Computing chemistry on the web

Igor V. Tetko

Institute of Bioorganic & Petrochemistry, Murmanskaya 1, 02094 Kiev, Ukraine and Institute for Bioinformatics, GSF, Neuherberg, Germany, http://www.vcclab.org

pre-print of the article published in:

Drug Discovery Today, 2005, vol. **10**, (22), 1497-1500.

Address for correspondence: Igor V. Tetko
Institute for Bioinformatics
GSF - Forschungszentrum fuer Umwelt und Gesundheit, GmbH
Ingolstaedter Landstrasse 1,
D-85764 Neuherberg, Germany
Telephone: +49-89-3187-3575
Fax: +49-89-3187-3585
e-mail: itetko@vcclab.org

Running head: Virtual Computational Chemistry Laboratory
Keywords: on-line analysis, physico-chemical property predictions, indices calculation, model generation and validation, comparison of web technologies, bioinformatics, chemoinformatics
Teaser: The development of on-line software tools is changing the way we traditionally perform our analysis in drug design, but will chemoinformatics be forever behind bioinformatics in this development?

This manuscript contains 10 pages including 1 Table and 1 Figure.

Despite the dramatic growth of the Internet, the number of practical applications in drug design, particular for prediction ADME/T properties, which are available on-line remains limited. For example, the number of methodological publications about lipophilicity predictions has gradually increased over the last ten years and more than 25 these articles are expected to be published in different journals in 2005.[1] At the same time, the number of programs available for on-line prediction of this important property is in range of ten applications [2]. Publicly available tools for predicting many other important physico-chemical and biological properties simply do not exist. Thus there is a need to develop new Web applications to boost drug design and chemoinformatics studies. In this article we describe why publishing programs on the Internet is important and beneficial for the authors and its users, describe several example of such sites and critically discuss why this field remains underdeveloped compared to its nearest rival, bioinformatics.

Why Develop Web Services?

A typical research activity of a computational chemist includes preparation of data, optimization of molecular structures, calculation of indices, selection of the most important indices and deriving property-activity correlation using statistical methods. The calculated models are then revised and poorly predicted compounds are analyzed to get some insight into which new indices are required to improve the overall model or to detect errors and inconsistency in data preparation. However, if a new method has been developed, it might be interesting to compare it with previously existing approaches. This could be difficult if the algorithm is only referenced in a paper. But, even if the method is available as source or binary code, a dedicated computer platform, particular operating system, configuration of system parameters, compiler, libraries, etc, could be required to perform such a study. A distribution of binary code may unintentionally propagate viruses or spy software. There could be conflict of interests, e.g. if the authors plan to commercialize it and substantial efforts may be required by the authors to support it.

On the contrary, publishing on the Web allows a program to be executed in the same environment it was developed. This software is easy to maintain, release updates and any Web user can access and use it in his/her work. The Web applications also

provide much better dissemination of information about the algorithm thanks to powerful search engines, e.g. Google or Yahoo.

Some examples of Web services

An increasing number of diverse tools for performing data analysis in chemistry on the Internet is available for users (Table 1) as also reviewed elsewhere.[2-5] The Virtual Computational Chemistry Laboratory (VCCLAB, http://www.vcclab.org)[6] is one of the most comprehensive resources for data analysis in chemistry on the Web. The front-end of the laboratory visible to the users is represented by applets (Figure 1). They are used to start, control and display results of different tasks ranging from the generation of descriptors to the development of predictive models. The Java-based system seamlessly integrates programs running on computers and various operating systems in five countries in Europe. For example, the unsupervised forward selection algorithm[7] is implemented on Silicon Graphics in Portsmouth (UK), Corina[8] is executed on a Linux machine in Erlangen (Germany), ALOGPS[9] runs on a MacOsX system, calculation of Dragon,[10] E-state and fragment–based descriptors are performed on Windows in Milano, Kiev and Moscow, respectively. A simple click of the mouse starts a sequence of tasks to be executed on computers located thousands of miles one from one another. Currently, the site calculates hundreds of tasks per day and has more than 600 registered users, including 46% with PhD. The USA contributed the largest number, 115, of the users and the second largest number, 81, is from India. The 207 users from the European Union are headed by the United Kingdom (31) and Germany (27). The number of academic users (76%) is followed by industrial (17%) and governmental users (7%).

