J Comput Aided Mol Des (2006) 20:249–261
DOI 10.1007/s10822-006-9048-0

ORIGINAL PAPER

ADAAPT: Amgen's data access, analysis, and prediction tools

Sung Jin Cho · Yaxiong Sun · William Harte

Received: 15 December 2005 / Accepted: 7 May 2006 / Published online: 21 June 2006
© Springer Science+Business Media B.V. 2006

The Amgen's Data Access Analysis Prediction Tools (ADAAPT) system is a desktop decision support tool developed to provide flexible access and analysis of chemical and biological data. The system is platform independent, adaptable, easily deployed, and scalable. It consists of four main modules: access, analysis, prediction, and tools. The access module contains numerous user interfaces designed to retrieve data easily. The analysis module provides standard computational tools to perform property calculation, QSAR/QSPR, and statistical analyses. The prediction module contains in-house models to calculate a drug-likeness score and absorption index. Finally, the tools module provides a wide array of features that are of general interest to our scientists.

Introduction

Rapid developments in chemistry, screening and automation in the last decade have led to the generation of large amounts of data for a typical drug discovery process. As the amount of chemical and biological information increases, the success of drug discovery and development, now more than ever, depends on how well information is integrated throughout the research processes.

S. J. Cho (✉) · Y. Sun · W. Harte
Molecular Structure, Amgen Inc., MS 29-M-B, One Amgen Center Drive, Thousand Oaks, CA 91320, USA
e-mail: scho@amgen.com

The first step in dealing with this complex problem is to build a flexible data warehouse to collect and store the vast amount of disparate data, which includes chemical structures and properties, genome sequences, biological activities, and ADMET information. In addition, the warehouse must be able to handle any future data depository needs. The second step in dealing with this problem, once the data is in place, is to build an adaptable system to access and analyze the data. These data accessing and analysis tools must be disseminated to the research scientists and management in a timely and efficient manner to support their decision-making. The ultimate goal is to associate data with empirical rules to allow automated inferences and provide proactive guidances to end-users. That is, the system will not only support (reactive) but influence (proactive) the decision making process. For example, at the lead optimization stage, one can envision a system where existing SAR information can be retrieved and used by chemists to decide what compounds to make next. This is an example of reactive decision support system. On the other hand, one can also envision in a situation where, upon synthesizing and testing a set of compounds, a system provides a list of attractive candidates based on all the existing information in the data warehouse. This is an example of proactive decision support system. The problem with the proactive decision support system is that, especially in the pharmaceutical industry, questions one asks are multilayered and multifaceted, and most of the time data we store in the warehouse is not enough to answer those questions, thus requiring constant human intervention.

General concepts of such drug discovery information system have been discussed previously [1–4] and

papers describing specific modules [5–7] as well as a complete information management system [8] have been published. Previous works, however, have dealt primarily with answering the first part of the problem, the construction of an efficient data warehouse system. In this paper, we present Amgen's Data Access Analysis Prediction Tools (ADAAPT) system, which is our attempt to deal with the second problem, building an adaptable system to access and analyze the data. ADAAPT is a desktop decision support tool developed for chemists, biologists, HTS scientists, and computational chemists. ADAAPT is an evolving endeavor designed to change with the users' needs. The ADAAPT client consists of four main modules: access, analysis, prediction, and tools. The access module contains numerous user interfaces designed to easily retrieve information. The analysis module provides a number of standard computational tools to perform property calculation, QSAR/QSPR and statistical analyses. The prediction module contains in-house models to calculate a drug-likeness score and absorption index. Finally, the tools module provides many useful features that do not belong to any of the other three modules, including visualization tools to handle common plotting needs. The design philosophy, architecture, and detailed description of each module will be discussed.

Design philosophy

During the design stage, we set forth a number of key requirements that the ADAAPT system must deal with: platform independence, adaptability, easy of deployment, and scalability. One might argue that the platform independence, especially in the pharmaceutical industry, is probably less important than the other three criteria, since the majority of ADAAPT system's end users work in the Microsoft Windows [9] environment. Nevertheless, we decided to add this as one of the key requirements because our users include computational chemists and crystallographers (a small but important group of research scientists) who also work in other computing environments such as UNIX. Platform independence also makes our system adaptable, which is our second criterion. Adaptability is one of key requirements because our workflow is dynamic and, most of the time, it is difficult to foresee what is the best approach one should take. While a traditional software design approach attempts to understand and address all of needs prior to building the information management system, this paradigm rarely succeeds in meeting the objectives, because the demands evolve iteratively over time as the software develops. Rather than gathering all possible questions that the users are ever likely to ask and building a system to accommodate everyone's needs, we wished to make the system flexible enough so features could be added and changes made in a timely and efficient manner. Such a system requires many changes and releases, so it is necessary that the system is easily deployed to users without interrupting their work. Finally, the system must be able to handle growing numbers of users across multiple sites.

Method

The ADAAPT system utilizes a three-tiered architecture: client, middle tier, and databases. The client and middle tiers were both written using C/C++ and consist of over 84,000 lines of code. In addition, several C programs utilizing the Daylight [10] toolkit and numerous c-shell and Perl [11] scripts were written for various chemoinformatic tasks. The OpenGL [12] module in the Qt [13] tool kit was used for development of the visualization tools. The Qt toolkit is also used to provide platform independence. Currently, the ADAAPT client is compiled on Windows 2000 [9], Red Hat Linux [14], and Sgi Irix [15]. For Windows 2000 users, deployment of the client is simplified by creating a standalone executable and placing it and necessary Dynamic Linked Libraries (DLLs) on a network drive. Users can then access the system via a shortcut, thereby eliminating any installation of the client on a local machine, and making any new release of the client transparent to the users by simply copying over the newly compiled version onto the network drive. Users using the older versions are not affected because they still exist; the older version of ADAAPT is simply renamed. When users close existing ADAAPT client sessions and relaunch the client, they automatically access the new version of the client. There is no back-compatibility issue with ADAAPT itself because ADAAPT is an environment containing different modules. Modules, on the other hand, can be refined, or a whole new one can be created, but typically never get deleted. Linux and SGI users only need to add a path and set up ADAAPT environment variables in their initialization environment file (e.g., *.cshrc*). Most analysis tasks and any operation which does not require additional information from the data warehouse are performed via the ADAAPT client. This is done to preserve server resource. The role of middle tier is to provide information requested by the client. The information can be retrieved directly from the data warehouse, or from specially formatted flat

files that are constructed to expedite data access and enhance user experience. The middle tier also increases the adaptability of the system because it is not tied to a specific system. It is designed to handle any third party software and databases in order to provide necessary information to users. The communication between the client and middle tier is accomplished by creating a port between them, and sending and receiving formatted strings. In fact, any application can connect to the middle tier via specifying a proper IP address and port number. Rather than using one port number to handle all communication, each port number is assigned to a specific task or group of tasks that can then be easily distributed over many servers. Currently, the ADAAPT middle tier runs on two Linux-based servers, although it is not limited to this environment. The only limitation is the platform dependence of third party software and databases.

Property calculation

Cerius2 [16] and ACD/PhysChem [17] are used to calculate default properties that are used in the ADAAPT system.

User authentication

The Lightweight Directory Access Protocol (LDAP) [18] server provides authentication services, and users can logon to the ADAAPT system using their Windows networking username and password.

Results and discussion

The ADAAPT client is a multiple document interface (MDI) application. A MDI application has a single main window and any number of child windows, displayed within the main window. It provides an excellent way to organize many windows generated in an application such as the ADAAPT client. The Windows 2000 version of the ADAAPT client is shown in Fig. 1. The main window features a number of buttons to provide easy access to many commonly used tasks.

