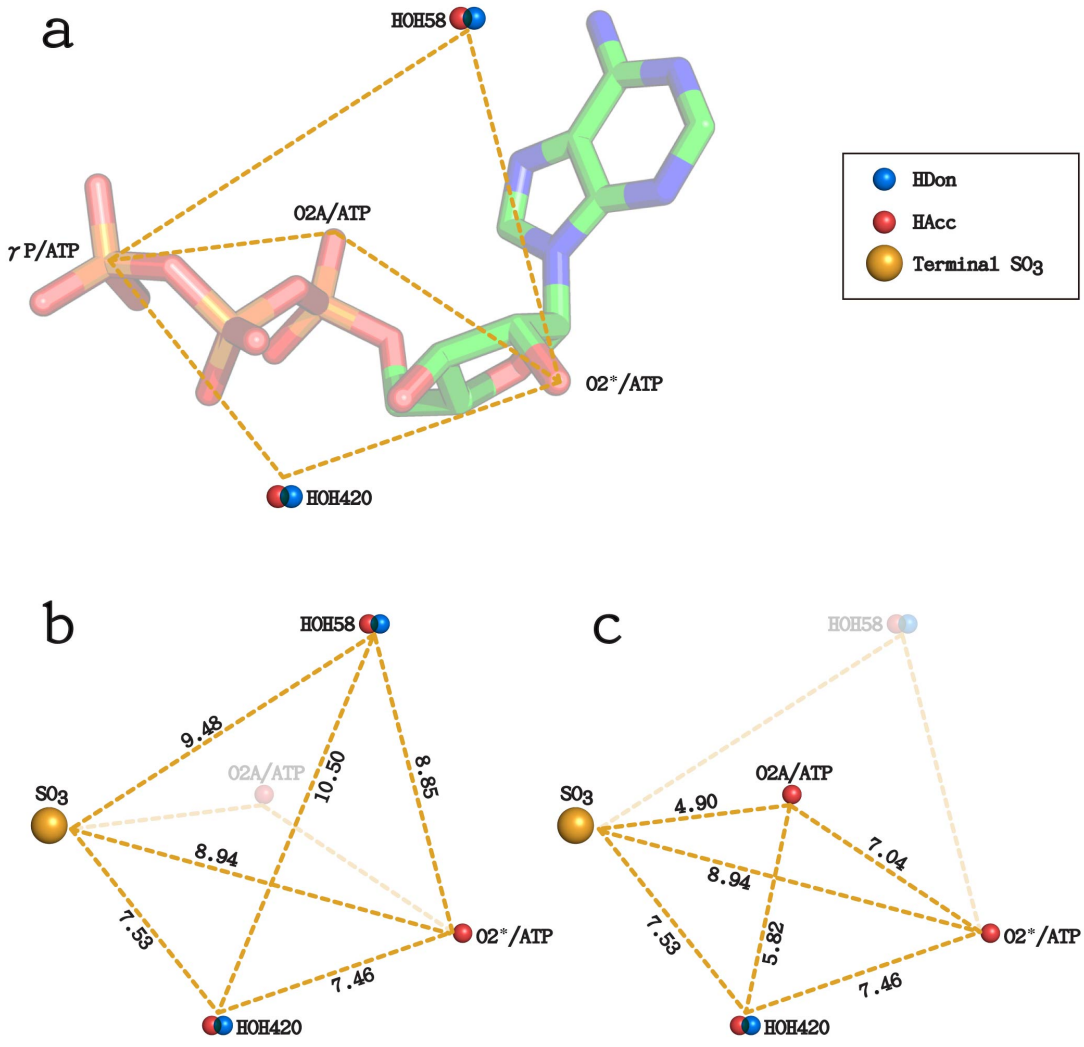


6.2.2.1 Design of screening criteria

We aim to find the compounds which are able to mimic the interactions of ATP and F-2,6-BP molecules in the two binding sites. Figure 6.7 (a) and Figure 6.8 (a) schematically show the interesting interatomic distances found in ATP and F-2,6-BP molecules, respectively. In the case of ATP (Figure 6.7), the interatomic contacts and distances between O2*/ATP, O2A/ATP, γP/ATP, HOH58 and HOH420 form the screening criteria to search for the preferred candidates, but we select a terminal sulphate ($-SO_3^-$) to substitute for the position of γP/ATP in order to continue the study from stage 1. The designs of the screening motifs have been topologically outlined in Figure 6.7 (b) and (c). The first screening operation extracts compounds from the EDULISS 2.0 collection which possess one or more terminal $-SO_3^-$, and then we further investigate their interatomic distances between specified atom types which have to geometrically match the motifs shown as the figures. The two atoms in the compounds to mimic HOH58 and HOH420 could be either hydrogen bond donor or acceptor. A series of tolerances have been given for each interatomic distance from 10 to 25 %.

Although the interface of EDULISS 2.0 has provided the facility to perform ACD, it can only ensure that the hit compounds possess the specified interatomic distances but not the conjunctions. An in-house java script was written using the CDK java tool kit and was used to accomplish these multiple requirements. This screening approach has also been applied in the cases of F-2,6-BP molecules (Figure 6.8) which illustrates the relevant interaction of the bound F-2,6-BP molecules in the effector site and the motifs for screening, shown as Figure 6.8 (b) to (f).

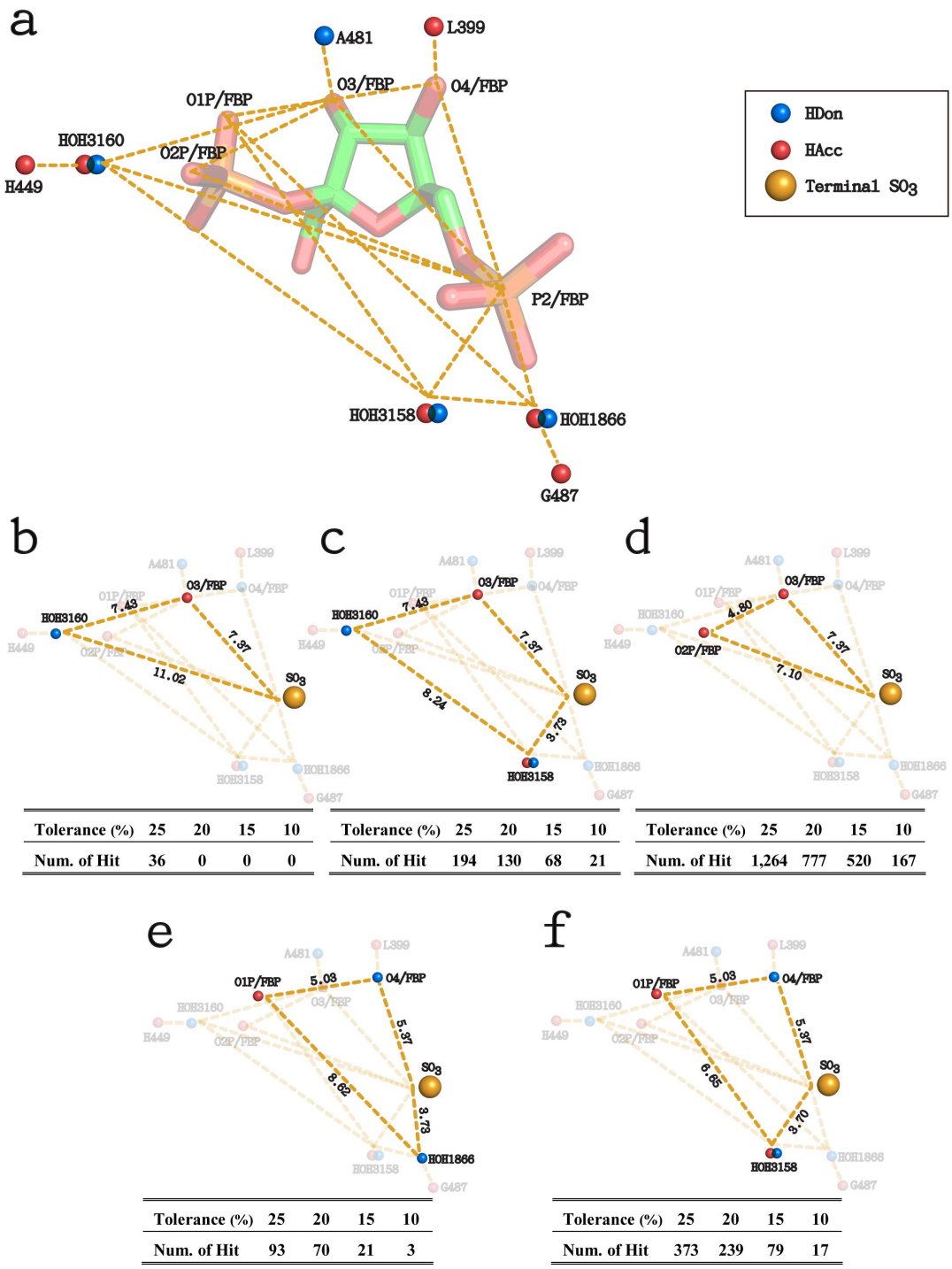
Figure 6.7. Schematic diagrams of the interesting interactions of ATP in the active site. (a) represents the selected interatomic conjunctions between three ATP atoms and two waters. (b) and (c) topologically outline the interatomic conjunctions and distances between the selected atoms, and form the screening motifs. The hydrogen bond donors (HDon) and acceptors (HAcc) are coloured by blue and red, and labelled by their residue names and IDs, respectively. The position of γP/ATP has been replaced by a terminal sulphate (SO_3). The tables below (b) and (c) list the number of hit compounds based on the motifs and a series of tolerances for the interatomic distances.



Tolerance (%)	25	20	15	10
Num. of Hit	125	49	25	11

Tolerance (%)	25	20	15	10
Num. of Hit	143	65	33	3

Figure 6.8. Schematic diagrams of the relevant interactions of F-2,6-BP (FBP) in the effector site. (a) represents the selected interatomic conjunctions between five atoms in F-2,6-BP and three waters. The description of (b) to (f) is the same as described in Figure 6.7 (b) and (c). The position of P2/FBP has been replaced by a terminal sulphate (SO_3).



Tolerance (%)	25	20	15	10
Num. of Hit	36	0	0	0

Tolerance (%)	25	20	15	10
Num. of Hit	194	130	68	21

Tolerance (%)	25	20	15	10
Num. of Hit	1,264	777	520	167

Tolerance (%)	25	20	15	10
Num. of Hit	93	70	21	3

Tolerance (%)	25	20	15	10
Num. of Hit	373	239	79	17

6.2.2.2 Screening results

The screening result shows that in the EDULISS 2.0 collection 5,365 compounds possess one or more terminal SO_3. The numbers of hit compounds in a range of tolerances are tabulated below the figures of each case in Figure 6.7 and 6.8, respectively. Since the screening has effectively excluded the unfit compounds, the numbers of hits are acceptable for further detailed inspection to select the preferred candidates one by one. Also there is a set of compounds which appears to be hit throughout the various cases consistently. Considering the reasons mentioned above and the compound solubility, we selected 8 compounds to be test candidates for further experimental assay. The chemical structures of the 8 compounds are shown as Figure 6.9 in 2D together with their common properties. As the application of ACD aims to screen compounds which possess the specified interatomic conjunctions and distances regardless of the conventional rule-based screenings, only one compound (Comp. 4) fits the Lipinski's rule of five.

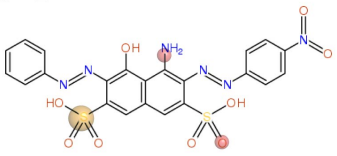
6.2.2.3 Bioassay and crystallisation results

Table 6.2 summarises the examined inhibition of the *Lm*PYK enzyme activity caused by the 8 compounds. The examination is based on the pH change monitored over time as a direct result of *Lm*PYK enzyme activity. In the uninhibited state, the pH value increases from 6.3 to 7.3 as *Lm*PYK converts ADP and PEP into ATP and pyruvate. The pH value change is caused by the loss or gain of a hydroxyl group. All of the assays were done in duplicate and the values are listed in Rate (1) and (2) columns. As a result, Comp. 3 and Comp. 5 show 33.34 and 100 % inhibition in *Lm*PYK enzyme activity, respectively.

Figure 6.9. Chemical structures of the 8 selected compounds whose ACDs fitted the screening motifs shown in Figure 6.7 (b) to (c) and Figure 6.8 (b) to (f). The common chemical properties are tabulated for each compound. The orientated 3D structures are the representative if the compound fits two or more screening motifs shown in Figure 6.7 and Figure 6.8. (continued on next page)

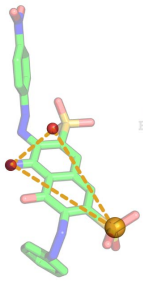
Comp. 1

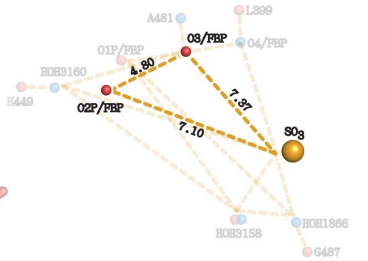
Fit Fig 6.8(d)



Naphthol Blue Black
SPH1-106-635 / N9002

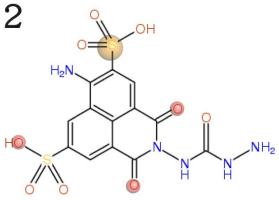
MW	MLogP	nHDon	nHAcc	nCIC[a]	RBN[b]	PSA[c]
572.58	1.10	5	14	4	7	267.01





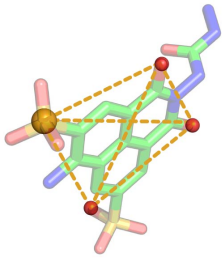
Comp. 2

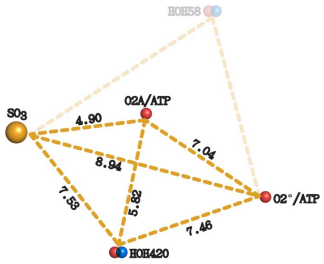
Fit Fig 6.7(c)



Lucifer Yellow CH dipotassium salt
SPH1-105-237 / L0144

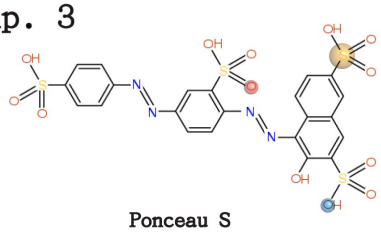
MW	MLogP	nHDon	nHAcc	nCIC	RBN	PSA
445.43	-0.48	8	14	3	3	257.74





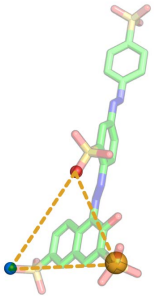
Comp. 3

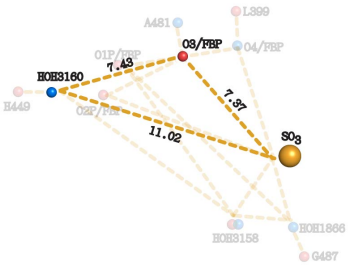
Fit Fig 6.8(b)



Ponceau S
SPH1-107-090 / P3504

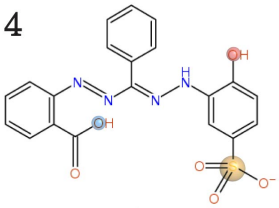
MW	MLogP	nHDon	nHAcc	nCIC	RBN	PSA
672.70	-0.20	5	17	4	8	320.67





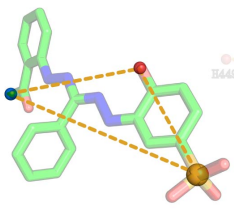
Comp. 4

Fit Fig 6.8(b)



Zincon sodium salt
SPH1-171-370 / 201332

MW	MLogP	nHDon	nHAcc	nCIC	RBN	PSA
439.46	2.73	3	10	3	7	172.22



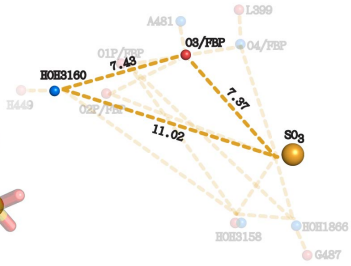
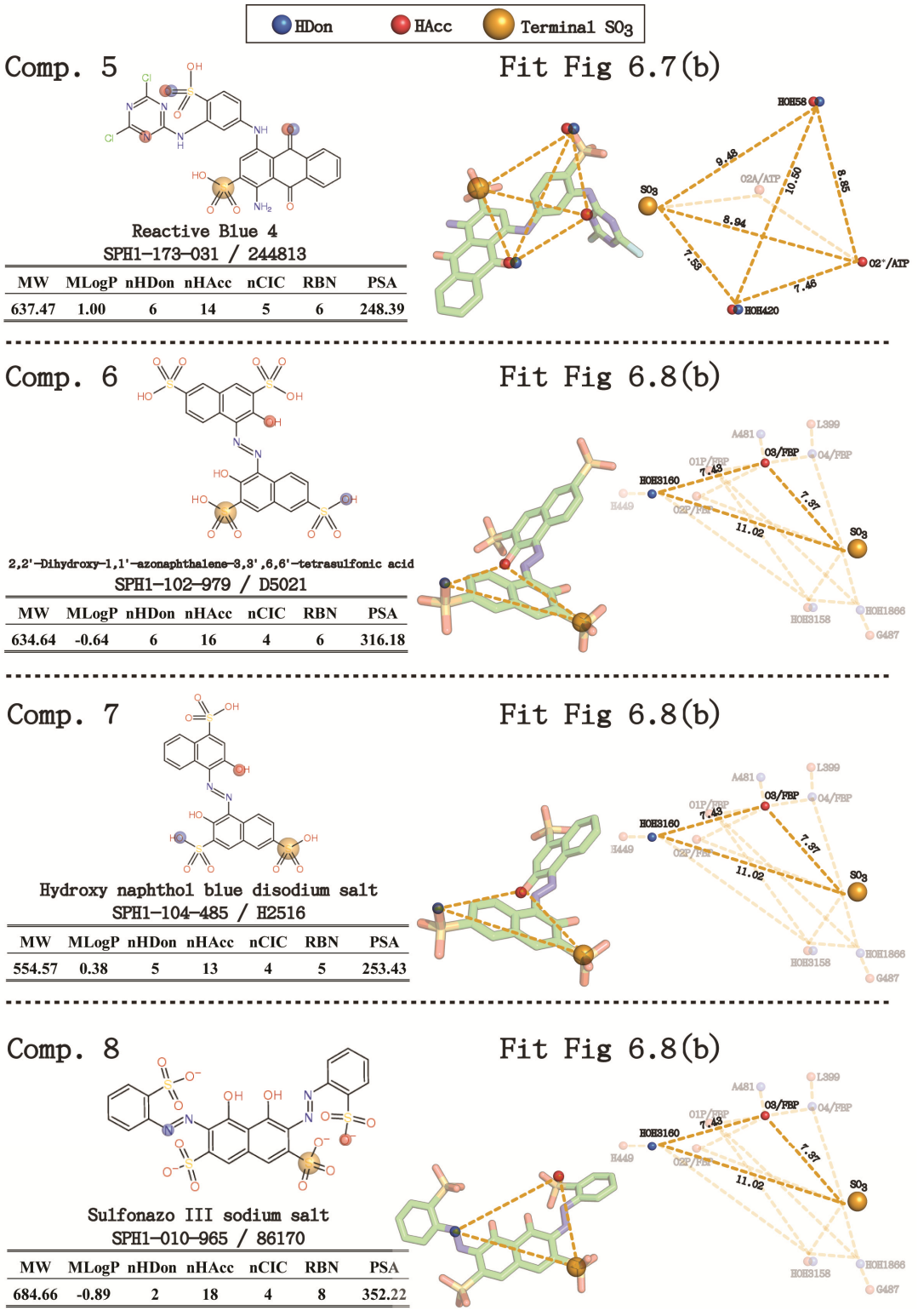


Figure 6.9. Continued.



a: Number of rings; **b**: Number of rotatable bonds; **c**: Topological polar surface area using N,O,S,P polar contributions.

Table 6.2. Summary of all *Lm*PYK inhibition data caused by the 8 compounds listed in Figure 6.9 [a].

Compounds	Rate (1)	Rate (2)	Average	Activity (%)	Inhibition (%)	Activation (fold)
Control [b]	21.21	22.27	21.74	100.00	-	-
Comp. 1	31.51	27.96	29.74	136.80	-	1.4
Comp. 2	65.25	61.36	63.31	291.19	-	2.9
Comp. 3	16.18	12.76	14.47	66.66	33.34	-
Comp. 4	46.24	48.60	47.42	218.12	-	2.2
Comp. 5	0	0	0	0	100	-
Comp. 6	-	49.25	49.25	226.54	-	2.3
Comp. 7	79.00	84.88	81.94	376.91	-	3.8
Comp. 8	NA	NA	NA	NA	-	-

a: All assays were done in duplicate and the values are listed in Rate (1) and (2) columns.
b: No ligand added.

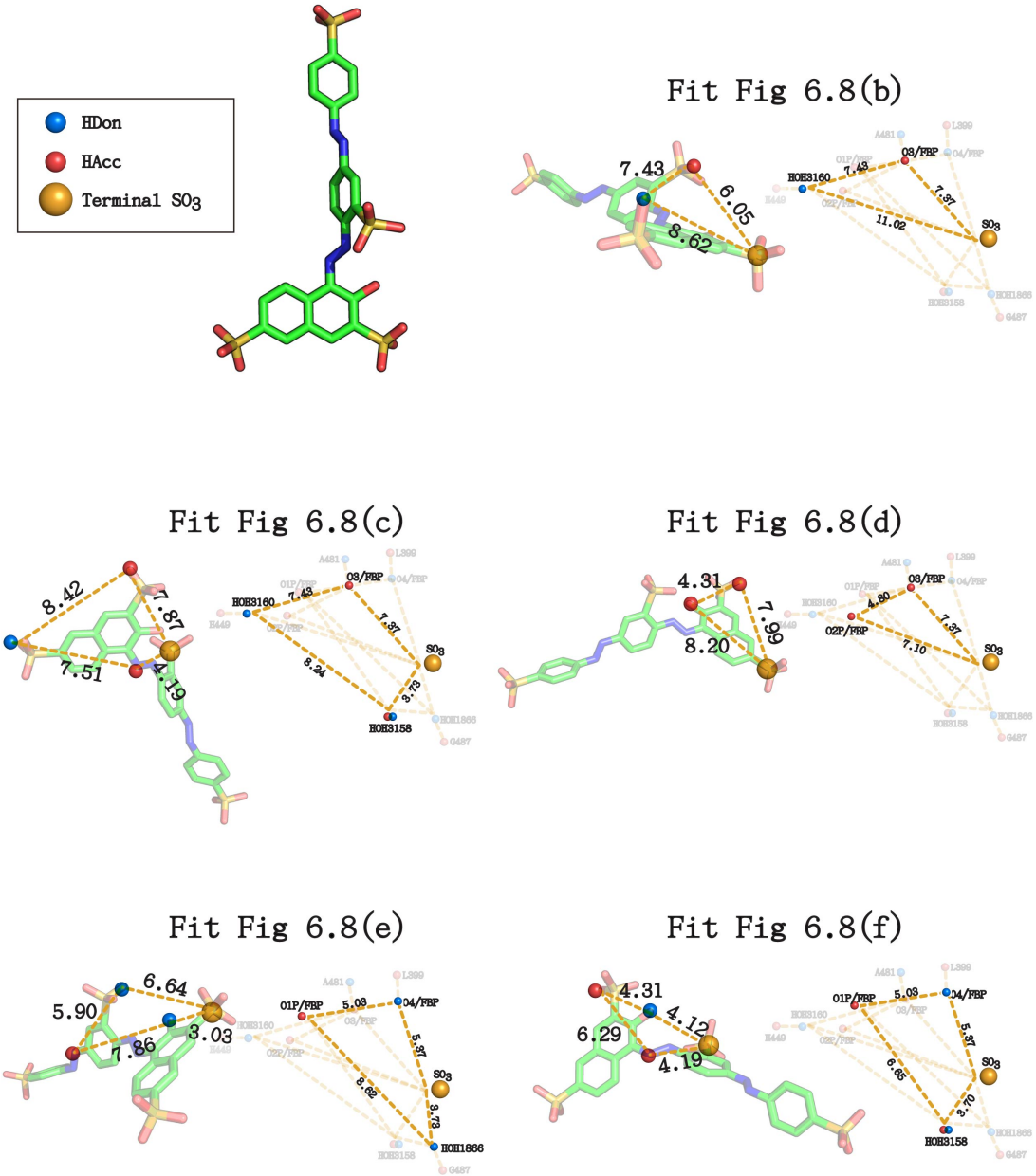
Comp. 3 / Ponceau S is predicted to bind the effector site as it has been found in the hit of each screening motif for the effector site. The schematic diagrams of the fitted ACDs, i.e. interatomic conjunctions and distances between the selected atoms, are shown as Figure 6.10. The most well fitted part is in its disulfonic naphthol fragment. The Comp. 3 bound complex structure has been crystallised successfully at a resolution of 2.7 Å and its structure model is shown as Figure 6.11 (a) where the four subunits of the tetramer conformation are coloured by yellow, gray, green and red respectively. The blue subunit belongs to the neighbouring symmetry unit. Comp. 3, marked by the red frame, binds the effector site of the red subunit and is partially placed in the active site of the blue subunit, illustrated as Figure 6.11 (b). There are four residues involved in the binding interactions in the effector site, including H483,

K486 and T402 at a potential hydrogen bond length (≤ about 3.2 Å) and V485 forming hydrophobic contact.

As F-2,6-BP (FBP) and three waters, including HOH3158, HOH3160 and HOH1866, form a part of the screening motifs, these four molecules have been placed by overlapping the domain A and C of the *Lm*PYK/ATP/FBP complex structure (aligning the residues 1-86 and 206-498; RMSD fit: 0.334) in order to be contrasted with the binding position of Comp. 3, shown as Figure 6.11 (c). Although the binding position of Comp. 3 does not appear to exactly overlap the four motif molecules, the disulfonic naphthol fragment of Comp. 3 is very close to a phosphate group of the FBP and one of the waters at the distance ≤ 3.2 Å. It occupies a part of FBP binding site. As the result, the protein loop from residue A481 to G487 forming the allosteric binding pocket is accordingly rotated to accommodate the binding of Comp. 3. The binding of the allosteric effector molecule FBP is therefore no longer possible as the site has been blocked and cannot accommodate FBP binding.

As the effector site residues mostly interact with the disulfonic naphthol fragment of Comp. 3, we further tested 2-naphthol-3,6 disulfonic acid, i.e. the core of the fragment. The crystal of *Lm*PYK in the complex with 2-naphthol-3,6 disulfonic acid has been obtained but the extremely long cell edge (about 500 Å) is problematic for the collection of processable data. The modelling and refinement process is still underway at this time.

Figure 6.10. Schematic diagrams of Comp. 3 / Ponceau S in different orientations to show whose ACDs, i.e. interatomic conjunctions and distances between the selected atoms, are fitted to the five screening motifs shown in Figure 6.8 (b) to (f).



Comp. 5 / Reactive Blue 4 is expected to bind the active site as it fits the motif for the active site shown as Figure 6.7 (b) at a tolerance 15 %. The schematic diagrams of the fitted ACDs are shown as Figure 6.12. Although Comp. 5 shows a good inhibition in *Lm*PYK enzyme activity according to the bioassay result, a crystal of its *Lm*PYK complex bound at the active site could not be obtained successfully.

Previous studies indicate that a series of anthraquinone dyes, including Comp. 5, are able to mimic the adenine nucleotides acting as ligands to interact with nucleotide-binding sites of some relevant enzymes and receptors such as hexokinase (Puri 1994; De Moliner, Moro et al. 2003; Breier, Bohacova et al. 2006). We therefore modified the screening technique and used MCS (Maximum Common Subgraph, described in chapter 3) to select other analogues of Comp. 5 containing the core structure anthraquinone. The additionally selected candidates are Reactive Blue 2, Acid Blue 25 and Acid Blue 80, and their chemical structures with the common properties are shown as Figure 6.13. The core structure, i.e. anthraquinone, is coloured red.

Only Acid Blue 80 bound complex crystal has been obtained at a resolution of 2.3 Å. The model of the tetramer structure is shown as Figure 6.14 (a) where the bound Acid Blue 80, marked by the red frames, can be found in the active sites of the two gray subunits. The Acid Blue 80 molecule binds in the active site via potential hydrogen bond interactions with five residues including the side chains of R49, N51, Y304 (~3.5 Å) and K335 (~3.7 Å) and the backbone nitrogen of G331, illustrated as Figure 6.14 (b).

In order to contrast the geometrical binding position of Acid Blue 80 with the position of ATP, Figure 6.14 (c) shows the binding positions of the two molecules by

overlapping the domain A and C of the *Lm*PYK/ATP/FBP complex structure (aligning the residues 1-86 and 206-498; RMSD fit: 0.413). One of the sulphate groups in Acid Blue 80 almost overlaps the α-phosphate of ATP. As Acid Blue 80 occupies the active site, it consequently prevents the ADP molecule from binding.

6.3 Summary

The facilities of EDULISS 2.0 provide a convenient interface allowing users to perform a series of queries. It is most useful in the early stages of experimental ligand binding projects for initial screening and to indicate which initial compounds should be purchased. Its large collection of small-molecule structure files also offers a higher probability to hit candidates, particularly when using complex screening criteria. The application of ACD gives a good scope for the ligand design as well as helping to reduce the number of candidates so that researchers can directly inspect and further select the preferred compounds based on empirical bio- or physicochemical knowledge. The screening approaches used in this study have shown success in selecting ligands that have been shown by X-ray crystallography to bind to the template protein. The crystal structures of these complexes provide insight into the interactions occurring in ligand binding site and into the conformation changes that can be induced in the complex.

Figure 6.11. Structure model of Comp. 3 / Ponceau S bound *Lm*PYK at resolution 2.7 Å. (a) illustrates the tetramer conformation where the four subunits are coloured by yellow, gray, green and red respectively. The blue subunit belongs to the neighbouring symmetry unit. Comp. 3 / Ponceau S binds the effector site of the red subunit and is partially placed in the active site of the blue subunit marked by the red frame. (b) enlarges the Comp. 3 binding site and shows the interactions between Comp. 3 and four residues in the effector site via hydrogen bond and hydrophobic interactions. (c) contrasts the geometrical binding position of Comp. 3 with the location of F-2,6-BP (FBP) and three water molecules located by overlapping the domain A and C of the *Lm*PYK/ATP/FBP complex structure (aligning the residues 1-86 and 206-498; RMSD fit: 0.334).

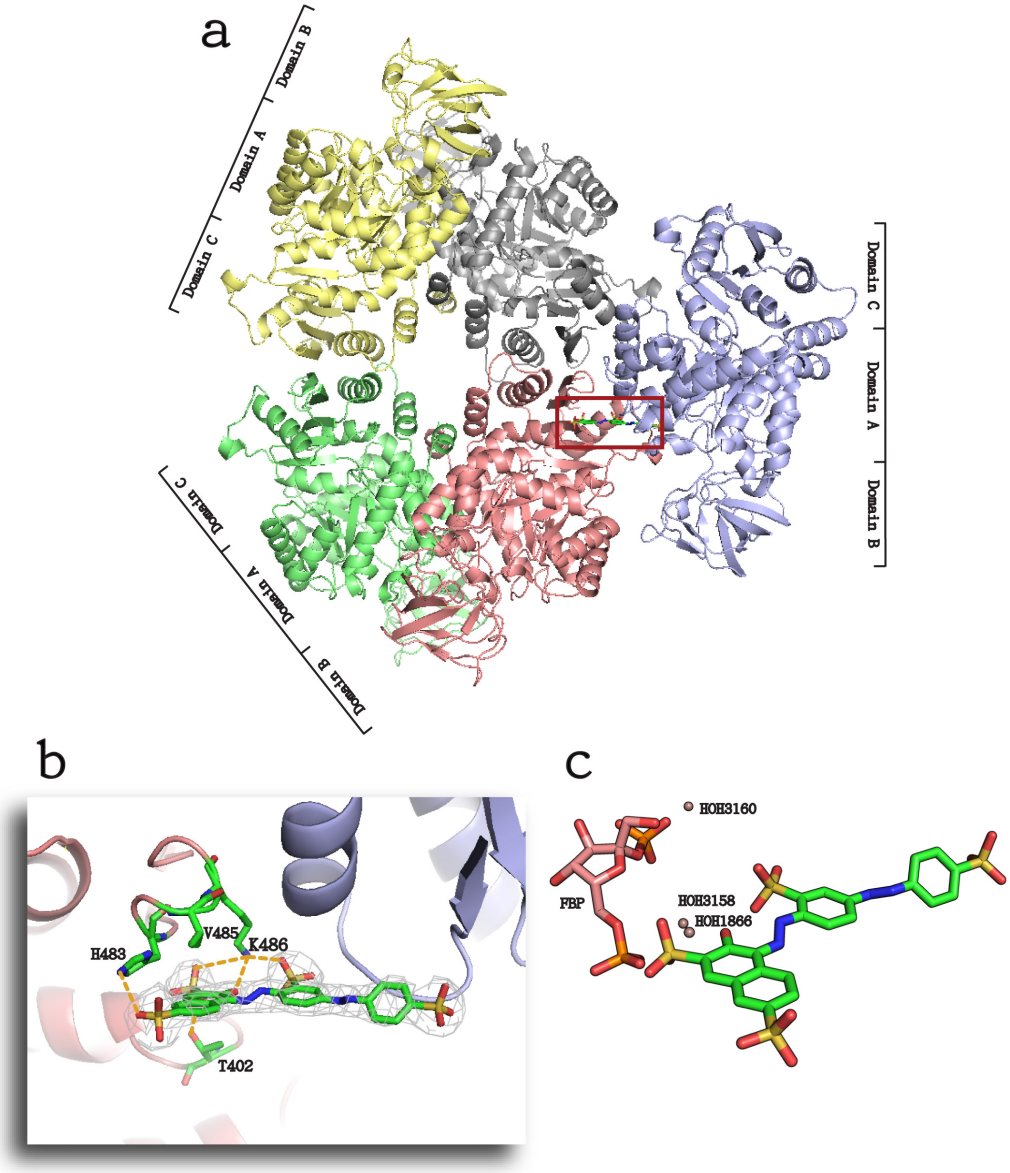


Figure 6.12. Schematic diagrams of Comp. 5 / Reactive Blue 4 in different orientations to show whose ACDs, i.e. interatomic conjunctions and distances between the selected atoms, are fitted the screening motifs shown in Figure 6.7 (b) and (c).