Similar projects to provide comprehensive resources on the Web also exist in industry[11] or in collaborative researches between academia and industry.[12] The LINK3D project[12] developed tools and 3D software for synchronous collaboration in the field of drug design, in particular the virtual meeting software. Novartis system supports more than a thousand users with molecular modelling and molecular processing tasks, including the calculation of molecular and substituent properties, property-based virtual screening, visualization of molecules, bioisosteric design, etc.[11] However, these systems are usually not available publicly. The success of the VCCLAB site clearly indicates that there is also a demand for development of public versions of such systems.

The chemical community is also actively involved in the development of new protocols for the Internet. The Chemical Markup Language (CML, http://www.xml-cml.org) was one of the first XML-based standards for scientific exchange of information developed by the UK group.[13] The same group is actively involved in the popularization of the Semantic Web and new ways of scientific publishing on the Web.[14,15] This is, without a doubt, very important work. However, it looks like that a large part of chemical community is not yet actively involved in this development and the number of Web services in the field remains limited. To this extent it is interesting to compare chemo- and bioinformatics. The latter is definitely leading in Web developments.

Why chemo- and bioinformatics are different in respect to publishing on the Web?

One of the explanations for the dissimilarity in both fields is a dramatic difference in the amount of bioinformatics and chemoinformatics data and computers resources required to store and process them. For example, the human genome data are stored as 2.7 GB of zipped files at Ensembl (http://www.ensembl.org). Moreover, the data and methods to analyze them are frequently changed, e.g. four updates of the human genome have already been released only in 2005. The analysis of bioinformatics data is very time consuming, e.g., InterPro domain[16] calculations take several days to be completed on one CPU even for a single human chromosome. The clustering of sequences is also computationally expensive and has stimulated the development of specialized methods.[17] The human genome is, however, just one of more than 400 genomes that are annotated and publicly available, e.g. at MIPS[18], and their number continues to grow. These apparent computational difficulties and huge amounts of data strongly encourage cooperation between different research centers and boost development of web technologies.

The situation in drug design and chemoinformatics is different. The largest commercial database of chemical compounds, the iResearch library (http://www.chemnavigator.com), comprises less than 15 million unique SMILES. All these structures can be stored in a 50 MB compressed file. The public database ZINC[19] offers a smaller set totaling just a few million unique compounds. All these data can be

processed in just a few hours by, e.g. the ALOGPS program. These are, however, very general databases that have only limited applications in the field. Moreover, unlike bioinformatics where the genome sequence and its position on the chromosome provides a lot of information and can be used in many studies, the structure of a molecule should be accompanied by measured physical or biological activities. Such measurements are extremely expensive. Therefore, the databases for the development of new methods, e.g. physico-chemical properties, are of orders of magnitude smaller. One of the largest databases in the field, Physical Properties Database (PHYSPROP)[20], comprises slightly more than 25,000 compounds. Moreover, this is a commercial and not a public database. The largest datasets of biological properties, such as Blood Brain Barrier[21] or intestinal absorption[22] comprise just a few hundred compounds. Larger collections of these and other ADME/T properties exist in industry, but they are not available for public developments due to privacy issues. Thus an inadequate amount of data and their limited availability significantly slow down development of chemoinformatics compared to bioinformatics. Of course, some other differences like legacy, data complexity, etc., also contribute to the problem.[15]

Another important aspect is a motivation to develop such resources. One of the main outcomes of academic activity is the publication of articles. But the leading chemical journals do not yet have experience of accepting and publishing articles describing web resources and the number of such publications is limited. At the same time the ratio of regular articles to application notes (usually describing Web applications) in Bioinformatics journal is about 5:3 for the first half of 2005. It is quite common to publish a methodological article that will be accompanied a few months later by an application note in this journal. The apparent success of this strategy is illustrated by the impact factor of Bioinformatics, which jumped from 3.4 to 6.7 in just two years according to the Institute for Scientific Information (http://www.isinet.com).