Access module

The data access module consists of three different kinds of views, which depend on a user's need: assay

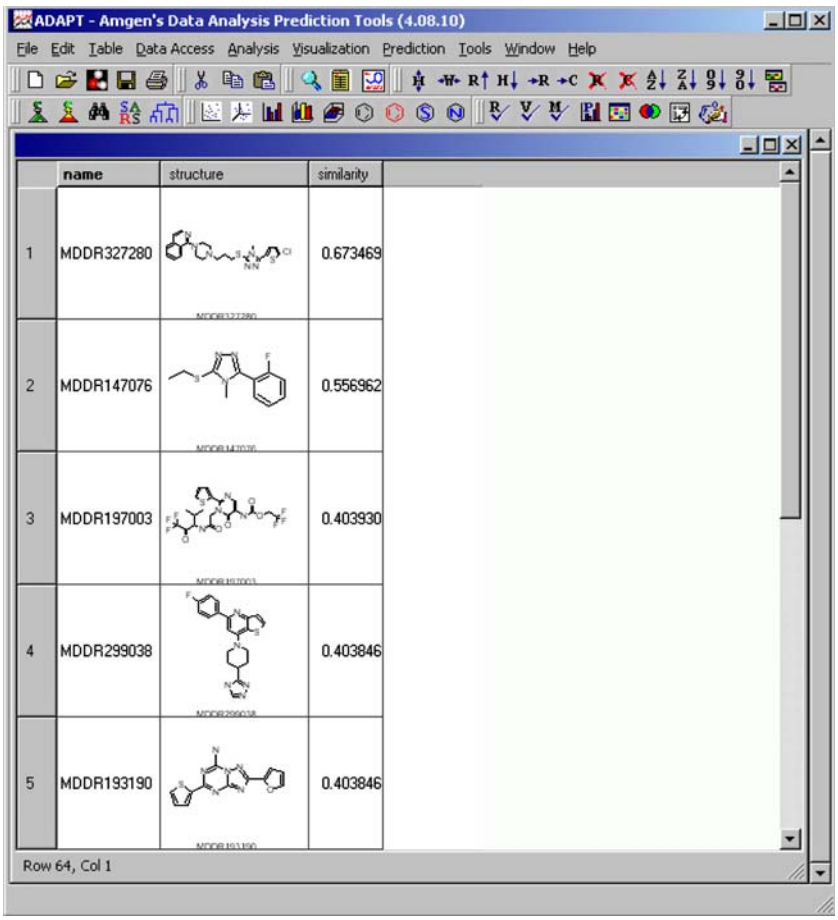


Fig. 1 The Windows 2000 version of the ADAAPT client with a table containing MDDR compounds and their structures

centric view, compound centric view, and project centric view. Each view is an intuitive dialog box designed to easily access information stored in the database. The assay centric view is used to access data via assay codes. An assay code is an identification code given to an assay when it is registered. It is a compact way to represent an assay system but lacks description. The assay centric view provides a convenient way to search for the assay code. A user can simply type in a keyword, and the dialog box will narrow down the list of codes in a drop down list box. When a specific code is selected, associated result type will be available for a selection. Optional date and value cutoffs can be entered in order to reduce the search results. Multiple assay codes may be selected. The query can be also saved for later use. The search can be performed to extract all compounds tested against selected assays, or a list of compounds can be specified. The search results contain the Amgen compound ID, assay code, and result type, value, unit, description, and date. Once the raw data is received from the server, pivoting of the data is performed automatically. The compound centric view is a way to access data using a compound ID rather than the assay code. This view retrieves all stored information pertaining to the specified compound. Results are conveniently grouped into HTS, project data, PK/ADME Data, Property, and Toxicological Data. A user has an option to extract all results, or just those of a specified result type.

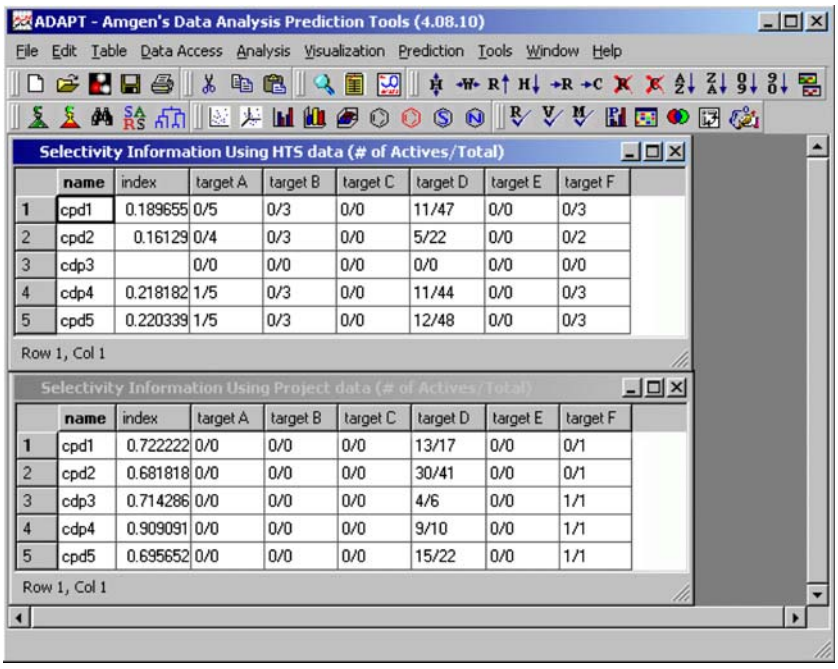
The project centric view is a way to access data using a project name. Unlike the other two views, the project centric view requires a project leader to create a project information page by adding appropriate assays and representative compounds. The project leader can also choose to add a contact person, critical issues, additional descriptions of assay codes, and desired compound profiles. The goal is to create a one-page project summary, so that anyone interested in the project can get a current status of the project easily. Once the page is created, all compounds tested for each project are retrieved and cells in the project summary page are automatically populated. This eliminates multiple, similar database queries requested by many scientists working in a same project.

Other useful data access features include protein structure information and selectivity information. The protein structure information page is a graphical user interface to access in-house protein ligand complexes. Searches can be performed using the protein name and/or compound identification number. Selected protein and/or protein ligand complexes can be selected and imported directly to the DS Modeling Viewer (version 5.0) [16]. Multiple structures can be imported. Because similar protein-ligand complexes are pre-aligned before being deposited into the database, users can quickly examine how various ligands bind to the same protein. A similarity search can be also performed against ligands complexed in proteins, enabling a user to examine how a new compound might bind to a protein. The selectivity information is similar to the compound centric view described earlier, in that all assay results are associated with the specified list of compounds. The difference is that rather than looking at the individual assay results, they can be grouped together into different target classes and examined for their selectivity. Figure 2 shows a screen shot of the selectivity information of five compounds. Because compounds in the database are routinely screened, it is difficult to keep track of what their selectivity profile looks like. With this feature, users can easily identify which target classes compounds were screened against and how many times these compounds were considered as a hit. Answers to such questions do exist in our data warehouse and need to be asked more.

Analysis module

The data analysis module consists of three main submenus: property calculation, database search, and computational tools. Property calculations can be performed on compounds in the data warehouse as well as new compounds. The structure can be copied to the table in ADAAPT client via copying and pasting ChemDraw [19] or ISIS/draw [20] structures directly. Users also have an option to input structures as SMILES [21], and SD files can be also read in directly (SMILES are generated automatically when the files are read). Users have an option to select and unselect types of properties to be calculated or use the default selection. The properties of selected structures are appended adjacent to the selected list and are color-coded according to predetermined rules (Fig. 3); green, yellow, magenta, and red colors are used to indicate good, warning, unreliable, and bad, respectively. Most of our compounds are preprocessed, and their calculated properties can be displayed extremely fast. Compounds not in our database, however, can take a few seconds to several minutes depending on the number of compounds and types of selected properties.

