Open Access

Internet in Drug Design and Discovery

Igor V. Tetko*

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute for Bioinformatics and Systems Biology, Neuherberg, D-85764, Germany and Institute of Bioorganic & Petrochemistry, Ukrainian Academy of Sciences, Kyiv-94, 02660, Ukraine

Abstract: The development of Internet technologies and the WWW has dramatically influenced all aspects of modern life. Is their development also beneficial for drug discovery? This article reviews advantages of web technologies and their influence on the drug discovery process.

INTRANET TOOLS IN PHARMA COMPANIES

Currently, major large pharma companies use Web technology for the Intranet data analysis of small molecules [1-5]. There are at least several clear advantages why companies are investing in the development of web technologies.

- **Available everywhere.** The Web tools are accessible on different platforms thus allowing the users to work in the environment they prefer. This eliminates need to spend resources for training of end-users to work with program installed on specialized systems such as Linux, Unix, etc.
- **Decreased support costs.** The web technologies dramatically decrease costs for installation, support and upgrade of the software. The most recent versions of the software tools become immediately and simultaneously available for all users within the corporate Intranets following the upgrade of the software at the central server.
- **Decreased license costs.** The licensing of the software for web use within corporate Intranet can be much cheaper compared to multiple licenses required to install stand-alone versions at multiple computers.
- **Software integration.** Software tools developed using different programming languages (Perl, Python, C++, Java) and running on different platforms (Windows, Mac, Windows, Linux) can be easily integrated. This also includes support of specialized/legacy software. The old versions, which can run only under outdated versions of systems (i.e., Mac Os 9, Silicon Graphics) can be easily integrated into the Web tools and be available to all users.
- **Easy monitoring.** The Web tools allow an easy monitoring of the usage of each particular tool and may help to decide which products should be further supported and which ones are not required/have limited use and thus can be removed or/and changed to single licenses.
- **Grid calculations.** Time demanding calculations, such as docking or quantum chemical calculations can be sent in parallel to multiple computers over Intranet thus efficiently using resources and increasing productivity.

The training of employers to use Web tools requires less time. The users are usually well familiar to use Internet and browsers. There is also a minimal danger to the misuse of the software and to damage the system or the program by an accident or virus infection. The Web drug discovery tools covers a large range of applications from molecular property calculations, bioisosteric design to virtual screening and docking as summarized in ref [1].

The wide-spread usage of Web tools in the Industry does not however, imply, that the development of such tools is easy. The interface of the Web-tool should be simple and intuitive, have only a minimum number of input parameters and clear logic of data processing as well as are documented in help files. Complex interfaces will increase the possibility of user mistakes and may lead to invalid and nonsense results. All these issues should be carefully considered when developing the Web interfaces.

Since the development of Web tools may require considerable resources, it is not surprisingly that companies offering such services appear. One example of such companies is the Pipeline Pilot from SciTegic (http://accelrys.com). It provides a platform for an easy data integration and publishing *via* Webports.

Of course, because of the proprietary nature of drug discovery, researchers in pharmaceutical industry are usually not allowed submission of their structures and datasets or/and post questions that can disclosure their work to Web outside of the company intranet. Because of these limitations, the scientists cannot easily use public services to post their questions and exchange information. However, it does not prevent them to implement private versions of useful public tools on the in house Intranet, as it is the case with Pfizer's version of Wikipedia, Pfizerpedia [6].

