D1042–D1048 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 4 November 2010*
doi:10.1093/nar/gkq878

EDULISS: a small-molecule database with data-mining and pharmacophore searching capabilities

Kun-Yi Hsin, Hugh P. Morgan, Steven R. Shave, Andrew C. Hinton, Paul Taylor and Malcolm D. Walkinshaw*

The Centre for Translational and Chemical Biology, The University of Edinburgh, King's Buildings, Edinburgh, UK

Received May 30, 2010; Revised September 3, 2010; Accepted September 17, 2010

ABSTRACT

We present the relational database EDULISS (EDinburgh University Ligand Selection System), which stores structural, physicochemical and pharmacophoric properties of small molecules. The database comprises a collection of over 4 million commercially available compounds from 28 different suppliers. A user-friendly web-based interface for EDULISS (available at http://eduliss.bch.ed.ac.uk/) has been established providing a number of data-mining possibilities. For each compound a single 3D conformer is stored along with over 1600 calculated descriptor values (molecular properties). A very efficient method for unique compound recognition, especially for a large scale database, is demonstrated by making use of small subgroups of the descriptors. Many of the shape and distance descriptors are held as pre-calculated bit strings permitting fast and efficient similarity and pharmacophore searches which can be used to identify families of related compounds for biological testing. Two ligand searching applications are given to demonstrate how EDULISS can be used to extract families of molecules with selected structural and biophysical features.

INTRODUCTION

The high throughput screening regimes of the past 20 years led by big pharma and more recently developed by screening centres through the Molecular Libraries Roadmap program are providing increasing amounts of publicly available biological information. The bioassay and compound databases in PubChem (1) contain information on over 25 million structures and on over 60 million data points from thousands of assays. Smaller but well annotated databases like ChEMBLdb with over 500 000 entries provide information on the properties and activities of drug-like molecules and their targets (2). This explosion of data linking compounds to biological activity should provide a means for predicting new biological effects for large numbers of classes of small drug-like molecules using bioinformatic and database mining approaches (3).

In order to test such *in silico* predictions it is important to have databases of available compounds. It is only relatively recently that searchable interactive small molecule databases have become available to non-commercial research groups. One such resource is ChemDB (4), a searchable chemical database containing nearly 5 million small molecules with their stereoisomers. Interactive databases like ZINC (5) provide large and well annotated collections with some searching capacity. Such databases can contain a variety of structurally related information stored as SMILES strings, InChI or Daylight fingerprints (6). 3D coordinates may also be used as input for structure-based virtual screening (7–9) or pharmacophore searching (10). The idea of relating the activity of a molecule to the spatial distribution of a number of functional groups (11) has been widely used in QSAR (12) and structure-based studies as implemented in programs like GRID (13), LigandScout (14) and Catalyst (15).

The EDULISS database stores 3D atomic coordinates for each molecule along with over 1600 calculated molecular properties. These so called molecular descriptors provide a numerical profile for each molecule consisting of calculated values such as molecular weight, surface area and number of rotatable bonds. By using a selection of descriptors it is possible to rapidly select small related families of molecules from the database. An extension of this selection procedure provides a very efficient way of

*To whom correspondence should be addressed. Tel: 00441316507056; Fax: 00441316507055; Email: m.walkinshaw@ed.ac.uk

identifying unique compounds. The database also stores a range of interatomic distances between various atom types for each molecule. The overall statistics of interatomic distances is used in an ultrafast shape searching algorithm (16). A specific subset of interatomic distances between all hydrogen bond donor and acceptor atoms, halogens, phosphorous and sulphur atoms provide what we call the Interatomic Pharmacophore Profile (IPP). All such distance information is stored for each molecule in pre-calculated bit-strings which provide the basis of a wide range of pharmacophore searching routines and also in the identification of similarly shaped molecules. The EDULISS database is therefore a useful tool for identifying commercially available molecules based on similarity or pharmacophore searches. It is distinguished from other web resources by having over 1600 descriptors for each compound and the ability to carry out unique 3D and 2D searches. There are also convenient links for a subset of compounds to the PubChem database allowing easy access to biological data.