Is there a light at the end of the tunnel?

There is a hope that the situation in the chemoinformatics field will change. The recent PubChem initiative of the National Institute of Health[23] will house both compound information from the scientific literature as well as screening and probe data from molecular libraries screening centers. This may provide large amounts of high

quality data with biological and ADME/T activity of chemical compounds for drug design studies. The availability of these data may dramatically change the field and boost development of web resources similar to those available in bioinformatics. Considering the expense of drug failures at last stages of development due to unsatisfactory ADME/T properties,[24] there is an increasing motivation for large pharmaceutical companies to release some of their data to promote development of new chemoinformatic methods. Since the privacy of molecular structures is of paramount importance for the success of the drug industry, a development of approaches to release data but not the underlying molecule structures is actively explored in the field.[25] The attitude towards the publishing of Web services is also changing and one of the top publications in the field, Journal of Chemical Information and Modeling, plans to publish a dedicated issue on Web services in 2006 (W. Warr, personal communications).

Conclusions

Given the benefits brought to bioinformatics by Web applications, it is attractive to encourage the development of these technologies in the cheminformatics field. The publishing of data/methods on the Web allows other researchers to avoid duplication, to reuse and to validate the results of previous studies in a new development. The Web servers increase awareness about the existing software and may increase citation of the article. The appearance of new protocols and standards for data sharing on the WWW makes development of new applications easier and straightforward. The VCCLAB can be used as a prototype to develop such projects. The developed technology allows integration of new third-party applications, which could be made available to the worldwide community.

Acknowledgement

I thank Johann Gasteiger, Roberto Todeschini, Peter Ertl, David Livingstone, Vladimir Palyulin, Vsevolod Tanchuk, Alexander Makarenko and members of their teams for their contributions, testing and development of the VCCLAB site and Louise Riley for valuable remarks. This work was partially supported by INTAS 00-0363 project.

References

1 Tetko, I.V. and Livingstone, D.J. (2006) Rule-based systems to predict lipophilicity. In *Comprehensive Medicinal Chemistry II: In silico tools in ADMET* (Vol. 5) (Testa, B. and van de Waterbeemd, H., eds.), pp. in press, Elsevier

2 Tetko, I.V. (2003) The WWW as a tool to obtain molecular parameters. *Mini Rev. Med. Chem.* 3 (8), 809-820.

3 Van de Waterbeemd, H. and De Groot, M. (2002) Can the Internet help to meet the challenges in ADME and e-ADME? *SAR QSAR Environ. Res.* 13 (3-4), 391-401.

4 Marchand-Geneste, N. and Carpy, A.J. (2004) e-Quantum chemistry free resources. *SAR QSAR Environ. Res.* 15 (1), 43-54.

5 Jonsdottir, S.O. et al. (2005) Prediction methods and databases within chemoinformatics: emphasis on drugs and drug candidates. *Bioinformatics* 21 (10), 2145-2160.

6 Tetko, I.V. et al. (2005) Virtual Computational Chemistry Laboratory (VCCLAB) http://www.vcclab.org. In *QSAR2004* (Vol. in press)

7 Whitley, D.C. et al. (2000) Unsupervised forward selection: a method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* 40 (5), 1160-1168.

8 Sadowski, J. et al. (1994) Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *J. Chem. Inf. Comput. Sci.* 34 (4), 1000-1008.

9 Tetko, I.V. and Tanchuk, V.Y. (2002) Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program. *J. Chem. Inf. Comput. Sci.* 42 (5), 1136-1145.