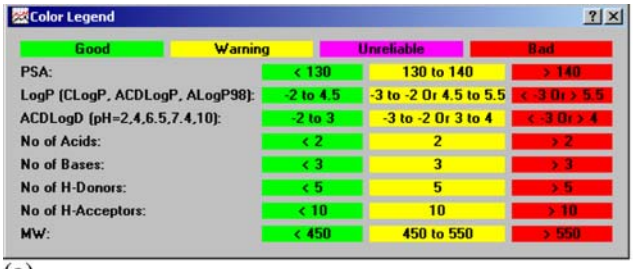
The Database Search menu provides a convenient way to perform similarity and substructure searches against various databases or lists of compounds. The search can be done using a compound ID, SMILES, or SMARTS [22]. A bit string similarity search can be performed using Daylight fingerprint [23], atom pair



	name	index	target A	target B	target C	target D	target E	target F
1	cpd1	0.189655	0/5	0/3	0/0	11/47	0/0	0/3
2	cpd2	0.16129	0/4	0/3	0/0	5/22	0/0	0/2
3	cdp3		0/0	0/0	0/0	0/0	0/0	0/0
4	cdp4	0.218182	1/5	0/3	0/0	11/44	0/0	0/3
5	cpd5	0.220339	1/5	0/3	0/0	12/48	0/0	0/3

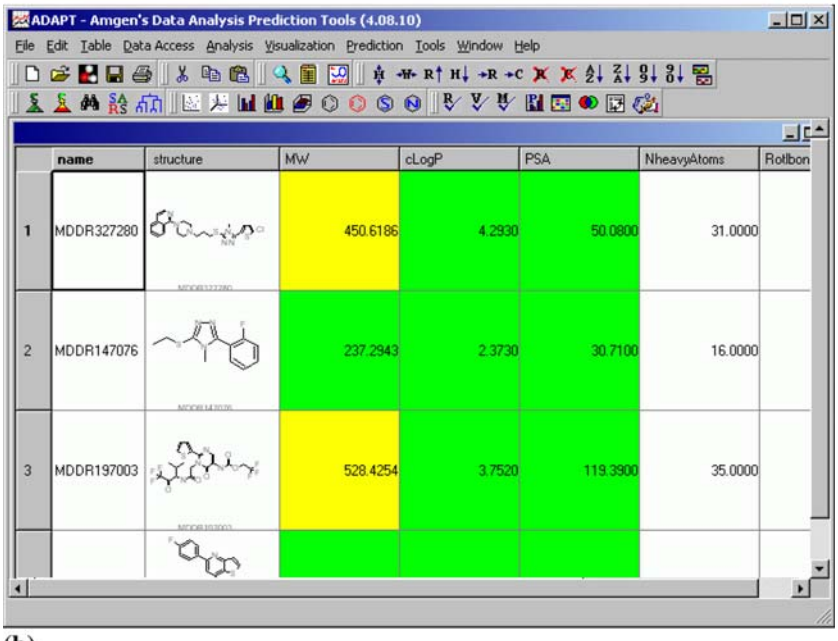
	name	index	target A	target B	target C	target D	target E	target F
1	cpd1	0.722222	0/0	0/0	0/0	13/17	0/0	0/1
2	cpd2	0.681818	0/0	0/0	0/0	30/41	0/0	0/1
3	cdp3	0.714286	0/0	0/0	0/0	4/6	0/0	1/1
4	cdp4	0.909091	0/0	0/0	0/0	9/10	0/0	1/1
5	cpd5	0.695652	0/0	0/0	0/0	15/22	0/0	1/1

Fig. 2 Get Selectivity Information dialog box. Two tables containing HTS and project selectivity data of five compounds are shown



	Good	Warning	Unreliable	Bad
PSA:	< 130	130 to 140		> 140
LogP (CLogP, ACDLogP, ALogP98):	-2 to 4.5	-3 to -2 Or 4.5 to 5.5		< -3 Or > 5.5
ACDLogD (pH=2,4,6.5,7.4,10):	-2 to 3	-3 to -2 Or 3 to 4		< -3 Or > 4
No of Acids:	< 2	2		> 2
No of Bases:	< 3	3		> 3
No of H-Donors:	< 5	5		> 5
No of H-Acceptors:	< 10	10		> 10
MW:	< 450	450 to 550		> 550

(a)



	name	structure	MW	cLogP	PSA	NheavyAtoms	RotBon
1	MDDR327280		450.6186	4.2930	50.0800	31.0000	
2	MDDR147076		237.2943	2.3730	30.7100	16.0000	
3	MDDR197003		528.4254	3.7520	119.3900	35.0000	

(b)

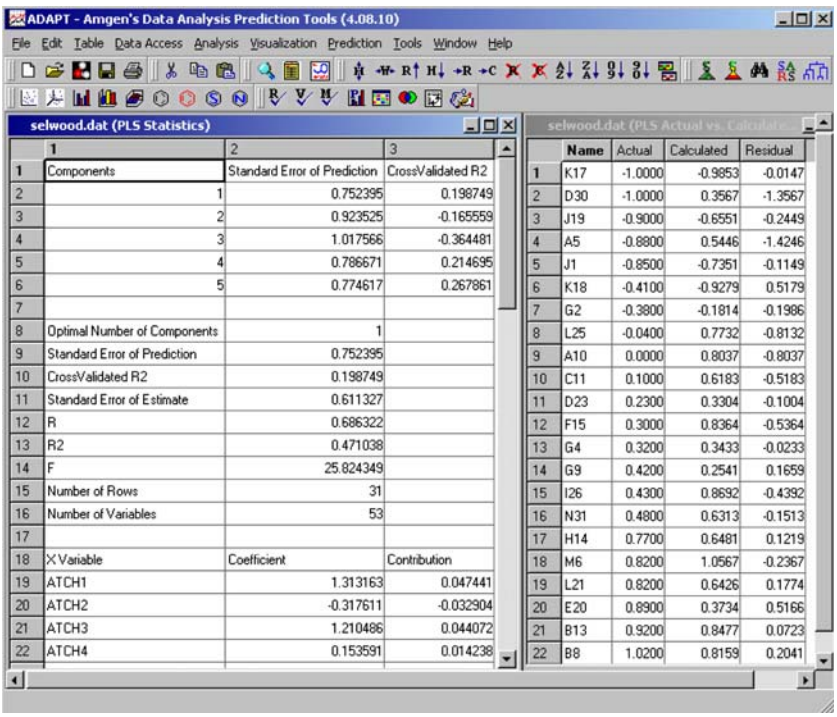
Fig. 3 Color Legend window (**a**) and a table containing structures and colored calculated properties (**b**)

[24], or binding property pair [25, 26] descriptors. Again structures can be copied directly from ChemDraw or ISIS draw. In order to expedite the similarity search, all new compounds are searched against all databases as the part of daily preprocessing procedure.

The Computational Tools menu contains tools to perform many useful functions such as generating similarity and correlation matrices, performing multivariate statistical analyses such as principal component analysis (PCA) [27], multiple linear regression (MLR) [28], and partial least squares (PLS) [29, 30] performing clustering such as Jarvis-Patrick [31] and hierarchical clustering [32], performing sampling such as diverse ordering and diverse selection, and structure activity relationship (SAR) sensitivity analysis. Detailed description and application of each method is beyond the scope of this paper. Instead, how such tools are used in ADAAPT will be illustrated.