PROGRAM AND DATABASE DESCRIPTION

Database description

Currently, EDULISS stores over 5.5 million (over 4 million unique) compounds in total, containing data from 28 different commercial and other smaller specialist compound catalogues (Supplementary Data S1). 2D and 3D coordinates for each molecule are stored with over 1600 topological, geometrical, physicochemical and toxicological descriptors per compound. In this database, over 3.9 million compounds fit the Lipinski's rule of five (17) and a total of 3.4 million fit the Oprea lead-like criteria (18): that is molecular weight $\leq$460, number of rotatable bonds $\leq$10, calculated Log P between -4 and 4.2, number of hydrogen bond acceptors $\leq$9, number of hydrogen bond donors $\leq$5 and number of rings $\leq$4. The database also contains over 520 000 compounds with molecular weight $<$250 Da and potentially fitting the needs of fragment-based screening (19).

The biological properties of a subset of 291 000 compounds stored in EDULISS has been retrieved from four other databases, including PubChem, BindingDB (20), ChemBank (21) and DrugBank (22), by identifying identical molecules using the Maximum Common Subgraph algorithm (23). The identity of these compounds in the external databases has been obtained and stored in the EDULISS database. A direct link between EDULISS and the external database has been implemented on the search result pages. Once a particular compound which is identical to one of the PubChem compounds has been hit by either 3D/2D similarity or molecule ID search, the link in the 'Chemical Properties' box can lead users to the appropriate PubChem web page. Certain catalogues (e.g. the National Cancer Institute) contain many compounds for which there are a lot of biological data and most hits will have links to the relevant PubChem bioassay summary page.

The EDULISS database is held in a MySQL server. The web-based interface of EDULISS 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 (Figure 1). The web site utilizes Apache Tomcat as the web server and the runtime environment for Java technologies mentioned above. For the molecule drawing and visualizing, JME (http://www.molinspiration.com/jme/) and Jmol are utilized which are applications written in Java providing interactive features and have been included in the EDULISS web pages. On the query result page, the users can download the SDfile of hit compounds with their descriptor values. To date, this database has been used freely by the researchers from over 20 countries via its web-based interface.

Treatment of compound structure data files, SDfiles

Regardless of the source of catalogues, the compounds used for EDULISS were entirely collected as 2D SDfile formats then converted into 3D atomic coordinates using CONCORD software. After the conversion process, the molecules were processed by DRAGON 5.4 (http://www.talete.mi.it/) and DEREK (http://www.lhasalimited.org) software calculating 1664 physicochemical and potential toxicity properties for each compound.

DATABASE APPLICATIONS

Recognition of unique compounds

As EDULISS holds millions of compounds from various suppliers, it is useful to be able to determine the number of unique compounds in the collection. A 2D graph theory algorithm, Maximum Common Subgraph, MCS (23), has been implemented. Although the MCS is able to precisely identify isomorphous compounds, the number of pairwise comparisons increase as $N \times (N-1)$ where N is the number of compounds and the run time grows dramatically from 1 h to 1 day when the dataset increases from 800 to 3200 compounds (Supplementary Data S2). Thus, it is impossible to go through the whole EDULISS collection using this method.

We have developed a method to efficiently identify unique compounds by clustering according to specific descriptor values (molecular properties). Using this approach the required graphical comparisons can be considerably decreased. Preliminary studies using molecular weight and atom type were not very useful as only 6% of the compounds in EDULISS could be uniquely identified. However a number of other molecular descriptors show much better discrimination; W3D [Wiener 3D index (24)], Whete [Wiener-type index from electronegativity weighted distance matrix (25)] and Vu [a molecular size descriptor which is one of the Weighted Holistic Invariant Molecular descriptors (26)]. The combination of these three descriptors alone was sufficient to identify 3 117 625 unique compounds (out of a total of 4 011 697 unique compounds present in EDULISS). The remaining 2 million compounds were grouped using the three descriptors (W3D, Whete and Vu) into 845 193 clusters. The compounds in these clusters with identical descriptors were then compared using MCS. This procedure reduces