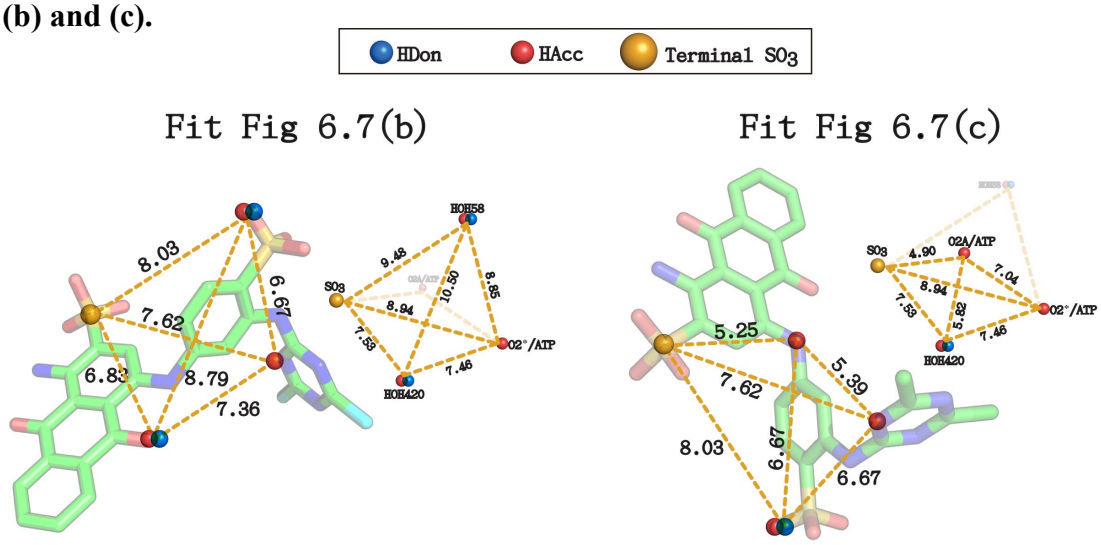
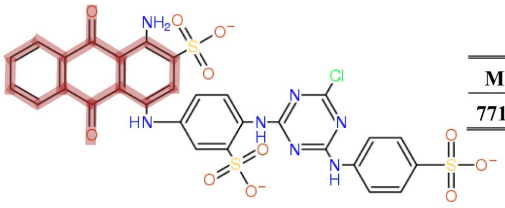
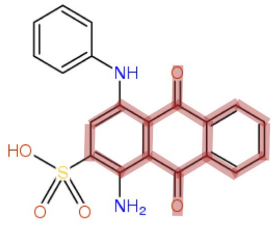


Figure 6.13. Chemical structures of the 3 additionally selected compounds which possess the same core, i.e. Anthraquinone marked by red, as Comp. 5 / Reactive Blue 4. The common chemical properties are tabulated for each compound. The labels represent the compound's name, SPH number and the ID assigned by vendor.



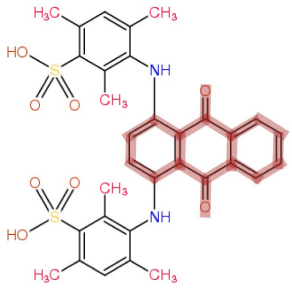
Reactive Blue 2
SPH1-197-618 / R115

MW	MLogP	nHDon	nHAcc	nCIC	RBN	PSA
771.19	0.62	5	18	6	9	331.66



Acid Blue 25
SPH1-171-641 / 210684

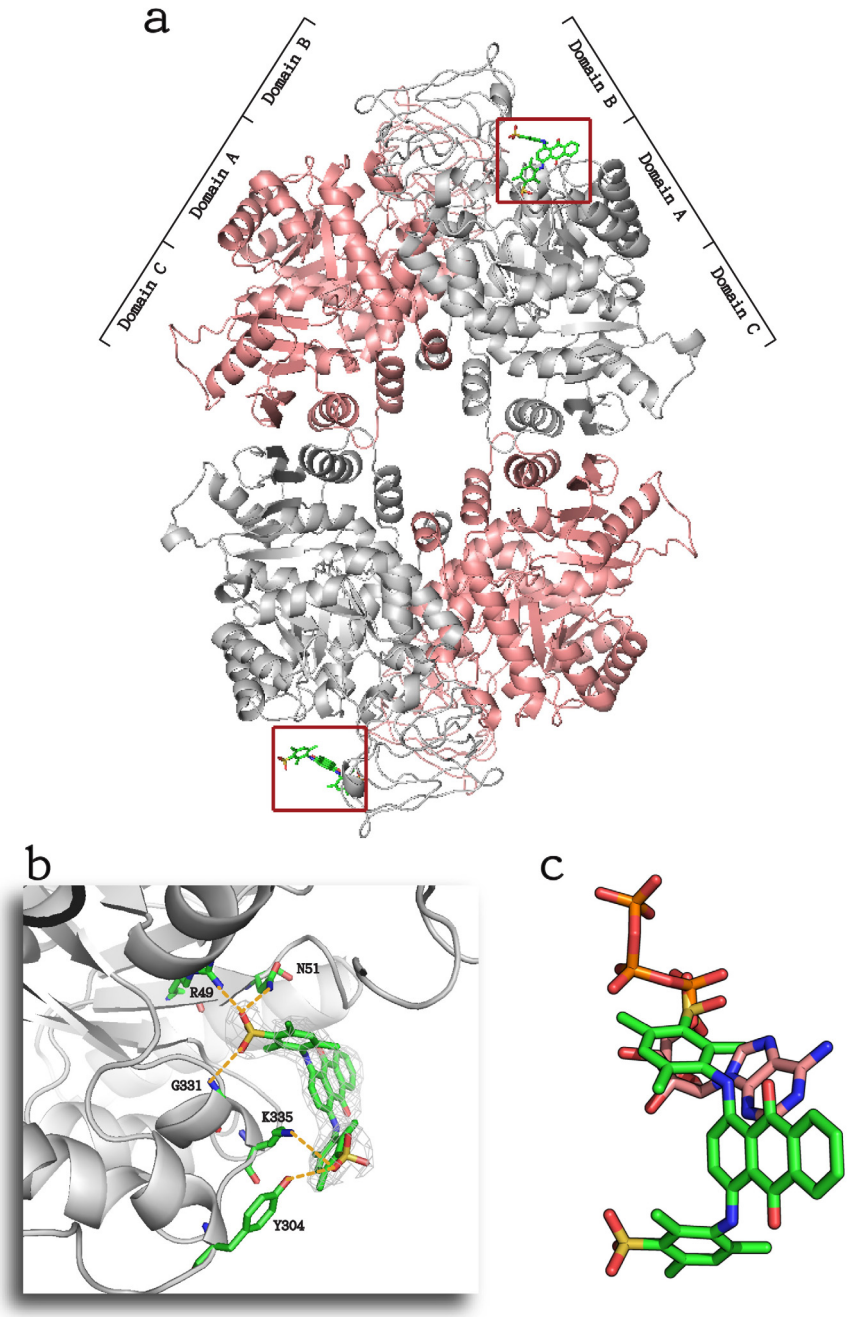
MW	MLogP	nHDon	nHAcc	nCIC	RBN	PSA
394.43	1.77	4	7	4	3	134.94



Acid Blue 80
SPH1-171-628 / 210323

MW	MLogP	nHDon	nHAcc	nCIC	RBN	PSA
634.78	2.88	4	10	5	6	183.70

Figure 6.14. Structure model of Acid Blue 80 bound *Lm*PYK at the resolution 2.3 Å. (a) illustrates the tetramer conformation where the bound Acid Blue 80 molecules, marked by the red frames, can be found in the active sites of the two gray subunits. (b) enlarges the Acid Blue 80 binding site and shows the potential hydrogen bond interactions between Acid Blue 80 and five residues in the active site. (c) contrasts the geometrical binding position of Acid Blue 80 with the location of ATP located by overlapping the domain A and C of the *Lm*PYK/ATP/FBP complex structure (aligning the residues 1-86 and 206-498; RMSD fit: 0.413).



6.4 Reference:

Breier, A., V. Bohacova, et al. (2006). "Inhibition of (Na (+)/K (+))-ATPase by Cibacron Blue 3G-A and its analogues." General Physiology and Biophysics **25**(4): 439-53.

De Moliner, E., S. Moro, et al. (2003). "Inhibition of Protein Kinase CK2 by Anthraquinone-related Compounds a structural insight" Journal of Biological Chemistry **278**(3): 1831-1836.

Inglese, J., D. S. Auld, et al. (2006). "Quantitative high-throughput screening: A titration-based approach that efficiently identifies biological activities in large chemical libraries." Proceedings of the National Academy of Sciences **103**(31): 11473.

Lovell, S. C., A. H. Mullick, et al. (1998). "Cooperativity in Bacillus stearothermophilus pyruvate kinase." Journal of Molecular Biology **276**(4): 839-851.

Matte, A., L. W. Tari, et al. (1998). "How do kinases transfer phosphoryl groups?" Structure **6**(4): 413-419.

Mattevi, A., M. Bolognesi, et al. (1996). "The allosteric regulation of pyruvate kinase." FEBS Letters **389**(1): 15-19.

Mattevi, A., G. Valentini, et al. (1995). "Crystal structure of Escherichia coli pyruvate kinase type I: molecular basis of the allosteric transition." Structure **3**(7): 729-741.

Mesecar, A. D. and T. Nowak (1997). "Metal-ion-mediated allosteric triggering of yeast pyruvate kinase. 2. A multidimensional thermodynamic linked-function analysis." Biochemistry **36**(22): 6803-13.

Oria-Hernandez, J., N. Cabrera, et al. (2005). "Pyruvate Kinase Revisited: the activating effector of K^+." Journal of Biological Chemistry **280**(45): 37924.

Puri, R. N. (1994). "Inactivation of yeast hexokinase by Cibacron Blue 3G-A: spectral, kinetic and structural investigations." Biochemical Journal **300**(1): 91-98.

Rigden, D. J., S. E. V. Phillips, et al. (1999). "The structure of pyruvate kinase from Leishmania mexicana reveals details of the allosteric transition and unusual effector specificity." Journal of Molecular Biology **291**(3): 615-635.

Schaftingen, E., F. R. Opperdoes, et al. (1985). "Stimulation of Trypanosoma brucei pyruvate kinase by fructose 2, 6-bisphosphate." European Journal of Biochemistry **153**(2): 403-406.

Tulloch, L. B., H. P. Morgan, et al. (2008). "Sulphate Removal Induces a Major Conformational Change in Leishmania mexicana Pyruvate Kinase in the Crystalline State." Journal of Molecular Biology **383**(3): 615-626.

Valentini, G., L. Chiarelli, et al. (2000). "The Allosteric Regulation of Pyruvate Kinase: a site-directed mutagenesis study." Journal of Biological Chemistry **275**(24): 18145-18152.

Yu, S., L. L. Y. Lee, et al. (2003). "Effects of metabolites on the structural dynamics of rabbit muscle pyruvate kinase." Biophysical Chemistry **103**(1): 1-11.

7. Development of a database and web-based interface for novel metalloprotein descriptors and their application

7.1 Introduction

Previous chapters have discussed the physicochemical descriptors, including the atomic characteristic distance (ACD) between specified atom types for small molecules. This chapter discusses a different set of descriptors that describe the geometry of various metal sites in metalloproteins, provisionally named unorthodox metalloprotein descriptors. The work in this study mainly develops a relational database system and web-based interface to store and apply these descriptors, called MESPEUS (**ME**tal **S**ites in **P**roteins at **E**dinburgh **U**niver**S**ity).

The investigation of metal sites in metalloproteins has for some time been a major interest and research area of Dr. Marjorie M Harding of the School of Biological Sciences (http://tanna.bch.ed.ac.uk/ for more details). Since her research was first published in 1999 (Harding 1999), she has constantly investigated the newly added crystallographic structures and updates her results, while also discussing them in a series of publications (Harding 2000; Harding 2001; Harding 2002; Harding 2004; Harding 2006). The subjects of the study mainly involve the validation of the coordination number of the specified metal sites, the coordination shapes, the donor types, the metal-donor distances and the architecture of metal coordination groups etc. The study manages to provide a useful reference and potential recommendations for building or validating a protein structure model during refinement, particularly for structures where resolution is limited. The compilation of structural data is also useful for interpretation of electron-density maps of new protein structures.

As the volume of data has increased considerably, a database system and web-based interface become appropriate in order to manage the relevant information efficiently and share it with the crystallography community. My colleague Yi-gong Sheng has also contributed to write a set of perl scripts for the work of establishing the database. The following sections present the procedures for creation and manipulation of MESPEUS with some demonstrations and applications. These were published in 2008 (Hsin, Sheng et al. 2008).

7.2 Materials and methods

7.2.1 Details of information stored

The geometry information of metal in the MESPEUS database is derived from metalloprotein structures which were deposited in the Protein Data Bank (PDB) not later than 1 January 2007 (Berman, Henrick et al. 2007). We choose the PDB files, including protein and nucleic acid crystal structures, which contain one or more metal atoms and are determined by diffraction methods at a resolution of 2.5 Å or better. When more than one copy of the protein molecule with metal site(s) is present in the crystal asymmetric unit, all of the metal sites are stored in the database. The metal atoms selected are biologically common, including Fe, Ni, Mn, Ca, Cu, Na, Mg, K, Co and Zn.

Some of the information is directly extracted from the content of PDB files, including 1) the name of protein, 2) the class and title of protein given in the HEADER of the PDB file, 3) the structure resolution in Å, 4) the crystal space group and cell dimensions, 5) the refinement programme used, 6) the R factor and R_{free}, 7)

the *B* values of atoms involved in metal sites, 8) the occupancy values of donor atoms.

The other stored information regarding the geometry of metal sites is extracted and then evaluated for each PDB file by a programme called MP (Harding 2001). The programme performs the identification of metal sites in PDB files, the confirmation of donor atoms in residues or ligands, the measurement of metal-donor distances in metal sites, and the difference between the actual distance and the target distance (Harding 2006) is then obtained. The names of all metal and donor atoms are stored as they appear in the PDB files. In order to describe the distortion of the shape of a coordination group, the MP programme estimates the r.m.s deviation between the actual interbond angles around the metal atom and the values for ideal geometry (Harding 2000). The following subsections give the essential descriptions for the evaluation mentioned above.

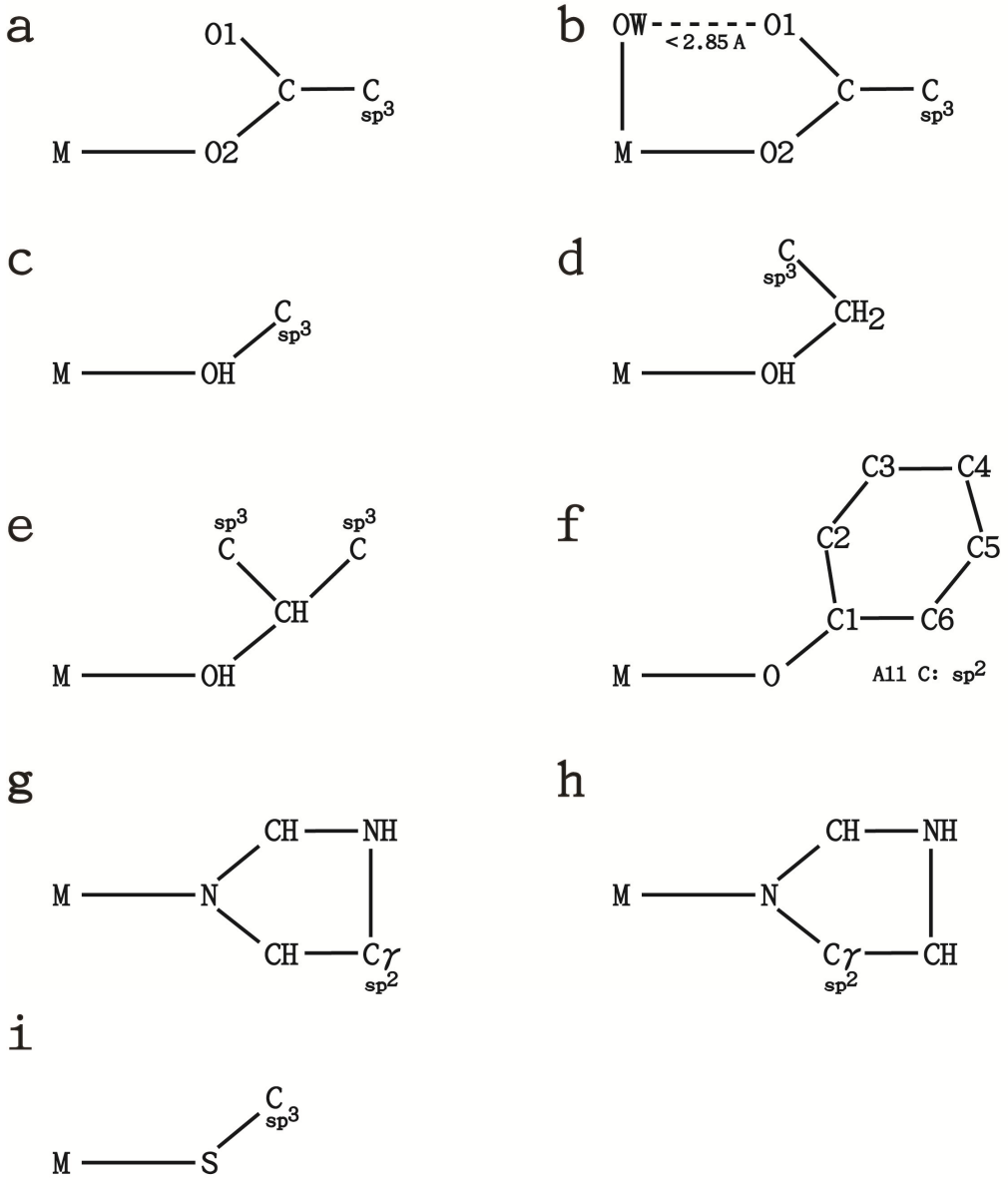
7.2.1.1 Target distance

The target distance in the present study is a set of predicted distances between donor groups and specified metals, which are obtained from a series of analyses of metal sites in protein structures using the PDB deposited coordinates combined with the information from analogous metal-coordination compounds in the Cambridge Structural Database, CSD (Allen 2002; Allen and Taylor 2004). The CSD was used through the UK Chemical Database Service at Daresbury Laboratory (Fletcher, McMeeking et al. 1996), and stores the results of over 400,000 determinations at present, including small organic and metal-organic molecules. The structures deposited in the CSD have been determined by X-ray or neutron diffraction and are,

by marcromolecular standards, very well resolved and the resulting structures have been applied in various fields (Allen and Motherwell 2002; Orpen 2002; Taylor 2002).

Initially, the study of target distances only used the metal complexes in the CSD as it provided a better accuracy than that of most protein structure determinations. All of the small molecules used in the study were required to have *R* factor ≤ 0.065 and contain specified metals interacting with ligands which are analogues to the amino acid side chains commonly found in proteins. The analogues mimic the carboxylate groups, alcohols, phenolates, thiolates, imidazole groups of amino acid side chains and water molecules. The motifs of metals interacting with these analogues are shown in Figure 7.1, schematically quoted from Harding 1999. Programmes QUEST and VISTA (Allen, Davies et al. 1991; Bruno, Cole et al. 2002) were used for the motif search and analysis. The observed distances in different metal-donor combinations were then subjected to a series of statistical analyses giving a set of objectively reliable values called target distance. Since the PDB is continually adding new structures, similar analysis to that mentioned above was then applied to the PDB data (determined at structure resolution ≤ 1.25 Å), so that target distances have been refined by evaluating the agreement between the values obtained from PDB and CSD. The revised target distances have been tabulated in Harding 2006. Atoms, excluding carbon and phosphorus, are identified as part of the metal site, i.e. donors, if they are within the target distance plus a tolerance 0.75 Å to the metal atom.

Figure 7.1. Motifs of M-analogues in CSD for the search queries where the letter M represents a metal coordinated with an analogue which mimics the amino acid side chain. Both (a) and (b) illustrate coordinations between metals and carboxylate groups; the bonds of M-O2, C-O1 and C-O2 can be any type; in (b) a hydrogen bond is involved in the coordination occurred in many carboxylates where the OW is a water molecule. (c), (d) and (e) represent the alcohol groups coordinated to metals. (f) shows a phenolate groups coordinated to a metal; the bonds of M-O shown in (c), (d), (e) and (f) can be any type. (g) and (h) illustrate imidazole groups coordinated to metals in where all bounds could be any type but the Cγ atoms should be sp^2 carbon. (i) shows a thiolate group coordinated to metal where the bond of M-S could be type. The full descriptions of these motifs can be found in Harding 1999.



7.2.1.2 Deviation of the shape of a coordination group

The distortion of the shape of a coordination group can be described as the r.m.s deviation, δ, when the actual interbond angles are compared with the values of a regular (or ideal) geometry. The present study mainly considers the cases of metal coordination number 4, 5, and 6. The ideal geometries of coordination groups are shown in Figure 7.2, schematically quoted from Harding 2000. The geometries are regardless of the bond type. The r.m.s deviation of the coordination number 4, 5, and 6 can be obtained by following equation:

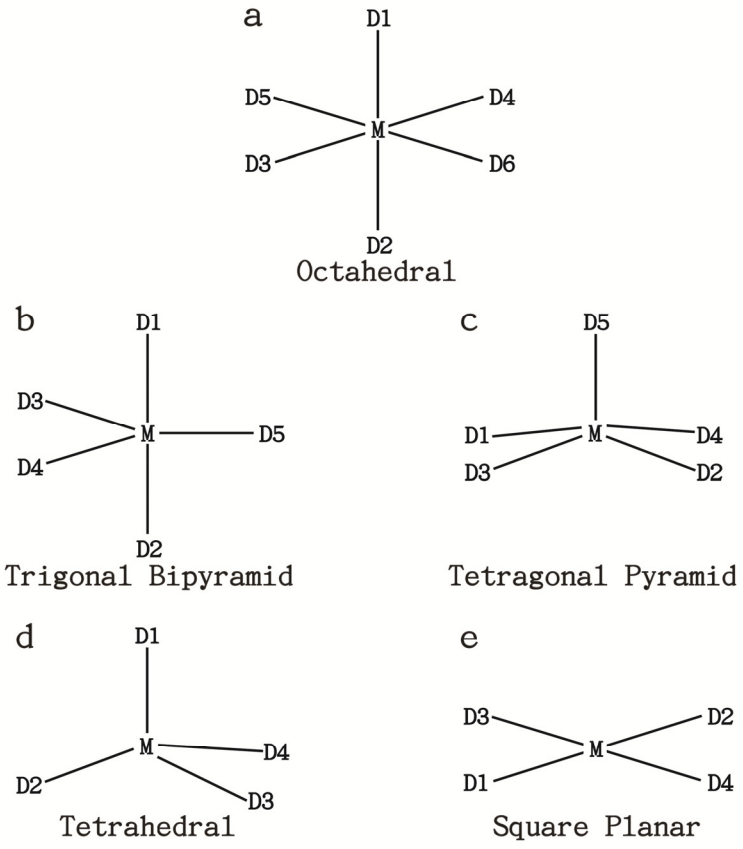
$$\delta_{shape} = \left[\sum_{i=1}^{n} (\alpha_i - \alpha_{ideal})^2 / n \right]^{1/2} \qquad eq.\ 7.1$$

where the α_i is the actual value of an angle measured from the interbonds in the coordination group and the α_{ideal} stands for the ideal bond angle in a regular geometry. The *n* is the number of angles which are needed to provide a full description for each coordination shape and it could be deduced from *CN*(*CN*-1)/2 where *CN* is the coordination number (Howard, Copley et al. 1998). For instance, there are 10 angles needed to describe a trigonal bipyramid or tetragonal pyramid sphere as their coordination number is 5, that is $n = 5(5-1)/2 = 10$. The smaller the value of δ the closer is the actual shape to the ideal geometry.

Other than determining the distortion of a coordination geometry, the r.m.s deviation can also be an indicator describing the nearest shape when the coordination group tends to form multi-geometries alternatively. As shown in Figure 7.2, the shape may be either trigonal bipyramid or tetragonal pyramid when the coordination number is 5. In order to determine a nearest shape for this case, the r.m.s deviations of the two coordination geometries can be calculated respectively and then the nearest

description of shape can be assigned by comparing their deviations. For example, the nearest shape is recommended as trigonal bipyramid when $\delta_{\text{trigonal bipyramid}} < \delta_{\text{tetragonal pyramid}}$, otherwise it is assigned as tetragonal pyramid. Similarly, this judgement can be applied to the case of coordination number 4 determining the shape if tetrahedral or square planar.

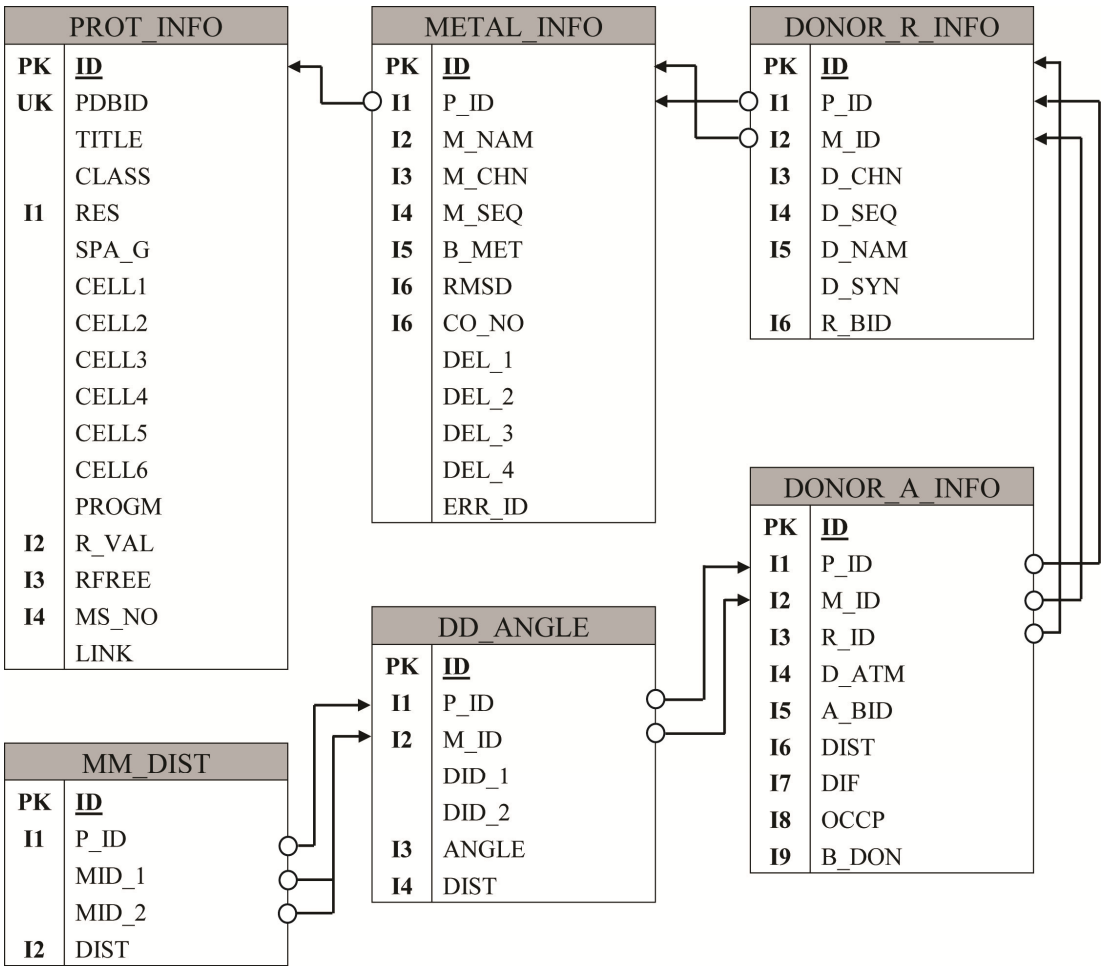
Figure 7.2. Regular geometries (shapes) of coordination groups where the letter M represents a metal coordinated with donor atoms D. (a) illustrates a metal coordinates with 6 donor atoms and the ideal shape of the coordination group is a octahedral. Octahedral bond angles are 90° so that for example, the ideal angles of D1-M-D4 and D1-M-D2 are 90° and 180°, respectively. (b) and (c) show the geometries when the coordination number is 5 and the shape could be either trigonal bipyramid or tetragonal pyramid. The largest bond angles are D1-M-D2 in both two shapes. (d) and (e) are the regular geometries for the coordination number 4. In (d), all of the ideal bond angles in a tetrahedral are equivalent, e.g. D1-M-D2: 109.5°, whereas the ideal bond angles in a square planar all are 90°.



7.2.2 Construction of the database

As described in section 2.2.4, a relational database is the most reasonable choice for the data storage as it is fast, efficient and compatible with various scripts. Similarly to EDULISS 2.0, we chose a MySQL database for the present study and designed a set of proper tables to relationally organise data. The major schema of MESPEUS database design is shown in Figure 7.3, in which a number of keys, i.e. primary key (PK), unique key (UK) and index (I), have been assigned to the specified columns in order to manage the data integrally and enhance the query performance.

Figure 7.3. Major schema of MESPEUS database design.



PK: Primary key; **UK**: Unique key; **I**: Index

The root of the tables is the "PROT_INFO" table which contains the fundamental geometry information of collected protein structures, including the information extracted from the content of PDB files directly and the number of metal sites found in a structure. The columns of CELL1 to CELL6 are for the cell dimensions, i.e. a, b, c, α, β, γ (Å and degrees), and the LINK column provide the link of the particular protein page shown on the PDB website.