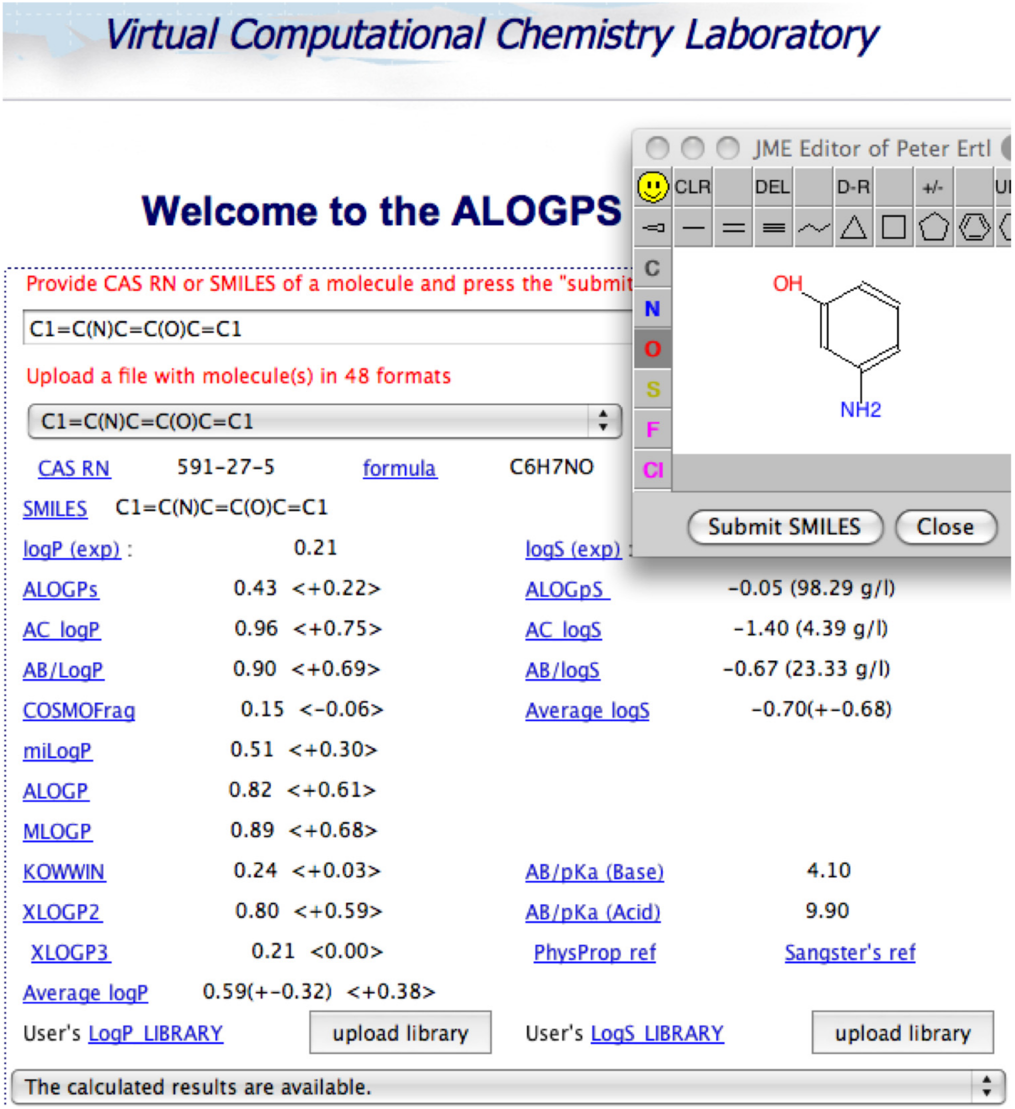
*Address correspondence to this author at the Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH), Institute for Bioinformatics and Systems Biology, Neuherberg, D-85764, Germany and Institute of Bioorganic & Petrochemistry, Ukrainian Academy of Sciences, Kyiv-94, 02660, Ukraine; E-mail: itetko@vcclab.org

PUBLIC WWW RESOURCES

There is a growing interest in developing public and free Web tools for drug discovery on the Internet. These developments are equally supported both by Academia and commercial firms. The publishing or making the software publicly available on the Internet does not contradict the idea of its successful marketing. Indeed, considering the extremely importance of privacy of data in Industry, no commercial users will ever use public resources for analysis of their compounds. Thus, by making software freely available online commercial firms do not loose potential customers. However, freely available accessible tools offer a great possibility of testing the algorithms. If users are satisfied with the results, they can decide to buy the software.

The need for confidentiality is much less in Academia. Of course, academic studies do lead to the development of new drugs. However, the main target of these studies is usually not to develop a final drug molecule but to understand mechanisms of action of compounds. The molecules used in such studies can easily have poor ADME/T properties and are unlikely good lead candidates. Therefore, the value of the molecular structure by itself is not so important in these studies and there are no strict limitations to analyze these molecules at public servers.