Figure 4 shows a PLS analysis performed using the Selwood [33] dataset. Unlike previous computational tools, which are performed on the client side, Jarvis-Patrick and hierarchical clustering are performed on the server side in order to simplify the clustering setup. The selected compounds are sent to the middle layer, where Daylight fingerprints are extracted and necessary clustering algorithms are applied. The clustering results are then sent back to the client. The clustering results can be sorted according to either their original order or clustering order. For hierarchical clustering results, three different levels (0.5, 0.6, and 0.7 similarity cutoffs) are shown by default, to eliminate the difficulty analyzing the full tree. An interactive hierarchical clustering result viewer is available to examine the full tree. However, we found that users tend to be overwhelmed with the number of nodes and levels found in the tree, and we decided to simplify the hierarchical clustering results by creating three different lists using 0.5, 0.6, and 0.7 similarity cutoffs. These three cutoffs typically produce representative lists. If they require further analysis, the interactive hierarchical clustering result viewer can be used. The cluster can also be ranked and sorted according to the most active compounds found in each cluster when such activity information exists. The diverse ordering and selection methods [34], which are also performed in the middle layer, provide the means to select diverse sets of compounds. The major difference between two approaches is that the diverse selection maintains the list order, whereas the diverse ordering does not; rather the diverse selection selects a list based on a specified similarity cutoff.

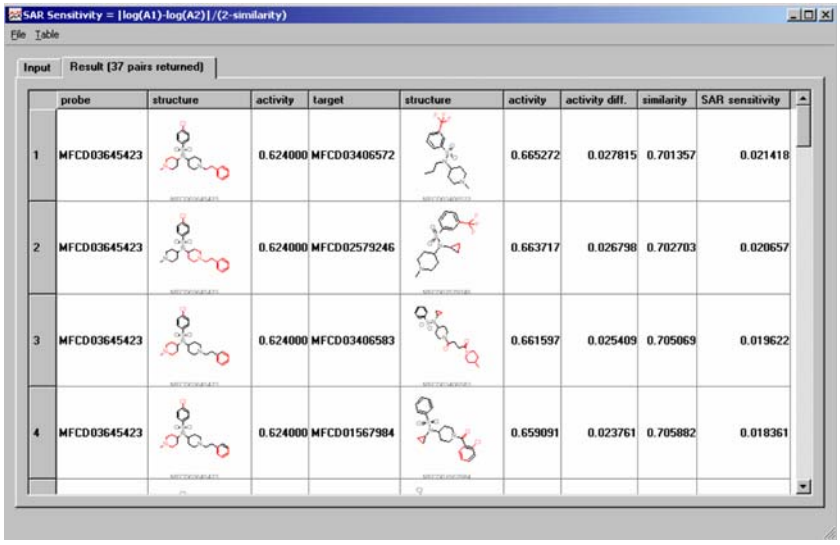
The SAR Sensitivity menu provides a way to visually inspect what structural differences give rise to activity differences. This can be performed using one or multiple probes against many targets, ranging from moderately active to inactive compounds. When the calculation button is pressed, the user has the option to specify what similarity cutoff to use. The default value is 0.7, meaning any pair with a similarity greater than 0.7 will be saved, and a maximum common substruc-



	1	2	3
1	Components	Standard Error of Prediction	CrossValidated R2
2	1	0.752395	0.198749
3	2	0.923525	-0.165559
4	3	1.017566	-0.364481
5	4	0.786671	0.214695
6	5	0.774617	0.267861
7			
8	Optimal Number of Components	1	
9	Standard Error of Prediction	0.752395	
10	CrossValidated R2	0.198749	
11	Standard Error of Estimate	0.611327	
12	R	0.686322	
13	R2	0.471038	
14	F	25.824349	
15	Number of Rows	31	
16	Number of Variables	53	
17			
18	X Variable	Coefficient	Contribution
19	ATCH1	1.313163	0.047441
20	ATCH2	-0.317611	-0.032904
21	ATCH3	1.210486	0.044072
22	ATCH4	0.153591	0.014238

	Name	Actual	Calculated	Residual
1	K17	-1.0000	-0.9853	-0.0147
2	D30	-1.0000	0.3567	-1.3567
3	J19	-0.9000	-0.6551	-0.2449
4	A5	-0.8800	0.5446	-1.4246
5	J1	-0.8500	-0.7351	-0.1149
6	K18	-0.4100	-0.9279	0.5179
7	G2	-0.3800	-0.1814	-0.1986
8	L25	-0.0400	0.7732	-0.8132
9	A10	0.0000	0.8037	-0.8037
10	C11	0.1000	0.6183	-0.5183
11	D23	0.2300	0.3304	-0.1004
12	F15	0.3000	0.8364	-0.5364
13	G4	0.3200	0.3433	-0.0233
14	G9	0.4200	0.2541	0.1659
15	I26	0.4300	0.8692	-0.4392
16	N31	0.4800	0.6313	-0.1513
17	H14	0.7700	0.6481	0.1219
18	M6	0.8200	1.0567	-0.2367
19	L21	0.8200	0.6426	0.1774
20	E20	0.8900	0.3734	0.5166
21	B13	0.9200	0.8477	0.0723
22	B8	1.0200	0.8159	0.2041

Fig. 4 PLS analysis example using the Selwood dataset. Tables containing PLS statistics and actual, calculated, and residual values are shown



	probe	structure	activity	target	structure	activity	activity diff.	similarity	SAR sensitivity
1	MFCD03645423		0.624000	MFCD03406572		0.665272	0.027815	0.701357	0.021418
2	MFCD03645423		0.624000	MFCD02579246		0.663717	0.026798	0.702703	0.020657
3	MFCD03645423		0.624000	MFCD03406583		0.661597	0.025409	0.705069	0.019622
4	MFCD03645423		0.624000	MFCD01567984		0.659091	0.023761	0.705882	0.018361

Fig. 5 SAR sensitivity example using a set of ACD compounds. Input dialog box is used to set probe and target compounds. Similarity cutoff value is specified to reduce the number of comparison pairs. The results are summarized in a table

ture of each pair is then determined. The structural difference is highlighted in red. Structurally similar pairs with a large activity difference will give a high SAR sensitivity score (Fig. 5).

Prediction module

The prediction module consists of three selections: calculate a drug-likeness score, calculate an absorption index, and perform predictions using an QSAR equation. The drug-likeness score and absorption index are our implementation of Xu and Stevenson's drug-like index (DLI) [35] and Egan et al.'s 95% and 99% confidence ellipses [36], respectively. Figure 6 shows several sample drug-likeness scores for Available Chemicals Directory (ACD)[1] compounds; higher score indicates greater drug-likeness. Figure 7 depicts how the absorption index is calculated. Compounds located inside of the 99% confidence ellipse (inner ellipse) are given a value of 0. Compounds located between the 99% and 95% confidence ellipses are given a value of 1, and compounds located outside of the 95% confidence ellipse are given a value of 2. The last selection under the prediction module utilizes existing QSAR models generated using ADAAPT's PLS or MLR functions to perform a prediction. Figure 8 shows the user interface, which can be used to specify a table containing new compounds and a QSAR model.