The table "METAL_INFO" stores the properties of metal atoms found in PDB structures. The metal name, residue number and chain letter are saved as they appear in the atom list in PDB files. The RMSD column is for the r.m.s difference between each actual metal-donor distances found in the metal site and the target values. It is a useful quality indicator for the reported distances and geometry of a metal site. A large r.m.s difference may simply be due to large coordinate errors in a low resolution structure, or to an error in identification of donor atoms or interpretation of the site. The columns of DEL_1 to DEL_4 are for the four indicators of metal stereochemistry. As mentioned in section 7.2.1.2, the distortions of the coordination shapes have been measured and described as the deviation (δ) for the cases of coordination number 4 to 6. For coordination number 4, the DEL_1 and DEL_2 column store the r.m.s deviations from an ideal tetrahedral and square planar respectively. The deviation from bar 4 (a '–' symbol upon 4) symmetry is stored in the column DEL_3. The bar 4 symmetry in a regular tetrahedron is illustrated in Figure 7.4. Consider it is a regular tetrahedron composed of four donors A, B, C and D around metal M and all of the interbond angles, i.e. AMB etc, are equal. The dashed line is one of the bar 4 axes and a regular tetrahedron has three bar 4 axes, which bisects the angles AMD and BMC. If the tetrahedron is squashed along the

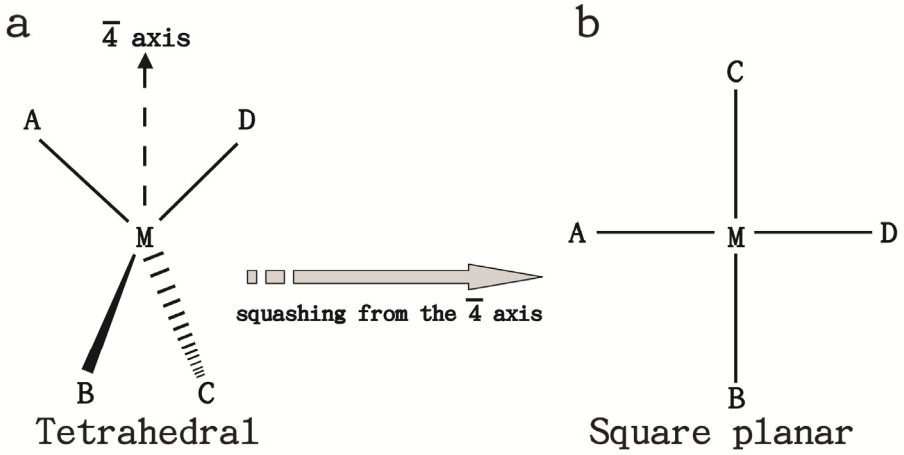
dashed line, i.e. the bar 4 axis, eventually it becomes square planar with ABCD all in the same plane as M. At all stages of this squashing, the symmetry is still bar 4, and the angles should be AMD=BMC and AMB=AMC=DMC=DMB. The calculation of the r.m.s deviation of bar 4 symmetry in this study is defined by Marjorie Harding and the steps are shown as following:

bm1: the mean of AMD and BMC

bm2: the mean of AMB, AMC, DMC, DMB

dbar4: one sixth of the sum of the squares of AMD-**bm1**, BMC-**bm1**, AMB-**bm2**, AMC-**bm2**, DMB-**bm2**, DMC-**bm2**

Figure 7.4. Illustration of bar 4 symmetry in a regular tetrahedron for the r.m.s. deviation calculation, which is composed of four donors A, B, C and D, and a metal M. (a) A, M and D are in the plane of the paper, B is in front and C is behind. The dashed line is a bar 4 axis which bisects the angles AMD, BMC (the other two bar 4 axes should be able to be observed between AMB and CMD). (b) when the tetrahedral is squashed from the bar 4 axis, it becomes a square planar and the characteristic of angles should be AMD=BMC and AMB=AMC=DMC=DMB.



This is then calculated for each of the three possible positions of the bar 4 axis (bisecting AMD, AMB and CMD) and the minimum value is taken. The value of

column DEL_4 is recorded as 0 when the nearest shape is tetrahedral, whereas it is 1 for square planar. For coordination number 5, the column DEL_1 and DEL_2 are for the deviations from trigonal bipyramid and tetragonal pyramidal, respectively. When the nearest shape is trigonal bipyramid, the DEL_4 is 0, whereas > 0 for tetragonal pyramidal. The deviations are stored in the DEL_1 column for the case of coordination number 6.

In some cases, two alternative positions for a donor molecule or metal itself have been detected during refinement. The atoms of the distorted molecule are labelled by A and B according to their occupancy values as higher occupancy is labelled as A and B for the lower. The labels are recorded as the last character of the atom names in the PDB file. Disorder like this is sometimes reported in poorly determined structures, but it is certainly seen in some structures determined at high resolution. Figure 7.5 illustrates four instances in which the disorders occur in either the coordination group or the metal atom. Figure 7.5 (a) demonstrates the two alternative positions for a carbonate ion which is coordinated to an iron atom found in the 1A8E protein at a structure resolution of 1.6 Å. In this type disorder, the atoms with lower occupancy, i.e. labelled as B, have been recognised and excluded from storage. The second case (Figure 7.5 (b)) shows that the zinc atom is too close to the sulphur atom at a distance of 2.36 Å and the nearest oxygen atom is at a distance of 1.75 Å (PDB ID: 1C1N; resolution: 1.4 Å), so that the atoms of zinc and SO_4 molecule have been labelled as A and B in its PDB file. The case shown in Figure 7.5 (c) reveals that a disorder due to the alternative positions for the Ca (colored by yellow) and Mg (colored by green) atoms found in the 1S83 protein structure at a resolution of 1.25 Å. The occupancy values of the Ca and Mg atoms

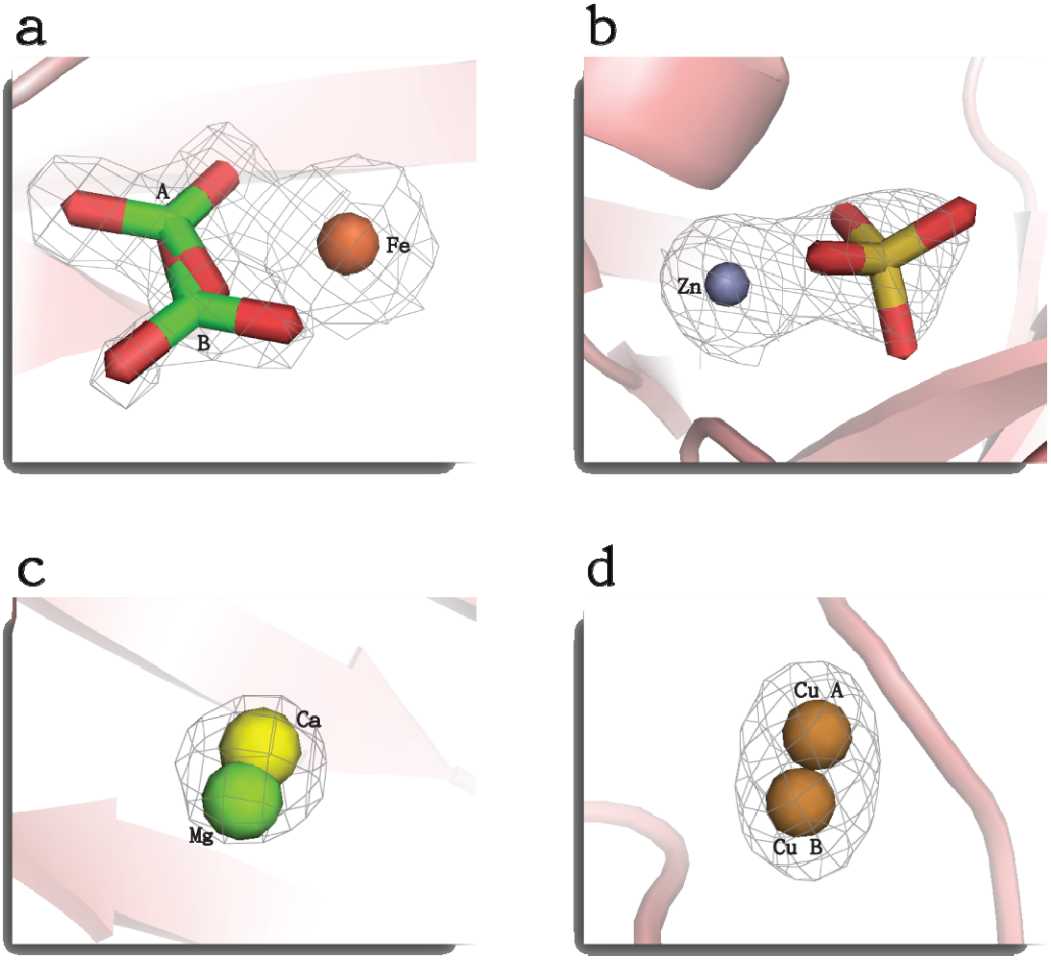
are both 0.5 since they share the occupancy in the refinement. In Figure 7.5 (d) is the case found in 2GBA (resolution: 0.92 Å) where two identical copper atoms are disordered and labelled as A and B, respectively. The two atoms are very close to each other as the distance between them is only 0.67 Å. All of the disorders illustrated in Figure 7.5 have been given an error flag for the associated metal atom and recorded in column ERR_ID of table "METAL_INFO".

The "DONOR_R_INFO" table mainly stores the identifiers of donor residues as shown in PDB files. When the metal site involves the oxygen atoms of carboxylate from GLU and ASP and the oxygen atoms are both within the target distances plus the tolerance or if carboxylate groups bridge two metal atoms, the interaction will be defined as bidentate and labelled as "b" in the R_BID column, otherwise blank. The identifiers of donor atoms with the values of atomic occupancy and *B*-factors as shown in PDB files can be found in table "DONOR_A_INFO". The actual metal-donor distance and the difference from target distance are stored in the DIST and DIF column, respectively. In the same way, the donor atoms from carboxylate groups forming bidentate interaction with metals are labelled as "b" in the A_BID column. The interbond angles in metal sites and the distances between donor atoms are stored in the "DD_ANGLE" table. The "MM_DIST" table contains the distances between metals found in each protein structure if the protein possesses two or more metal sites.

All of the tables have been relationally organised by a set of specified columns. For example every P_ID column refers to the ID column of root table "PROT_INFO" and the M_ID columns all refer to the ID column of "METAL_INFO". The

description of the columns in each table is also available at the MESPEUS website, http://eduliss.bch.ed.ac.uk/MESPEUS/images/MESPEUS_table.pdf.

Figure 7.5. Disorders in coordination group and in metal atom. (a) illustrates the two alternative positions for a carbonate ion coordinated with an iron atom (PDB ID: 1A8E). The CO_3 at the position with higher occupancy (0.65) is labelled as A, the other with lower occupancy (0.35) labelled as B and excluded from storage. (b) shows the distortion of zinc and SO_4 molecules with occupancy 0.38 and 0.48, respectively (PDB ID: 1C1N). The zinc atom is very close to the sulphur and nearest oxygen atom at the distance of 2.36 and 1.75 Å, respectively. (c) reveals a disorder due to the alternative positions for the Ca (yellow) and Mg (green) atom (PDB ID: 1S83). (d) shows two identical disordered ions, i.e. copper A and B, found in 2GBA. The two copper atoms are very close at the distance 0.67 Å. An error flag has been given for these distortions.



7.2.3 Construction of the web-based interface

In the same way as EDULISS 2.0, the web-based interface of MESPEUS is built by using the Java Servlet technology and JavaServer Pages and is hosted at an Apache Tomcat web server. The software architecture is constructed as Model-View-Controller (MVC) model in order to enable extensibility in functionality and for ease of maintenance (described in section 2.4.1). For representing the molecular structures in 3D on web pages, this web site uses Jmol as the viewer (described in section 2.4.2.1). An open source application for displaying statistical charts, called JFreeChart (http://www.jfree.org/), has also been applied, which provides the charts with drilldown enabled function allowing users to examine charts to the desired level of detail. The web-based interface allows users to set series of query criteria to access the MESPEUS database without requiring any knowledge of SQL. It displays the accessible information of metal coordination groups and renders the individual metal site with distances, angles and coordination geometry etc. The home page of MESPEUS web site is shown as Figure 7.6 and is available at http://eduliss.bch.ed.ac.uk/MESPEUS/. A detailed demonstration for its manipulation is given in section 7.4.1.

7.3 The statistical profile of metal sites in MESPEUS

At present, the database contains 10,919 metalloprotein structures extracted from the PDB. A significant number of metal atoms are listed in whole PDB files which do not appear to be within chemical bonding distance, i.e. target distance plus 0.75 Å, of any appropriate atoms. In many cases, there are not even any appropriate atoms within 3.6 Å. These cases do not make chemical sense, and can only be regarded as

incomplete structure determinations and these metal atoms are not included in the MESPEUS database. The numbers of metal sites are listed in Table 7.1 grouped by metal types. The listed numbers include the cases in which the metal sites are present in the asymmetric unit, so that the number of distinct metal sites should be fewer than the number listed. Further, the PDB contains many groups of very similar proteins within which the metal sites are similar or identical. Some metal or donor atoms are listed with low occupancy and for the disordered pairs like those shown in Figure 7.5 (a), only the one with higher occupancy has been retained. Coordinate errors can be large when atom positions have low occupancy and so metal donor distances derived from them may have large errors; the presence of disordered sites gives meaningless results for coordination number. Sites with low occupancy can be identified in the database as described in section 7.2.2 and excluded from searches, leaving the observations with occupancy 1.0 separated into column (2) of Table 7.1. Column (3) and (4) are further subdivided from column (2) as not all of the metal atoms interact with protein directly. For example, there is a large number of cases for Mg which are simply hydrated ions, such as $Mg(OH_2)_6^{2+}$. Alternatively, some metal atoms are only interacting with ligand molecules, or with DNA or RNA. A survey shows that there are 312 of the PDB files used for MESPEUS database described as DNA or RNA in their PDB HEADER and 308 of them are simply DNA or RNA structures as none of the amino acid residues can be found in the PDB files. Figure 7.7 gives a good instance (PDB ID: 1DPL; resolution: 0.83 Å) to show a structure deposited in PDB which is a DNA structure without the presence of protein. The coordination group of the Mg site in the structure is composed of six water molecules forming a typical octahedral geometry, i.e. $Mg(OH_2)_6^{2+}$. As the structure

has been well determined, the occupancy values of the donor atoms are 1.0. Given the result shown in column (3), the most common metals in metalloprotein are Fe, Ca, Zn and Mg, whereas there are relatively fewer Co and Ni containing proteins.

Figure 7.6. Home page of MESPEUS web-based interface.

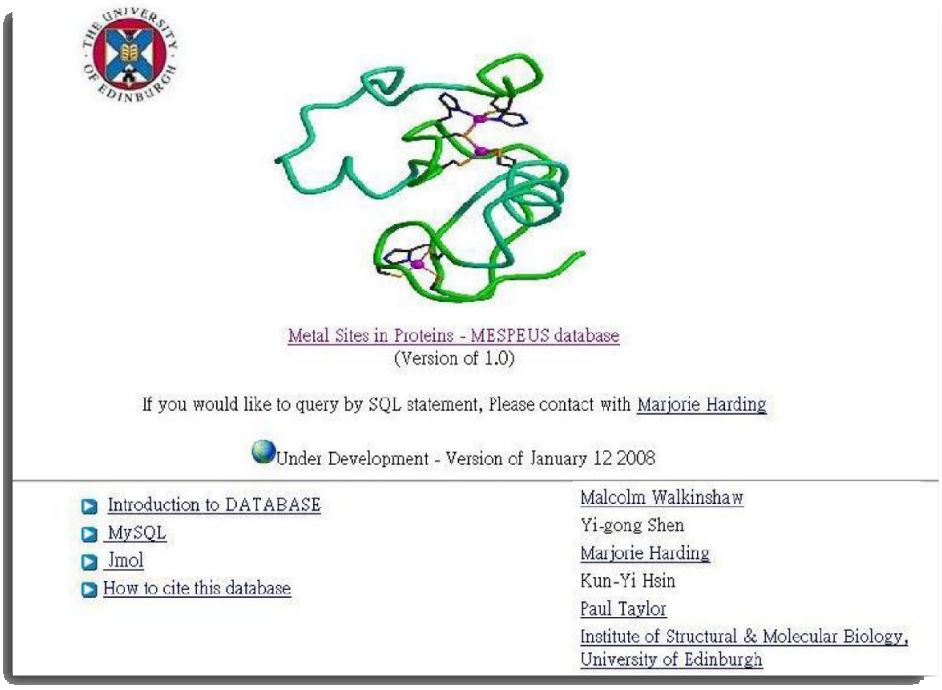
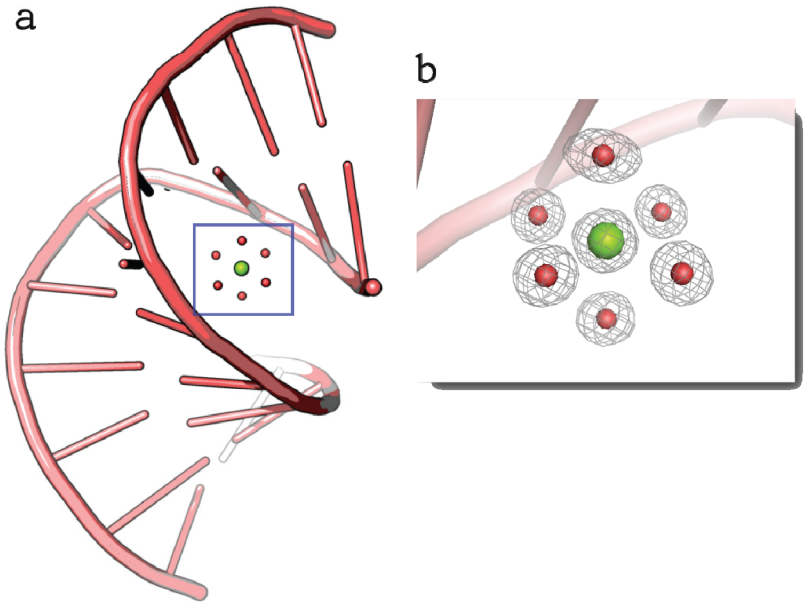


Table 7.1. Numbers of metal sites in the MESPEUS database.

Metal	(1) All sites in database	(2) Sites with metal and donor occupancy = 1.0	(3) Sites from column (2) with metal-protein interactions	(4) Sites from column (2) without metal-protein interactions
Na	2,372	2,069	1,768	301
Mg	5,561	4,953	3,676	1,277
K	1,424	1,259	1,108	151
Ca	7,120	6,425	6,018	407
Mn	2,118	1,810	1,688	122
Fe	7,763	7,226	7,003	223
Co	637	506	356	150
Ni	534	398	366	32
Cu	1,327	1,105	1,088	17
Zn	6,031	5,073	4,979	94
Total	34,896	30,824	28,050	2,774

Figure 7.7. DNA structure with no protein present (PDB ID: 1DPL; resolution: 0.83 Å) where the Mg site is composed of six water molecules (marked by blue frame). The values of donor occupancy are 1.0 and the site is formed as a typical octahedral geometry shown in (b).



7.4 Applications of MESPEUS database and web interface

7.4.1 Fundamental manipulation of web interface

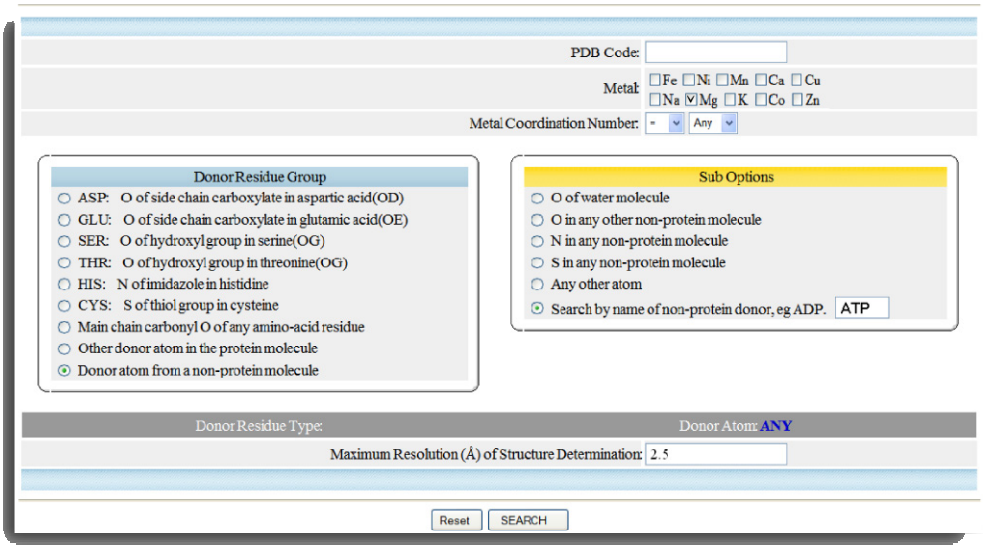
The MESPEUS web interface has been designed to allow straightforward searching for particular kinds of interactions, with the possibility of selecting only higher resolution structures. Alternatively, users can simply give a PDB ID to yield all of the metal sites in that protein. The query criteria can be set up via the main query page as shown in Figure 7.8. The page provides a list of convenient options, as tabulated below the figure, which are the particular types of donor residue groups and atoms found in database. As the selection shown in Figure 7.8, the page has been filled for a query about Mg links to ATP with the maximum structure resolution set at 2.5 Å, then the first result is a list of all the metal-donor atom distances

satisfying the search query, together with information on coordination number of the metal, shape, atom numbers and so on, as shown in Figure 7.9. The distributions of the reported Mg-O_{ATP} distances and structure resolutions of hit proteins can be displayed by histograms together with the general statistical profiles, such as the numbers of observations, the means of reported distances and structure resolutions and their value ranges. The histograms are drilldown enabled allowing users access to the desired level of detail. For example, if the bar marked by * is clicked upon, the page will only list the data of the samples whose Mg-O_{ATP} distances are within the range of ≥ 1.8 and < 1.9 Å. The page also allows downloading the query result as a tab-separated file and entries can be removed from the list by clicking the "Delete" button before calculating distributions.

Users can click the link of metal name printed in the list as marked by ** in Figure 7.9 to open a page like Figure 7.10 showing the detail of the specified metal site. Alternatively, a page like Figure 7.11 shows the whole metal-containing protein when users click the link of PDB ID as marked by ***. Both pages use Jmol to present the molecular structures in 3D with a set of buttons below the pictures allowing users to conveniently manipulate the structures, such as dynamically to centre the specified metal, rotate the views, measure the angles or distances, reveal the location(s) of metal site(s) in protein, or add other information. Users can obtain the desired geometry information of metal sites and inspect them via these two pages directly.

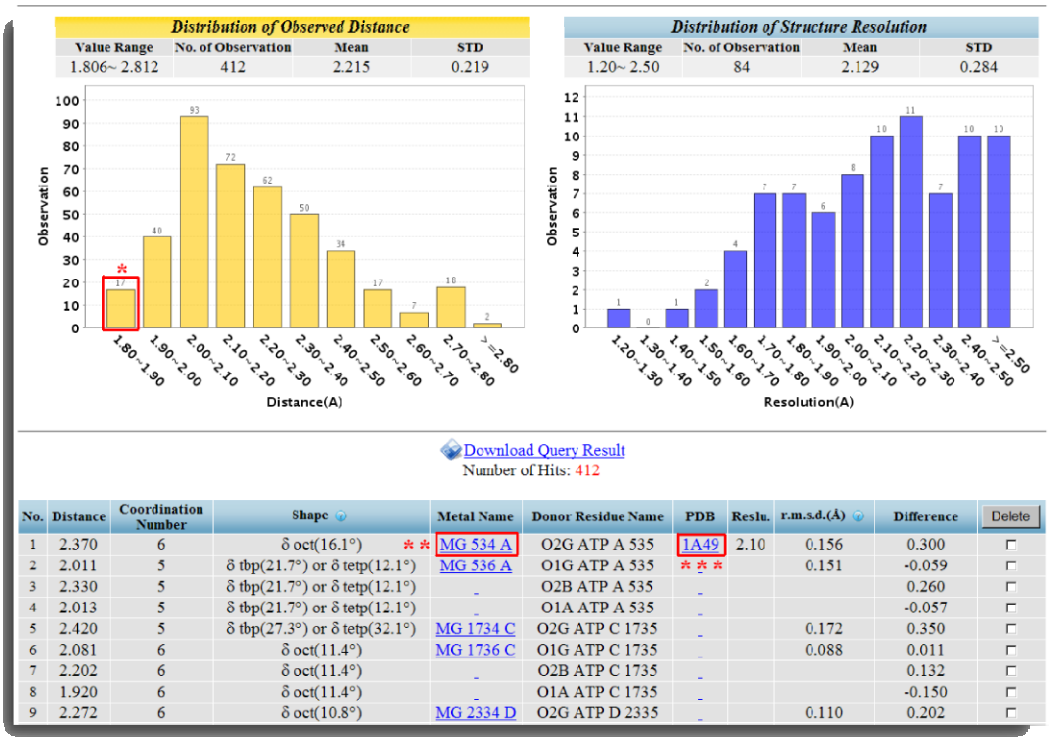
In order to serve the researchers who are familiar with SQL, a page allowing input of SQL statements and direct access of the MESPEUS database has been set up and is available at http://eduliss.bch.ed.ac.uk/MESPEUS/query_SQL.jsp.

Figure 7.8. Main query page for the MESPEUS web interface. From the top, users can simply give a PDB ID to yield all of the metal sites in that protein. Alternatively, users can set a variety of query criteria, including to select metal of interest (multiple choices allowed); to specify the metal coordination number as any or 1 to 12 with 6 different logical operators; to choose the type of donor residue group (the list on the left) and atom (on the right), or input a name of non-protein donor, such as ATP, which should be identical with the residue code recorded in the PDB file; to determine the maximum structure resolution (Å). The options for donor groups and atoms are tabulated as blow.



	Donor Residue Group		Sub Options
1	ASP	O of side chain carboxylate in aspartic acid (OD)	Any Monodentate Bidentate
2	GLU	O of side chain carboxylate in glutamic acid (OE)	Any Monodentate Bidentate
3	SER	O of hydroxyl group in serine (OG)	-
4	THR	O of hydroxyl group in threonine (OG)	-
5	HIS	N of imidazole in histidine	ND and NE N of imidazole in histidine (ND) N of imidazole in histidine (NE)
6	CYS	S of thiol group in cysteine	-
7	Main chain carbonyl O of any amino-acid residue		-
8	Other donor atom in the protein molecule		O of amide group of asparagine (OD) O of amide group of glutamine (OE) O of phenolate group of tyrosine (OH) S of methionine side chain (SD) Main Chain N, Any AA (Rare) N in LYS side chain (Very Rare) N in ARG side chain (Very Rare) Any other atom
9	Donor atom from a non-protein molecule		O of water molecule O in any other non-protein molecule N in any non-protein molecule S in any non-protein molecule Any other atom Search by name of non-protein donor

Figure 7.9. Result page of the search for Mg linked to ATP with the maximum structure resolution at 2.5 Å. The search yielded 412 examples in 84 proteins or DNA/RNA structures. The distributions of observed Mg-O_{ATP} distances and the structure resolutions of hit proteins can be displayed as histograms together with their general statistical profiles, such as the numbers of observations, the means of reported distances and structure resolutions and their value ranges. The list from the left side gives the actual Mg-O_{ATP} distance measured, coordination number, shape with the r.m.s deviation from the ideal geometry, metal name, donor group, PDB ID with resolution, the r.m.s. difference between metal-donor atom distances and the target distances, and the difference between the two distances. The list can be downloaded as a tab-separated file and the entries can be removed by clicking the "Delete" button. *: Clicking here selects this distance range only. **: Links to a page like Figure 7.10. *: Links to a page like Figure 7.11 showing the whole metal-containing protein.**



No.	Distance	Coordination Number	Shape	Metal Name	Donor Residue Name	PDB	Reslu.	r.m.s.d.(Å)	Difference	Delete
1	2.370	6	δ oct(16.1°) * *	MG 534 A	O2G ATP A 535	1A49	2.10	0.156	0.300	☐
2	2.011	5	δ tbp(21.7°) or δ tetp(12.1°)	MG 536 A	O1G ATP A 535	* * *		0.151	-0.059	☐
3	2.330	5	δ tbp(21.7°) or δ tetp(12.1°)	-	O2B ATP A 535	-			0.260	☐
4	2.013	5	δ tbp(21.7°) or δ tetp(12.1°)	-	O1A ATP A 535	-			-0.057	☐
5	2.420	5	δ tbp(27.3°) or δ tetp(32.1°)	MG 1734 C	O2G ATP C 1735	-		0.172	0.350	☐
6	2.081	6	δ oct(11.4°)	MG 1736 C	O1G ATP C 1735	-		0.088	0.011	☐
7	2.202	6	δ oct(11.4°)	-	O2B ATP C 1735	-			0.132	☐
8	1.920	6	δ oct(11.4°)	-	O1A ATP C 1735	-			-0.150	☐
9	2.272	6	δ oct(10.8°)	MG 2334 D	O2G ATP D 2335	-		0.110	0.202	☐

Figure 7.10. Page for specified metal site found in PDB ID: 1F2U. When selecting the metal name in the results listed in Figure 7.9, the metal site can be displayed by Jmol, as well as its relation to the whole protein molecule together with the information of the site geometry and a list of donor resides. Here, the metal, Mg (cyan coloured), is coordinated by the O atoms from SER and GLN, two phosphate O atoms (one β, one γ) of ATP, and two water molecules. The buttons below the pictures allow users to centre and rotate the view and add other information.

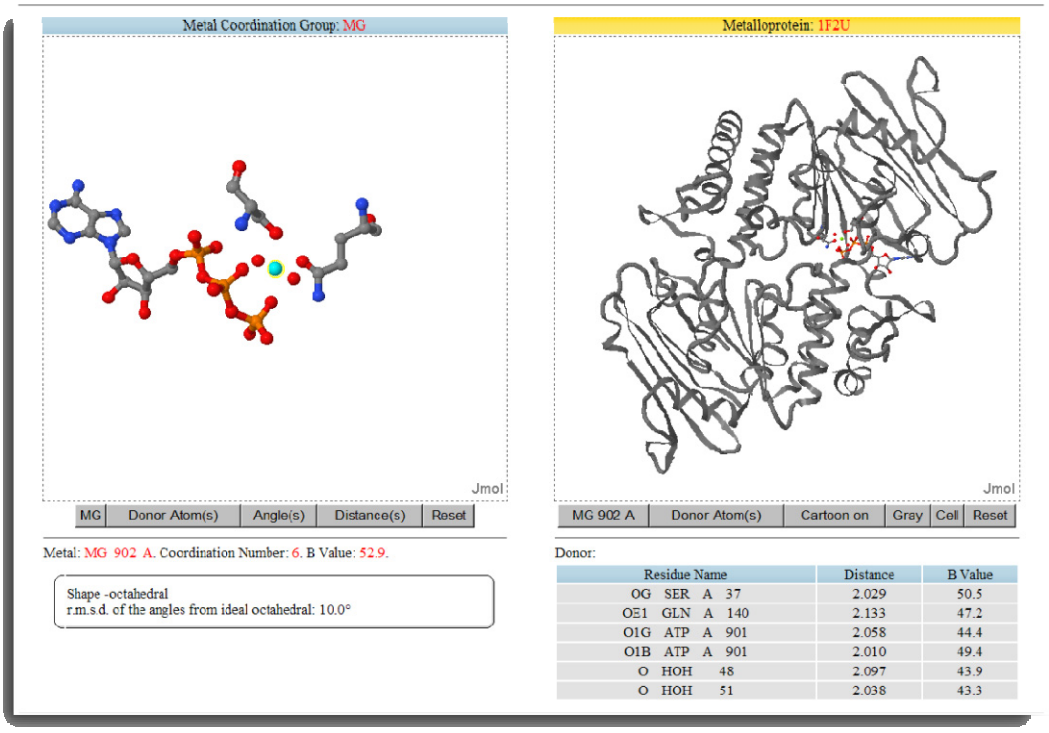
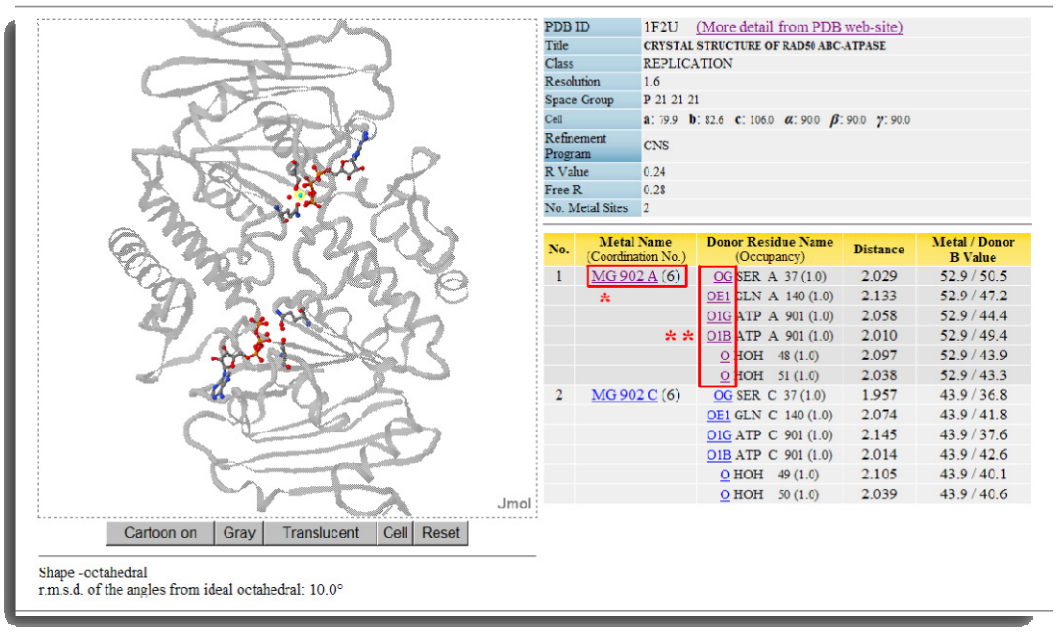


Figure 7.11. Page for specified metal containing protein (PDB ID: 1F2U). When the PDB ID is selected in the results listed in Figure 7.9, the metalloprotein can be displayed by Jmol, as well as its crystallographic information and the details of each metal site found in the structure. This page also allows directly linking to the PDB web site for more details of this protein. Here, the protein has two chains, A and B, with one Mg-ATP site on each. *: A dynamic link to centre this metal site in the view on left. **: Links to a page like Figure 7.10 showing the specified metal site.