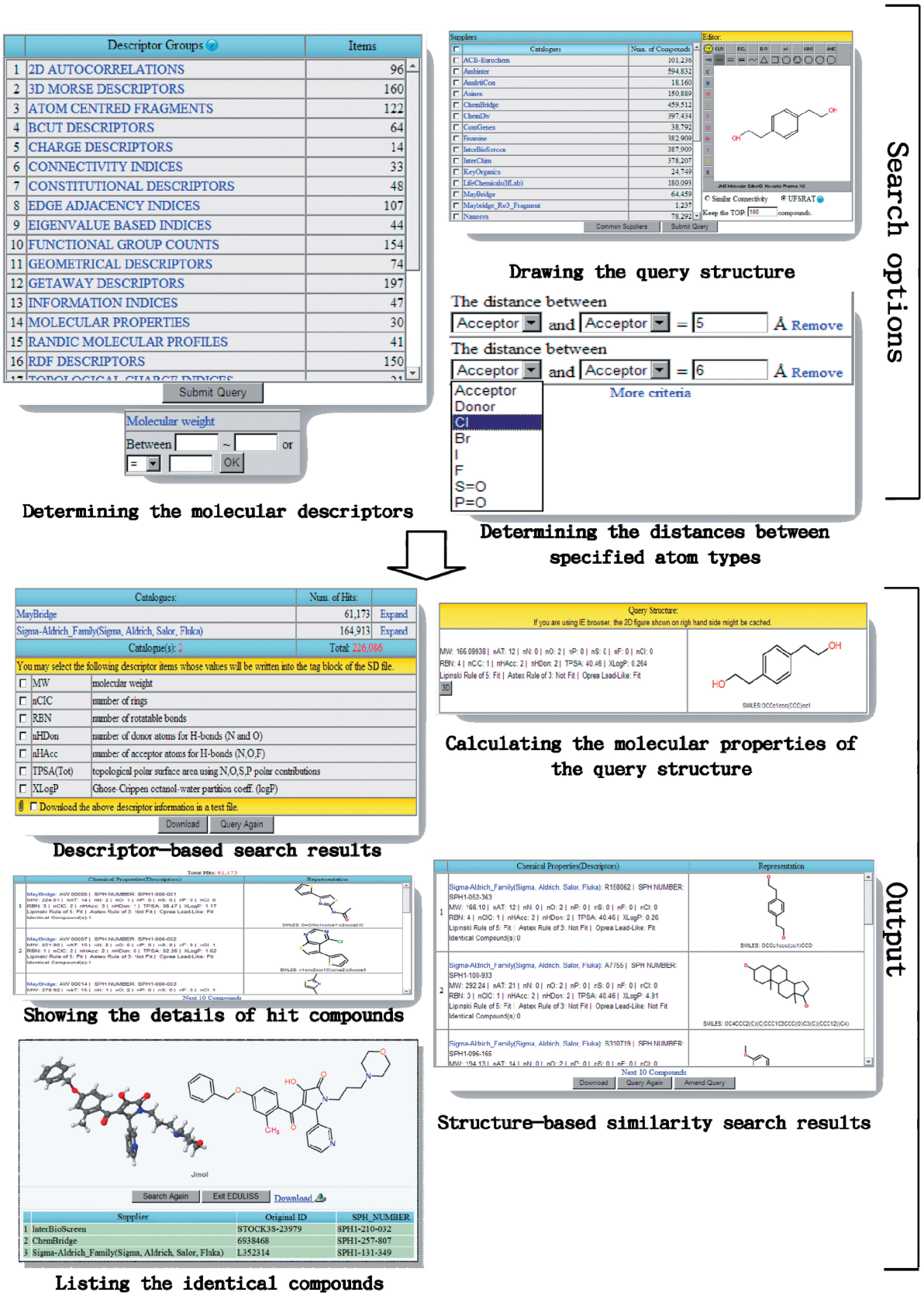


Figure 1. The EDULISS web-based interface provides four search options, including descriptor-based searches, structure-based similarity searches, IPP searches and search by molecule ID. On the query result pages, the users can download the SDfile of hit compounds with their descriptor values.

the number of required pair-wise comparisons using MCS down to 6 495 096 which can be carried out in 20 h.