Tools module

The difference between some of the features in the data analysis and tool modules is somewhat arbitrary since both modules require compound information and/or assay results. Both perform analyses whether it is a structure activity/property relationship analysis or a structural analysis. Most of the time, however, the tools module contains various experimental tools, which might be useful to end users. It is a convenient way to interact with users to explore novel ideas. Examples include data visualization, compound profiling, R-group decomposition, and table manipulation tools. For data visualization, 2D scatter, 2D bar, 3D scatter, 3D bar, and 3D surface plots are available (Fig. 9). All plots can be generated by simply highlighting rows and columns in a table. When a data point represents a compound, its 2D structure can be easily displayed by clicking it with a mouse. Data points can be also selected from the plot and copied to a new table. For the 2D scatter plot, any two columns in a table can be easily selected to be plotted against each other (Fig. 10). It is also possible to represent the size of data points with the values in the third column of the table. All 3D plots can be rotated, translated, and scaled to interactively explore and find a convenient orientation for the visual inspection of the data. Compound profiling is an interactive filtering tool, using properties (PSA, ClogP, H-Donor, H-Acceptor, molecular weight, and rotatable bonds) and chemotypes (23 bad groups, 90 functional groups, 51 reactive groups) to narrow down a list of compounds (Fig. 11). Vertical and horizontal bars, representing the property and chemotype distributions, change in real time as compounds are included or excluded from the list. Modifying the PSA range, for example, changes the distribution of other properties as well as the chemotypes. The 2D structures of compounds can be displayed by clicking a vertical or a horizontal bar

[1] Available Chemicals Directory, MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577

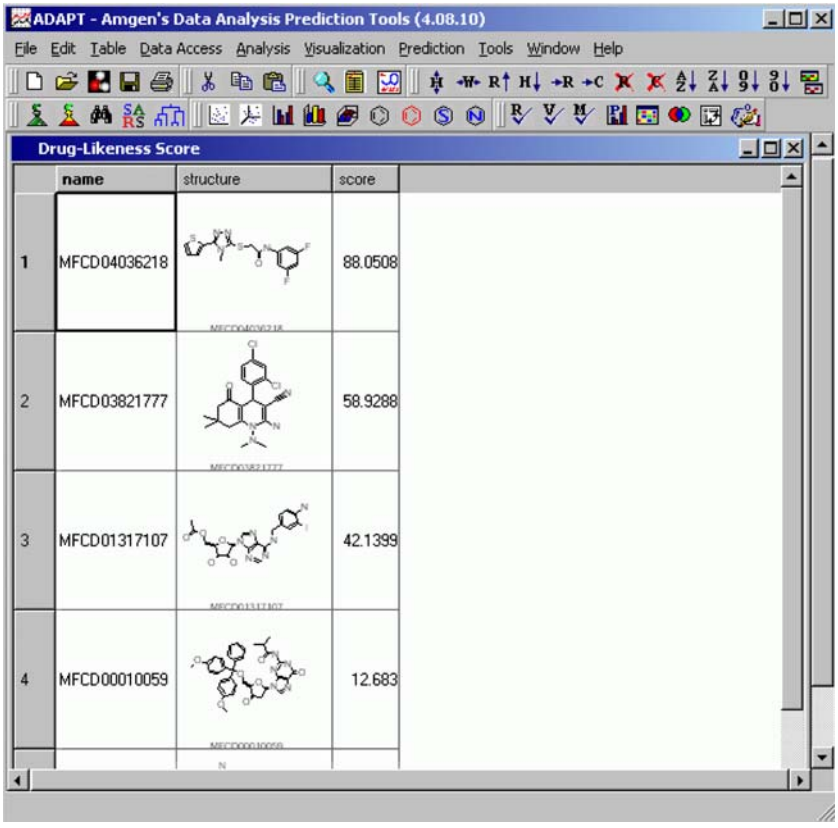


Fig. 6 Drug-likeness score calculation example using a set of ACD compounds

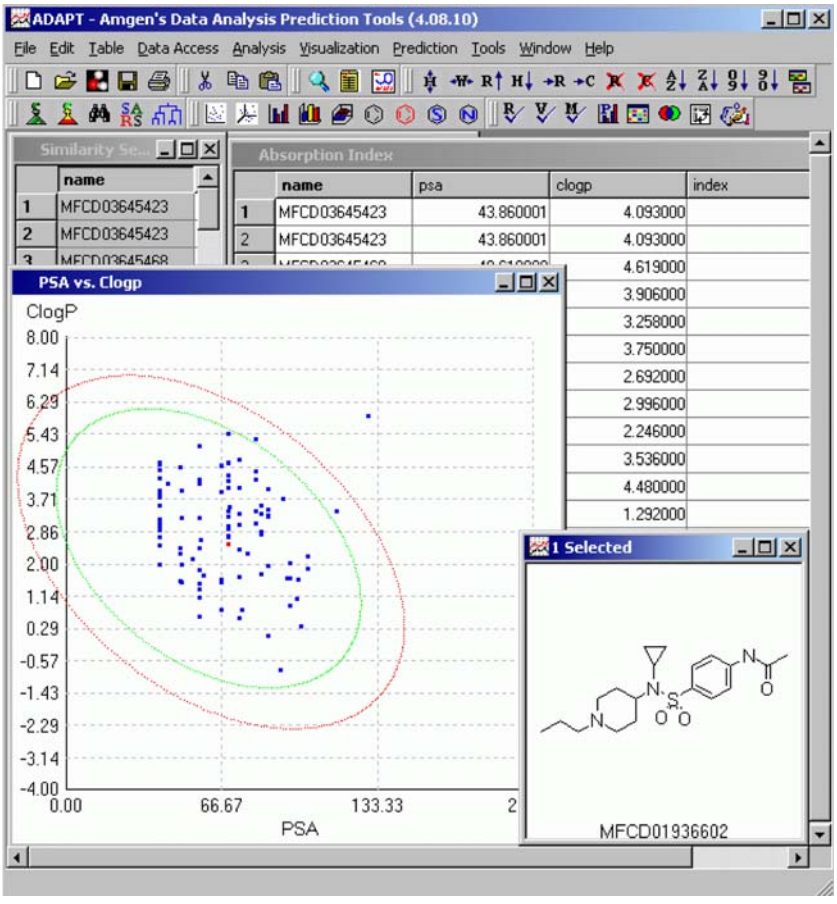
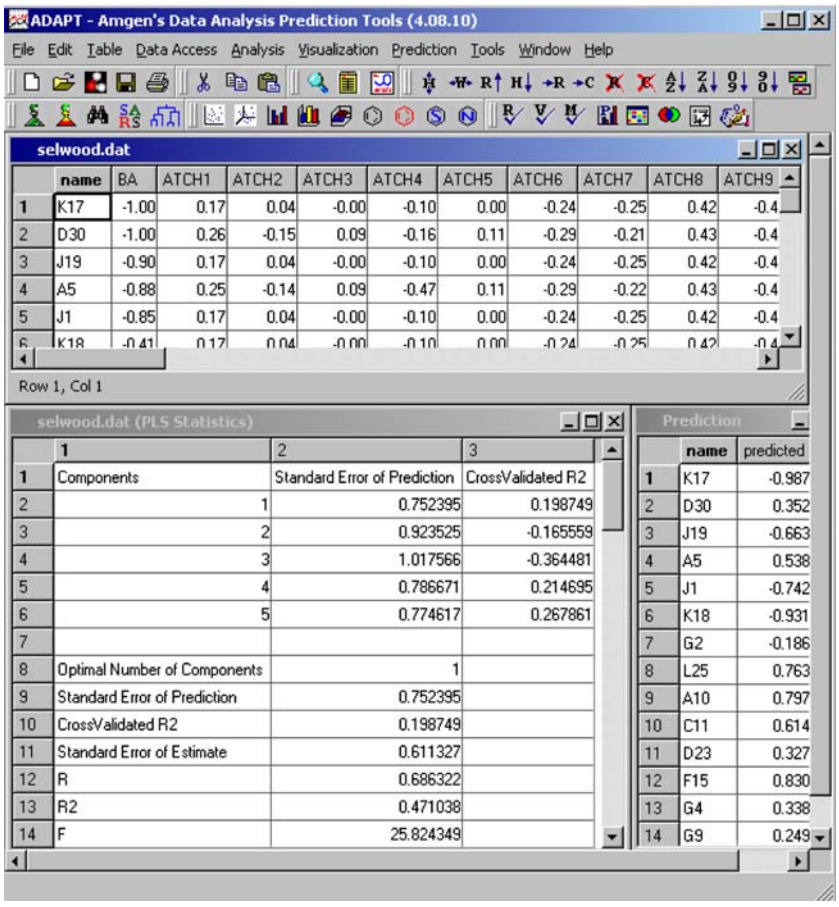