7.4.2 Importance of near atomic resolution in deriving mean bond distances

We have been concerned to derive the best average values for distances from metal atoms to different kinds of donor atoms in structures deposited in the PDB, for use in validation, electron density map interpretation, model fitting or for restrained refinement. The average interaction distance can be evaluated from all the data in the database, or from the higher resolution structure determinations only. The average found for a particular distance is often slightly smaller when only higher

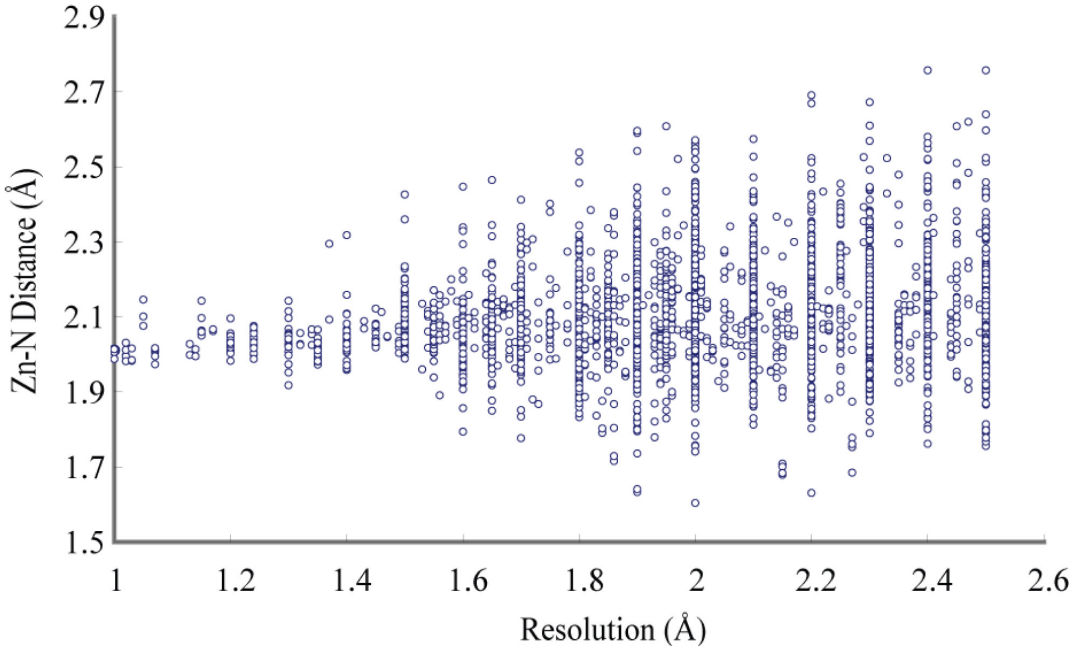
resolution observations are used than when all results are used. Table 7.2 gives an example of Zn-N of histidine with coordination number 4, in which the observations listed include the interaction from Nδ and Nε of imidazole side chain (coordination groups with the donor atom occupancy ≤ 0.9 and geometrical disorder as described in section 7.2.2 have been excluded). It shows that the mean distance can be fairly different when the observations are various in the level of resolution determinations and the standard deviation, representing the scatter of results, is smaller at higher resolution. The most precise averages, i.e. those with smallest standard deviation, are from X-ray structures with resolution better than 1.3 Å (i.e. near atomic resolution).

This result agrees with a previous study which applied CSD data where the average for the equivalent interactions there is 2.00 (0.02) Å for 25 observations (Harding, 2006). Figure 7.12 illustrates the wide scatter of reported distances from poor to good resolution structures. It shows that good consistency of observations only occurs at a fairly high resolution. The standard deviations in the level of high resolution determination suggest that very little variation in the particular distance is expected, e.g. < 0.02 Å, but the wide variation above must be due to coordinate errors remaining after structure refinement. According to these results, it is strongly recommended that for such averages only the highest resolution results practicable be used, consistent with a reasonable number of observations for statistical requirements.

Table 7.2. Effect of resolution on mean distance found for Zn-N of histidine, with Zn coordination number 4. It excludes the coordination groups with the occupancy of donor atoms ≤ 0.9 and geometry disorder. The Nδ and Nε of histidine imidazole side chain are both included as there is no significant difference in their means if the two Nδ and Nε are kept separate.

Max. resolution (Å)	No. observations	Mean (s.d.) (Å)
2.5	3,583	2.10 (0.13)
1.8	1,007	2.07 (0.09)
1.5	261	2.06 (0.06)
1.3	88	2.03 (0.04)
1.1	21	2.02 (0.04)

Figure 7.12. Reported values of Zn-N_{His} distances for zinc coordination number. The observations include the interaction of both Nδ and Nε of imidazole side chain.



7.4.3 Application: a survey of interactions between Mg and adenosine triphosphate (ATP)

Phosphate-magnesium interactions are crucial in many enzyme mechanisms, such as in the kinases mentioned in chapter 6. The interactions are essential in most ATP catalysis (Williams, Oren et al. 1993; Hansson and Kannangara 1997; Mesecar and Nowak 1997; Zhang, Yang et al. 2009). There is a particularly rich source of structural information available in the PDB for metal-ATP interactions and this survey provides an overview of the interaction geometries found from the database for protein-Mg-ATP complexes. The MESPEUS web interface can find all links from Mg to O of ATP as the case shown in Figure 7.9 and present the Mg site as in Figure 7.10 and 7.11. For the detailed analysis, it was more convenient to use SQL statements to access the database to find all Mg with ATP links and then to apply perl scripts for further investigation. There are 197 Mg-$O_{ATP\text{-}phosphate}$ sites in 84 proteins found in the database, in which the occupancy values of donor atoms are > 0.9 without geometrical disorders. Table 7.3 shows the mean distance of Mg-$O_{ATP\text{-}phosphate}$ using structures solved at different resolutions. The observations exclude the distances > 2.6 Å as they have been considered as outliers. Seven observations in structures at resolution < 1.5 Å give a mean Mg-$O_{ATP\text{-}phosphate}$ distance of 2.05 (0.07) Å, but for a reliable average more observations are desirable.

Table 7.4 shows the distribution of coordination number for Mg-ATP sites in different level of structure resolution. The most common is coordination number 6, which is consistent with the result of searching the coordination number found of all Mg sites in structures regardless of the donor types (in the research only structures

determined at resolution 1.8 Å or better were used, and 63.9 % of Mg sites have a coordination number of 6). Some observations from lower resolution structures appear to have coordination number < 4 which suggests that the reliability of these data is questionable and some of the analyses may be incomplete or the atoms identified as Mg may actually be water molecules as these molecular sizes are quite similar.

Table 7.3. Mean distance of Mg-$O_{ATP\text{-}phosphate}$ in different level of structure resolutions. The observations listed here have excluded the samples with distance > 2.6 Å.

Max. resolution (Å)	No. observations	Mean (s.d.) (Å)	No. proteins
2.5	385	2.18 (0.18)	82
2.3	253	2.17 (0.18)	61
2.0	126	2.17 (0.18)	34
1.8	74	2.13 (0.15)	21
1.5	7	2.05 (0.07)	3

Table 7.4. Coordination numbers found for Mg-ATP sites in different level of structure resolutions.

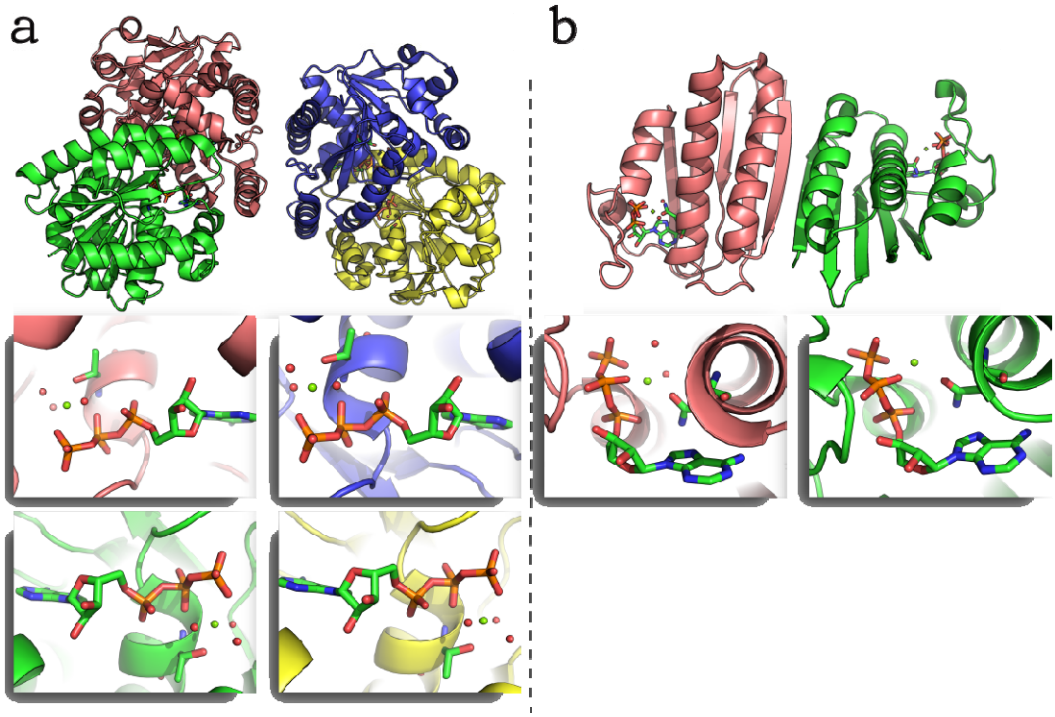
Max. resolution	Coordination numbers						
(Å)	1	2	3	4	5	6	7
2.5	3	10	20	26	21	114	3
2.3	-	4	8	10	11	92	1
2.0	-	1	4	4	4	48	1
1.8	-	-	2	2	4	26	1
1.5	-	-	-	1	-	2	-

After removal of identical sites within crystal asymmetric units (e.g. the examples shown in Figure 7.13), 66 Mg-$O_{ATP\text{-}phosphate}$ remained and are listed in Appendix 2

together with the information of donors and some statistics. It shows that the most common protein types are transferases. Mg is normally linked to protein through one or two amino acid side chains, mainly ASP, GLU, ASN, GLN or SER, THR, but there are two examples with three links to protein, and three with links to main chain carbonyl oxygen (marked by a superscript **a**).

Previous experimental studies have tried to address the coordination pattern, i.e. the preferred interaction, between Mg and the oxygen atoms of ATP phosphates, but the results are equivocal. Some of them propose that the Mg^{2+} ion may be preferably coordinated with the oxygen atoms from each of the three ATP phosphates (Kuntz 1973; Mildvan 1987), but some suggest that the usual coordination pattern should involve two oxygen atoms only, i.e. the oxygen from γ- and β-phosphate groups respectively (Auerbach, Huber et al. 1997; Liao, Sun et al. 2004). The present survey tends to agree with the latter. ATP may be linked to Mg through one, two or all three phosphate groups. The most common pattern (23 examples) is linkage through the β- and γ-phosphate groups, but all other possibilities also occurred. There are only two cases where linkage is through two O atoms of the same phosphate group (marked by a superscript **b**). The Mg coordination group may also include one to four water molecules and occasionally another small molecule like oxalate as a bidentate ion.

Figure 7.13. Removal of identical sites within crystal asymmetric units for Appendix 2. All Mg atoms and water molecules are shown as green and red spheres, respectively. The interacting amino acids and ATP are shown as stick style. (a) illustrates that there are four identical Mg-ATP sites found in each chain of 2BEK, all composed of one O_{THR}, two $O_{ATP\text{-}phosphate}$ ($O_{\beta\text{-}}$ and $O_{\gamma\text{-}phosphate}$) and three water molecules. Only the Mg site located in chain A, colored by red, has been retained and listed in Appendix 2. (b) shows that two Mg sites in 1TID have been listed in the table, because the site found in chain A (colored by red) is composed of one O_{ASN}, three $O_{ATP\text{-}phosphate}$ ($O_{\alpha\text{-}}$, $O_{\beta\text{-}}$ and $O_{\gamma\text{-}phosphate}$) and two water molecules, whereas the site located in chain B (colored by green) only involves one O_{ASN}, two $O_{ATP\text{-}phosphate}$ ($O_{\beta\text{-}}$ and $O_{\gamma\text{-}phosphate}$).



7.5 Discussion and summary

Unlike the descriptors for small molecules, the study of the unorthodox metalloprotein descriptors, i.e. the geometry information of metal sites, mainly concerns the inter-molecular interactions as well as factors in the determination of the protein structure. The storage for these descriptors should not only include complete crystallographic information of global protein structures but also record the evaluated deviation from ideal geometries of metal sites in order to provide a precise observation and application.

The MESPEUS database with its user-friendly web interface allows immediate identification and display of the metal sites in any specified protein whose structure is in the PDB, determined at resolution 2.5 Å or better. Alternatively, the web interface can identify all metal sties with a particular kind of contact, giving distance, coordination number, and atoms names for each as well as average distances and structural resolutions and their distributions. Further, as shown in the cases of applications, SQL queries to the database can extract additional information about the interactions of different metals with proteins.

7.6 Reference:

Allen, F. H. (2002). "The Cambridge Structural Database: a quarter of a million crystal structures and rising." Acta Crystallographica Section B: Structural Science **58**(3): 380-388.

Allen, F. H., J. E. Davies, et al. (1991). "The development of versions 3 and 4 of the Cambridge Structural Database System." Journal of Chemical Information and Computer Sciences **31**(2): 187-204.

Allen, F. H. and W. D. S. Motherwell (2002). "Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry." Acta Crystallographica Section B: Structural Science **58**(3): 407-422

Allen, F. H. and R. Taylor (2004). "Research applications of the Cambridge structural database (CSD)." Chemical Society Reviews **33**(8): 463-475.

Auerbach, G., R. Huber, et al. (1997). "Closed structure of phosphoglycerate kinase from Thermotoga maritima reveals the catalytic mechanism and determinants of thermal stability." Structure **5**(11): 1475-1483.

Berman, H., K. Henrick, et al. (2007). "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data." Nucleic Acids Research **35**(Database issue): D301.

Bruno, I. J., J. C. Cole, et al. (2002). "New software for searching the Cambridge Structural Database and visualizing crystal structures." Acta Crystallographica Section B: Structural Science **58**(3): 389-397

Fletcher, D. A., R. F. McMeeking, et al. (1996). "The United Kingdom chemical database service." Journal of Chemical Information and Computer Sciences **36**(4): 746-749.

Hansson, M. and C. G. Kannangara (1997). "ATPases and phosphate exchange activities in magnesium chelatase subunits of Rhodobacter sphaeroides". National Acad Sciences **94:** 13351-13356.

Harding, M. M. (1999). "The geometry of metal-ligand interactions relevant to proteins." Acta Crystallographica Section D: Biological Crystallography **55**(8): 1432-1443.

Harding, M. M. (2000). "The geometry of metal-ligand interactions relevant to proteins. II. Angles at the metal atom, additional weak metal-donor interactions." Acta Crystallographica Section D: Biological Crystallography **56**(7): 857-867.

Harding, M. M. (2001). "Geometry of metal-ligand interactions in proteins." Acta Crystallographica Section D: Biological Crystallography **57**(3): 401-411.

Harding, M. M. (2002). "Metal-ligand geometry relevant to proteins and in proteins: sodium and potassium." Acta Crystallographica Section D: Biological Crystallography **58**(5): 872-874.

Harding, M. M. (2004). "The architecture of metal coordination groups in proteins." Acta Crystallographica Section D: Biological Crystallography **60**(5): 849-859.

Harding, M. M. (2006). "Small revisions to predicted distances around metal sites in proteins." Acta Crystallographica Section D: Biological Crystallography **62**(6): 678-682.

Howard, J. A. K., R. C. B. Copley, et al. (1998). "Systematic analysis of metal coordination sphere geometry from crystallographic data: a general method for detecting geometrical preferences, deformations and interconversion pathways." Chemical Communications **1998**(20): 2175-2176.

Hsin, K., Y. Sheng, et al. (2008). "MESPEUS: a database of the geometry of metal sites in proteins." Journal of Applied Crystallography **41**(5): 963-968.

Kuntz, G. (1973). "Contrasting structures of magnesium and calcium adenosine-triphosphate complexes as studied by nuclear relaxation." Federation Proceedings **32**: 546-552.

Liao, J. C., S. Sun, et al. (2004). "The conformational states of Mg.ATP in water." European Biophysics Journal **33**(1): 29-37.

Mesecar, A. D. and T. Nowak (1997). "Metal-Ion-Mediated Allosteric Triggering of Yeast Pyruvate Kinase. 1. A Multidimensional Kinetic Linked-Function Analysis?" Biochemistry **36**(22): 6792-6802.

Mildvan, A. S. (1987). "Role of magnesium and other divalent cations in ATP-utilizing enzymes." Magnesium **6**(1): 28-33.

Orpen, A. G. (2002). "Applications of the Cambridge Structural Database to molecular inorganic chemistry." Acta Crystallographica Section B: Structural Science **58**(3): 398-406.

Taylor, R. (2002). "Life-science applications of the Cambridge Structural Database."

Wang, G. and R. L. Dunbrack (2003). "PISCES: a protein sequence culling server". Oxford University Press **19:** 1589-1591.

Williams, R. L., D. A. Oren, et al. (1993). "Crystal Structure of Myxococcus xanthus Nucleoside Diphosphate Kinase and its Interaction with a Nucleotide Substrate at 2 A Resolution." Journal of Molecular Biology **234**(4): 1230-1247.

Zhang, J., P. L. Yang, et al. (2009). "Targeting cancer with small molecule kinase inhibitors." Nature reviews. Cancer **9**(1): 28

8. Summary, conclusions and future work

In drug discovery projects, research importantly focuses on ligand-protein interactions. To discover a new drug, the general intent is to find a drugable small molecule from tested compounds (or drug candidates), which should be able to specifically interact with the binding site of the preselected target protein resulting in the disruption of the native ligand-protein interactions and inhibiting the protein's activity. In modern drug discovery, several approaches have been used to aid the small-molecule screening, such as high-throughput screening (HTS), database mining or virtual high-throughput screening (VHTS) as mentioned in Chapter 1. In order to carry out the process efficiently, those approaches iteratively utilise databases (or libraries) which are seen as a powerful and convenient tool in storing and retrieving relevant information.

This project applies structure-based and database mining approaches to study ligand-protein interactions. The major aims of this project include: the development of a small-molecule database which stores a vast number of commercially purchasable compound structures and their molecular properties (i.e. the descriptors) with a web-based interface for the database; the discovery of correlations between compound activity and molecular descriptors using the known bio-assay outcomes; the application of designed structure-based mining to discover ligands (inhibitors) for targeted protein structures. Finally, another area of ligand-protein interactions considered in this project is the metal-binding sites in metalloproteins. The present work undertook some surveys of various metal sites found in metalloprotein

structures as well as the development of a database system and web-based interface to store and apply the geometry information of those metal sites.

8.1 Development of small-molecule database, EDULISS 2.0

In order to support the studies of structure-activity relationship (SAR) and structure-based mining, we have developed a relational database system called **ED**inburgh **U**niversity **L**igand **S**election **S**ystem (EDULISS 2.0) which primarily stores the structure-data files (SDfiles) of +5.5 million commercial purchasable small molecules collected from 28 prominent supplier chemical catalogues and over 1,500 various calculated molecular properties (descriptors) for each compound. The tables of this database have been designed to allow good accessibility for other applications or scripts and a well planed approach to implement future updates. A user-friendly web-based interface of EDULISS 2.0 has been established and is available at http://eduliss.bch.ed.ac.uk/ providing a series of straightforward facilities to satisfy the need of data-mining for drug discovery.

The present work develops a very efficient procedure to identify unique compounds from the vast collection using three high discrimination descriptors (W3D: Wiener 3D index; Whete: Wiener-type index from electronegativity weighted distance matrix; Vu: V total size index, unweighted WHIM descriptors) to distinguish the compounds prior to performing the structure comparisons. This strategy considerably reduces the number of required pair-wise structure comparisons from 3×10^{13} down to 6×10^{6}, thus the recognition of unique compound in EDULISS 2.0 can be done very efficiently. Using this approach we could show that the total number of unique compounds present is 4,011,697.

8.2 Applications of Structure-Activity Relationship (SAR) in ligand-protein binding study

PubChem bioassay data, an NMR based screening assay for a human FKBP12 protein (PubChem AID: 608) was utilised together with a series of molecular descriptors to construct a prediction model using a Logistic Regression approach. This model reveals the molecular descriptors which are correlated with ligand activity, and predicts the probability of ligand-protein binding. The result is that 38 descriptors are found to be good predictors (listed in Table 5.2): they are mainly calculated from three-dimensional representation of a molecule (21 out of 38), and are the classes of geometrical descriptors, including RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors and GETAWAY descriptors. This result seems to indicate that the 3D-based descriptors have higher potential effectiveness in such an SAR modelling study. The model fitting exercise also reveals the presence of some molecular functional groups and fragments that are clearly associated with the prediction of compound activity. The numbers of functional groups nRCONR2, nRSR and nThiazoles show a statistically significant difference between active and inactive compounds measured by the analysis of the ANOVA procedure. In the prediction of ligand-protein binding probability, however, the built models are barely satisfactory to predict the activity of compounds accurately in the 20 stricter validation tests.

The present work applies a neural network technique called SOMs to visualise the PubChem compound similarity based on the 38 predictive descriptors. The analysis of SOM succeeds in aggregating the 36 % active compounds (16 out of 44) in a cluster and discriminate them from the 95 % inactive compounds as only 172 (out of

3,724) inactive compounds were found in the cluster, and successfully visualises the result as shown in Figure 5.6. Most of the active compounds aggregated in the cluster show strong binding affinities and possess two common structural patterns including the amide (N–C=O) and sulfide (C–S–C) groups with both groups adjacent showing a higher probability to bind with FKBP12.

This SAR study succeeds in revealing the potentially important molecular descriptors and essential sub-structures involved in the binding interaction but the built model only predicts the activity of compounds accurately for fewer cases in the stricter validation tests. The failure of prediction may be caused by the fact that the descriptors we used for the modelling exercise represent the properties of a whole molecular structure, but the active compounds may only partially interact with the binding site, e.g. only a fragment of the molecule involved in the interaction. Moreover, some descriptors are highly correlated to others and in some cases the descriptors belonging to the same group present high correlations between themselves as demonstrated in Figure 4.1 and 4.3. The large number of redundant or similar variables in representing the molecular characteristics may distort the modelling evaluation and cause a situation that the selected descriptors in the repeated validation tests are often different but they are actually very similar. Too many descriptors also cause over-fitting in the modelling process as the more the descriptors used, the greater the likelihood to obtain a correlation. The large number of descriptors also increases the requirements of computing and storage. The selection of proper descriptors prior to modelling exercise should be considered in the future work, such as only selecting the representation of the descriptors which are

similar or show high correlations between themselves. Similarly, to provide a full the representation of the descriptors, it is also necessary to consider the presence of bioisosteres which are molecular functional groups or atomic types that have physicochemical similarities, including the size, shape, hydrophobicity, electronic distribution and so on, and show similar biological properties (Kraus 1983). For example, the groups of CO_2, N_2O, N_3^-, and CNO^- are found to be similar in physical properties and are considered as a group of isosteres, as well as the Cl, Br, I, SH and PH_2 because these atomic types have the same number (in this case 7) of peripheral electrons (Patani and LaVoie 1996). A full description is given by Silverman 2004.

8.3 Application of Atomic Characteristic Distance (ACD) and EDULISS 2.0 for ligand discovery

We have developed a molecular descriptor called ACD (Atomic Characteristic Distance) to profile the distribution of specified atom types in a compound, including halogens, sulphur, phosphorus, and hydrogen bond donors and acceptors. This molecular descriptor is encoded into a set of bit strings using the bit-wise algorithm (illustrated in Figure 2.8) which are ideal for storage, manipulative facility and high-speed screening. The present work applies the ACD and the facility of EDULISS 2.0 including its small-molecule collection and web-based interface to find possible inhibitors for pyruvate kinase. The selected candidate compounds were tested for the inhibition of protein catalysis activity, and some of the protein-ligand complex structures have also been crystallised.

In stage 1, we designed a mining question based on the positions of the five sulphate ions found in the active and effector sites of the sulphate-bound PK structure (i.e.

*Lm*PYK 3E0V) to find the compounds possessing two or more sulphate groups. The sulphur atoms are at the distances fitting the distances between the five sulphate ions. We selected 1,3,6,8-pyrenetetrasulfonic acid via the interface of EDULISS for co-crystallising the ligand-protein complexes. The result is that the complex has been crystallised successfully at the resolution 2.1 Å and its structure model is shown as Figure 6.5. Four pyrene-like compounds, i.e. 1,3,6,8-pyrenetetrasulfonic acid, were found in the active sites of each subunit; one additional compound was bound between subunit B of adjacent model in the lattice. The four pyrene-like compounds found in the active sites appear to bind the active sites weakly. At the binding position of the odd one, the compound potentially enhances the packing arrangement in the crystal unit cells thus it promotes crystallisation and provides a better structure resolution. We also observed that this compound can only bind the edge of subunit B but not in the active sites at low concentration, hence this compound has been used in lower concentration for some consequent crystallographic experiments in order to improve the structure resolution.

In stage 2, we used an *Lm*PYK/ATP/FBP complex structure as the target and the screening took both the interatomic distances between various atom types and the hydrogen bond interactions observed in the binding sites into account. We applied the idea of ACD and wrote a java script to accomplish these multiple requirements. The design of screening criteria and the result of each screening are illustrated and tabulated in Figure 6.7 and 6.8. The screens are seen to be able to effectively exclude the unfit compounds allowing further detailed inspection to select the preferred candidates one by one. We selected 8 compounds (shown as Figure 6.9)

and two of them (Comp. 3 / Ponceau S and Comp. 5 / Reactive Blue 4) show 33.34 and 100 % inhibition in *Lm*PYK enzyme activity, respectively (summarised in Table 6.2). The Comp. 3 bound complex structure has been crystallised successfully at a resolution of 2.7 Å and its structure model is shown as Figure 6.11. Comp. 3 occupies a part of FBP (the allosteric effector) binding site resulting in the inhibition of the kinase activity. Although Comp. 5 shows a good inhibition in *Lm*PYK enzyme activity, a crystal of its *Lm*PYK complex bound at the active site could not be obtained successfully. We therefore modified the screening technique to select three analogues of Comp. 5 (Reactive Blue 2, Acid Blue 25 and Acid Blue 80) containing the core structure anthraquinone shown as Figure 6.13. The Acid Blue 80 bound complex crystal has been obtained at a resolution of 2.3 Å (shown as Figure 6.14), which occupies the active site and consequently prevents the ADP molecule from binding and disturbs the kinase's allosteric function.

The web-based interface of EDULISS 2.0 is most useful in the early stages of experimental ligand binding projects for initial screening. As EDULISS contains a large collection of commercially purchasable small molecules, it offers a higher probability to hit candidates, particularly when using complex screening criteria, and allows researchers to purchase the preferred compounds directly without chemical synthesis. The application of ACD gives a good scope for the ligand design and has shown success in selecting ligands. In future work, ACD could be developed to include information about rings in a compound. Rings affect molecular hydrophobicity which is seen to be important in the binding interaction. For example, we could flag the geometrical centre of an interesting ring and then determine the

distances between the centre and the specified atom types, so that the characteristic of the compound can be structurally profiled in more detail and the ligand design can take the interaction involving the hydrophobic pocket found in the binding site into account. At present, bit strings of ACD can only ensure that the compounds possess the specified interatomic distances but not the conjunctions of the specified atoms. This remaining requirement should be accomplished in the future work. Finally, the EDULSS database and interface should be continually updated, stabilised and enhanced with new elements to extend its utility for other applications. For example, the collected small molecules can be further decomposed into fragments using e.g. Murcko method (Bemis and Murcko 1996) for the requirement of fragment-based drug design (Rees, Congreve et al. 2004; Hajduk 2006). This approach is increasingly used for drug discovery in this decade and has been successful in finding potent inhibitors for various targets (Hajduk and Greer 2007). Alternatively, the fragments could be applied to virtual high-throughput screening to simplify the computational analysis for binding interaction.

8.4 Development of a database and web-based interface for novel metalloprotein descriptors and their application

Other than the use of physicochemical descriptors, including the atomic characteristic distance (ACD), we also discuss a different set of descriptors that describe the geometry of various metal sites to look into the metal-binding sites in metalloproteins for the study of ligand-protein interaction. The geometric information of metals is derived from metalloprotein structures which were deposited in the Protein Data Bank (PDB) not later than 1 January 2007, including protein and

nucleic acid crystal structures. The structures should contain one or more metal atoms and are determined by diffraction methods at a resolution of 2.5 Å or better. The metal atoms selected are biologically common, including Fe, Ni, Mn, Ca, Cu, Na, Mg, K, Co and Zn. The present work has developed a relational database system and web-based interface to store and apply these descriptors, called MESPEUS (**ME**tal **S**ites in **P**roteins at **E**dinburgh **U**niver**S**ity) and it is available at http://eduliss.bch.ed.ac.uk/MESPEUS/.

This study discusses the importance of near atomic resolution in deriving mean bond distances between observed metal and the donor atoms. As an example, the observed mean distance for Zn-N of histidine with coordination number 4 is given in Table 7.2. The mean distance can be fairly different depending on the resolution of the structure determination. The standard deviation, representing the scatter of results, is smaller at higher resolution; good consistency of observations only occurs at a fairly high resolution as illustrated in Figure 7.12. A survey of interactions between Mg and adenosine triphosphate (ATP) indicates that the most common protein types which contain Mg-$O_{ATP\text{-}phosphate}$ sites are transferases and Mg is normally linked to protein through one or two amino acid side chains, mainly ASP, GLU, ASN, GLN, SER or THR. As the preferred interaction of Mg-$O_{ATP\text{-}phosphate}$, the most common pattern is linkage through the β- and γ-phosphate groups, but all other possibilities also occurred (detailed in Appendix 2). Mg is commonly coordinated with 6 donors and other than amino acids and ATP the coordination group may also include one to four water molecules and occasionally another small molecule like oxalate as a bidentate ion.

The MESPEUS database with its user-friendly web interface provides a good capability to immediately identify and display the metal sites. The stored geometry information of metal sites manages to provide a useful reference and potential recommendations for building or validating a protein structure model during refinement as well as useful for interpretation of electron-density maps in new protein structures. In some metal site studies, researchers are widely interested in the donors which indirectly interact with the metal atom through hydrogen bonds to one of the direct donors, i.e. the so-called second shell (Christianson and Fierke 1996; Karlin, Zhu et al. 1997). Second shell donors are defined as any non-amino acid molecule or residue whose distance to the non hydrogen atoms of the direct donor groups is within 3.5 Å (Karlin and Zhu 1997). The second shell donor groups have shown a contribution to the energetic stabilisation of the metal complex (Dudev, Lin et al. 2003). The future work of MESPEUS can extend its scope to involve the geometry information of these second shell donor groups allowing researchers to understand their characteristic or the preference of donor group types, and the information can also be applied in the structure refinement.

8.5 Reference:

Bemis, G. W. and M. A. Murcko (1996). "The properties of known drugs. 1. Molecular frameworks." Journal of Medicinal Chemistry **39**(15): 2887-2893.

Christianson, D. W. and C. A. Fierke (1996). "Carbonic anhydrase: evolution of the zinc binding site by nature and by design." Accounts of Chemical Research **29**(7): 331-339.

Dudev, T., Y. Lin, et al. (2003). "First- Second Shell Interactions in Metal Binding Sites in Proteins: A PDB Survey and DFT/CDM Calculations." Journal of the American Chemical Society **125**(10): 3168-3180.

Hajduk, P. J. (2006). "Fragment-based drug design: How big is too big?" Journal of Medicinal Chemistry **49**(24): 6972-6976.

Hajduk, P. J. and J. Greer (2007). "A decade of fragment-based drug design: strategic advances and lessons learned." Nature Reviews Drug Discovery **6**(3): 211-219.

Karlin, S. and Z. Y. Zhu (1997). "Classification of mononuclear zinc metal sites in protein structures". National Acad Sciences **94:** 14231-14236.

Karlin, S., Z. Y. Zhu, et al. (1997). "The extended environment of mononuclear metal centers in protein structures." National Acad Sciences **94:** 14225-14230.

Kraus, J. L. (1983). "Isoterism and molecular modification in drug design: New n-dipropylacetate analogs as inhibitors of succinic semi aldehyde dehydrogenase." Pharmacological Research Communications **15**(2): 119-129.