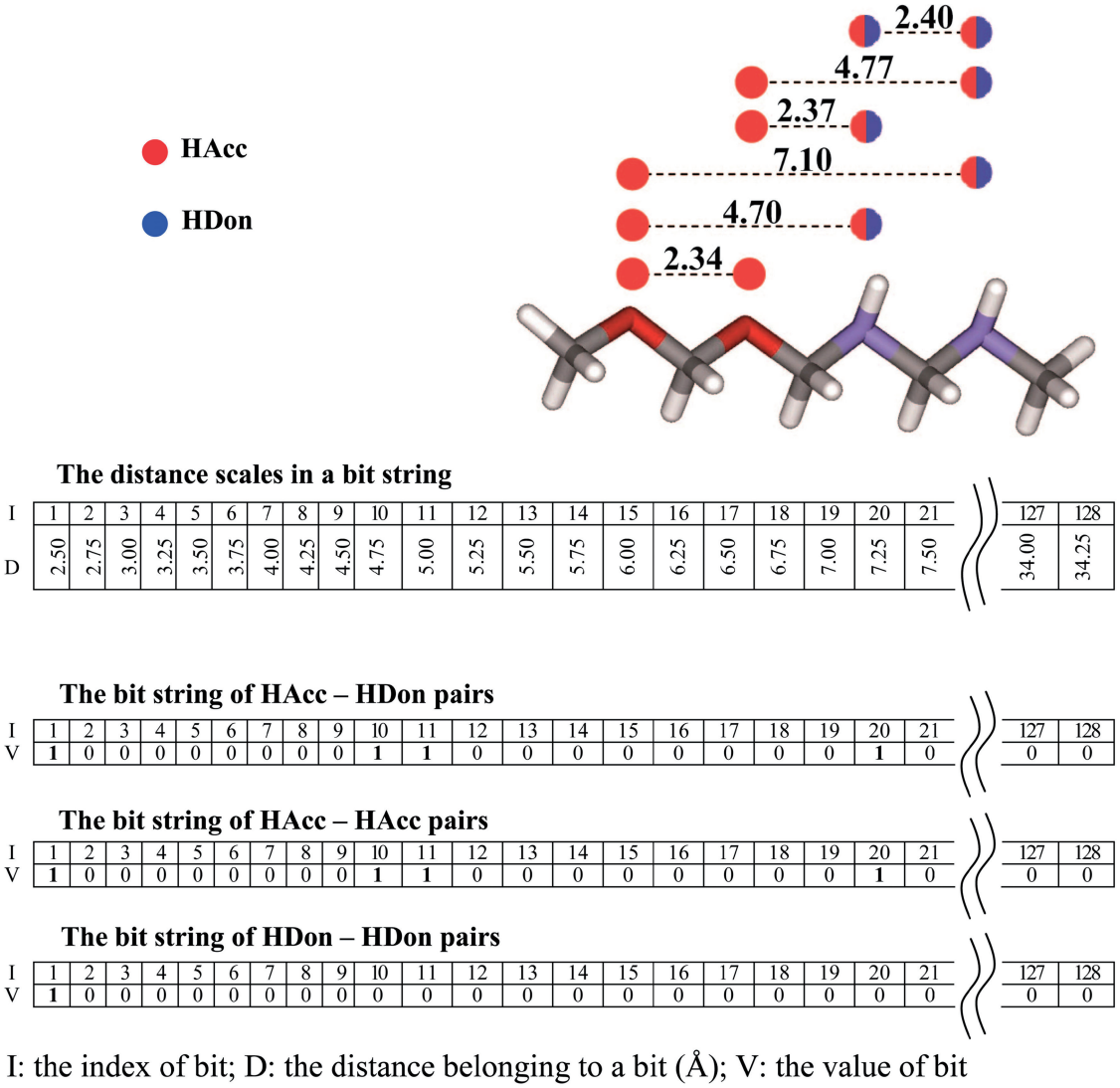
Similarity searching

EDULISS stores more than 1600 molecular descriptors for each compound and users can select a series of descriptor items as a query to identify a subset of molecules which will share common properties. Molecular descriptors are primarily organized into 20 groups according to their attributes, so that the users can conveniently choose and set preferred values for the query. For example it is a simple matter to extract from the Sigma-Aldrich catalogue the 164 913 out of 199 492 compounds that fit the Lipinski rule of five and the 142 660 that comply with the Oprea lead-like criteria.

EDULISS also provides geometrical similarity searches based on a 3D similarity measurement called Ultra Fast Shape Recognition with Atom Types (UFSRAT). UFSRAT uses pre-generated geometric descriptors for molecules within EDULISS to discriminate between both the overall geometric, hydrophobic and electrostatic shape of molecules.