Fig. 7 Absorption index calculation example using a set of ACD compounds. PSA vs. ClogP plot is shown, and the structure of a data point colored in red is displayed



	name	BA	ATCH1	ATCH2	ATCH3	ATCH4	ATCH5	ATCH6	ATCH7	ATCH8	ATCH9
1	K17	-1.00	0.17	0.04	-0.00	-0.10	0.00	-0.24	-0.25	0.42	-0.4
2	D30	-1.00	0.26	-0.15	0.09	-0.16	0.11	-0.29	-0.21	0.43	-0.4
3	J19	-0.90	0.17	0.04	-0.00	-0.10	0.00	-0.24	-0.25	0.42	-0.4
4	A5	-0.88	0.25	-0.14	0.09	-0.47	0.11	-0.29	-0.22	0.43	-0.4
5	J1	-0.85	0.17	0.04	-0.00	-0.10	0.00	-0.24	-0.25	0.42	-0.4
6	K18	-0.41	0.17	0.04	-0.00	-0.10	0.00	-0.24	-0.25	0.42	-0.4

	1	2	3
1	Components	Standard Error of Prediction	CrossValidated R2
2	1	0.752395	0.198749
3	2	0.923525	-0.165559
4	3	1.017566	-0.364481
5	4	0.786671	0.214695
6	5	0.774617	0.267861
7			
8	Optimal Number of Components	1	
9	Standard Error of Prediction	0.752395	
10	CrossValidated R2	0.198749	
11	Standard Error of Estimate	0.611327	
12	R	0.686322	
13	R2	0.471038	
14	F	25.824349	

	name	predicted
1	K17	-0.987
2	D30	0.352
3	J19	-0.663
4	A5	0.538
5	J1	-0.742
6	K18	-0.931
7	G2	-0.186
8	L25	0.763
9	A10	0.797
10	C11	0.614
11	D23	0.327
12	F15	0.830
13	G4	0.338
14	G9	0.249

Fig. 8 Perform Prediction example using the Selwood dataset. A table containing the predicted values is shown

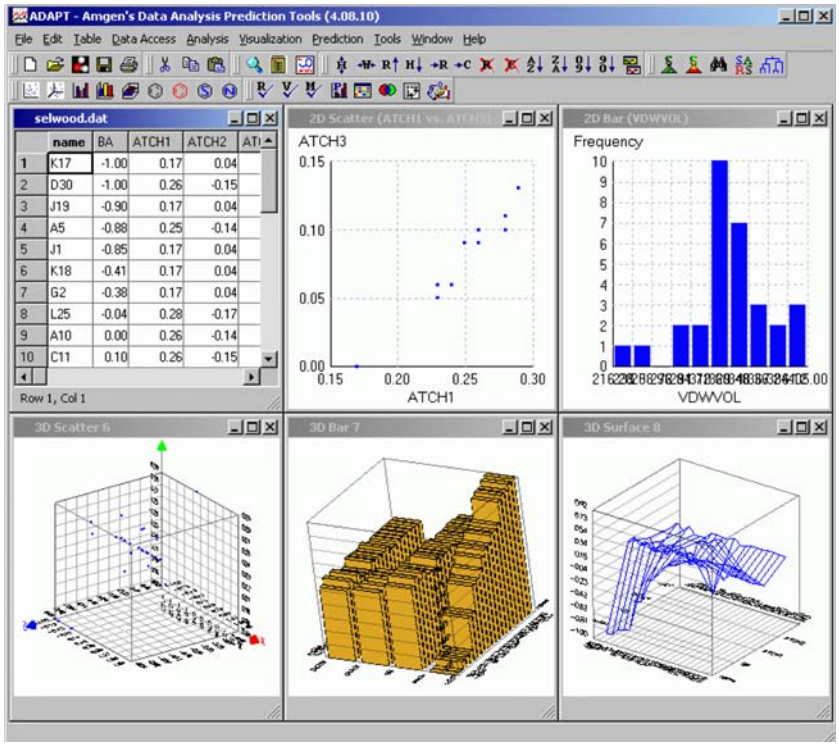


Fig. 9 Data visualization examples using the Selwood dataset. 2D scatter, 2D bar, 3D scatter, 3D bar, and 3D surface plots are shown