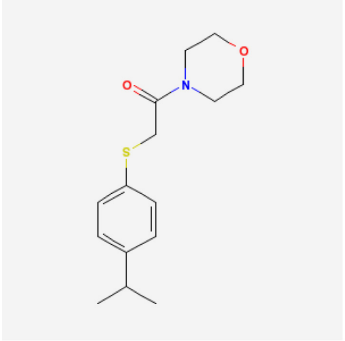
Patani, G. A. and E. J. LaVoie (1996). "Bioisosterism: A Rational Approach in Drug Design." Chemical reviews **96**(8): 3147.

Rees, D. C., M. Congreve, et al. (2004). "Fragment-based lead discovery." Nature Reviews Drug Discovery **3**(8): 660-672.

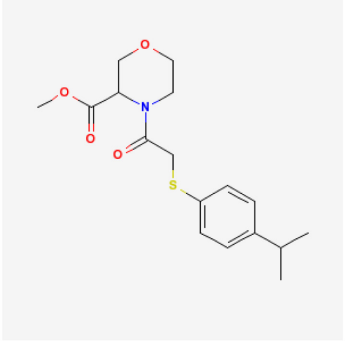
Silverman, R. B. (2004). The organic chemistry of drug design and drug action,

Academic Press**:** P29-34.

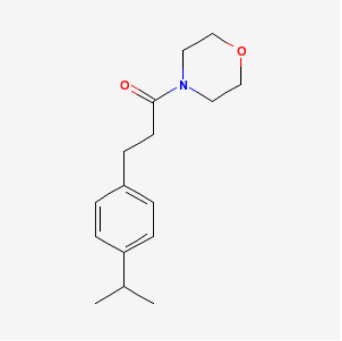
Appendix 1. Chemical structures of the 44 active compounds in the PubChem bioassay (PubChem AID: 608). The labels represent the compound's ID, estimated Kd value (μM) and the rank in the overall bioassay. (continued on next three pages)



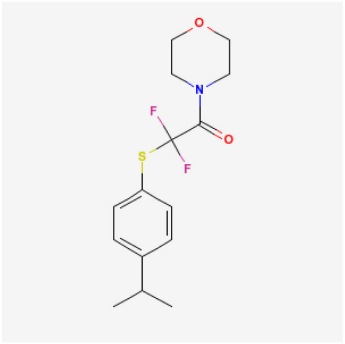
16725062/0.05/1



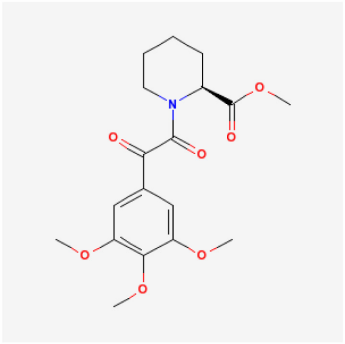
16725059/0.3/2



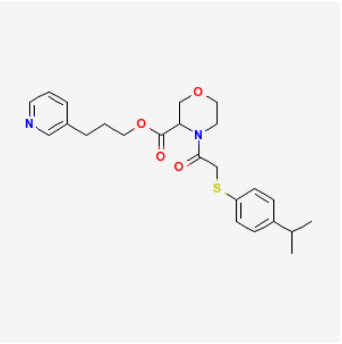
16725066/2.3/3



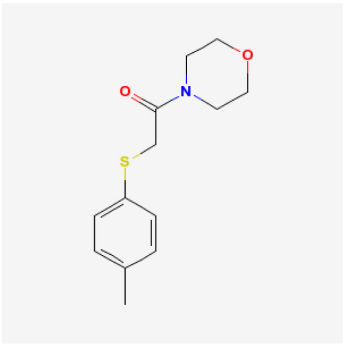
16725067/2.8/4



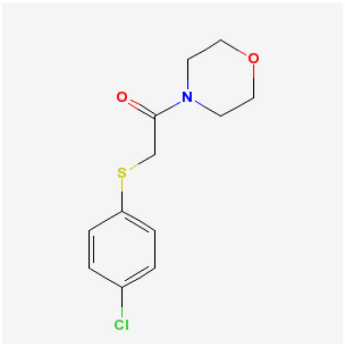
16725057/3/5



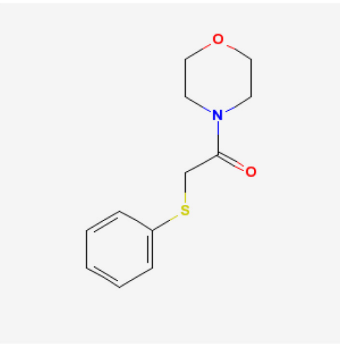
16725061/5.6/6



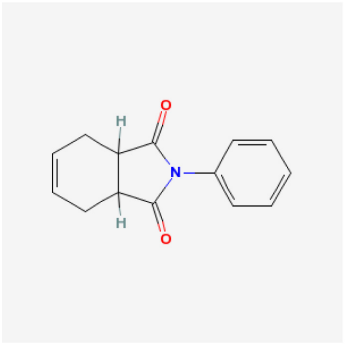
875760/12/7



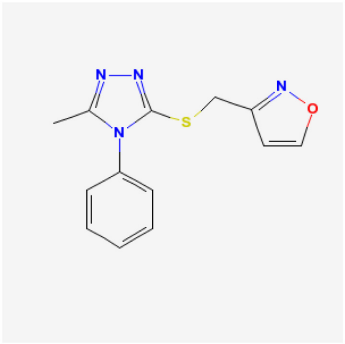
879920/19/8



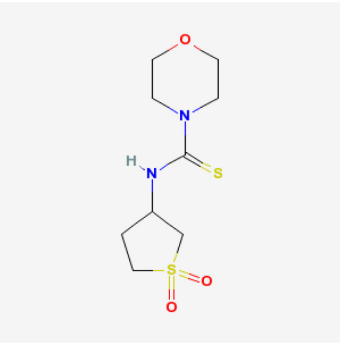
532039/23.5/9



2843508/26.1/10

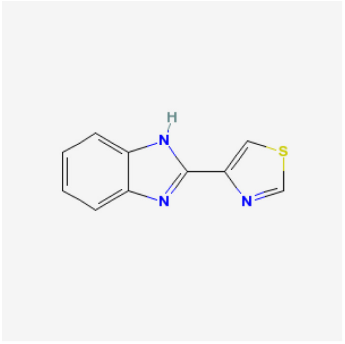


712009/38.5/11

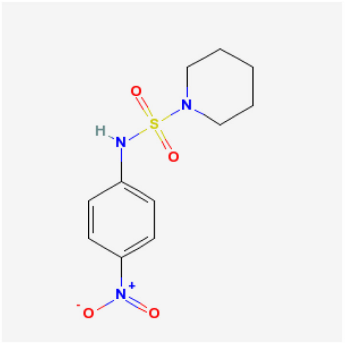


2906052/47.4/12

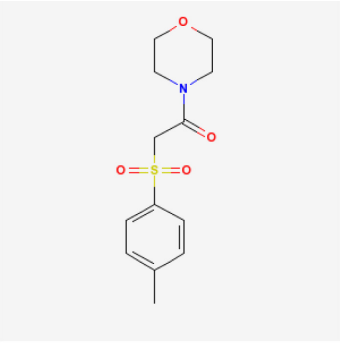
Appendix 1. Continued.



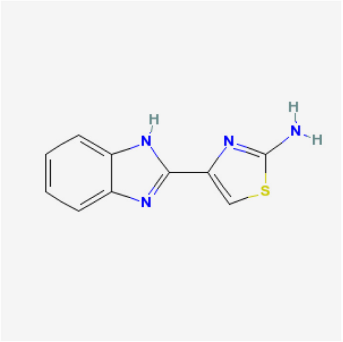
5430/56/13



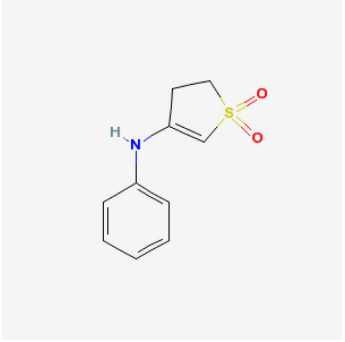
2926730/60/14



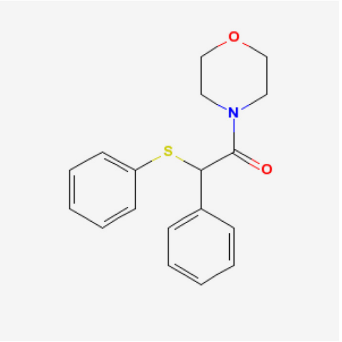
285619/95/15



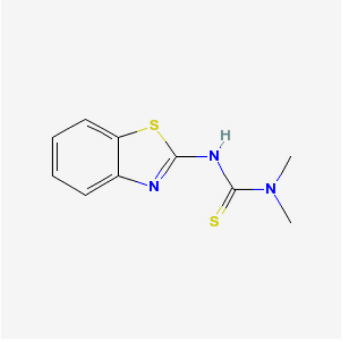
776692/98/16



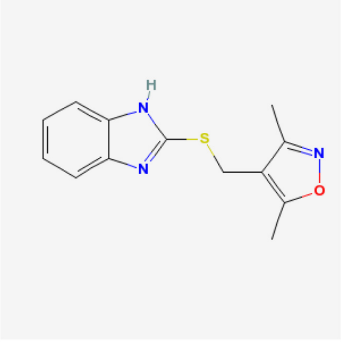
1380383/135/17



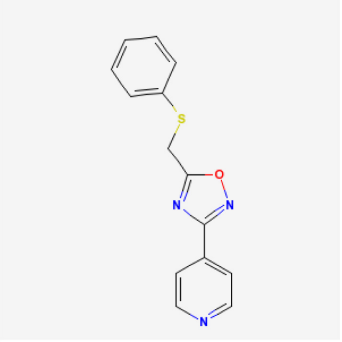
2886349/146/18



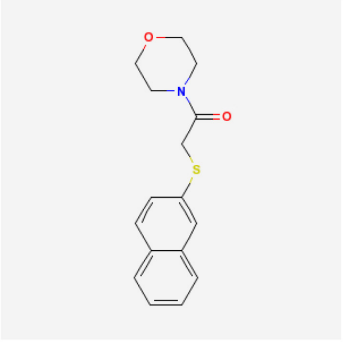
946666/152.1/19



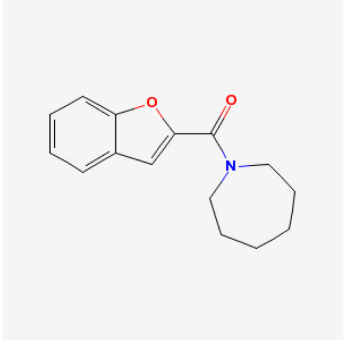
975504/157/20



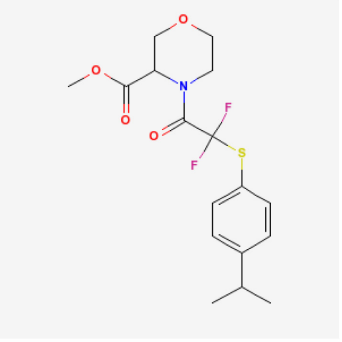
894766/167/21



2200849/180/22

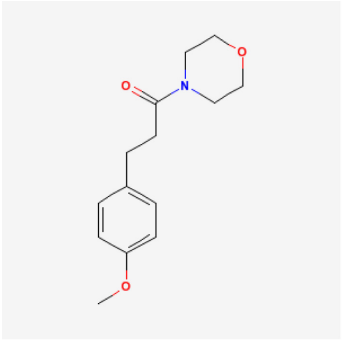


977283/184/23

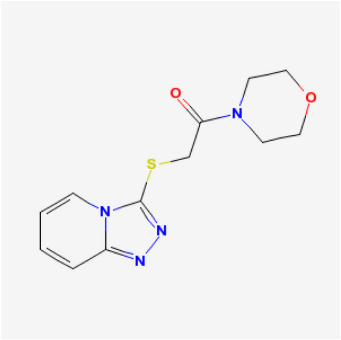


16725065/190/24

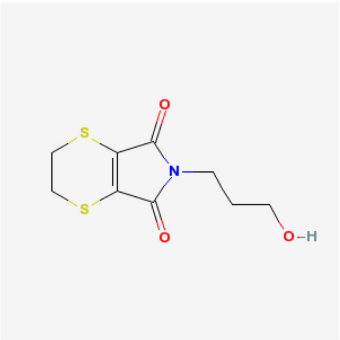
Appendix 1. Continued.



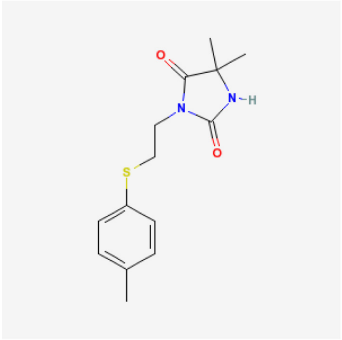
3397179/204/25



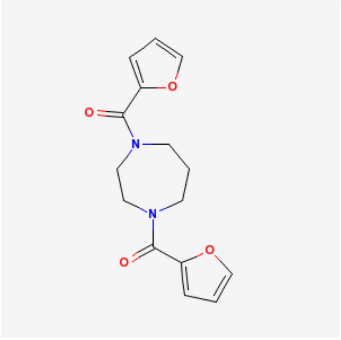
752201/221/26



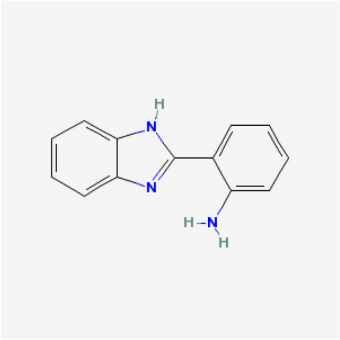
2227556/250.8/27



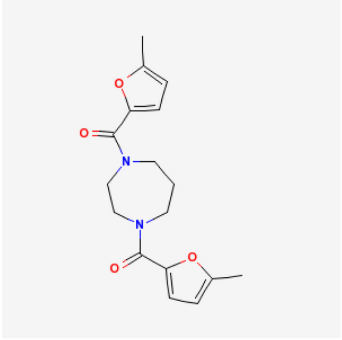
973322/270/28



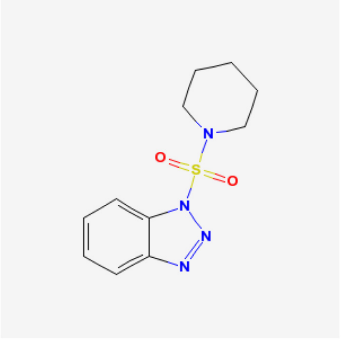
750345/277/29



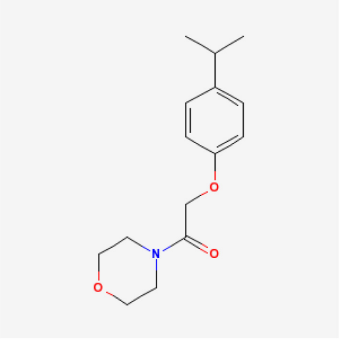
79869/286/30



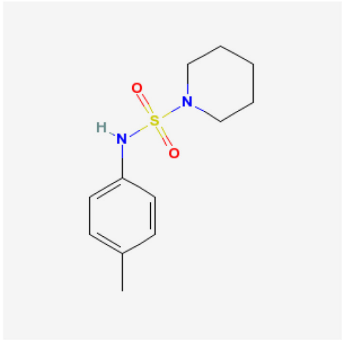
753179/374/31



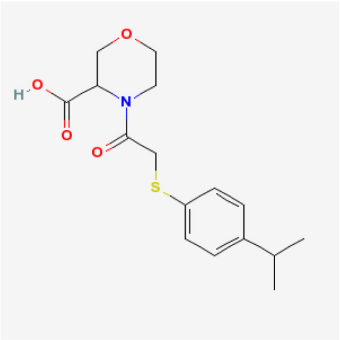
2216937/385/32



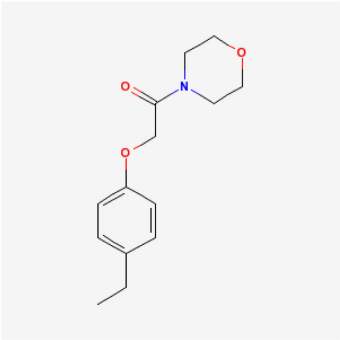
797980/394/33



224941/425/34

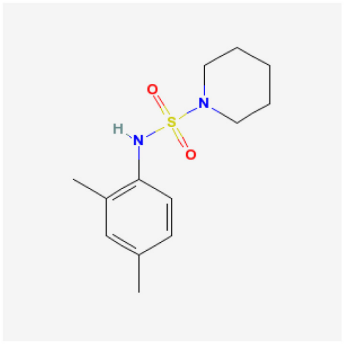


16725060/426/35

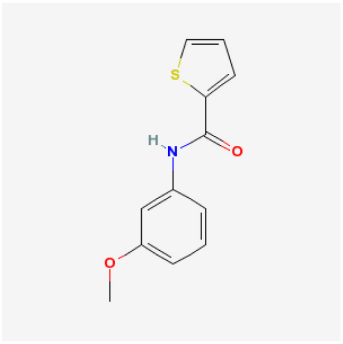


806693/435/36

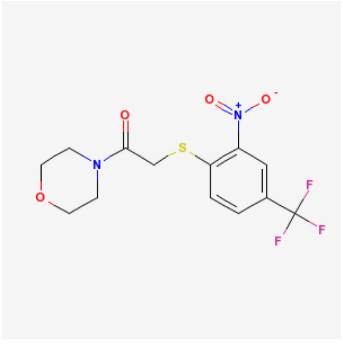
Appendix 1. Continued.



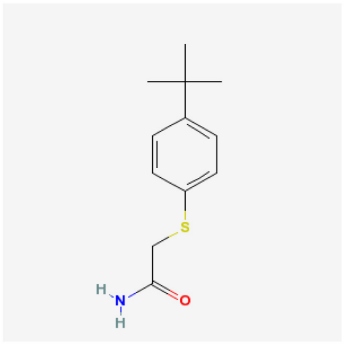
2988699/436/37



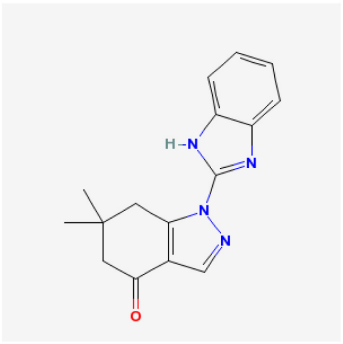
790310/440/38



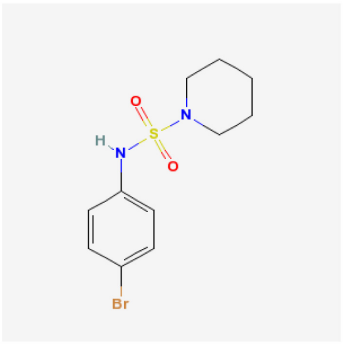
1321581/448/39



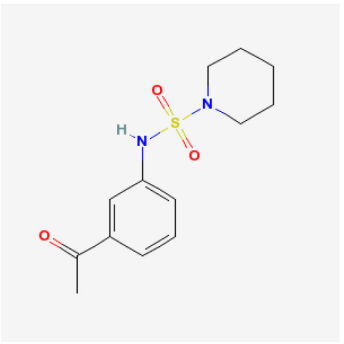
975595/455/40



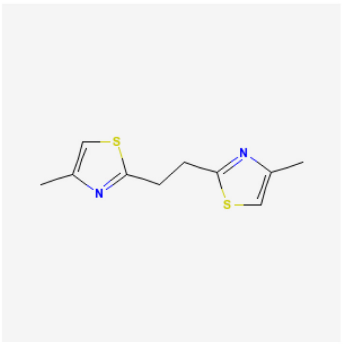
972660/491/41



2140092/669/42



1258596/686/43



969581*

*: Data didn't be provided by PubChem.

Appendix 2. Mg sites with links to ATP and to protein. Column cngp lists the one letter code of amino acid side chain donor groups and O represents main chain carbonyl oxygen; sd1 and sd2 are sequence separations of first and second donors, second and third donors (-1 for nil); CN is coordination number; rms is r.m.s. deviation of metal-donor atoms distance from target distance (described in section 7.2.1.1) to be a quality indicator; resln is structure resolution; npd_1 to npd_5 are full names from PDB files of all non-protein donor groups; class is from PDB header. Only three sites link to main chain carbonyl oxygen (marked by a superscript 'a'), and two cases where the linkage is through two O atoms of the same phosphate group (marked by a superscript 'b'). (continued on next three pages)

cngp	sd1	sd2	CN	rms	resln	PDB	npd_1	npd_2	npd_3	npd_4	npd_5	Class
D	-1	-1	3	0.39	2.50	2AQX	ATP_O3G_2462_	ATP_O2A_2462_	_	_	_	TRANSFERASE
D	-1	-1	6	0.09	1.60	1KJ9	ATP_O1G_1_	HOH_O_123_	HOH_O_129_	HOH_O_133_	HOH_O_252_	TRANSFERASE
D	-1	-1	4	0.14	2.50	2AQX	ATP_O3G_1462_	ATP_O2A_1462_	HOH_O_105_	_	_	TRANSFERASE
D	-1	-1	4	0.19	2.50	2AQX	ATP_O1G_2462_	ATP_O1B_2462_	_	_	_	TRANSFERASE
D	-1	-1	6	0.23	2.25	1Y8Q	ATP_O2G_801_	ATP_O2B_801_	HOH_O_220_	HOH_O_414_	HOH_O_727_	LIGASE
D	-1	-1	3	0.26	2.40	1DER	ATP_O1B_1_J	ATP_O2A_1_J	_	_	_	CHAPERONIN
D	-1	-1	6	0.36	2.10	1U5R	ATP_O2B_411_	HOH_O_18_	HOH_O_46_	HOH_O_316_	_	TRANSFERASE
D	-1	-1	2	0.37	2.30	1Q97	ATP_O3B_485_	_	_	_	_	TRANSFERASE
D	-1	-1	4	0.49	2.30	1Q97[b]	ATP_O1G_485_	ATP_O2G_485_	ATP_O1A_485_	_	_	TRANSFERASE
DD	2	-1	6	0.35	2.00	1DY3	ATP_O2B_200_A	ATP_O1A_200_A	HOH_O_78_Z	HOH_O_140_Z	_	PYROPHOSPHORYLASE
DD	2	-1	6	0.47	2.00	1DY3	ATP_O1G_200_A	ATP_O2B_200_A	87Y_O16_201_A	HOH_O_141_Z	_	PYROPHOSPHORYLASE
DD	98	-1	6	0.12	2.11	1MB9[b]	ATP_O2G_702_	ATP_O2A_702_	ATP_O3A_702_	HOH_O_655_	_	HYDROLASE
DE	16	-1	3	0.54	2.20	2CJA	ATP_O2A_1507_B	_	_	_	_	LIGASE
DE	99	-1	5	0.37	2.40	1N56	ATP_O1A_1803_	HOH_O_648_	HOH_O_1023_	_	_	TRANSFERASE/DNA
DOD[a]	1	97	7	0.35	2.40	1N56	ATP_O1G_1803_	ATP_O1A_1803_	HOH_O_666_	_	_	TRANSFERASE/DNA

cngp	sd1	sd2	CN	rms	resln	PDB	npd_1	npd_2	npd_3	npd_4	npd_5	Class
E	-1	-1	6	0.10	1.60	1KJ9	ATP_O3G_1_	ATP_O2B_1_	EDO_O1_15_	HOH_O_35_	_	TRANSFERASE
E	-1	-1	3	0.39	2.15	1YFR	ATP_O1B_1500_	ATP_O1A_1500_	_	_	_	LIGASE
E	-1	-1	3	0.61	2.40	2I4O	ATP_O3G_402_	ATP_O2B_402_	_	_	_	LIGASE
ED	2	-1	5	0.28	2.20	1TFW	ATP_O3G_1501_	ATP_O2B_1501_	ATP_O1A_1501_	_	_	TRANSFERASE/RNA
ED	24	-1	6	0.11	2.10	1A49	OXL_O1_2333_D	OXL_O4_2333_D	ATP_O2G_2335_D	HOH_O_6439_	_	TRANSFERASE
ED	24	-1	5	0.17	2.10	1A49	OXL_O1_1733_C	OXL_O4_1733_C	ATP_O2G_1735_C	_	_	TRANSFERASE
EE	12	-1	6	0.11	1.60	1KJ9	ATP_O2G_5_	ATP_O2A_5_	HOH_O_420_	_	_	TRANSFERASE
EQ	3	-1	4	0.41	2.40	2I4O	ATP_O2B_402_	ATP_O2A_402_	_	_	_	LIGASE
EQ	3	-1	5	0.39	2.40	2I4O	MG_MG_302_	ATP_O2B_401_	ATP_O2A_401_	_	_	LIGASE
HN	163	-1	3	0.39	2.00	1YUN	ATP_O1B_1220_	_	_	_	_	TRANSFERASE
N	-1	-1	6	0.15	2.40	2BU2	ATP_O2G_1386_A	ATP_O2B_1386_A	ATP_O1A_1386_A	HOH_O_40_Z	HOH_O_87_Z	TRANSFERASE
N	-1	-1	6	0.31	2.20	1TC0	ATP_O1B_601_	ATP_O2A_601_	HOH_O_75_	HOH_O_109_	HOH_O_110_	CHAPERONE
N	-1	-1	3	0.32	2.50	1TID	ATP_O1G_201_	ATP_O1B_201_	_	_	_	TRANSCRIPTION
N	-1	-1	6	0.43	2.50	1TID	ATP_O1G_200_	ATP_O1B_200_	ATP_O1A_200_	HOH_O_2_	HOH_O_49_	TRANSCRIPTION
N	-1	-1	3	0.45	2.00	1CSN	ATP_O1G_299_	ATP_O2A_299_	_	_	_	PHOSPHOTRANSFERASE
ND	13	-1	6	0.08	2.40	1S9J	ATP_O1B_535_	ATP_O2A_535_	HOH_O_97_	HOH_O_98_	_	TRANSFERASE
ND	13	-1	6	0.11	2.20	1QMZ	ATP_O2G_381_A	ATP_O3B_381_A	ATP_O2A_381_A	HOH_O_111_Z	_	COMPLEX(PROTEINKINASE/CYCLIN)
ND	13	-1	6	0.37	2.10	1U5R	ATP_O2G_412_	ATP_O2A_412_	HOH_O_31_	HOH_O_225_	_	TRANSFERASE
O[a]	-1	-1	3	0.49	1.70	1XNG	ATP_O1G_304_	ATP_O2B_304_	_	_	_	LIGASE
O[a]	-1	-1	6	0.34	1.70	1XNG	ATP_O1G_303_	ATP_O2B_303_	ATP_O2A_303_	HOH_O_55_	HOH_O_161_	LIGASE
R	-1	-1	3	0.46	2.50	1XDP	ATP_O2A_704_	HOH_O_94_	_	_		TRANSFERASE

cngp	sd1	sd2	CN	rms	resln	PDB	npd_1	npd_2	npd_3	npd_4	npd_5	Class
RR	30	-1	4	0.40	2.50	1XDP	ATP_O1G_701_	_	_	_	_	TRANSFERASE
S	-1	-1	6	0.03	1.85	2IYW	ATP_O1G_201_A	ATP_O1B_201_A	HOH_O_10_Z	HOH_O_200_Z	HOH_O_204_Z	TRANSFERASE
S	-1	-1	6	0.06	1.70	1E2Q	ATP_O3G_302_A	ATP_O1B_302_A	HOH_O_21_Z	HOH_O_22_Z	HOH_O_23_Z	THYMIDYLATEKINASE
S	-1	-1	6	0.10	2.27	1W7A	ATP_O2G_1801_A	ATP_O2B_1801_A	HOH_O_2101 _A	HOH_O_2110 _A	HOH_O_2132 _A	DNABINDING
S	-1	-1	6	0.14	2.50	1XEF	ATP_O3G_800_	ATP_O2B_800_	HOH_O_160_	HOH_O_165_	HOH_O_171_	TRANSPORTPROTEIN
S	-1	-1	6	0.28	2.50	2C8V	ATP_O3G_5292_A	ATP_O1B_5292_A	HOH_O_4_Z	HOH_O_13_Z	HOH_O_97_Z	XIDOREDUCTASE
S	-1	-1	2	0.69	2.40	1KVK	ATP_O3G_535_	_	_	_	_	TRANSFERASE
SQ	28	-1	6	0.06	1.50	2CBZ	ATP_O1G_1873_A	ATP_O2B_1873_A	HOH_O_187_Z	HOH_O_278_Z	_	TRANSPORT
SQ	103	-1	6	0.04	1.60	1F2U	ATP_O1G_901_A	ATP_O1B_901_A	HOH_O_48_	HOH_O_51_	_	REPLICATION
SQ	105	-1	6	0.10	2.50	1XEX	ATP_O1G_1001_	ATP_O1B_1001_	HOH_O_30_	HOH_O_109_	_	CELLCYCLE
T	-1	-1	5	0.08	2.40	1E79	ATP_O2G_600_A	ATP_O2B_600_A	HOH_O_72_Z	HOH_O_73_Z	_	ATPPHOSPHORYLASE
T	-1	-1	6	0.16	1.80	2BEK	ATP_O3G_500_A	ATP_O2B_500_A	HOH_O_62_Z	HOH_O_183_Z	HOH_O_321_Z	CHROMOSOMESEGREGATION
T	-1	-1	6	0.16	1.94	1SVM	ATP_O1G_800_A	ATP_O2B_800_A	HOH_O_5_	HOH_O_803_	HOH_O_804_	VIRUS/VIRALPROTEIN
T	-1	-1	6	0.19	2.40	1II0	ATP_O2G_1591_	ATP_O1B_1591_	HOH_O_2073_	HOH_O_2102_	HOH_O_2209_	HYDROLASE
T	-1	-1	6	0.22	2.20	1YTM	ATP_O3G_1541_	ATP_O2B_1541_	HOH_O_9_	HOH_O_229_	HOH_O_351_	LYASE
T	-1	-1	6	0.30	1.80	1OS1	ATP_O3G_541_	ATP_O2B_541_	HOH_O_674_	HOH_O_716_	HOH_O_738_	LYASE
TDE	38	61	6	0.16	1.80	1A82	ATP_O3G_802_	ATP_O1B_802_	HOH_O_444_	_	_	BIOTINBIOSYNTHESIS
TE	86	-1	6	0.13	2.10	1G64	ATP_O2B_1000_	ATP_O2A_1000_	HOH_O_1380_	HOH_O_1381_	_	TRANSFERASE
TE	86	-1	6	0.14	2.10	1G64	ATP_O2B_999_	ATP_O2A_999_	HOH_O_1331_	HOH_O_1332_	_	TRANSFERASE
TQ	28	-1	6	0.07	2.05	2BBS	ATP_O2G_1_	ATP_O2B_1_	HOH_O_7_	HOH_O_202_	_	TRANSPORTPROTEIN
N	-1	-1	6	0.18	2.20	1TC0	ATP_O1B_301_	ATP_O2A_301_	HOH_O_38_	HOH_O_50_	HOH_O_125_	CHAPERONE

cngp	sd1	sd2	CN	rms	resln	PDB	npd_1	npd_2	npd_3	npd_4	npd_5	Class
The following proteins have ≥ 75 % sequence identity to proteins in the list above, as established with PISCES (Wang and Dunbrack 2003) and the sites appear very similar.												
E	-1	-1	6	0.16	1.60	1KJ8	ATP_O3G_1_	ATP_O2B_1_	HOH_O_29_	HOH_O_231_	_	TRANSFERASE
E	-1	-1	3	0.51	1.60	1KJ8	ATP_O3G_5_	HOH_O_726_	_	_	_	TRANSFERASE
EE	12	-1	6	0.08	1.60	1KJ8	ATP_O2G_1_	ATP_O2A_1_	HOH_O_22_	_	_	TRANSFERASE
EE	12	-1	7	0.41	1.60	1KJ8	ATP_O2G_5_	ATP_O3B_5_	ATP_O2A_5_	HOH_O_539_	_	TRANSFERASE
ND	13	-1	6	0.32	2.10	1U5R	ATP_O3G_411_	ATP_O2A_411_	HOH_O_218_	HOH_O_261_	_	TRANSFERASE
T	-1	-1	4	0.26	2.30	1XMJ	ATP_O2G_1_	ATP_O2B_1_	HOH_O_105_	_	_	MEMBRANEPROTEIN,HYDROLASE
TQ	28	-1	6	0.13	2.25	1XMI	ATP_O2G_4_	ATP_O2B_4_	HOH_O_687_	HOH_O_709_	_	MEMBRANEPROTEIN,HYDROLASE
TQ	28	-1	4	0.15	2.35	1R0Z	ATP_O2G_404_	ATP_O2B_404_	_	_	_	TRANSPORTPROTEIN
TQ	28	-1	5	0.17	2.20	1R0X	ATP_O2G_4_	ATP_O2B_4_	HOH_O_378_	_	_	TRANSPORTPROTEIN

Papers by the candidate

Hsin, K., Y. Sheng, et al. (2008). "MESPEUS: a database of the geometry of metal sites in proteins." Journal of Applied Crystallography **41**(5): 963-968.