Public or freely downloadable resources includes visualization tools (Rasmol http://www.openrasmol.org, Jmol http://jmol.sourceforge.net), software development tools (CDK http://cdk.sourceforge.net), docking software (ArgusLab http://www.arguslab.com), molecular property prediction and calculation of descriptors (PreADME http://preadme.bmdrc.org) are comprehensively reviewed in refs [14-17]. The Virtual Computational Chemistry Laboratory (VCC LAB http://www.vcclab.org) represents an example of such



Welcome to the ALOGPS

Provide CAS RN or SMILES of a molecule and press the "submit

C1=C(N)C=C(O)C=C1

Upload a file with molecule(s) in 48 formats

C1=C(N)C=C(O)C=C1

CAS RN	591-27-5	formula	C6H7NO
SMILES	C1=C(N)C=C(O)C=C1		
logP (exp) :	0.21	logS (exp) :	
ALOGPs	0.43 <+0.22>	ALOGpS	-0.05 (98.29 g/l)
AC logP	0.96 <+0.75>	AC logS	-1.40 (4.39 g/l)
AB/LogP	0.90 <+0.69>	AB/logS	-0.67 (23.33 g/l)
COSMOFrag	0.15 <-0.06>	Average logS	-0.70(+-0.68)
miLogP	0.51 <+0.30>		
ALOGP	0.82 <+0.61>		
MLOGP	0.89 <+0.68>		
KOWWIN	0.24 <+0.03>	AB/pKa (Base)	4.10
XLOGP2	0.80 <+0.59>	AB/pKa (Acid)	9.90
XLOGP3	0.21 <0.00>	PhysProp ref	Sangster's ref
Average logP	0.59(+-0.32) <+0.38>		
User's LogP LIBRARY	upload library	User's LogS LIBRARY	upload library

The calculated results are available.

Fig. (1). Free interactive calculation of physico-chemical properties at Virtual Computational Chemistry Laboratory site (http://www.vcclab.org). The results of ALOGPS [7] as well as of several popular public (KOWWIN [8], XLOGP2 [9] and XLOGP3 [10]) and commercial programs, including Molinspiration (http://www.molinspiration.com), Pharma Algorithms (http://www.ap-algorithms.com), Cosmologic GmBH (http://www.cosmologic.de), Actelion property explorer (http://www.actelion.com), ALOGP [11] and MLOGP [12] (implemented in Dragon software [13]) are displayed at one page. The on-line access to PHYSPROP database (http://www.syrres.com/esc/physdemo.htm) is also used to provide experimental values (if available) of the analyzed compound.

services for drug discovery [18]. It includes tools for physico-chemical property prediction (see Fig. **1**), descriptor generation and data analysis using linear (Partial Least Squares) and non-linear (Associative and Polynomial Neural Networks) computational approaches. Carolina Exploratory Center for Cheminformatics Research (CECCR) (http://ceccr.unc.edu), Cheminformatics Modeling Laboratory (http://eccr.stat.ncsu.edu/ChemModLab), Carolina Cheminformatics Workbench (C-ChemBench http://ceccr.ibiblio.org) are other examples of currently developing on-line tools for QSAR and drug discovery studies.

CHEMO- AND BIOINFORMATICS DATABASES

A large number of chemoinformatics and bioinformatics databases important for drug discovery are available on the Internet. This includes major commercial databases, such as Beilstein (http://www.beilstein.com) and Gmelin databases (http://info.crossfiregmelin.com), which comprises literature data on activities and physico-chemical properties of 8.2 million compounds and 10 million reactions or iResearchLibrary (http://www.chemnavigator.com), which tracks over 46 million compounds available from > 240 chemistry suppliers. These databases can be queried on-line to detect compounds with particular activity or properties to be used, e.g. in structure-activity relationship studies or patent search.

There is also a growing number of free online databases which offering some similar functionality. For example, ZINC database (http://zinc.docking.org/) offers 4.6 million compounds in 3D format that can be used in docking. The PubChem database (http://pubchem.ncbi.nlm.nih.gov) provides information on biological activities of small molecules as part of the NIH Molecular Libraries Roadmap Initiative. It also has a growing repository of about 20 million compounds collected from different sources. All these data can be freely downloaded and used in the drug discovery process. Another example is http://www.chemspider.com, which provides on-line access to about 18 million compounds and also include literature search and property prediction software. The other databases such as NIST databases (http://www.nist.gov/srd/), FDA Drug Database (http://www.accessdata.fda.gov/scripts/cder/drugsatfda), Drug Bank (http://www.drugbank.ca/), Distributed Structure-Searchable Toxicity (DSSTox) Database (http://www.epa.gov/ncct/dsstox) represent more specialized resources that can be used for search of drugs and their properties.