Pharmacophore searching using EDULISS

The IPP for each molecule in the database consists of interatomic distances calculated between 8 different atom classes; namely hydrogen bond donor atoms (HDon), hydrogen bond acceptor atoms (HAcc), halogens (fluorine, chlorine, bromine and iodine), sulphur and phosphorus atoms. This gives rise to 15 possible types of interatomic distance for each molecule. Distances are stored in strings 128 bits long as Boolean values (1, true; 0, false). The first bit represents a distance less than or equal to 2.50 Å, the next bit is 0.25 Å longer (i.e. >2.50 and ≤2.75 Å) and so forth until the last bit which represents any distance >34.00 Å. Thus, there are 15 bit strings for each molecule representing the 15 types of possible distance pairs. Figure 2 illustrates the



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

Figure 2. Examples of the bit string composition of a virtual compound. Hydrogen bond acceptor (HAcc) atoms are coloured red and hydrogen bond donor (HDon) atoms are coloured blue.

composition of three bit strings showing distances between HAcc and HDon.

This facility enables compounds to be identified that have a specific geometric arrangement of atoms ('pharmacophore') as defined by pair-wise distances of hydrogen bond donors, acceptors, halogens, phosphorous or sulphur atoms (where the searches are restricted to S and P atoms that form double bonds to oxygen). For pharmacaphore searching, a bit string is generated for the user-defined query distances which are then compared to that of each compound in the database. If a specific true bit in the query matches, the distance criterion is met. A user can perform a multi-distance query in a single search. Apart from very efficient storage, bit strings also provide a very fast searching method as the necessary Boolean operations can be carried out very quickly. Users can specify the query by defining preferred distances between selected atom types using the web-based interface. The results are then displayed and hits may be downloaded.

Case study 1: identifying cyclin dependent kinases inhibitors. As a test for the pharmacaphore searching routine we used the eight available structures of CDK complexes stored in the PDBbind database (27) with PDB codes 1AQ1, 1DI8, 1DM2, 1E1V, 1E1X, 1FVV, 1UNH and 2A4L. Three nitrogen atoms of the adenine ring of ATP were used as a template to search for ATP-analogues (Figure 3). Applying the three interatomic distance criteria as shown in Figure 3b, four out of the eight ligands (1DM2, 1E1V, 1E1X and 1UNH) were identified as illustrated in Figure 3c–f, (The other four ligands were not recognized as they do not have the adenine-like pharmacophore)

Case study 2: identifying pyruvate kinase inhibitors. The glycolytic enzyme pyruvate kinase (PYK) is a drug target against trypanosomatid infection (28). Fructose-2,6-bisphosphate (F-2,6-BP) acts as an allosteric activator (29). We are interested in identifying analogue molecules which interfere with allosteric regulation. Figure 4a schematically shows selected interatomic