Taylor P., E Blackburn et al. (2007). "Ligand discovery and virtual screening using the program LIDAEUS." British Journal of Pharmacology **153**: S55-S67.

Journal of
Applied Crystallography

ISSN 0021-8898

Received 17 May 2008
Accepted 1 August 2008

MESPEUS: a database of the geometry of metal sites in proteins

K. Hsin, Y. Sheng, M. M. Harding,* P. Taylor and M. D. Walkinshaw

Centre for Translational and Chemical Biology, University of Edinburgh, Michael Swann Building, Mayfield Road, Edinburgh EH9 3JR, UK. Correspondence e-mail: marjorie.harding@ed.ac.uk

A database with details of the geometry of metal sites in proteins has been set up. The data are derived from metalloprotein structures that are in the Protein Data Bank [PDB; Berman, Henrick, Nakamura & Markley (2006). *Nucleic Acids Res.* **35**, D301–D303] and have been determined at 2.5 Å resolution or better. The database contains all contacts within the crystal asymmetric unit considered to be chemical bonds to any of the metals Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu or Zn. The stored information includes PDB code, crystal data, resolution of structure determination, refinement program and *R* factor, protein class (from PDB header), contact distances, atom names of metal and interacting atoms as they appear in the PDB, site occupancies, *B* values, coordination numbers, information on coordination shapes, and metal–metal distances. This may be accessed by SQL queries, or by a user-friendly web interface which searches for contacts between specified types of atoms [for example Ca and carboxylate O of aspartate, Co and imidazole Nδ of histidine] or which delivers details of all the metal sites in a specified protein. The web interface allows graphical display of the metal site, on its own or within the whole protein molecule, and may be accessed at http://eduliss.bch.ed.ac.uk/MESPEUS/. Some applications are briefly described, including a study of the characteristics of Mg sites that bind adenosine triphosphate, the derivation of an average Mg—$O_{phosphate}$ distance and some problems that arise when average bond distances with high precision are required.

1. Introduction

Metal atoms occur in many proteins and may be essential for catalytic function, or for the maintenance of structure, or for as yet unidentified reasons. Metal sites can be well characterized by X-ray crystal structure determination. In the relational database MESPEUS (metal sites in proteins at Edinburgh University), we have assembled geometric information for ten biologically common metals from files in the Protein Data Bank (PDB; Berman *et al.*, 2006). The database now includes sites in crystal structures determined at resolution 2.5 Å or better, by diffraction methods, for the metals Na, Mg, K, Ca, Mn, Fe, Co, Ni, Cu and Zn, and in the PDB at 1 January 2007. There are 34 896 metal sites in 10 919 structures; nucleic acid structures are included, as well as proteins. Geometric information can be useful (i) in the interpretation and fitting of models to electron-density maps, (ii) in the validation of structures or in restrained structure refinement when only low-resolution diffraction data are available, and (iii) for defining a 'docking pocket' for virtual screening in structure-based drug design (see Taylor *et al.*, 2008). Furthermore, when a metal site has been identified in a new structure, related sites in known structures can be compared quickly. Na and K sites are included because, although their interactions with protein are better described as electrostatic interactions than as bonds, knowledge of the geometry of the sites can be useful. When they occur in metalloproteins, the metals V, Mo and W are normally present in anionic forms, such as MoO_4^{3-}, and often in association with cofactors; their interaction with the protein structure and its amino-acid side chains is quite different from that of the metals that are present as cations, and so they are not included in the database.

Previous work has extracted characteristic bond distances and metal-site geometry from the PDB (Harding, 2000, 2001, 2006) and made comparisons with small molecule structures in the Cambridge Structural Database (CSD; Allen, 2002). These publications did not give access to the individual observations. MESPEUS now contains these observations from structures in the PDB at 1 January 2007. Castagnetto *et al.* (2002) described a different database, Metallo-Scripps MDB, also available on the Web and concerned with metals in protein structures, but new material does not appear to have been added to this since 2004.

For each metal atom, interactions with surrounding atoms which are within chemically bonding distances were extracted using the program *MP* described by Harding (2000); this distance is the 'target' distance, already characterized (Harding, 2006), plus a tolerance, 0.75 Å, to allow for coordinate errors in structure determination. These 'donor' atoms in the metal coordination group are often O, N or S from functional groups in the amino-acid side chains of the protein, or of the main chain carbonyl group, but they may also be from water molecules, substrate analogues or other non-protein molecules present in the crystal. Thus, for each metal site, a coordination number and a description of the geometry could be derived and stored (see the next section for details of the information stored).

The web interface for this MESPEUS database provides the possibility of searching for different combinations of metal and donor atom and retrieving the distances. The mean distance (and standard deviation) can be evaluated; the maximum crystal structure resolution used can be restricted; individual metal sites can be examined by graphical display, with or without the surrounding protein structure.

2. Methodology

2.1. Details of information stored

Some information has been extracted directly from PDB files and some with the program *MP* (Harding, 2000, where a fuller description of some of the terms can be found). For each protein or nucleic acid structure in the PDB, determined by diffraction methods at a resolution of 2.5 Å or better, and containing one or more metal atoms, the PDB code is stored, together with the name of the protein, the class of protein (HEADER in PDB file), the data resolution (Å), the space group and cell dimensions, the refinement program used, the *R* factor, and $R_{\rm free}$. Within each structure, the names of all metal and donor atoms are stored as they appear in the PDB file; for metal atoms the coordination number, *B* value and information on the shape of the site (*e.g.* the average deviation from tetrahedral or square planar for sites with coordination number 4) are stored, and for donor atoms their distances from the metal atom, their occupancies and *B* values. The r.m.s. difference between the actual metal donor atom distances and the target values is stored, as an indicator of the quality of the geometry at each metal site. Metal coordination numbers are stored, as evaluated by the program *MP* from the PDB file. Symmetry-related atoms are not stored in the PDB and no attempt has been made in this version of the MESPEUS database to generate them; occasionally the metal coordination group should include atoms in neighbouring symmetry-related units, *e.g.* when the metal atom lies on a two-, three- or fourfold symmetry axis in the crystal, and in these cases the value found for the coordination number will be too low. Also stored are the E.C. numbers for enzymes (where given in the PDB), metal–metal distances, indicators for bidentate carboxylate groups, angles between bonds at the metal sites, and error information, *e.g.* the presence of disorder at the site. All the information can be accessed by SQL queries [Structured Query Language, see http://dev.mysql.com/doc/refman/5.0/en/ or Chamberlin *et al.* (1976)]. Further details of the tables are available at http://eduliss.bch.ed.ac.uk/MESPEUS/query_SQL.jsp.

2.2. The construction of the database

The Fortran program *MP* (Harding, 2001) examined each PDB file in turn and gave a log file summarizing the geometric data for each metal site found. Using Perl scripts these data were written to the database tables. Information not given in the log files (*e.g.* structure refinement program used, *R* factors) was extracted directly from the PDB files. In cases where two alternative positions for a side chain have been detected in the refinement and indicated by *A* and *B* as the last character of the atom names, the *B* atoms (with lower occupancy) have been removed and an error flag set for the associated metal atom. About one-tenth of the metal sites are affected in this way. (At other metal sites where a number of low-occupancy donor atoms are reported there may also be disorder.)

2.3. Construction of the web interface

The web interface of MESPEUS was established utilizing a Java web-based approach. In order to enable changes in functionality and for ease of maintenance, the web site was constructed using Model-View-Controller (MVC) software architecture. MVC separates the data model, user phase and control logic into three individual components so that modifications to one component can be made with minimal impact to the others. The web interface allows the client to set series of query criteria to access the MESPEUS database without requiring any knowledge of SQL. It displays the accessible information of metal coordination groups and displays the individual metal site with distances, angles and coordination geometry.

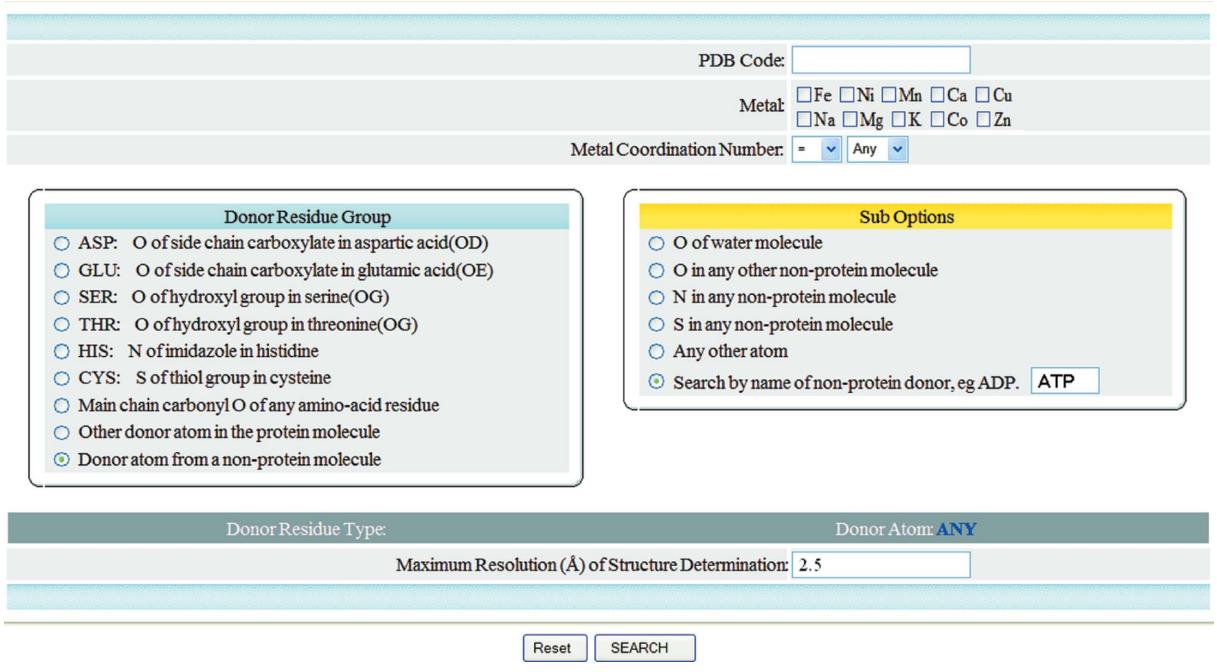


Figure 1
The search screen of the web interface, filled in for query about Mg links to adenosine triphosphate (ATP).

Table 1
Numbers of metal sites in the MESPEUS database.

Note that a significant number of metal atoms are listed in PDB files which do not appear to be within chemical bonding distance (target distance plus 0.75 Å) of any appropriate atoms; in many cases there are not even any appropriate atoms within 3.6 Å. This does not make chemical sense and can only be regarded as incomplete structure determination; these metal atoms are not included in the database.

Metal	(1) All sites in database	(2) Sites with metal and donor occupancy = 1.0	(3) Sites from column (2) with metal–protein interactions	(4) Sites from column (2) without metal–protein interactions
Na	2372	2069	1768	301
Mg	5561	4953	3676	1277
K	1424	1259	1108	151
Ca	7120	6425	6018	407
Mn	2118	1810	1688	122
Fe	7763	7226	7003	223
Co	637	506	356	150
Ni	534	398	366	32
Cu	1327	1105	1088	17
Zn	6031	5073	4979	94
Total	34 896	30 824	28 050	2774

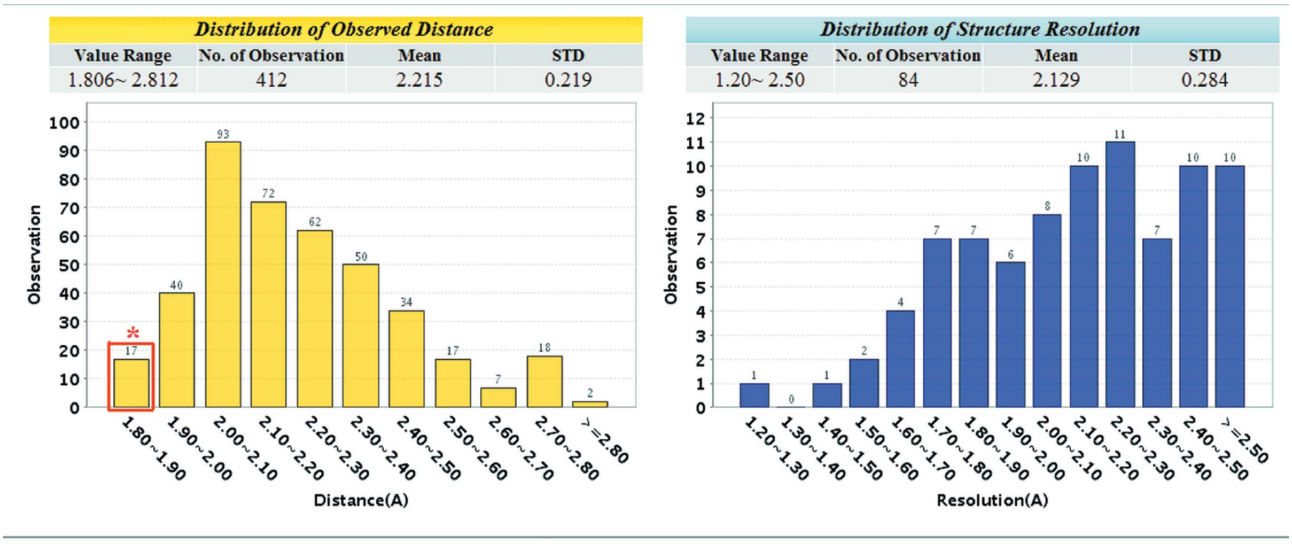
3. Contents of database

At present, the database contains material from 10 919 structures in the PDB, with numbers of metal sites as shown in Table 1. Some metal or donor atoms are listed with low occupancy, and in some cases there were pairs of disordered sites only one of which has been retained. Coordinate errors can be large when atom positions have low occupancy, and the presence of disordered sites gives meaningless results for coordination numbers. These sites can be identified in the database and excluded from searches, leaving the numbers shown in column (2) of Table 1 (and except where stated otherwise all the applications below use fully occupied atom positions). Not all the metal sites interact directly with protein. The number of these is particularly large for Mg [see Table 1, column (4)]. Some are simply hydrated ions such as $Mg(OH_2)_6^{2+}$; others interact only with ligand molecules, or with RNA or DNA (about 300 of the PDB files used for our database describe RNA or DNA structures without protein).

When more than one copy of the protein molecule with metal site(s) is present in the crystal asymmetric unit, all are listed in the database – for metal to donor atom distances they should represent separate observations. For this reason the number of distinct metal sites is only *circa* 60% of the total number of sites listed. Furthermore, the PDB contains many groups of very similar proteins within which the metal sites are similar or identical.

3.1. The web interface

The MESPEUS web interface has been designed to allow straightforward searching for particular kinds of interactions, with the possibility of selecting only higher-resolution structures, as shown in Fig. 1. Alternatively, a PDB code can be given to yield all the metal sites in that protein. The first result is a list of all the metal–donor



Download Query Result
Number of Hits: 412

No.	Distance	Coordination Number	Shape	Metal Name	Donor Residue Name	PDB	Reslu.	r.m.s.d.(Å)	Difference	Delete
1	2.370	6	δ oct(16.1°) **	MG 534 A	O2G ATP A 535	1A49	2.10	0.156	0.300	☐
2	2.011	5	δ tbp(21.7°) or δ tetp(12.1°)	MG 536 A	O1G ATP A 535	***		0.151	-0.059	☐
3	2.330	5	δ tbp(21.7°) or δ tetp(12.1°)	_	O2B ATP A 535	_			0.260	☐
4	2.013	5	δ tbp(21.7°) or δ tetp(12.1°)	_	O1A ATP A 535	_			-0.057	☐
5	2.420	5	δ tbp(27.3°) or δ tetp(32.1°)	MG 1734 C	O2G ATP C 1735	_		0.172	0.350	☐
6	2.081	6	δ oct(11.4°)	MG 1736 C	O1G ATP C 1735	_		0.088	0.011	☐
7	2.202	6	δ oct(11.4°)	_	O2B ATP C 1735	_			0.132	☐
8	1.920	6	δ oct(11.4°)	_	O1A ATP C 1735	_			-0.150	☐
9	2.272	6	δ oct(10.8°)	MG 2334 D	O2G ATP D 2335	_		0.110	0.202	☐

Figure 2
Results of the search for Mg linked to ATP; this yielded 412 examples in 84 proteins or DNA/RNA structures. Average Mg—O distances and their distributions can be displayed. *: Clicking here selects this distance range only. **: Links to a page like Fig. 3 showing details of the specified metal site. ***: Links to a page like Fig. 4 showing the whole metal-containing protein.

atom distances satisfying the search query, together with information on coordination number of the metal, shape, atom names *etc.* as shown in Fig. 2. Mean distances and distributions can be obtained, or the list may be downloaded for other use. Any metal site selected can be displayed graphically and its position in the whole protein structure shown (Figs. 3 and 4).

4. Applications of database and web interface

4.1. The importance of near atomic resolution in deriving mean bond distances

We have been concerned to derive the best average values for distances from metal atoms to different kinds of donor atoms in structures deposited in the PDB, for use in validation, electron-density map interpretation, model fitting or restrained refinement. The average bond distance can be evaluated from all the data in the database, or from the higher-resolution structure determinations only. We and others (*e.g.* Meyer Klaucke, 2007) have noticed that often the average found for a particular distance, for example Zn—N of histidine with coordination number 4, shown in Table 2, is slightly smaller when only high-resolution structure determination results are used than when all results are used. The standard deviation (s.d.), representing the scatter of results, is, as expected, smaller too.

The most precise averages, *i.e.* those with smallest standard deviation, are those for the highest-resolution results, better than 1.3 Å at least, *i.e.* near atomic resolution; these agree best with the CSD average for equivalent interactions, 2.00 (2) Å for 25 observations (Harding, 2006). Fig. 5 shows the wide scatter of reported distances from poor-resolution structures; it is only at fairly high resolution that there is good consistency of observations. On chemical grounds very little variation in this particular distance is expected (say < 0.02 Å); nearly all the variation above must be due to coordinate errors remaining after structure refinement. (Other kinds of distance, *e.g.* M—$O_{carboxylate}$, can show more chemical variation.) It is strongly recommended that for such averages only the highest-resolution results practicable be used, consistent with a reasonable number of observations.

Table 2
Effect of resolution on mean distance found for Zn—N of histidine, with Zn coordination number 4 (Nδ and Nε of histidine are both included; if they are kept separate, no significant difference in their means is found).

Maximum resolution (Å)	No. of observations	Mean (s.d.) (Å)
2.5	3583	2.10 (13)
1.8	1007	2.07 (9)
1.5	261	2.06 (6)
1.3	88	2.03 (4)
1.1	21	2.02 (4)

4.2. Application: a survey of interactions between Mg and adenosine triphosphate (ATP)

Phosphate–magnesium interactions are crucial in many enzyme mechanisms. There is a particularly rich source of structural infor-

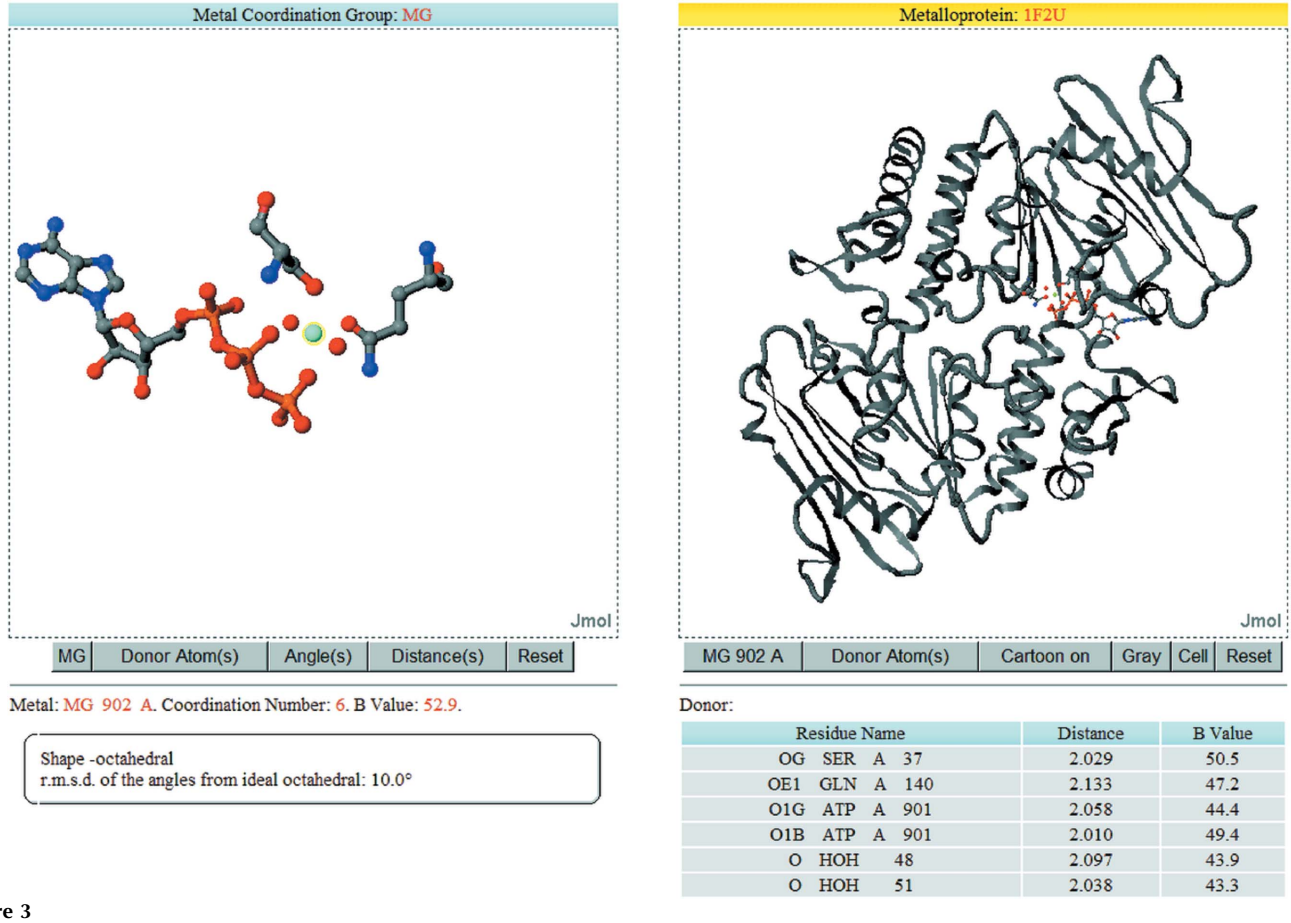
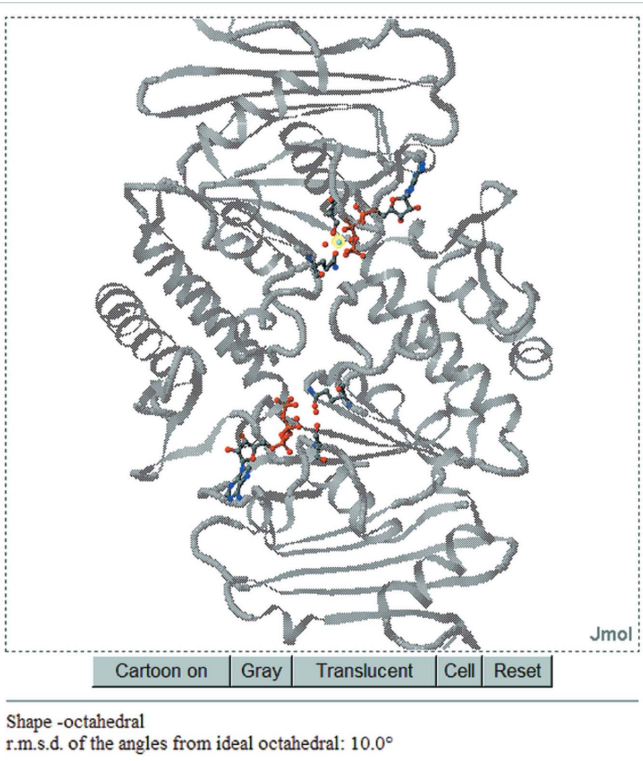


Figure 3
By clicking on the metal name in the results shown in Fig. 2, the metal site can be displayed by *Jmol* (http://jmol.sourceforge.net/), as well as its relation to the whole protein molecule. Here, the metal, Mg (cyan coloured), is coordinated by the O atoms from Ser and Gln, two phosphate O atoms (one β, one γ) of ATP, and two water molecules. The buttons below the pictures allow users to centre and rotate the view and add other information.



PDB ID	1F2U (More detail from PDB web-site)
Title	CRYSTAL STRUCTURE OF RAD50 ABC-ATPASE
Class	REPLICATION
Resolution	1.6
Space Group	P 21 21 21
Cell	a: 79.9 b: 82.6 c: 106.0 α: 90.0 β: 90.0 γ: 90.0
Refinement Program	CNS
R Value	0.24
Free R	0.28
No. Metal Sites	2

No.	Metal Name (Coordination No.)	Donor Residue Name (Occupancy)	Distance	Metal / Donor B Value
1	MG 902 A (6)	OG SER A 37 (1.0)	2.029	52.9 / 50.5
	*	OE1 GLN A 140 (1.0)	2.133	52.9 / 47.2
		O1G ATP A 901 (1.0)	2.058	52.9 / 44.4
	**	O1B ATP A 901 (1.0)	2.010	52.9 / 49.4
		O HOH 48 (1.0)	2.097	52.9 / 43.9
		O HOH 51 (1.0)	2.038	52.9 / 43.3
2	MG 902 C (6)	OG SER C 37 (1.0)	1.957	43.9 / 36.8
		OE1 GLN C 140 (1.0)	2.074	43.9 / 41.8
		O1G ATP C 901 (1.0)	2.145	43.9 / 37.6
		O1B ATP C 901 (1.0)	2.014	43.9 / 42.6
		O HOH 49 (1.0)	2.105	43.9 / 40.1
		O HOH 50 (1.0)	2.039	43.9 / 40.6

Figure 4
By clicking on the protein name in the results shown in Fig. 2, information about the protein is tabulated, as well as details of each metal site; the whole protein molecule is displayed by *Jmol*. Here the protein has two chains, *A* and *C*, with one Mg–ATP site on each. *: A dynamic link to centre this metal site in the view on the left. **: Links to a page like Fig. 3 showing the specified metal site.

mation available in the PDB for metal–ATP interactions and this survey provides an overview of the interaction geometries found from the database for protein–Mg–ATP complexes. The web interface can find all links from Mg to O of ATP. For the detailed analysis, it was more convenient to use one SQL query to find all Mg with ATP links, and a second to pick up all the other atoms, protein or otherwise, linked to each Mg; Perl scripts were then used to present the results more conveniently. This yielded 136 Mg sites from 46 proteins; transferases were the commonest protein types. After removal of identical sites within crystal asymmetric units, 66 sites remained and are listed, together with some statistics, in Table D (deposited[1]). In these, Mg is normally linked to protein through one or two amino-acid side chains, most commonly Asp, Glu, Asn, Gln, Ser and Thr, but there are two examples with three links to protein, and three with links to main chain carbonyl oxygen. ATP may be linked to Mg through one, two or all three phosphate groups; the commonest pattern (23 examples) is linkage through the β- and γ-phosphate groups, but all other possibilities occur, although there are only two cases where linkage is through two O atoms of the same phosphate group. The Mg coordination group may also include one to four water molecules, and occasionally another small molecule like oxalate (as a bidentate ion). More than half have total coordination number 6, the expected coordination number for Mg. A surprisingly large number, 20, appear to have coordination number ≤ 4; this suggests that the reliability of some of the data is questionable – some of the analyses may be incomplete, or some of the atoms identified as Mg may in fact be water molecules. Figs. 3 and 4 show an example of an Mg site with ATP. Seven observations in structures at resolution ≤ 1.5 Å give a mean Mg$-$O$_{\rm phosphate}$ distance of 2.05 (7) Å, but for a reliable average more observations are desirable.

4.3. Application: Mg interactions with other phosphates

To establish an Mg$-$O$_{\rm phosphate}$ distance from a larger number of higher-resolution observations, it was necessary to include other phosphate ligands. Identifying phosphates from the atom names in the PDB is not straightforward, but by selecting with SQL queries O atoms whose residue names (het group names, see 'het groups' section of PDBSUM, http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/) include the letter P and whose atom names include one of the letters A, B or G, or O atoms whose atom names include the letter P, over 2000 Mg$-$O$_{\rm phosphate}$ links were found, 75 of which are from

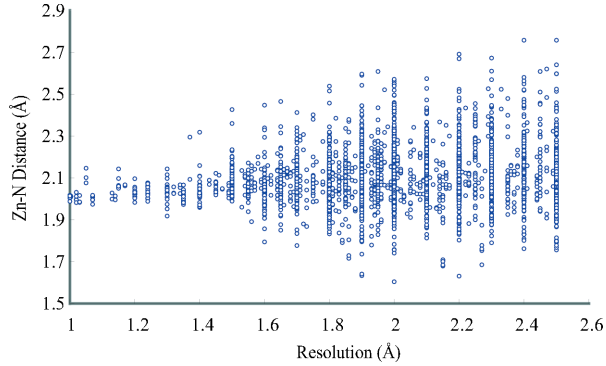


Figure 5
Reported values of Zn $-$N$_{\rm his}$ distances for Zn coordination number 4, as a function of resolution.

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: KK5027). Services for accessing these data are described at the back of the journal.

structures with resolution 1.3 Å or better. These have an average Mg—O distance of 2.06 (8) Å, in good agreement with the value above [and also close to the target distance given for $Mg—O_{carboxylate}$ (Harding, 2006)]. The commonest links found are to adenosine diphosphate (ADP; 453 links to Mg), ATP (410), atrial natriuretic peptide (ANP; 299), and guanosine disphosphate (GDP; 219), imidotriphosphate (GNP; 195) and triphosphate (GTP; 113). It is again clear that linkage through two O atoms of one $–PO_4–$ group is very rare, and that for the triphosphates ATP, GTP and CTP (cytidine) the linkage is often bidentate, through the β- and γ-phosphate groups, not the α group.

4.4. Application: Mg compared with Mn and Co

DNA polymerases, RNAses, transposases and integrases all require one or two Mg ions as an integral part of their enzymatic mechanism (Steitz & Steitz, 1993). The effect of substituting Mn or Co for Mg frequently results in modified activities, either partial or complete inhibition of activity or in some cases enhancement of activity and broadening of substrate specificity (Frank & Woodgate, 2007). It may be possible to relate at least some of these differences in enzyme activity to differences in coordination geometry of the bound metal in the active site. Mg and Co divalent ions are very similar in size, while Mn^{2+} is significantly larger, *e.g.* Mg—O and Co—O are ∼2.07 Å while Mn—O is ∼2.17 Å. They have somewhat different preferences for amino-acid donors; Co and Mn are much less likely than Mg to interact with serine, threonine or main chain carbonyl oxygen, and more likely to interact with histidine [see http://tanna.bch.ed.ac.uk/, item 4, '..frequency of occurrence..', or Dokmanić *et al.* (2008)]. Here, coordination numbers (Table 3) and shapes have been compared in structures with resolution 1.8 Å or better. For each metal, more than half of the metal sites have coordination number 6, and the next most important coordination number is 5. Coordination numbers higher than 6 are almost always associated with multidentate ligands, and coordination numbers less than 4 or 5 may represent incomplete structure determinations, or cases where additional donor atoms are present but are in neighbouring asymmetric units not listed in the PDB files, or, particularly in the case of Mg, unreliable data in the PDB, because of the difficulty of locating Mg or distinguishing it from O of water. When the coordination number is 6 the shape is octahedral, but the average distortion from ideal octahedral is considerably less for Mg, 7 (5)°, and Co, 8 (8)°, than for Mn, 11 (6)° [see Harding (2000) for a discussion of distortions]. When the coordination number is 5 the ideal shapes may be described as trigonal bipyramidal or tetragonal pyramidal; real metal sites are often between these two, and the description is based on which ideal version it is least distorted from. For Mg the tetragonal pyramid is strongly favoured (more than 95% of observations) with average distortion 8 (6)°; for Mn and Co there are near equal numbers of each shape with greater distortions. (All this information can be gathered from repeated queries of the web interface and some manual averaging, or it can be found very efficiently with SQL queries of the database.)