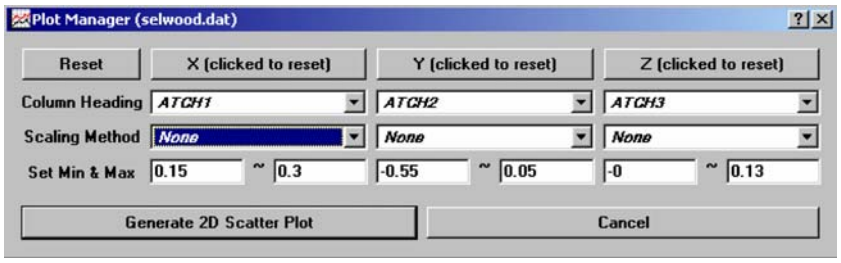


Fig. 10 Plot manager example using the Selwood dataset

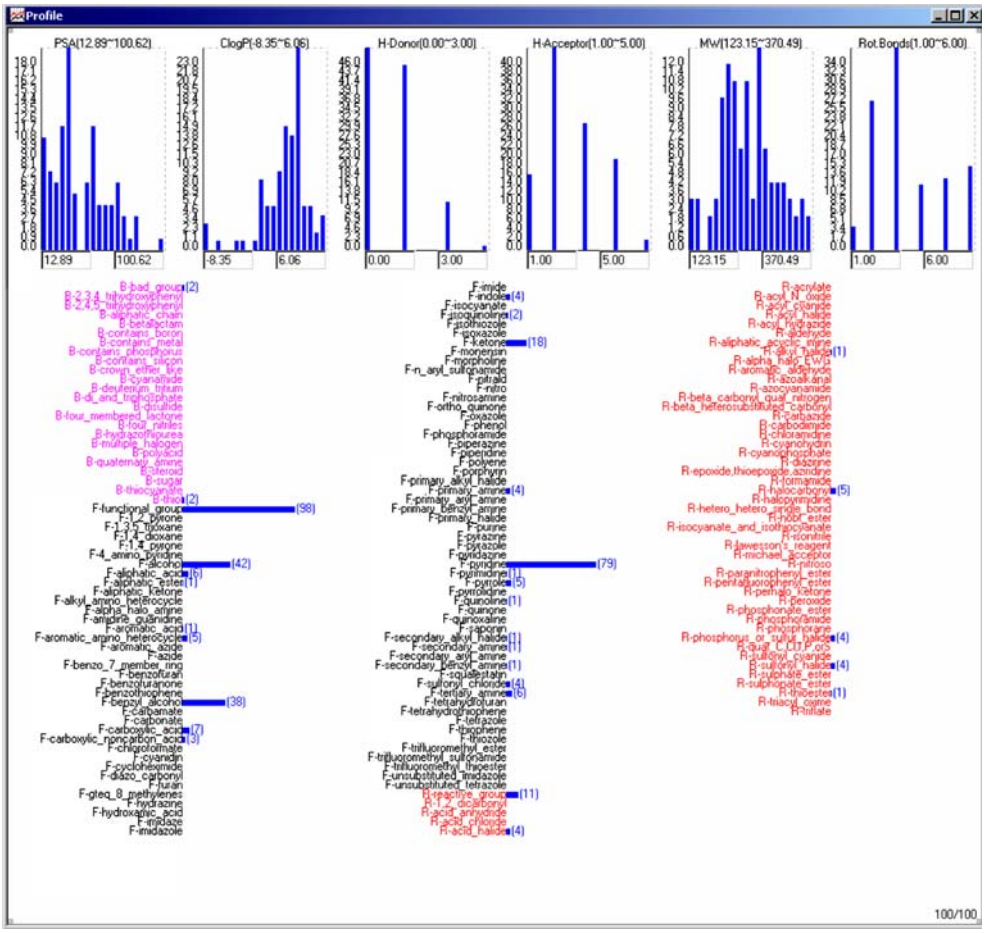


Fig. 11 Compound profiling example using a set of ACD compounds. The compounds are profiled according to their properties (PSA, ClogP, H-Donor, H-Acceptor, molecular weight, and rotatable bonds) and chemotypes (23 bad groups, 90 functional groups, 51 reactive groups)

with a mouse. The final list of compounds can then be copied back to the ADAAPT table. R-group decomposition is another experimental tool available in Tool module (Fig. 12). Once a scaffold and certain attachment points are defined, this tool decomposes selected compounds according to this fragmenting rule (Fig. 12a). Figure 12b shows the R-group decomposition results. Finally, table manipulation tools include a number of simple but extremely useful features such as filtering, coloring, and managing lists. The goal is to provide an easy solution to a common task that are frequently encountered by our scientists. Filtering selects rows in a table based on the number of criteria. This tool is similar to the ''find'' feature available in a typical Windows application but is designed to deal with the table format appearing in our work flow. ''Not Blank'' filter mode, for example, selects rows in a table which do not contain any blank cells and creates a new table. Table coloring is used when a user wishes to emphasize a set of values in a table using color. This is especially useful when a table contains multiple columns containing different screening results. Good and bad screen results, for example, can be colored in green and red, respectively. Up to six different colors are available. The list manager provides an easy way to perform Boolean-

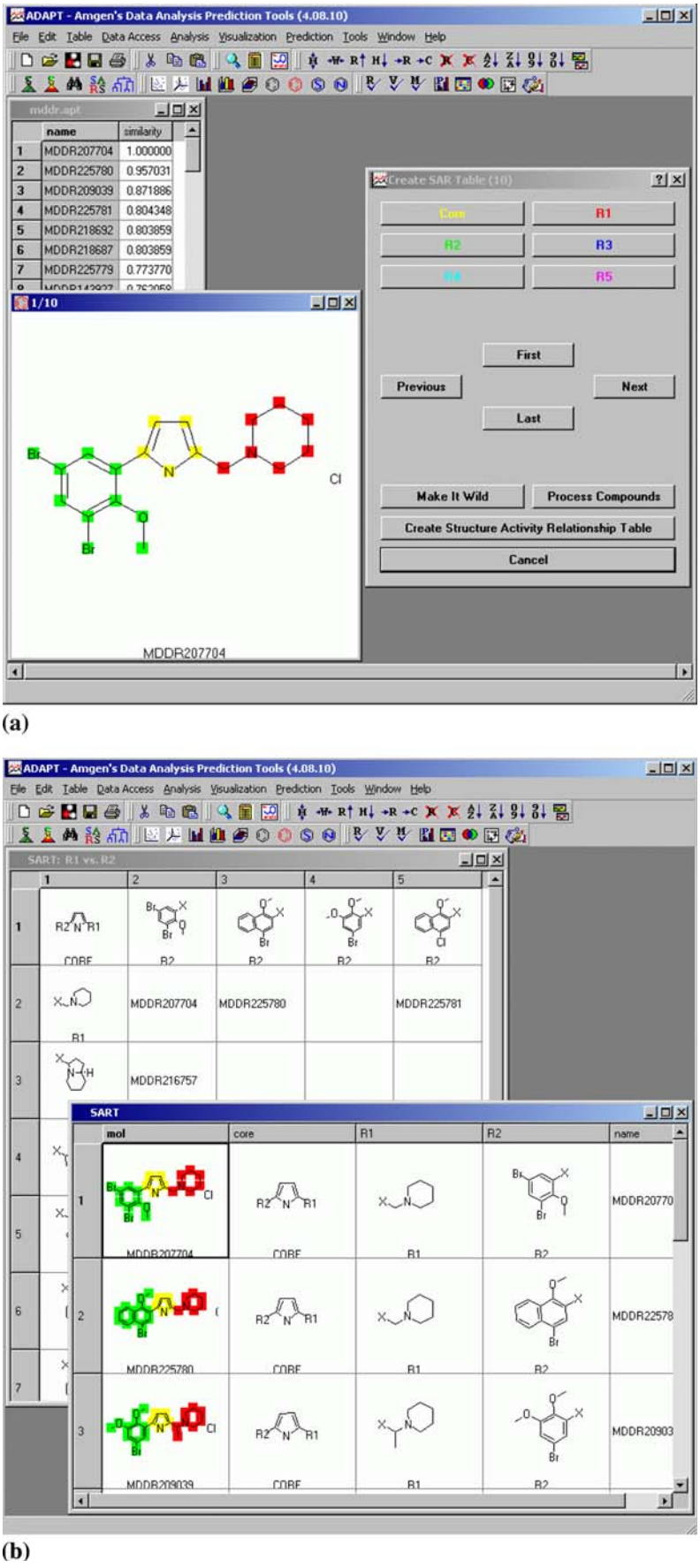


Fig. 12 R-group decomposition example using a set of MDDR compounds. Decomposition rule generator (**a**) is used to assign R-groups. Decomposed compounds are shown in two different ways (**b**)