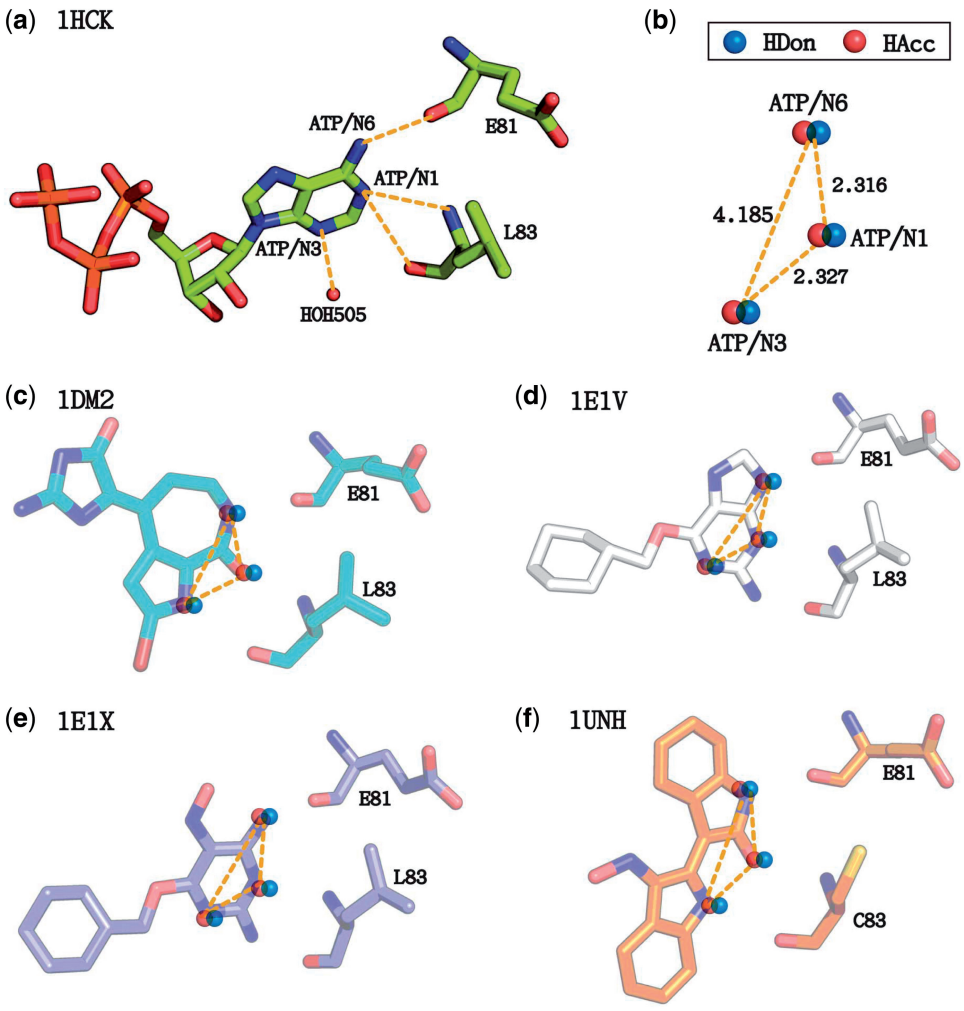
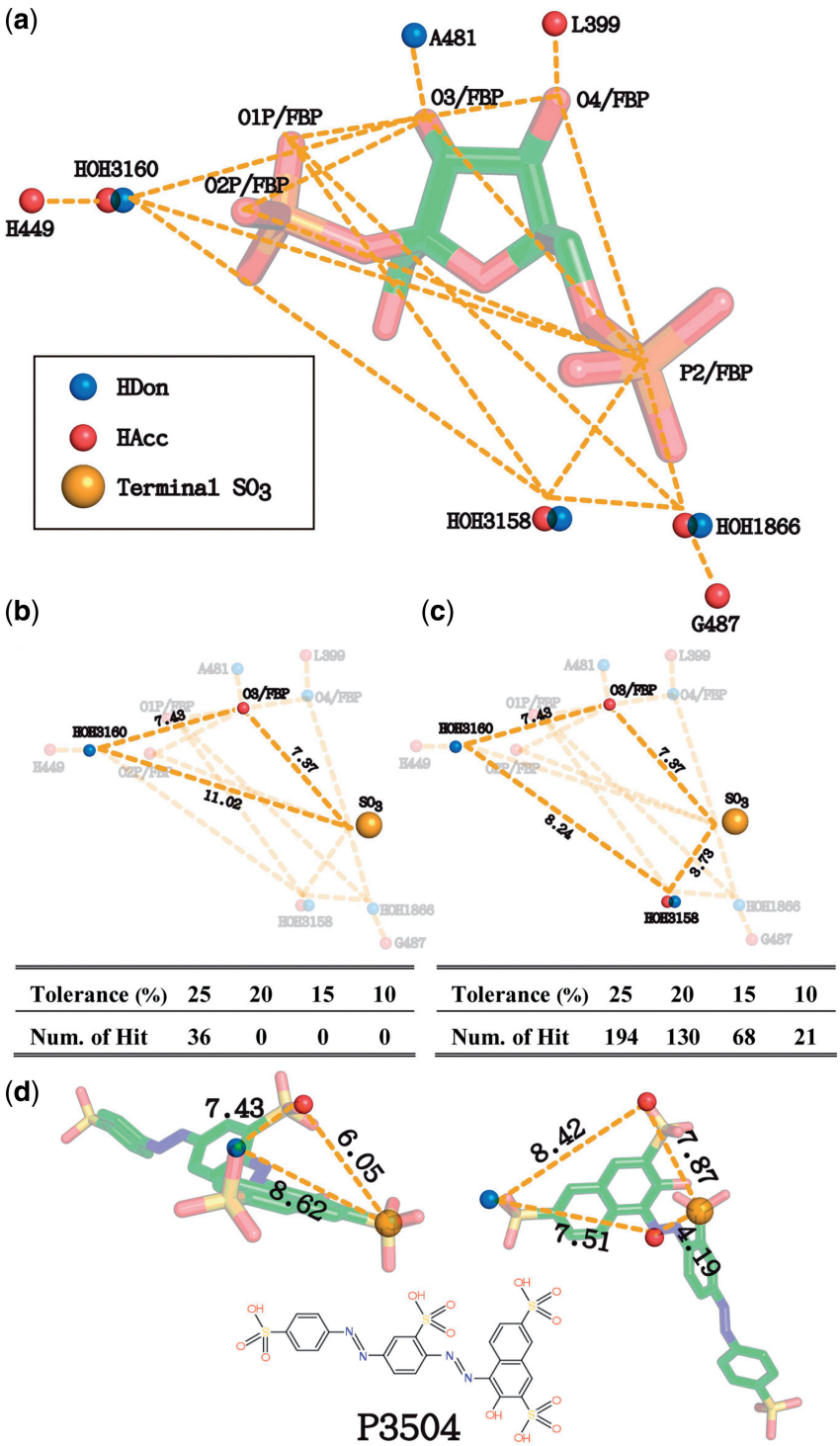