4.5. Application: listing of metal coordination groups

These lists were described by Harding (2004). Metal coordination groups (metal sites) were categorized by the sequence of amino-acid donors and the relative positions of these amino acids in the polypeptide chain, using one-letter codes for the amino acids, and the differences between the successive residue numbers. For example, CHCC Zn 2 18 3 describes a coordination group in which Zn is coordinated to the thiolate S of Cys(n), where n is the amino-acid residue number, an imidazole N of His(n + 2), and the thiolate S atoms of Cys(n + 2 + 18) and Cys(n + 2 + 18 + 3). Coordination number and information on other non-protein donors present are also given. The new database has allowed the preparation of new lists including all metal sites in the PDB at January 2007. For each metal atom in the database, the full names of all its donor atoms were extracted with an SQL query; Perl scripts were used to convert these to one-letter amino-acid codes, check that all donor atoms have site occupancy > 0.7, work out sequence separations *etc.*, as well as remove duplicate coordination groups within the asymmetric unit of any one crystal. The lists are available at http://tanna.bch.ed.ac.uk (item 3, 'New lists..'). They allow quick identification of other proteins with coordination groups identical or similar to a specified one, and could provide a resource for the study of patterns of metal coordination and protein evolution. A direct link to this database is planned.

Table 3
Coordination numbers found for Mg, Mn and Co in structures determined at resolution 1.8 Å or better.

	1	2	3	4	5	6	7	8	9
Mg	42	84	91	101	195	956	25	2	1
Mn	3	6	8	26	78	246	19	3	
Co	2	5	3	11	52	131			

5. Conclusions, summary

The database, with its user-friendly web interface, allows immediate identification and display of the metal sites in any specified protein whose structure is in the PDB, determined at resolution 2.5 Å or better. Alternatively, the web interface can identify all metal sites with a particular kind of contact, giving distance, coordination number and atom names for each, as well as average distances and distribution of distances. Furthermore, as shown in several applications, SQL queries to the database can extract additional information about the interactions of different metals with proteins.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2006). *Nucleic Acids Res.* **35**, D301–D303.
Castagnetto, J. M., Hennessy, S. W., Roberts, V. A., Getzoff, E. D., Tainer, J. A. & Pique, M. E. (2002). *Nucleic Acids Res.* **30**, 379–382.
Chamberlin, D. D., Astrahan, M. M., Eswaran, K. P., Griffiths, P. P., Lorie, R. A., Mehl, J. W., Reisner, P. & Wade, B. W. (1976). *IBM J. Res. Dev.* **20**, 560–575.
Dokmanić, I., Šikić, M. & Tomić, S. (2008). *Acta Cryst.* D**64**, 257–263.
Frank, E. G. & Woodgate, R. (2007). *J. Biol. Chem.* **282**, 24689–24696.
Harding, M. M. (2000). *Acta Cryst.* D**56**, 857–867.
Harding, M. M. (2001). *Acta Cryst.* D**57**, 401–411.
Harding, M. M. (2004). *Acta Cryst.* D**60**, 849–859.
Harding, M. M. (2006). *Acta Cryst.* D**62**, 678–682.
Meyer Klaucke, W. (2007). Personal communication.
Steitz, T. A. & Steitz, J. A. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 6498–6502.
Taylor, P., Blackburn, E., Sheng, Y. G., Harding, S., Hsin, K. Y., Kan, D., Shave, S. & Walkinshaw, M. D. (2008). *Br. J. Pharmacol.* **153**, S55–S67.

British Journal of Pharmacology (2008) 153, S55–S67
 0007–1188/08 $30.00
www.brjpharmacol.org

REVIEW

Ligand discovery and virtual screening using the program LIDAEUS

P Taylor, E Blackburn, YG Sheng, S Harding, K-Y Hsin, D Kan, S Shave and MD Walkinshaw

The Centre for Translational and Chemical Biology, The University of Edinburgh, Michael Swann Building, King's Buildings, Mayfield Road, Edinburgh, UK

This paper discusses advances in docking and scoring approaches with examples from the high-throughput virtual screening program LIDAEUS. We describe the discovery of small molecule inhibitors for the immunophilin CypA, the cyclin-dependent kinase CDK2 and the cyclapolin series of potent Polo-like kinase inhibitors. These results are discussed in the context of advances in massively parallel computing and in the development of annotated databases.
British Journal of Pharmacology (2008) **153,** S55–S67; doi:10.1038/sj.bjp.0707532; published online 26 November 2007

Keywords: virtual screening; LIDAEUS; EDULISS; cyclophilin; cyclin-dependent kinase; Polo-like kinase

Abbreviations: CDK, cyclin-dependent kinase; CypA, human cyclophilin-A; CLogP, the octanol-water partition coefficient, calculated using the Biobyte program (http://biobyte.com.index.html) developed by Hansch and Leo; EDULISS, EDinburgh University LIgand Selection System; LIDAEUS, LIgand Discovery at Edinburgh UniverSity; MlogP, the octanol-water partition coefficient calculated as described by Moriguchi *et al*. (1992); RMSD, root mean square distance

Virtual screening overview: tools and approaches

Ligand discovery can be regarded as a simple matching problem: we would like to find a small molecule (ligand) with the appropriate shape and charge properties to bind effectively to a target protein of interest. High-throughput screening (HTS) provides one possible experimental route to a solution and libraries consisting of over 1 million compounds can be tested in days. Computational screening provides a complementary approach and with massively parallel processing, millions of compounds per week can be tested. Estimates of the number of potential small molecule drug-like compounds vary between 10^{18} and beyond 10^{63} (Lipinski and Hopkins, 2004). Consequently, for any specific target protein, even if the results from each assay and each docking run were totally reliable (which is not the case), it would still be impossible to test binding for every potential ligand. The commonly accepted Lipinski criteria (Lipinski *et al*., 1997) for orally active drug-like molecules set physicochemical property limits to increase the probability of good drug bioavailability. Drug-like molecules are expected to have a molecular weight (MW) $\leqslant$500 Da, $\leqslant$5 hydrogen bond donors (HBDs), $\leqslant$10 hydrogen bond acceptors (HBAs) and a CLog P $\leqslant$5 (the octanol-water partition coefficient calculated as described by Moriguchi *et al*. (1992) (MLogP) $\leqslant$4.15). More stringent criteria have been proposed for initial searches. For example, Lead likeness restricts MW to $<$350 Da and CLogP (the octanol-water partition coefficient, calculated using the Biobyte program (http://biobyte.com.index.html) developed by Hansch and Leo) to $<$3 (Teague *et al*., 1999). Even these more stringent cutoffs do little to reduce the astronomical numbers of potential ligands and exploring such a large-scale-matching problem will require imaginative computational and experimental approaches.

Protein targets

Recent reviews have attempted to estimate the number of druggable proteins in the Protein Data Bank (PDB) (Berman *et al*., 2000). Druggable proteins have structural features that facilitate binding to drug-like molecules. For proteins to progress from intrinsic druggability to becoming a target requires drug binding to modulate the biological role of the protein and to bring about therapeutic benefit (Fishman and Porter, 2005). Currently available literature identifies 1300 studied protein drug targets from humans and infective

Correspondence: Professor MD Walkinshaw, The Centre for Translational and Chemical Biology, The University of Edinburgh, Michael Swann Building, King's Buildings, Mayfield Road, Edinburgh, EH9 3JR, UK.
E-mail: m.walkinshaw@ed.ac.uk
Received 29 July 2007; revised 27 September 2007; accepted 4 October 2007; published online 26 November 2007

organisms (Hopkins and Groom, 2002; Russ and Lampel, 2005; Zheng *et al*., 2006). Estimates of the total number of druggable targets in the human genome have been made based on the number of disease genes; these give a total of up to 1500 targets out of 25 000 in the human genome (Hopkins and Groom, 2002). Bacterial and viral proteins also provide targets; published estimates of the number of targets from infective organisms are well over 1000. This suggests that there should be a pool of about 3000 drug targets in total (Zheng *et al*., 2006).

Of the 1300 currently studied targets, 44% are classified as enzymes, the most populated biochemical class. A total of 557 enzymes are current research targets and 134 have proved to be successful targets. Enzymes represent 50% of all successful targets (Zheng *et al*., 2006). A total of 280 research targets have experimentally determined structures with a specific drug-binding domain (represented by 107-folds), mainly by X-ray crystallography.

Within the PDB, there are about 250 uniquely different (that is $<10\%$ amino acid identity) well-determined structures in complex with 'peptide-ligands'. These represent a subset of protein–protein interactions where the interaction is controlled by a linear peptide on one side of the interface. Table 1 shows some examples of protein–peptide interactions. This group possibly represents the most druggable subset of protein–protein interactions. Short linear peptides are more amenable to replacement by small molecule mimetics. Modulating protein–protein interactions is particularly attractive due to the pivotal role of such interactions in cell signal transduction pathways and cell cycle progression (Fry and Vassilev, 2005; Chene, 2006).

There are a number of publicly accessible sources of protein–ligand binding affinities and web-based tools designed to aid the extraction of information from databases containing structural information on protein targets. For example, the BindingDB is a public, web-accessible database of measured binding affinities for biomolecules and contains data generated by isothermal titration calorimetry and enzyme inhibition (http://www.bindingdb.org/). The Relibase database (http://relibase.ebi.ac.uk/) is a web-based tool for the study of protein–ligand interaction. MSDmotif (http://www.ebi.ac.uk/msd-srv/msdmotif/) provides a tool for summarizing structural information on a database of over 6000 protein–ligand complexes found in the PDB.

Small molecule databases

A number of publicly available small molecule databases have been established over the last few years. The ligand.Info database (http://ligand.info) (Grotthuss *et al*., 2004) is a compilation of a number of publicly available databases providing a Meta-Database of over 1 million entries with calculated three-dimensional (3D) structures and some information about biological activity. The ZINC database (http://blaster.docking.org/zinc/) contains over 4.6 million commercially available compounds in various 2D and 3D formats (Irwin and Shoichet, 2005). Only compounds with MW $\leqslant$700 Da, and calculated LogP values between –4 and 6 are stored. Simple Lipinski filters or other discreet subsets of compounds can be selected.

An ambitious project financed by the National Institutes of Health has the goal of discovering sets of molecules that will specifically modulate the activities of the majority of gene product in the human and other organisms. Fast expanding databases are now being developed that contain results from a number of high-throughput screens, many of which use a set of over 100 000 chemically diverse molecules. These data are available at NCBI's database of small organic molecules at http://www.ncbi.nlm.nih.gov/sites/entrez?db=pcassay.

EDULISS, the EDinburgh University Ligand Selection System, is our in-house relational database that stores over 5 million available compounds, containing data from over 25 chemical catalogues. Of the 5.3 million compounds, 3.8 million are unique. 3D coordinates for each molecule are stored with over 1500 topological, geometric, physicochemical and toxicological descriptors per compound (Todeschini and Consonni, 2005). The descriptors can be used interactively to select subgroups of the database and also to provide profiling information. One approach to identify unique compounds is to compare the chemical graph of each compound with the graph of every other. This approach is extremely computationally expensive. An alternative method for identifying unique compounds in EDULISS's large collection has been developed. A small number of descriptors including a 3D-Wiener index, an electronegativity descriptor and a polarizability descriptor are used to group compounds. The resulting small groups of molecules with identical descriptors can then be compared using a

Table 1 Examples of protein-peptide interactions in the Protein Data Bank

PDB	*Protein*	*Peptide*	*Peptide sequence*	*Function*	*Reference*
1YCR	MDM2	p53	SQETFSDLWKLLPEN	Antitumour	(Vassilev *et al*., 2004)
1BXL (NMR)	Bcl-XL	Bak-BH3	GQVQRQLAIIGDDINR	Apoptosis	(Degterev *et al*., 2001)
1EBA	EPO	EPOR	GGTXSCHFGPLTWVCKPQGG	Anaemia	(Qureshi *et al*., 1999)
1EJ4, 1WKW	EiF4e	EiF4e-BP	RIIYDRKFLMECRN	Malignant transformation	(de and Graff, 2004)
1AXC	PCNA	p21	GRKRRQTSMTDFYHSKRRLIFS	Antitumour	(Gulbis *et al*., 1996)
1CKA	c-CRK	C3G	PPPALPPKKR	Oncogene	(Wu *et al*., 1995)
1GUX	Rb tumour suppressor	E7 peptide	DLYCYEQLN	Antitumour	(Lee *et al*., 1998)
1H9O	SH2	Penta -peptide	XVPML	Signal transduction, cancer	(Pauptit *et al*., 2001)
1QZ2	FKBP52	Hsp90	MEEVD	Steroid signalling pathways	(Wu *et al*., 2004)
1ELR	HOP	Hsp90	XMEEVD	Signalling pathways	(Scheufler *et al*., 2000)
1YVH	c-CBL	APS	GRARAVENQXSFY	Oncogene	(Hu and Hubbard, 2005)
1G3F (NMR)	XIAP-Bir3	Smac	AVPIAQKSE	Apoptosis	(Liu *et al*., 2000)

graph-matching program. A web-based interface for EDULISS has been developed; this provides a convenient way of extracting families of compounds with a user-defined set of properties.

Database profiling and compound selection

The EDULISS database comprises 25 different commercial and other smaller specialist compound collections. Of these, some 4.3 million fit the Lipinski 'rule of 5s' (Lipinski *et al*., 1997). A total of 3.2 million fit the Oprea lead-like criteria (Hann and Oprea, 2004). The more stringent Astex Rule of 3 is met by 230 000 compounds (Carr *et al*., 2005) (statistical profiles of some general descriptors are shown in Figure 1, descriptor ranges are shown in Table 2). A study by Oprea *et al*. (2007) investigated recent trends in the property space of leads, drugs and chemical probes. Leads are generally smaller, less complex and have lower LogP than drugs, due to the inevitable modifications involved in the medicinal chemistry optimization process.

It is desirable for a set of compounds for docking or assay to be selected considering both protein target and screening methodology. Solubility is of key importance for both bioavailability and 'screenability'. Experimentally derived aqueous solubility data are not available for the majority of compounds in the EDULISS database. Algorithms for predicting aqueous solubility from structure almost universally rely on a directly proportional relationship between LogP and solubility (Jorgensen and Duffy, 2002; Delaney, 2005). It might be appropriate to have a relaxed solubility requirement (MLogP $\leqslant$4.21) and to include relatively large compounds ($\leqslant$450 Da) with the aim of finding leads of high affinity and high specificity for the target. A greater range of molecular complexity may be explored with a higher MW cutoff (Schuffenhauer *et al*., 2006). However, the application of X-ray crystallography in lead discovery has different property requirements. The technique identifies fragments binding to significant regions of the target protein and then employs fragment growing or linking strategies to improve potency. Fragments are small molecules, 100–250 Da, with few functional groups (Rees *et al*., 2004; Carr *et al*., 2005; Hartshorn *et al*., 2005). In these techniques, virtual hit ligands are soaked into crystals. Relative protein concentrations are high, necessitating high ligand concentrations. Solubility problems can be compounded by the practice of soaking with a fragment cocktail to increase assay throughput. Virtual screening subsets designed for fragment screens

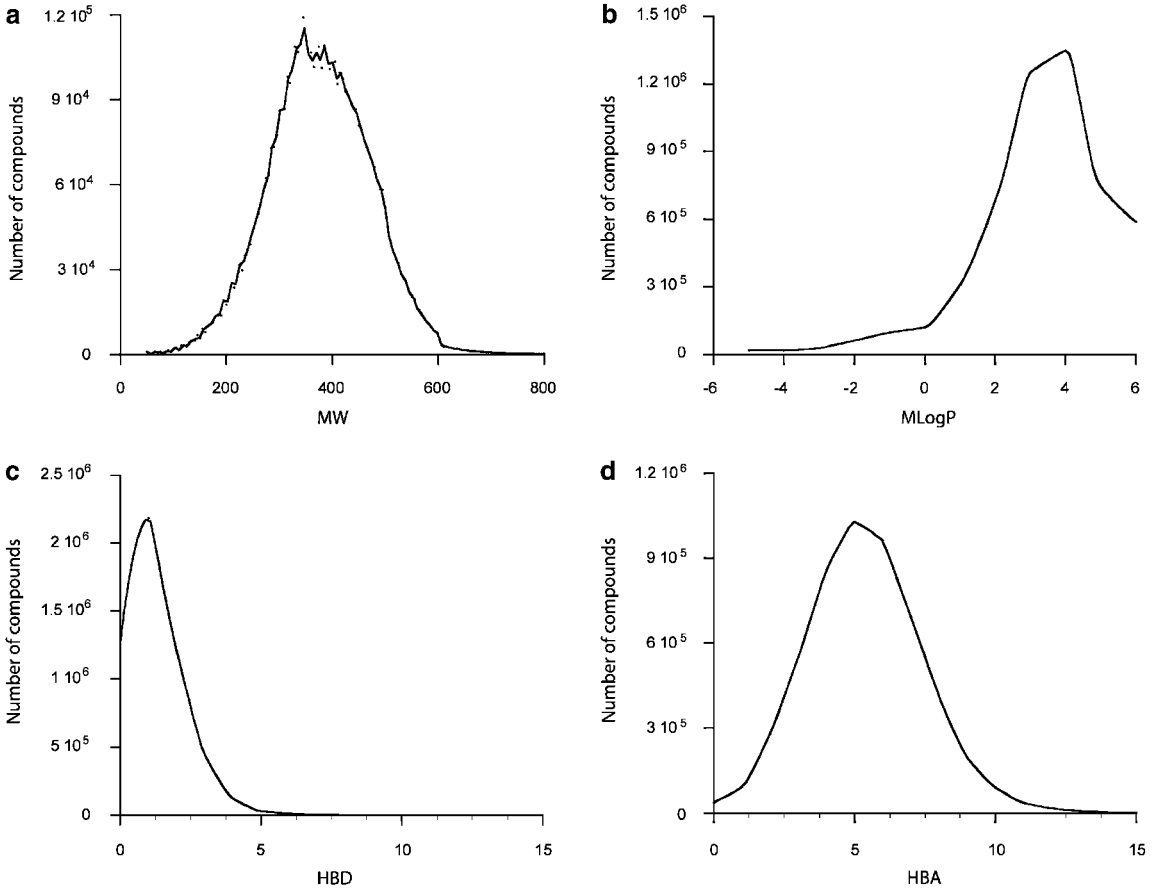


Figure 1 Molecular property profiles of 5.3 million compounds in the EDULISS database. (**a**) MW. (**b**) MLogP. (**c**) Number of HBDs. (**d**) Number of HBAs. Bin sizes for MWs are 5 Da and for MLogP, the number of HBDs and HBAs 1 U. EDULISS, EDinburgh University LIgand Selection System; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; MLogP, the octanol-water partition coefficient calculated as described by Moriguchi *et al*. (1992); MW, molecular weight

Table 2 Descriptor ranges for the EDULISS database of 5.3 million compounds

Descriptor	*Max*	*Average*	*Standard deviation*
Molecule weight	2180.59	374.28	95.77
Number of bonds	306	47.05	13.05
Number of aromatic bonds	69	13.42	5.97
Number of rings	18	3.14	1.18
Sum of atomic van der Waals volumes[a] (Å^3)	172.82	29.29	7.71
Number of rotatable bonds	77	5.31	2.73
MLogP	134.05	3.29	3.35
Topological polar surface area[b] (Å^2)	932.34	73.66	42.23

Abbreviations: EDULISS; LIgand Discovery at Edinburgh UniverSity; MLogP, the octanol-water partition coefficient calculated as described by Moriguchi (Moriguchi *et al*., 1992).
[a]Scaled on Carbon atom.
[b]Using N, O, S, P polar contributions (Todeschini and Consonni, 2005).

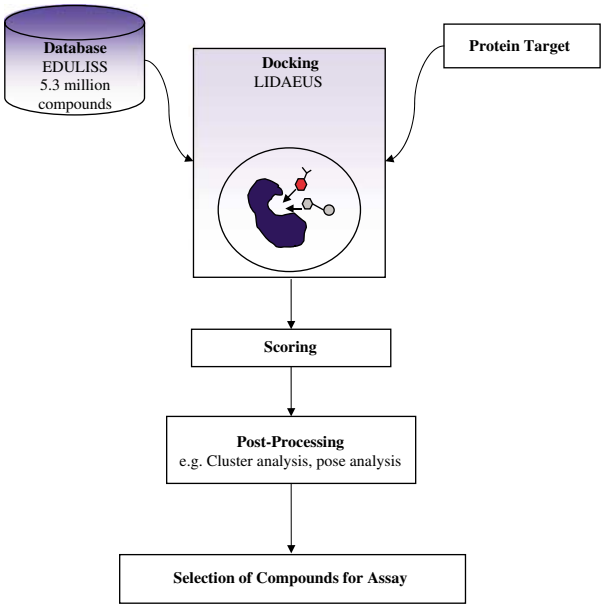


Figure 2 Steps involved in virtual screening using LIDAEUS. LIDAEUS, LIgand Discovery at Edinburgh UniverSity.

need to have stringent solubility requirements, MLogP $\leqslant 3.0$, while containing diverse scaffolds decorated with a broad range of functional groups (Moriguchi *et al*., 1992).

High-throughput virtual screening

High-throughput virtual screening achieves a high throughput of test ligands by using simplified non-quantum mechanical methods without the inclusion of complex molecular dynamics (Woo and Roux, 2005). Typically, the virtual screening process follows the steps outlined in Figure 2. A ligand is selected and positioned into the target protein-binding pocket in a given 'pose' (Muegge and Martin, 1999). The resultant complex is scored on the basis of intermolecular contacts to give a predicted strength of binding interactions (Woo and Roux, 2005). Flexible docking typically allows sampling of ligand and sometimes protein conformations during the docking procedure. Rigid body docking is however much less computationally expensive. Exploring the conformers of relatively simple molecules containing only three or four rotatable bonds (using a broad step size) requires the generation of over 200 starting conformations to be sampled in order to fully consider the majority conformational space (Guner *et al*., 2004). The most widely used flexible docking tools are GOLD (Genetic Optimization for Ligand Docking) (Jones *et al*., 1997), FlexX (Rarey *et al*., 1996; Kramer *et al*., 1999), DOCK (Ewing *et al*., 2001), AutoDock (Goodsell *et al*., 1996), Glide (Friesner *et al*., 2004; Halgren *et al*., 2004) and ICM (Internal Coordinate Mechanics) (Abagyan *et al*., 1994). A variety of different methods are used by the above tools to deal with ligand flexibility such as genetic algorithms, incremental construction, simulated annealing and Monte Carlo methods (Rosenfeld *et al*., 1995; Vieth *et al*., 1998). The diversity exhibited by scoring functions has been used in consensus scoring is implemented in, for example, X-SCORE (Wang *et al*., 2003). Using different but well-performing scoring functions, the accuracy of consensus methods can be greater than any individual scheme (Bissantz *et al*., 2000; Stahl and Rarey, 2001; Jacobsson *et al*., 2003; Raymond *et al*., 2004; Xing *et al*., 2004; Feher, 2006). However, 'artificial enrichment' is a potential pitfall, with scoring functions selected to perform well on a specific protein–ligand complex (Verdonk *et al*., 2004).

Perola *et al*. (2004) have reported that energy minimization can significantly improve the accuracy of docking poses found by GOLD (Jones *et al*., 1997) and ICM (Abagyan *et al*., 1994) programs. Our results also showed that there is better agreement between the docked pose and the crystallographic pose using rigid body refinement. A 'good fit' is defined as an root mean square distance (RMSD) $\leqslant 2$ Å between corresponding heavy atoms of the X-ray structure and the docked ligand pose.

Virtual screening has proved successful in a number of projects (Alvarez, 2004; Kitchen *et al*., 2004; Oprea and Matter, 2004; Ghosh *et al*., 2006) (Table 3). One of the major future challenges is to develop virtual screening methods capable of identifying ligands that will interrupt protein–protein interactions.

LIDAEUS as a tool for high-throughput virtual screening

LIDAEUS, LIgand Discovery at Edinburgh UniverSity, our in-house high-throughput virtual screening program (Wu *et al*., 2003) generates a grid of site points in the binding pocket of the target protein. Each site point is coloured: HBA, HBD or hydrophobic, depending on the preferred protein interaction (Figure 3).

Each molecule selected from the small molecule database is placed in the binding pocket and atoms of the docked molecule are matched to site points. An exhaustive fit of a given number of atoms from the docked molecule onto the site points is undertaken to identify reasonable poses. These are stored for subsequent rigid body energy minimization.

Table 3 Examples of hits from virtual screening experiments

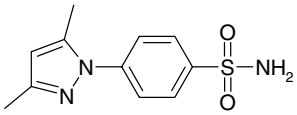
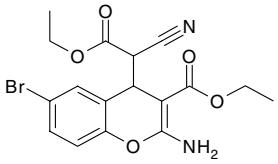
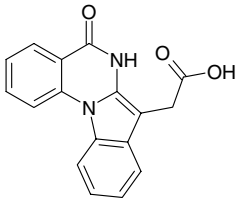
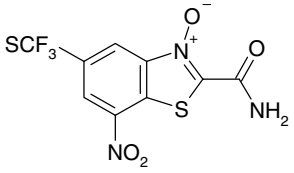
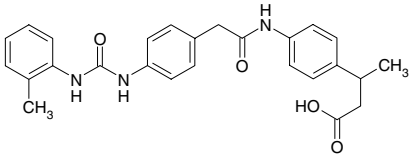
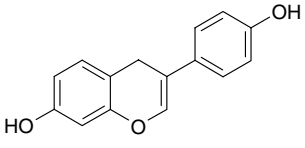
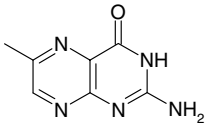
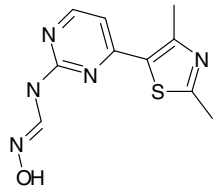
Target	*Virtual screening*	*Example structure**	*Reference*	*Assay*
Carbonic anhydrase II	FlexX (Rarey *et al.*, 1996)		(Gruneberg *et al.*, 2002)	IC_{50} Sub nanomolar
Bcl-2	DOCK (Ewing *et al.*, 2001)		(Wang *et al.*, 2000; Enyedy *et al.*, 2001)	IC_{50} 9 μM
CK2	DOCK (Ewing *et al.*, 2001)	* 	(Vangrevelinghe *et al.*, 2003)	IC_{50} 80 nM
Plk1	LIDAEUS (Wu *et al.*, 2003)		(McInnes *et al.*, 2006)	IC_{50} 20 nM
GPCR	Five targets 5-HT1A, 5-HT4, Dopamine D2, NK1, and CCR3	Compound PRX-93009 scored best for 5-HT1A, no structure shown	(Becker *et al.*, 2004)	Ki 1.0 nM
Integrin α4β1	Catalyst (Greene *et al.*, 1994)	* 	(Singh *et al.*, 2002)	IC_{50} 1.3 nM
ERβ	GOLD 2.0 (Jones *et al.*, 1997)		(Zhao and Brinton, 2005)	IC_{50} 0.68 μM
TGT	Unity/FlexX (Rarey *et al.*, 1996)		(Brenk *et al.*, 2003)	IC_{50} 0.25 μM

Table 3 *Continued*

Target	*Virtual screening*	*Example structure**	*Reference*	*Assay*
CDK2	LIDAEUS (Wu *et al*., 2003)		(Wu *et al*., 2003)	IC_{50} 2.2 μM

Structures marked with an asterisk do not represent those initially identified by *in silico* screening. Minor chemical modifications have been made and from these compounds the experimental data determined.

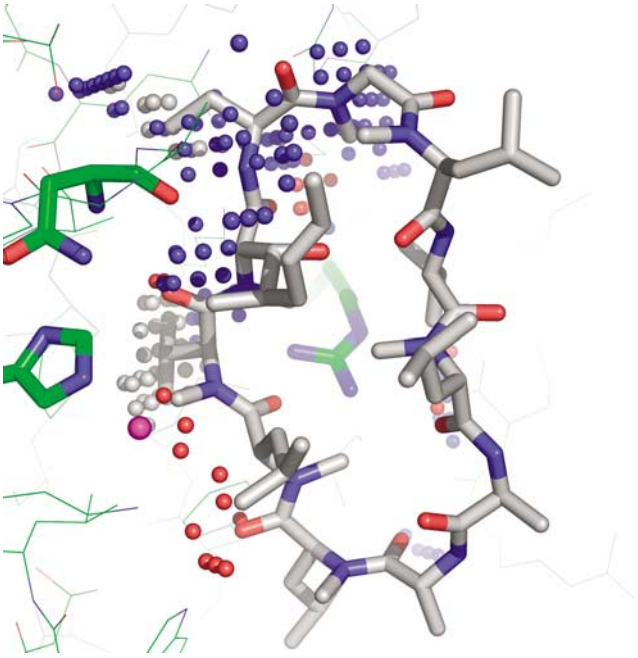


Figure 3 LIDAEUS site points in the binding pocket of CypA in complex with cyclosporine-A (PDB code 1cwa). Each site point is coloured depending on the preferred protein interaction (HBA, red; HBD, blue; hydrophobic, grey). The magenta sphere represents a key water molecule. Key residues are shown in bold. CypA, human cyclophilin-A; HBA, hydrogen bond acceptor; HBD, hydrogen bond donor; LIDAEUS, LIgand Discovery at Edinburgh UniverSity; PDB, Protein Data Bank.

There are various tunable parameters that influence how thoroughly the binding pocket is explored and hence the time required to dock a series of compounds. The precision with which an atom matches the site point, called 'resolution' in this program, is usually set between 0.02 and 0.06 Å and plays an important role in determining the number of allowed starting poses. Resolution values greater than 0.06 Å lead to an exponential growth in the number of starting poses. Increasing the number and density of the site points has a similar effect of dramatically increasing the number of allowed starting poses.

In LIDAEUS, there are two built-in scoring functions, a force field-based energy function and pose interaction profile (PIP) a knowledge-based function (Kan, 2007). The energy function is essentially a linear combination of van der Waals and hydrogen bonding energies. The geometry-dependent hydrogen bonding term incorporates salt bridges and obviates the need for calculating hydrogen atom positions. The program assigns fixed formal charges to identify ionizable groups.

The PIP score uses a protein interaction profile that can be assigned for specific regions of the binding pocket where explicit types of ligand interactions, for example, a particular hydrogen bond, are required. The PIP string is a hexadecimal code containing information about the interactions made between a given set of protein residues and the docked ligand. The target PIP string is usually based on a known X-ray crystal structure in which key features of the protein–ligand binding interaction have been identified. It is a very efficient process to match and score the bit strings of the target interactions against those calculated for the docked ligand pose. Thus, the PIP score can be used as a component of the final score to ensure that docked ligands have both favourable energies of interaction and satisfy specific interactions in their pose.

While LIDAEUS is broadly similar in function to many docking programs, it differs in two major areas: the extent to which the fitting protocol can be modified by the user and the modularity of the system. All definitions within the program in the way of atom and site point types are soft, that is, can be customized by the user. This happens at two levels: one can initially type individual atoms according to connectivity criteria and then group many or one of these types into colours used in the pose generation or scoring process. While using the default definitions allows for standard searches using atoms grouped into broad classes such as hydrophobic, HBA and HBD, it is possible to add specific restrictions.

LIDAEUS exists as a series of modules that run as a UNIX pipeline, so that initial typing of molecules, posing, scoring and sorting are all separate programs. It allows us to easily develop experimental modules and test different scoring methods. The program is being developed in two areas. The current flexible docking module is too slow to be used in a high-throughput mode and this is being addressed. Secondly, a front end is being written that allows intermediate users the ability to easily configure customizable features.

The examples discussed in the paper were run on a modest seven-node cluster. Run times are dependent on the target protein, the ligand complexity and the site point resolution set for LIDAEUS. However, representative times for a database of 50 000 ligands would be 6 h. Using an IBM Blue Gene/L supercomputer, run times have been reduced from 8 days on a seven-node cluster to 62 min on 1024 processors using a standard data set of 1.67 million small molecules.

Validation of LIDAEUS docking and scoring performance

The immunophilin proteins FKBP (FK506 Binding Protein) and human cyclophilin-A (CypA) have been used as test systems to develop and test the results from the database mining program LIDAEUS. Despite having similar biological profiles, the structure and active site of the two proteins are very different. Both proteins have peptidyl-prolyl isomerase activity and speed up the *cis–trans* equilibration of proline residues by lowering the barrier to rotation about the imide bond (Fischer *et al.*, 1993). Inhibition of the enzymatic turnover of an immunophilin substrate provides a functional assay for screening potential inhibitors (Fischer *et al.*, 1984). X-ray crystallographic, surface plasmon resonance, isothermal titration calorimetry and mass spectrometry results provide complementary techniques for characterizing ligand binding.