Contrary to chemoinformatics, bioinformatics databases are usually freely available. These are resources for sequence analysis and annotation, e.g. ExPASy, http://expasy.org, Ensembl, http://www.ensembl.org, MIPS, http://mips.gsf.de, metabolomic pathways, e.g. KEGG (http://www.genome.jp/kegg), drug metabolism, e.g. HMDB (http://www.hmdb.ca), Protein data Bank (http://www.wwpdb.org) and others [19]. All these resources can be useful in drug discovery projects for the identification of potential drug targets (e.g., detection of pathogenicity islands by comparison of genomes of pathogen and non-pathogen microorganisms) or analysis of resistance to drugs of microorganisms due to SNP (single nucleotide polymorphism) or analysis of ADME/T properties of investigated compounds.

LITERATURE DATABASES AND THE OPEN ACCESS INITIATIVE

The data mining of literature may also generate important knowledge for drug discovery. For example, by analyzing the literature one can determine new metabolic pathways or discover side effects of already known drugs, which can potentially lead to new drug targets. Such important knowledge, however, may be hidden in the article as secondary results.

Considering the growing mountain of life science data an automatic contextual text data mining can become very important in revealing these hidden pearls of information [20]. A systematic processing of abstracts available in PubMed (http://www.ncbi.nlm.nih.gov/entrez) or full text of articles provided by large publishers such as Elsevier (http://www.sciencedirect.com) can be used to make new discoveries "*in textico*". This field is very actively developing and it was unimaginable before the expansion of the Internet and the publication of articles in digital format. Unfortunately, the subscription to journals can be rather expensive or have limitations on the number of articles that can be automatically retrieved. This may change with the Open Access initiative (www.soros.org/openaccess) intended to provide free Web access to all published research articles and has lead to establishment of Public Library of Science (http://www.plos.org).

WEB SERVICES AND WORK FLOWS

Web publishing using static or dynamic HTML pages is convenient for interactive analysis of information. This type of publishing, however, is not convenient if one wants to do an automated data processing in a work-flow. The automation of HTML-based servers may require development of fragile screen-scraping integration scripts to extract information from one page and to introduce it to another. A small change in the output of Web pages may break the whole integration process. The cure against such integration is well known as Web services, i.e. Web pages specifically developed for automatic data processing. Development of Web services using XML-based protocols can to great degree solves problem of data inconsistency. A growing number of web services in life science were recently reviewed [21]. Moreover, the development of the Semantic Web (http://www.w3.org/2001/sw) may bring new ways of publishing scientific articles by including comprehensive experimental results in structured format. This will provide a new dimension for automatic processing of information.

The integration of Web services can still require some time and expertise. A new development in this field is Taverna [22] which provides a graphical interface for management and integration of Web Services. In addition to Taverna one can also use pipe-lining and workflow development using both commercial, e.g. previously mentioned Pipeline Pilot (http://accelrys.com), and open source providers such as KNIME (http://knime.org). A nice overview of different tools for development of WorkFlow technologies is provided at http://openwetware.org/wiki/Abhishek_Tiwari:Workflow_technology.

CONCLUSIONS

Interactive Web tools, on-line access to chemo and bioinformatics databases and on-line access to millions of articles provides resources that were unimaginable in the drug discovery field just 10-15 years ago. The famous law "A couple of months in the laboratory can save a couple of hours in the library" discovered last century by Harvard chemistry professor F.H. Westheimer should be rephrased by changing the word "library" to the "Internet". This is, however, only true if we sit in the right library, use the right tools and understand their limitations. Too much and noisy information can be more harmful than its absence. "It is much better to offer no answer with a clear explanation than to provide numbers without any meaning" was concluded by Ertl *et al.* [1]. However, even understanding this problem does not solve it, until there are no reliable methods to evaluate the quality of prediction [23].

Hopefully, the same Internet that brought us the ocean of information can also be used to wash it out by automatic detection and elimination of contradictory results. The adding of layers of intelligence to validate retrieved information and to separate gold from ash will be important to speeding up drug discovery in years to come.

REFERENCES

[1] P. Ertl, P. Selzer, and J. Mühlbacher, "Web-based cheminformatics tools deployed *via* corporate Intranets," *Drug Discov. Today: BIOSILICO,* vol. 2, pp. 201-207, 2004.

[2] Y. C. Martin, "What works and what does not: Lessons from experience in a pharmaceutical company," *QSAR Comb. Sci.,* vol. 25, pp. 1192-1200, Dec 2006.