Figure 3. Schematic diagrams illustrating the CDK inhibitor pharmacaphore search. (**a**) the key interaction of ATP-binding pocket in the CDK2–ATP complex (PDB id: 1HCK). (**b**) interatomic distances between the three atoms selected from the ATP adenine ring. The hydrogen bond donors (HDon) and acceptors (HAcc) are coloured blue and red and labelled by residue name and ID. (**c–f**) show the interesting interactions of the four hit CDK proteins whose ligand-bound interactions are similar to the pharmacaphore search model.



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

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

Figure 4. Schematic diagrams of the relevant interactions of F-2,6-BP (FBP) in the effector site of PYK. (**a**) shows the interatomic distances between five atoms in F-2,6-BP and three water molecules, selected for pharmacophore models. (**b** and **c**) Examples of pharmacaphore search models with hydrogen bond donors (HDon) and acceptors (HAcc) coloured blue and red. Sulfonate (SO_3) is coloured yellow. (**d**) illustrates the hit compound P3504 in different orientations to show that both pharmacophore profiles fit.

contacts and distances between five atoms in F-2,6-BP and three water molecules selected for pharmacophore searches. Two example search motifs are shown in Figure 4b and c. A series of tolerances have been given for each interatomic distance from 10 to 25%. The numbers of hit compounds in a range of tolerances are tabulated in Figure 4. We selected eight compounds (Sigma-Aldrich ID: N9002, L0144, P3504, 201332, 244813, D5021, H2516 and 86170) for further experimental assay based on visual inspection of the docked pose

and on calculated solubility. Of the eight selected compounds, five significantly affected the PYK enzyme kinetics. Figure 4d shows that both pharmacophore search models match atoms of the hit compound P3504 which showed 33% inhibition of enzyme activity. A complex of P3504 with *Leishmania mexicana* PYK (*Lm*PYK) has been crystallized and solved at a resolution of 2.7 Å (H. P. Morgan personal communication) showing the molecule binds at the effector site.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the staff at Centre of Translational and Chemical Biology.

FUNDING

The Wellcome Trust and the Scottish University Life Sciences Alliance for the use of Edinburgh Protein Production Facility. K.-Y.H. who did most of the work was supported by an Edinburgh University departmental scholarship. This work was not directly supported by any granting agencies or by commercial companies.

Conflict of interest statement. None declared.