A set of nine chemically related ligands of human CypA with IC_{50} values between 2 and 100 μM were used to test LIDAEUS docking performance (referred to as the test set). For each ligand, the X-ray structure is known and the RMSD between corresponding atoms of the X-ray structure and the ligand structure is used as a measure of fit. Correct docking poses (RMSD $\leqslant 2$ Å) of ligand 1 in the test set (Figure 4 shows the correlation of PIP score and energy score were E, with RMSD from X-ray structure for a ligand in the test set) were all scored >0.93 by the PIP function and their energy scores, $E < -11$ kcal mol^{-1}. PIP scores are normalized between 0 and 1: a high PIP score indicates conserved interactions between those in the X-ray structure and the docked pose. The other ligands in the CypA test set have similar results, showing that a combined total score of energy function and PIP (matching a defined pose iinteraction profile) ranks the correct docking binding mode higher than alternative poses (Equation 1).

$$S = W_1 E + W_2 \sum_i PIP_i \quad (1)$$

S, total score; *E*, force field-based energy score; *PIP*, knowledge-based PIP score; W_1, weighting factor specific to protein system; W_2, weighting factor specific to protein system.

For a set of nine related cyclophilin inhibitors, the effect of changing the weights W_1 and W_2 was examined by systematically trialling different values. For this series of compounds, the values that gave the best RMSD fit of the docked pose compared to the crystallographic pose were weights of 1 and -40 for W_1 and W_2, respectively. These values proved useful in identifying a new series of cyclophilin ligands (Kan, 2007).

The role of water in accurate docking

It has been reported in several studies that water-mediated protein–ligand interactions are an important factor in the docking process. Ligands can displace water in the active site or incorporate them as an extension of the protein surface. The presence of key water molecules can significantly improve docking performance (Pospisil *et al.*, 2002; Yang and Chen, 2004). Our results show that eight out of nine ligands in the test set were correctly docked into near-native positions by LIDAEUS, while re-docking of six of them was significantly improved when key water molecules were included in the protein–ligand binding system. The presence of the key waters enables the LIDAEUS program to identify several important interactions involved in the complex and construct the significant HBA or HBD site points at the binding atom locations. (Key waters are those that form bridging H bonds to both protein and ligand molecules). Ligand 7 (Figure 5) is an example where including water improved docking performance. The presence of the key waters enables the LIDAEUS program to construct the significant HBA or HBD site points at the binding atom

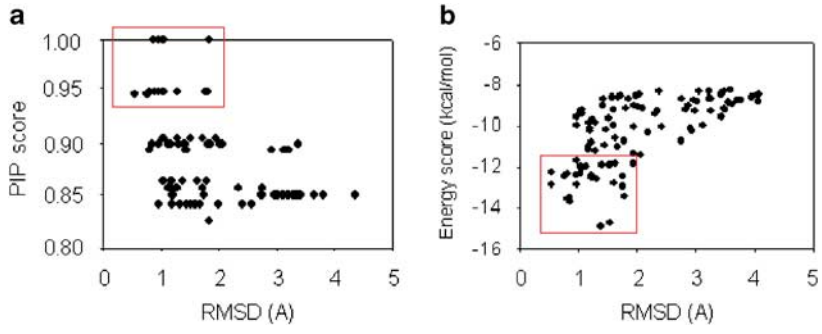


Figure 4 Correlation of PIP score and energy score, E, with RMSD from X-ray structure for ligand 1 of the CypA test set. (**a**) Plot of PIP score against RMSD of docking poses with respect to X-ray structure. (**b**) Plot of energy score, E, against RMSD with respect to X-ray structure. Red boxes highlight 'good poses' that meet scoring cutoffs: PIP score >0.93, energy score <-11 kcal mol^{-1}. CypA, human cyclophilin-A; PIP, pose interaction profile; RMSD, root mean square distance.

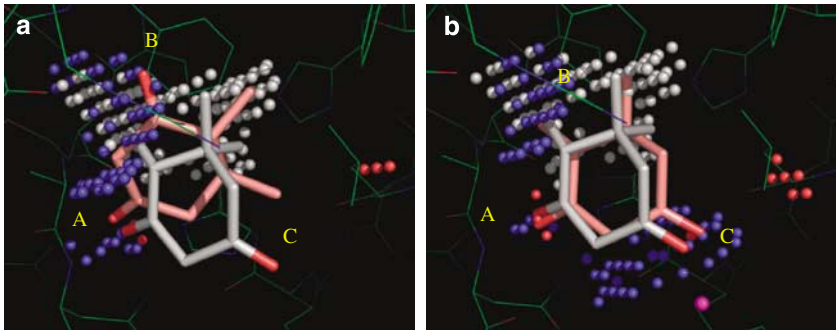


Figure 5 Including water in site point generation. Poses were improved when an essential water molecule was used in the calculation of energy maps and in site point construction. X-ray structure of the ligand is shown in white and docked poses in pink. (**a**) Illustrates how one oxygen atom (O1) from the ligand was put into the experimental position (position A), but the other oxygen atom (O2) was put into position B instead of position C, as revealed from the X-ray structure. (**b**) When the important water molecule (in magenta) is included in site point generation, competent site points set helped bring the oxygen atom to position C.

locations. Furthermore, energy maps contoured with the presence of key waters better represent the energy distribution in the local area. Thus, those maps would give the correct fits lower energy scores than maps generated without key waters.

Discovery of ligands for immunophilins

In a test to find CypA ligands, the ZINC database (Irwin and Shoichet, 2005) of 2 million compounds was used as an input for a LIDAEUS screen looking for compounds that would match site points in the active site of CypA (two parallel runs using 60 site points with a resolution of 0.06 Å and 170 site points with a resolution of 0.04 Å). The top 2000 poses were re-ranked, specifying that hydrophobic interactions with Phe113 and a hydrogen bond to Arg55 were satisfied, using PIP scoring (Figure 6).

The combined energy and PIP scores ranged from −164 to −80 (arbitrary units). The top 360 unique compounds were grouped according to chemical similarity (using molecular fingerprinting and Tanimoto coefficients) and binding mode (visually using Pymol). From this analysis, 14 compounds (all chemically distinct from known cyclophilin inhibitors) were purchased and tested for inhibition and binding by peptidyl-prolyl isomerase assay (Kofron *et al*., 1991). Eleven compounds showed a statistically significant reduction in peptidyl-prolyl isomerase activity. Six of the 14 compounds were 'hits' in the peptidyl-prolyl isomerase enzymatic assay; they inhibited CypA with IC_{50} values ranging from 27 to 135 μM. Subsequent isothermal titration calorimetry studies for the best three compounds gave K_d values of 2 to 8 μM.

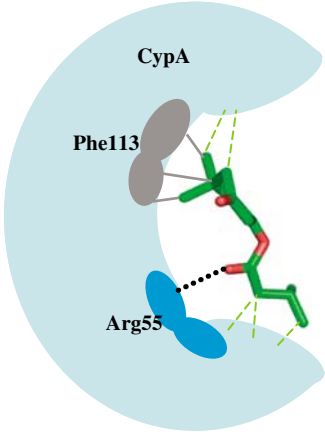


Figure 6 PIP used in the CypA virtual screening experiment. The top 2000 poses (rank ordered using the energy score, E,) were re-ranked, specifying that there were predicted hydrophobic interactions with Phe113 (grey lines) and a predicted hydrogen bond to Arg55 (black dotted lines) using PIP scoring. The diagram shows a molecule making interactions specified in the PIP. The green dashed lines are non-covalent interactions not specified in the PIP. CypA, human cyclophilin-A; PIP, pose interaction profile;

Virtual screening for CDK inhibitors using LIDAEUS

Cyclin-dependent kinases (CDKs) are key regulators in all steps of the cell cycle and as such are interesting targets for anticancer therapies. There are already a number of clinical trials underway with CDK2 and CDK4 inhibitors for a range of cancers (Collins and Garrett, 2005). The small molecule inhibitors, roscovotine (Seliciclib) and flavopiridol, are CDK2 inhibitors and show promising activity in lung cancer. These drugs target the ATP binding site of the CDKs. This is a problem in the design of CDK selective drugs, as all nine CDKs show some homology and most of the active site residues are well conserved. Another complicating factor in the design of specific inhibitors is that the active form of the kinase is induced by complex formation with a partner cyclin and phosphorylation of a specific threonine residue located on the T-loop of the kinase. These events cause subtle changes in active site geometry, which may be important for inhibitor design.

We used LIDAEUS to carry out a virtual screen of 50 000 commercially available compounds from the Maybridge catalogue (www.maybridge.com) docked into the active site of CDK2 (taken from the X-ray structure of the CDK2–staurosporine complex; PDB code 1AQ1). The predicted top 120 poses based on the docking score were screened at a fixed concentration of 30 μM using an assay to monitor the

inhibition of phosphorylation by CDK2/cyclinE. Twenty-nine percent of the compounds were classed as active by showing more than 30% inhibition. The most active four compounds all had a heteroaryl-2-amino-pyrimidine core and measured IC_{50} values between 0.9 and 17 μM (Table 4). X-ray crystal structures of the four hits were obtained and each was clearly identified in the ATP binding site. A comparison was made of the calculated docked pose (without any PIP influence) against the experimentally determined ligand structures. The four ligands were all found to dock in twisted conformations with a twist of 35° around the bond between the two aromatic rings. The RMSD atom against atom fit of the three top scoring docked ligands versus the experimental structure were 1.6, 1.58 and 3.42 Å with scores of −24, −23 and −20 kcal mol^{-1}, respectively. Despite the chemical similarity of these four ligands, they adopt different binding modes (Table 4) CYC1 and CYC2 form identical hydrogen-bond interactions to ATP: NH….O (Glu81), N….HN(Leu83) and CH….O(Leu83). When the amine group is substituted as in CYC3 and CYC4, the ATP binding mode is precluded and the ligand flips over to allow the bulky substituent to point out of the pocket. An alternative hydrogen bonding pattern is made CH…O (Glu81), N….HN(Leu83) and NH….O(Leu83). These four structures provided an excellent starting point for the design of chemical modifications. Over 40 related structures have been synthesized to optimize *in vivo* potency. The tightest binding ligand of this series, an amino derivative (CYC5), has a $K_i = 2$ nM and was shown to induce cell death in cultured HeLa cells (Wang *et al*., 2004a, b)

The importance of fine tuning a template structure in virtual screening

The shape and surface of the target pocket is clearly one of the most important factors in successful virtual screening runs. The search for CDK2-specific inhibitors highlighted the importance of understanding the biological role of the target protein. A crystallographic study was used to analyse the structures of six inhibitor ligands belonging to the thiazole-pyrimidine class, identified by LIDAEUS; both in complex with monomeric CDK2 and also with the binary CDK2/cyclinA active complex (Wu *et al*., 2003; Kontopidis *et al*., 2006). The activation of CDK2 by phosphorylation and cyclin binding causes significant loop and helix movements but leaves the shape of the ATP binding site relatively unchanged with a maximum side-chain movement between 1 and 2 Å for residues comprising this pocket. However, these small differences in pocket shape play a major role in the relative binding strengths of inhibitors. In some cases, the same ligands can adopt significantly different poses in the monomeric and active complexes. Binding enthalpies of the ligands have been estimated based on calculated van der Waals and hydrogen bond contacts measured in the crystal. The measured IC_{50} values correlate well with the calculated interaction energy (energy score) for the binary complex, but show poor correlation with the inactive complex. This fits with the way the assay has been carried out—using the active complex. It also suggests that the enthalpic energy-scoring scheme, using van der Waals and hydrogen bonding terms, provides a self-consistent measure of binding strength (Kontopidis *et al*., 2006).

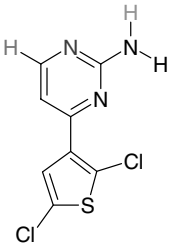
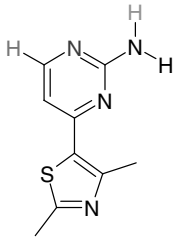
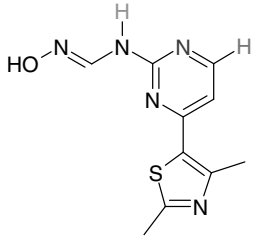
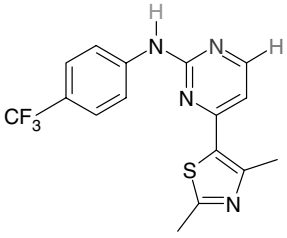
The discovery of cyclapolin, a potent Polo-like kinase inhibitor

Polo-like kinase 1 (Plk1) controls the G2/M transition of the cell cycle by phosphorylating a number of substrates that function in mitotic progression. Overexpression of Plk1 is frequently observed in tumours and in downregulation, using small interfering RNA, has been shown to inhibit cancer cell proliferation (McInnes *et al*., 2005). Small molecule Plk-specific inhibitors are valuable biological tools and can be used as leads for antitumour agents. A number of general kinase inhibitors, such as staurosporine and purvalanol, are known to inhibit Plk1 (McInnes *et al*., 2005). After years of intensive effort by academic and pharmaceutical research groups, the X-ray structure of the kinase domain has recently been published (Kothe *et al*., 2007) in complex with a pyrazole inhibitor (PHA-680626), which has an IC_{50} value of 0.5 μM (PDB code 2owb). Before this structure became available, we had developed a homology model of the kinase domain of Plk based on the staurosporine-bound conformation of protein kinase A, which has a 31% sequence identity (PDB code 1stc). The model was shown to be consistent with the known structure-activity properties of a series of ligands which were docked into the binding pocket in a similar manner to that found in CDK2. LIDAEUS was used to dock a library of 200 000 commercially available compounds into the modelled active site of Plk1. A total of 350 of the top-ranked compounds were then assayed by measuring inhibition of Plk1 phosphorylation of Cdc25C. A number of Plk1 inhibitors were identified with potencies ranging between 0.5 and 20 μM. A series of compounds (named the cyclapolins) based on the benziathole N-oxide core of the most active hit were synthesized and provide a consistent structure-activity relationships for the inhibition of Plk1 (Figure 7). The most active compound in this series showed significant improvement in potency and has an IC_{50} value of 2 nM. For this series, there is a good correlation between the docking score and potency. Treatment of HeLa cells with cyclapolin1 leads to mitotic cells that show severe spindle abnormalities (McInnes *et al*., 2006).

Outlook

The evolution of structure-based lead discovery has been guided by fashion and by some interesting technological advances. Twenty-five years ago, we had the first useful molecular graphics systems that could help medicinal chemists visualize molecular properties. This technology along with fast data collection and structure determination of protein X-ray structures opened the path to structure-based methods. Ironically, in the mid 1990s, just as this approach was beginning to bear fruit, the fashion swung to robotics and high-throughput screening, possibly spurred by the newly founded discipline of Combinatorial Chemistry,

Table 4 CDK2 inhibitors discovered using LIDAEUS

Compound	*Kinase inhibition (CDK2/ cyclin E) IC_{50} (μM)*	*Hydrogen-bonding pattern*	*Reference*
CYC1 	17	ATP hydrogen-bonding pattern: NH….O(Glu81), 2.96 Å, N….HN(Leu83), 3.64 Å, CH….O(Leu83), 3.36 Å	(Wu *et al*., 2003)
CYC2 	13	ATP hydrogen-bonding pattern: NH….O(Glu81), 2.86 Å, N….HN(Leu83), 3.30 Å, CH….O(Leu83), 3.25 Å	(Wu *et al*., 2003)
CYC3 	2.2	Alternative hydrogen bonding pattern: CH….O(Glu81), 3.31 Å, N….HN(Leu83),2.82 Å, NH….O(Leu83), 2.54 Å Ligand flips over to allow the bulky substituent to point out of the pocket	(Wu *et al*., 2003)
CYC4 	0.9	Alternative hydrogen bonding pattern: CH….O(Glu81), 3.31 Å, N….HN(Leu83), 2.92 Å, NH….O(Leu83), 2.58 Å Ligand flips over to allow the bulky substituent to point out of the pocket	(Wu *et al*., 2003)

Colour coding denotes atoms involved in key hydrogen-bonding interactions.

which made it possible to generate very large libraries of compounds. Now in larger organizations high-throughput screening, *in silico* and structure-based approaches are quite well integrated.

The main challenges in docking are still the old problems of how to efficiently model effects including dielectrics, entropy, water and flexibility. Advances in computer architectures may help tackle such problems. However, we also need to design methods that allow efficient simulation of these effects. Possibilities include using cliques of side chain conformations around the active site, and hybrid molecular modelling/quantum mechanical calculations. High-throughput virtual screening, using simplified methods (non-quantum mechanical or complex molecular dynamics), can already achieve docking rates of over 1 M compounds an hour (Shave *et al*., 2008).

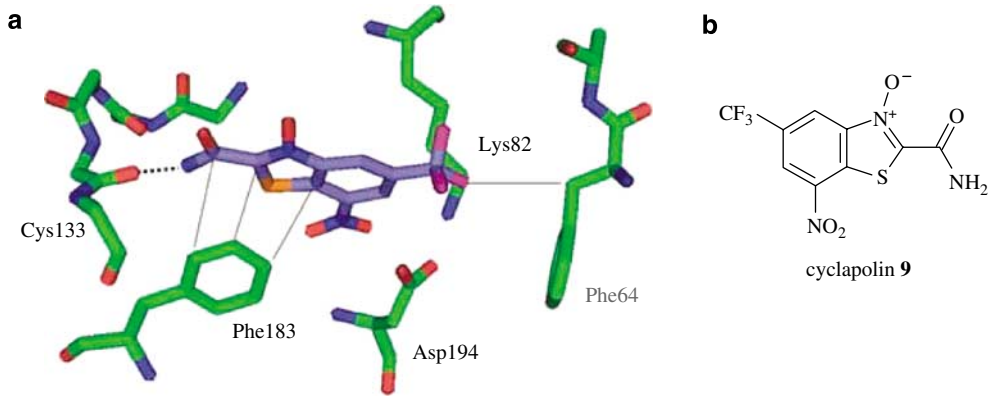


Figure 7 Cyclapolin 9 in the active site of Plk1. (**a**) Proposed binding mode of cyclapolin 9 in the ATP-binding site of Plk1, the black dotted line represents a hydrogen bonding interaction between cyclapolin and Cys133 and the grey lines hydrophobic interactions between cyclapolin and Phe183. Diagram reproduced from (McInnes *et al.*, 2006). (**b**) Cyclapolin 9 is the top hit from virtual screening, IC_{50} 500 nM. Plk1, Polo-like kinase 1.

Technical advances in miniaturization (396-well plates) and sensitive ligand-binding assays are already generating very large amounts of binding data, which contribute to structure-activity relationships. Accurate protein–ligand binding data can add to our understanding of how proteins recognize ligands. Identifying the key features of successful virtual screening calculations can only enhance the chances of discovering new ligands.

Acknowledgements

Elizabeth Blackburn acknowledges support from Organon, UK as part of a CASE studentship with the Biotechnology and Biological Sciences Research Council, UK (BBSRC). Simon Harding and Steven Shave acknowledge support from the BBSRC.

Conflict of interest

The authors state no conflict of interest.

References

Abagyan R, Totrov M, Kuznetsov D (1994). ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* **15**: 488–506.

Alvarez JC (2004). High-throughput docking as a source of novel drug leads. *Curr Opin Chem Biol* **8**: 365–370.

Becker OM, Marantz Y, Shacham S, Inbal B, Heifetz A, Kalid O *et al.* (2004). G protein-coupled receptors: *in silico* drug discovery in 3D. *Proc Natl Acad Sci USA* **101**: 11304–11309.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res* **28**: 235–242.

Bissantz C, Folkers G, Rognan D (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* **43**: 4759–4767.

Brenk R, Naerum L, Gradler U, Gerber HD, Garcia GA, Reuter K *et al.* (2003). Virtual screening for submicromolar leads of tRNA-guanine transglycosylase based on a new unexpected binding mode detected by crystal structure analysis. *J Med Chem* **46**: 1133–1143.

Carr RA, Congreve M, Murray CW, Rees DC (2005). Fragment-based lead discovery: leads by design. *Drug Discov Today* **10**: 987–992.

Chene P (2006). Drugs targeting protein–protein interactions. *ChemMedChem* **1**: 400–411.

Collins I, Garrett MD (2005). Targetting the cell division cycle in cancer. CDK and cell cycle checkpoint kinase inhibitors. *Curr Opin Pharmacol* **4**: 366–373.

de BA, Graff JR (2004). eIF-4E expression and its role in malignancies and metastases. *Oncogene* **23**: 3189–3199.

Degterev A, Lugovskoy A, Cardone M, Mulley B, Wagner G, Mitchison T *et al.* (2001). Identification of small-molecule inhibitors of interaction between the BH3 domain and Bcl-xL. *Nat Cell Biol* **3**: 173–182.

Delaney JS (2005). Predicting aqueous solubility from structure. *Drug Discov Today* **10**: 289–295.

Enyedy IJ, Ling Y, Nacro K, Tomita Y, Wu X, Cao Y *et al.* (2001). Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J Med Chem* **44**: 4313–4324.

Ewing TJ, Makino S, Skillman AG, Kuntz ID (2001). DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J Comput Aided Mol Des* **15**: 411–428.

Feher M (2006). Consensus scoring for protein–ligand interactions. *Drug Discov Today* **11**: 421–428.

Fischer G, Bang H, Mech C (1984). Determination of enzymatic catalysis for the *cis–trans*-isomerization of peptide binding in proline-containing peptides]. *Biomed Biochim Acta* **43**: 1101–1111.

Fischer S, Michnick S, Karplus M (1993). A mechanism for rotamase catalysis by the FK506 binding protein (FKBP). *Biochemistry (Moscow)* **32**: 13830–13837.

Fishman MC, Porter JA (2005). Pharmaceuticals: a new grammar for drug discovery. *Nature* **437**: 491–493.

Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT *et al.* (2004). Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **47**: 1739–1749.

Fry DC, Vassilev LT (2005). Targeting protein–protein interactions for cancer therapy. *J Mol Med* **83**: 955–963.

Ghosh S, Nie A, An J, Huang Z (2006). Structure-based virtual screening of chemical libraries for drug discovery. *Curr Opin Chem Biol* **10**: 194–202.

Goodsell DS, Morris GM, Olson AJ (1996). Automated docking of flexible ligands: applications of AutoDock. *J Comput Aided Mol Des* **9**: 1–5.

Greene J, Kahn S, Savoj H, Sprague P, Teig S (1994). Chemical function queries for 3D database search. *J Chem Inf Comput Sci* **34**: 1297–1308.

Grotthuss MV, Pas J, Koczyk G, Wyrwicz LS (2004). Ligand. Info small-molecule Meta-Database. *Comb Chem High Throughput Screen* **7**: 757–761.

Gruneberg S, Stubbs MT, Klebe G (2002). Successful virtual screening for novel inhibitors of human carbonic anhydrase: strategy and experimental confirmation. *J Med Chem* **45**: 3588–3602.

Gulbis JM, Kelman Z, Hurwitz J, O'Donnell M, Kuriyan J (1996). Structure of the C-terminal region of p21(WAF1/CIP1) complexed with human PCNA. *Cell* **87**: 297–306.

Guner O, Clement O, Kurogi Y (2004). Pharmacophore modeling and three dimensional database searching for drug design using catalyst: recent advances. *Curr Med Chem* **11**: 2991–3005.

Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT *et al*. (2004). Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem* **47**: 1750–1759.

Hann MM, Oprea TI (2004). Pursuing the leadlikeness concept in pharmaceutical research. *Curr Opin Chem Biol* **8**: 255–263.

Hartshorn MJ, Murray CW, Cleasby A, Frederickson M, Tickle IJ, Jhoti H (2005). Fragment-based lead discovery using X-ray crystallography. *J Med Chem* **48**: 403–413.

Hopkins AL, Groom CR (2002). The druggable genome. *Nat Rev Drug Discov* **1**: 727–730.

Hu J, Hubbard SR (2005). Structural characterization of a novel Cbl phosphotyrosine recognition motif in the APS family of adapter proteins. *J Biol Chem* **280**: 18943–18949.

Irwin JJ, Shoichet BK (2005). ZINC—a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**: 177–182.

Jacobsson M, Liden P, Stjernschantz E, Bostrom H, Norinder U (2003). Improving structure-based virtual screening by multivariate analysis of scoring data. *J Med Chem* **46**: 5781–5789.

Jones G, Willett P, Glen RC, Leach AR, Taylor R (1997). Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* **267**: 727–748.

Jorgensen WL, Duffy EM (2002). Prediction of drug solubility from structure. *Adv Drug Deliv Rev* **54**: 355–366.

Kan D (2007). Studies of protein-ligand interactions and the discovery of new cyclophilin inhibitors. PhD Thesis, University of Edinburgh, Edinburgh.

Kitchen DB, Decornez H, Furr JR, Bajorath J (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat Rev Drug Discov* **3**: 935–949.

Kofron JL, Kuzmic P, Kishore V, Colon-Bonilla E, Rich DH (1991). Determination of kinetic constants for peptidyl prolyl *cis–trans* isomerases by an improved spectrophotometric assay. *Biochemistry (Moscow)* **30**: 6127–6134.

Kontopidis G, McInnes C, Pandalaneni SR, McNae I, Gibson D, Mezna M *et al*. (2006). Differential binding of inhibitors to active and inactive CDK2 provides insights for drug design. *Chem Biol* **13**: 201–211.

Kothe M, Kohls D, Low S, Coli R, Cheng AC, Jacques SL *et al*. (2007). Structure of the catalytic domain of human polo-like kinase 1. *Biochemistry (Moscow)* **46**: 5960–5971.

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

Lee JO, Russo AA, Pavletich NP (1998). Structure of the retinoblastoma tumour-suppressor pocket domain bound to a peptide from HPV E7. *Nature* **391**: 859–865.

Lipinski C, Hopkins A (2004). Navigating chemical space for biology and medicine. *Nature* **432**: 855–861.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **23**: 3–25.

Liu Z, Sun C, Olejniczak ET, Meadows RP, Betz SF, Oost T *et al*. (2000). Structural basis for binding of Smac/DIABLO to the XIAP BIR3 domain. *Nature* **408**: 1004–1008.

McInnes C, Mazumdar A, Mezna M, Meades C, Midgley C, Scaerou F *et al*. (2006). Inhibitors of Polo-like kinase reveal roles in spindle-pole maintenance. *Nat Chem Biol* **2**: 608–617.

McInnes C, Mezna M, Fischer PM (2005). Progress in the discovery of polo-like kinase inhibitors. *Curr Top Med Chem* **5**: 181–197.

Moriguchi I, Hirono S, Liu Q, Nakagome Y, Matsushita Y (1992). Simple methods of calculating octanol water partition coefficient. *Chem Pharm Bull* **40**: 127–130.

Muegge I, Martin YC (1999). A general and fast scoring function for protein–ligand interactions: a simplified potential approach. *J Med Chem* **42**: 791–804.

Oprea TI, Allu TK, Fara DC, Rad RF, Ostopovici L, Bologa CG (2007). Lead-like, drug-like or 'Pub-like': how different are they? *J Comput Aided Mol Des* **21**: 113–119.

Oprea TI, Matter H (2004). Integrating virtual screening in lead discovery. *Curr Opin Chem Biol* **8**: 349–358.

Pauptit RA, Dennis CA, Derbyshire DJ, Breeze AL, Weston SA, Rowsell S *et al*. (2001). NMR trial models: experiences with the colicin immunity protein Im7 and the p85alpha C-terminal SH2-peptide complex. *Acta Crystallogr D Biol Crystallogr* **57**: 1397–1404.

Perola E, Walters WP, Charifson PS (2004). A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **56**: 235–249.

Pospisil P, Kuoni T, Scapozza L, Folkers G (2002). Methodology and problems of protein–ligand docking: case study of dihydroorotate dehydrogenase, thymidine kinase, and phosphodiesterase 4. *J Recept Signal Transduct Res* **22**: 141–154.

Qureshi SA, Kim RM, Konteatis Z, Biazzo DE, Motamedi H, Rodrigues R *et al*. (1999). Mimicry of erythropoietin by a nonpeptide molecule. *Proc Natl Acad Sci USA* **96**: 12156–12161.

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

Raymond JW, Jalaie M, Bradley MP (2004). Conditional probability: a new fusion method for merging disparate virtual screening results. *J Chem Inf Comput Sci* **44**: 601–609.

Rees DC, Congreve M, Murray CW, Carr R (2004). Fragment-based lead discovery. *Nat Rev Drug Discov* **3**: 660–672.

Rosenfeld R, Vajda S, DeLisi C (1995). Flexible docking and design. *Annu Rev Biophys Biomol Struct* **24**: 677–700.

Russ AP, Lampel S (2005). The druggable genome: an update. *Drug Discov Today* **10**: 1607–1610.

Scheufler C, Brinker A, Bourenkov G, Pegoraro S, Moroder L, Bartunik H *et al*. (2000). Structure of TPR domain–peptide complexes: critical elements in the assembly of the Hsp70–Hsp90 multichaperone machine. *Cell* **101**: 199–210.

Schuffenhauer A, Brown N, Selzer P, Ertl P, Jacoby E (2006). Relationships between molecular complexity, biological activity, and structural diversity. *J Chem Inf Model* **46**: 525–535.

Shave SR, Taylor P, Walkinshaw M, Smith L, Hardy J, Trew A (2008). Ligand discovery on massively parallel systems. *IBM Journal of Research and Development* **52** 1/2 January/March.

Singh J, Van VH, Liao Y, Lee WC, Cornebise M, Harris M *et al*. (2002). Identification of potent and novel alpha4beta1 antagonists using *in silico* screening. *J Med Chem* **45**: 2988–2993.

Stahl M, Rarey M (2001). Detailed analysis of scoring functions for virtual screening. *J Med Chem* **44**: 1035–1042.

Teague SJ, Davis AM, Leeson PD, Oprea T (1999). The Design of Leadlike Combinatorial Libraries. *Angew Chem Int Ed Engl* **38**: 3743–3748.

Todeschini R, Consonni V (2005). *Handbook of Molecular Descriptors*. Wiley-VCH: UK.

Vangrevelinghe E, Zimmermann K, Schoepfer J, Portmann R, Fabbro D, Furet P (2003). Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J Med Chem* **46**: 2656–2662.

Vassilev LT, Vu BT, Graves B, Carvajal D, Podlaski F, Filipovic Z *et al*. (2004). *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**: 844–848.

Verdonk ML, Berdini V, Hartshorn MJ, Mooij WT, Murray CW, Taylor RD *et al*. (2004). Virtual screening using protein–ligand docking: avoiding artificial enrichment. *J Chem Inf Comput Sci* **44**: 793–806.

Vieth M, Hirst JD, Dominy BN, Daigler H, Brooks CL (1998). Assessing search strategies for flexible docking. *J Comput Chem* **19**: 1623–1631.

Wang JL, Liu DX, Zhang ZJ, Shan SM, Han XB, Srinivasula SM *et al*. (2000). Structure-based discovery of an organic compound that binds Bcl-2 protein and induces apoptosis of tumor cells. *Proc Natl Acad Sci USA* **97**: 7124–7129.

Wang R, Lu Y, Wang S (2003). Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* **46**: 2287–2303.

Wang S, Meades C, Wood G, Osnowski A, Anderson S, Yuill R *et al*. (2004a). 2-Anilino-4-(thiazol-5-yl)pyrimidine CDK inhibitors: synthesis, SAR analysis, X-ray crystallography, and biological activity. *J Med Chem* **47**: 1662–1675.

Wang S, Wood G, Meades C, Griffiths G, Midgley C, McNae I *et al*. (2004b). Synthesis and biological activity of 2-anilino-4-(1H-pyrrol-3-yl) pyrimidine CDK inhibitors. *Bioorg Med Chem Lett* **14**: 4237–4240.

Woo HJ, Roux B (2005). Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc Natl Acad Sci USA* **102**: 6825–6830.

Wu B, Li P, Liu Y, Lou Z, Ding Y, Shu C *et al*. (2004). 3D structure of human FK506-binding protein 52: implications for the assembly of the glucocorticoid receptor/Hsp90/immunophilin heterocomplex. *Proc Natl Acad Sci USA* **101**: 8348–8353.

Wu SY, McNae I, Kontopidis G, McClue SJ, McInnes C, Stewart KJ *et al*. (2003). Discovery of a novel family of CDK inhibitors with the program LIDAEUS: structural basis for ligand-induced disordering of the activation loop. *Structure* **11**: 399–410.

Wu X, Knudsen B, Feller SM, Zheng J, Sali A, Cowburn D *et al*. (1995). Structural basis for the specific interaction of lysine-containing proline-rich peptides with the N-terminal SH3 domain of c-Crk. *Structure* **3**: 215–226.

Xing L, Hodgkin E, Liu Q, Sedlock D (2004). Evaluation and application of multiple scoring functions for a virtual screening experiment. *J Comput Aided Mol Design* **18**: 333–344.

Yang JM, Chen CC (2004). GEMDOCK: a generic evolutionary method for molecular docking. *Proteins* **55**: 288–304.

Zhao L, Brinton RD (2005). Structure-based virtual screening for plant-based ERbeta-selective ligands as potential preventative therapy against age-related neurodegenerative diseases. *J Med Chem* **48**: 3463–3466.

Zheng CJ, Han LY, Yap CW, Ji ZL, Cao ZW, Chen YZ (2006). Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol Rev* **58**: 259–279.