[3] R. Lewis, P. Ertl, E. Jacoby, *et al.* "Computational chemistry at Novartis," *Chimia,* vol. 59, pp. 545-549, 2005.

[4] M. Lobell, M. Hendrix, B. Hinzen, *et al.* "In silico ADMET traffic lights as a tool for the prioritization of HTS hits," *ChemMedChem,* vol. 1, pp. 1229-36, Nov 2006.

[5] M. D. Segall, A. P. Beresford, J. M. Gola, D. Hawksley, and M. H. Tarbit, "Focus on success: using a probabilistic approach to achieve an optimal balance of compound properties in drug discovery," *Expert Opin. Drug Metab. Toxicol.,* vol. 2, pp. 325-37, Apr 2006.

[6] R. Mulin, "Seeing the Forest at Pfizer - A radical knowledge-sharing initiative takes hold at the world's largest drugmaker," *C&EN,* vol. 85, pp. 29, 2007.

[7] I. V. Tetko, and V. Y. Tanchuk, "Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program," *J. Chem. Inf. Comput. Sci.,* vol. 42, pp. 1136-1145, Sep-Oct 2002.

[8] W. M. Meylan, and P. H. Howard, "Atom/fragment contribution method for estimating octanol-water partition coefficients," *J. Pharm. Sci.,* vol. 84, pp. 83-92, Jan 1995.

[9] R. Wang, Y. Gao, and L. Lai, "Calculating Partition Coefficient by Atom-Additive Method," *Perspect. Drug Discov. Des,,* vol. 19, pp. 47-66, 2000.

[10] T. Cheng, Y. Zhao, X. Li, *et al.* "Computation of octanol-water partition coefficients by guiding an additive model with knowledge," *J. Chem. Inf. Model,* vol. 47, pp. 2140-8, Nov-Dec 2007.

[11] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, and R. K. Robins, "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," *J. Chem. Inf. Comput. Sci.,* vol. 9, pp. 163-172, 1989.

[12] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita, "Simple method of calculating octanol/water partition coefficient," *Chem. Pharm. Bull.,* vol. 40, pp. 127-130, 1992.

[13] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, "Dragon software: An easy approach to molecular descriptor calculations," *Match-Commun. Math. Comput. Chem.,* vol. 56, pp. 237-248, 2006.

[14] I. V. Tetko, "The WWW as a tool to obtain molecular parameters," *Mini Rev. Med. Chem.,* vol. 3, pp. 809-820, Dec 2003.

[15] W. J. Geldenhuys, K. E. Gaasch, M. Watson, D. D. Allen, and C. J. Van der Schyf, "Optimizing the use of open-source software applications in drug discovery," *Drug Discov. Today,* vol. 11, pp. 127-32, Feb 2006.

[16] P. Ertl, and S. Jelfs, "Designing drugs on the internet? Free web tools and services supporting medicinal chemistry," *Curr. Top. Med. Chem.,* vol. 7, pp. 1491-501, 2007.

[17] I. V. Tetko, "Computing chemistry on the web," *Drug Discov. Today,* vol. 10, pp. 1497-1500, Nov 15, 2005.

[18] I. V. Tetko, J. Gasteiger, R. Todeschini, *et al.* "Virtual computational chemistry laboratory - design and description," *J. Comput. Aided Mol. Des.,* vol. 19, pp. 453-463, Jun 2005.

[19] D. S. Wishart, "Bioinformatics in drug development and assessment," *Drug Metab, Rev.,* vol. 37, pp. 279-310, 2005.

[20] M. Krallinger, R. A. Erhardt, and A. Valencia, "Text-mining approaches in molecular biology and biomedicine," *Drug Discov. Today,* vol. 10, pp. 439-45, Mar 15 2005.

[21] V. Curcin, M. Ghanem, and Y. Guo, "Web services in the life sciences," *Drug Discov. Today,* vol. 10, pp. 865-71, Jun 15 2005.

[22] D. Hull, K. Wolstencroft, R. Stevens, *et al.* "Taverna: a tool for building and running workflows of services," *Nucleic Acids Res.,* vol. 34, pp. W729-32, Jul 1 2006.

[23] I. V. Tetko, P. Bruneau, H. W. Mewes, D. C. Rohrer, and G. I. Poda, "Can we estimate the accuracy of ADME-Tox predictions?," *Drug Discov. Today,* vol. 11, pp. 700-707, Aug 2006.

Received: March 24, 2008 Revised: April 19, 2008 Accepted: April 20, 2008