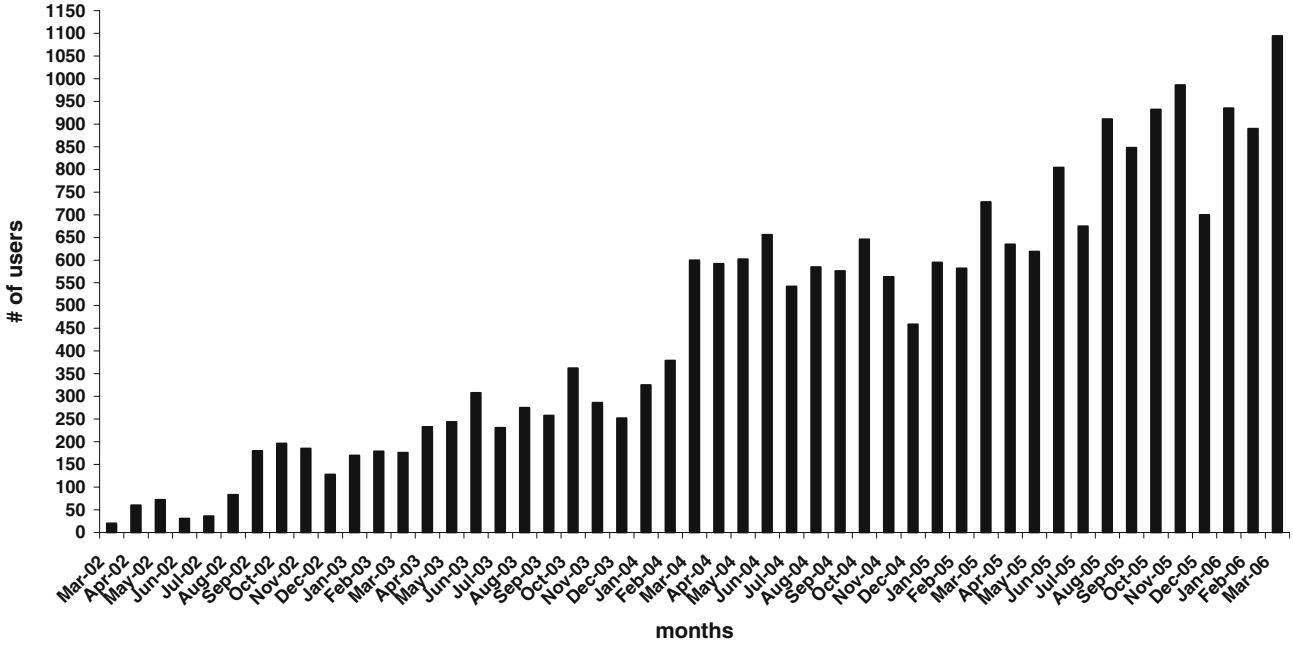


Fig. 13 ADAAPT usage

like ''and'' (find common rows), ''or'' (merge rows), or ''not'' (find rows not in the second table) operations on two tables, using specified key columns.

Usage

In a typical project, our chemistry project leader's workflow, for example, begins with creating a project information page, so that anyone interested in the project can get a current status of the project easily. A SAR table can be created from the project information page, and SAR sensitivity (Fig. 5) calculation and R-group decomposition (Fig. 12) are performed to visually inspect the biologically important structural features. A database search is performed to identify compounds structurally similar to actives. A QSAR model can be created by computing descriptors (Fig. 3) and performing PLS analysis (Fig. 4). Selectivity information (Fig. 2) is useful to identify what compounds to avoid. Proposed new structural changes can be prioritized by their drug-likeness score (Fig. 6), absorption index (Fig. 7), profiling (Fig. 11), and predicted activities (Fig. 8). This process is then repeated with more activity data. This is one of myriad ways ADAPT system can be used.

The success of a system such as ADAAPT does not depend on how many useful features it has or how easily it can be deployed. The success, in our view, is measured by how many managers and scientists utilize the system. A tool, computational or otherwise, with no users is of no value. ADAAPT usage statistics can be monitored, and Fig. 13 shows the number of users per months since ADAAPT was created. It is encouraging to see how ADAAPT usage has increased over time. This increase in usage reflects a number of factors, such as enhanced usability of features, increases in the number of available features, as well as the growth of our organization. A breakdown of users according to their departments shows that scientists from various groups are using the system on a daily basis (data not shown), and the fact that users now include non-scientists such as patent lawyers and business analysts is a real testament to its success.

Conclusion

We have designed and implemented the ADAAPT system to provide flexible access and analysis of chemical and biological data at Amgen. The system runs on multiple platforms and can be easily deployed. The features available in ADAAPT system can be easily modified, and new features can be added with a very short development cycle, without reinstallation of client-side software. In this paper, we have reported the four main modules in ADAAPT: access, analysis, prediction, and tools. They represent the ability to retrieve both structural and biological data, the ability to analyze and relate structural information to biological information, the ability to share QSAR/QSPR models, and the ability to enhance productivity. The list of

features available in each module is by no means complete and will continue to grow as users' needs evolve, making ADAAPT a truly adaptable system that evolves iteratively with its needs to provide proactive solutions.

Acknowledgement We are grateful for helpful discussions and feedback from other members of the modeling group at Amgen. We especially thank Michael Bartberger, Matthew Lee, and Mick Kappler for their critical reading of the manuscript.

Reference

1. Ahlberg C (1999) Drug Discov. Today 4:370
2. Fay N, Ullmann D (2002) Drug Discov. Today (information biotechnology suppl.) 7:S181
3. Claus BL, Underwood DJ (2002) Drug Discov. Today 7:957
4. Peakman T, Franks S, White C, Beggs M (2003) Drug Discov. Today 8:203
5. Trepalin SV, Yarkov AV (2001) J. Chem. Inf. Comput. Sci. 41:100
6. Ihlenfeldt W-D, Voigt JH, Bienfait B, Oellien F, Nicklaus MC (2002) J. Chem. Inf. Comput. Sci. 42:46
7. Adams N, Schubert US (2004) J. Comb. Chem. 6:12
8. Gobbi A, Funeriu S, Ioannou J, Wang J, Lee M-L, Palmer C, Bamford B (2004) J. Chem. Inf. Comput. Sci. 44:964
9. Microsoft Corporation; http://www.microsoft.com
10. Daylight Chemical Information Systems, Inc.; http://www.daylight.com
11. (a) Wall L Practical extraction and report language, version 5.005_03; http://www.perl.com. (b) Wall L, Christiansen T, Schwartz RL (1996) Programming Perl, 2nd edn. O'Reilly & Associates, Inc
12. Segal M, Akeley K (2003) The OpenGL graphic system: a specification (Version 1.5); Silicon Graphics, Inc. http://www.opengl.org/documentation/specs/version1.5/glspec15.pdf
13. Trolltech, Inc.; http://www.trolltech.com
14. Red Hat Inc.; http://www.redhat.com
15. Silicon Graphics, Inc.; http://www.sgi.com
16. Accelrys; http://www.accelrys.com
17. Advanced Chemistry Development, Inc.; http://www.acdlabs.com
18. Open LDAP Foundation; http://www.openldap.org
19. CambridgeSoft Corp., http://www.cambridgesoft.com
20. MDL Information System, http://www.mdli.com
21. Daylight Theory Manual, http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html
22. Daylight Theory Manual, http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html
23. Daylight Theory Manual, http://www.daylight.com/dayhtml/doc/theory/theory.finger.html
24. Carhart RE, Smith DH (1985) J. Chem. Inf. Comput. Sci. 25:64
25. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) J. Chem. Inf Comput. Sci. 36:118
26. Cho SJ, Shen CF, Hermsmeier MA (2000) J. Chem. Inf Comput. Sci. 40:668
27. Glen WG, Dunn WJ, Scott DR (1989) Tetrahedron Comput. Methodol. 2:349
28. Draper NR, Smith H (1966) Applied regression analysis. Wiley, New York
29. Wold S, Dunn WJ III (1983) J. Chem. Inf: Comput. Sci. 23:6
30. Collantes ER, Dunn WJ III (1995) J. Med. Chem. 38:2705
31. Jarvis RA, Patrick EA (1973) IEEE Trans. Comp. C-22:1025
32. Ward JH (1963) J. Am. Stat. Assoc. 58:236
33. Selwood DL, Livingstone DJ, Comley JCW, O'Dowd AB, Hudson AT, Jackson P, Jandu KS, Rose VS, Stables JN (1990) J. Med. Chem. 33:136
34. Higgs RE, Bemis KG, Watson IA, Wikel JH (1997) J. Chem. Inf. Comput. Sci. 37:861
35. Xu J, Stevenson J (2000) J. Chem. Inf Comput. Sci. 40:1177
36. Egan WJ, Merz KM Jr, Baldwin JJ (2000) J. Med. Chem. 43:3867