REFFERENCES

1. Geer,L.Y., Marchler-Bauer,A., Geer,R.C., Han,L., He,J., He,S., Liu,C., Shi,W. and Bryant,S.H. (2010) The NCBI BioSystems database. *Nucleic Acids Res.*, **38**, D492–D496.
2. EMBL-EBI. (2010) Open access drug discovery database launches with half a million compounds: www.ebi.ac.uk/chembldb/index.php. *EMBL-EBI Press Release*.
3. Miller,M.A. (2002) Chemical database techniques in drug discovery. *Nat. Rev. Drug Discov.*, **1**, 220–227.
4. Chen,J.H., Linstead,E., Swamidass,S.J., Wang,D. and Baldi,P. (2007) ChemDB update–full-text search and virtual chemical space. *Bioinformatics (Oxford, England)*, **23**, 2348–2351.
5. Irwin,J.J. and Shoichet,B.K. (2005) ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.*, **45**, 177–182.
6. 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. *J. Chem. Inf. Comput. Sci.*, **39**, 747–750.
7. Anderson,A.C. (2003) The process of structure-based drug design. *Chem. Biol.*, **10**, 787–797.
8. Ghosh,S., Nie,A., An,J. and Huang,Z. (2006) Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol.*, **10**, 194–202.
9. Lyne,P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today*, **7**, 1047–1055.
10. Fang,X. and Wang,S. (2002) A web-based 3D-database pharmacophore searching tool for drug discovery. *J. Chem. Inf. Comput. Sci.*, **42**, 192–198.
11. McGregor,M.J. and Muskal,S.M. (2000) Pharmacophore fingerprinting. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci.*, **40**, 117–125.
12. Deanda,F. and Stewart,E.L. (2004) Application of the PharmPrint methodology to two protein kinases. *J. Chem. Inf. Comput. Sci.*, **44**, 1803–1809.
13. Kastenholz,M.A., Pastor,M., Cruciani,G., Haaksma,E.E. and Fox,T. (2000) GRID/CPCA: a new computational tool to design selective ligands. *J. Med. Chem.*, **43**, 3033–3044.
14. Wolber,G. and Langer,T. (2005) LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.*, **45**, 160–169.
15. Patel,Y., Gillet,V.J., Bravi,G. and Leach,A.R. (2002) A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. *J. Comput. Aided Mol. Des.*, **16**, 653–681.
16. Ballester,P.J. and Richards,W.G. (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.*, **28**, 1711–1723.
17. Lipinski,C.A., Lombardo,F., Dominy,B.W. and Feeney,P.J. (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.*, **23**, 3–25.
18. Hann,M.M. and Oprea,T.I. (2004) Pursuing the leadlikeness concept in pharmaceutical research. *Curr. Opin. Chem. Biol.*, **8**, 255–263.
19. Hartshorn,M.J., Murray,C.W., Cleasby,A., Frederickson,M., Tickle,I.J. and Jhoti,H. (2005) Fragment-based lead discovery using X-ray crystallography. *J. Med. Chem.*, **48**, 403–413.
20. Liu,T., Lin,Y., Wen,X., Jorissen,R.N. and Gilson,M.K. (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.
21. Seiler,K.P., George,G.A., Happ,M.P., Bodycombe,N.E., Carrinski,H.A., Norton,S., Brudz,S., Sullivan,J.P., Muhlich,J. and Serrano,M. (2008) ChemBank: a small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.*, **36**, D351–D359.
22. Wishart,D.S., Knox,C., Guo,A.C., Cheng,D., Shrivastava,S., Tzur,D., Gautam,B. and Hassanali,M. (2007) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36(Suppl. 1)**, D901–D906.
23. Raymond,J.W. and Willett,P. (2002) Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput. Aided Mol. Des.*, **16**, 521–533.
24. Mihalic,Z. and Veljan,D. (1992) The distance martrix in chemistry. *J. Math. Chem.*, **11**, 223–258.
25. Ivanciuc,O., Ivanciuc,T. and Balaban,A.T. (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. *J. Chem. Inf. Comput. Sci.*, **38**, 395–401.
26. Todeschini,R., Lasagni,M. and Marengo,E. (1994) New molecular descriptors for 2D and 3D structures. Theory. *J. Chemom.*, **8**, 263–272.
27. Wang,R., Fang,X., Lu,Y. and Wang,S. (2004) The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.*, **47**, 2977–2980.
28. Mesecar,A.D. and Nowak,T. (1997) Metal-ion-mediated allosteric triggering of yeast pyruvate kinase. 2. A multidimensional thermodynamic linked-function analysis. *Biochemistry*, **36**, 6803–6813.
29. Morgan,H.P., McNae,I.W., Hsin,K.Y., Michels,P.A.M., Fothergill-Gilmore,L.A. and Walkinshaw,M.D. (2010) An improved strategy for the crystallization of Leishmania mexicana pyruvate kinase. *Acta Crystallogr. Sect. F Struct. Biol. Cryst. Commun.*, **66**, 215–218.