10 Todeschini, R. and Consonni, V. (2000) *Handbook of Molecular Descriptors*, WILEY-VCH

11 Ertl, P. et al. (2003) Web-based cheminformatics and molecular property prediction tools supporting drug design and development at Novartis. *SAR QSAR Environ. Res.* 14 (5-6), 321-328.

12 Pastor, M. et al. (2002) Distant collaboration in drug discovery: the LINK3D project. *J. Comput. Aided. Mol. Des.* 16 (11), 809-818.

13 Murray-Rust, P. and Rzepa, H.S. (1999) Chemical markup Language and XML Part I. Basic principles. *J. Chem. Inf. Comp. Sci.* 39, 928-942.

14 Murray-Rust, P. et al. (2004) Representation and use of chemistry in the global electronic age. *Org. Biomol. Chem.* 2 (22), 3192-3203.

15 Curcin, V. et al. (2005) Web services in the life sciences. *Drug Discov. Today* 10 (12), 865-871.

16 Mulder, N.J. et al. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.* 33, D201-205.

17 Tetko, I.V. et al. (2005) Super paramagnetic clustering of protein sequences. *BMC Bioinformatics* 6 (1), 82.

18 Riley, M.L. et al. (2005) The PEDANT genome database in 2005. *Nucleic Acids Res.* 33, D308-310.

19 Irwin, J.J. and Shoichet, B.K. (2005) ZINC--a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* 45 (1), 177-182.

20 The Physical Properties Database (PHYSPROP) is a trademark of Syracuse Research Corporation, www.syrres.com.

21 Platts, J.A. et al. (2001) Correlation and prediction of a large blood-brain distribution data set--an LFER study. *Eur. J. Med. Chem.* 36 (9), 719-730.
22 Zhao, Y.H. et al. (2001) Evaluation of human intestinal absorption data and subsequent derivation of a quantitative structure-activity relationship (QSAR) with the Abraham descriptors. *J. Pharm. Sci.* 90 (6), 749-784
23 Morrissey, S.R. (2005) NIH initiatives target chemistry. *Chem. Eng. News* 83 (1), 23-24.
24 Landers, P. (2003) Cost of Developing a New Drug Increases to About $1.7 Billion. In *The Wall Street Journal*.
25 Wilson, E.K. (2005) Is safe exchange of data possible? *Chem. Eng. News* 83 (17), 24-29.
26 Tetko, I.V. et al. (2001) Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J. Chem. Inf. Comput. Sci.* 41 (5), 1407-1421.
27 Wang, R. et al. (2000) Calculating Partition Coefficient by Atom-Additive Method. *Persp. Drug Discov. Des.* 19, 47-66.
28 Tetko, I.V. (2002) Neural network studies. 4. Introduction to associative neural networks. *J. Chem. Inf. Comput. Sci.* 42 (3), 717-728.
29 Tetko, I.V. (2002) Associative neural network. *Neur. Proc. Lett.* 16 (2), 187-199.
30 Aksyonova, T.I. et al. (2003) Robust Polynomial Neural Networks in Quantitative-Structure Activity Relationship Studies. *SAMS* 43 (10), 1331-1339.

Table 1. Some Examples of Free Computational Web Resources in Chemistry

no	name	provides	link
1	Corina	2D to 3D conversion of molecular structures	http://www2.chemie.uni-erlangen.de/software/corina/
2	Osiris	logP, solubility, toxicity, drug likeness	http://www.organic-chemistry.org/prog/peo
3	Petra	physico-chemical properties of compounds	http://www2.chemie.uni-erlangen.de/services/petra
4	Pre-ADME	molecular descriptors and various ADME/T properties	http://preadme.bmdrc.org/preadme
5	VCCLAB	molecular descriptors, physico-chemical properties and data analysis tools	http://www.vcclab.org

Extended comprehensive lists of other resources can be found elsewhere [2-5].

Figure legends

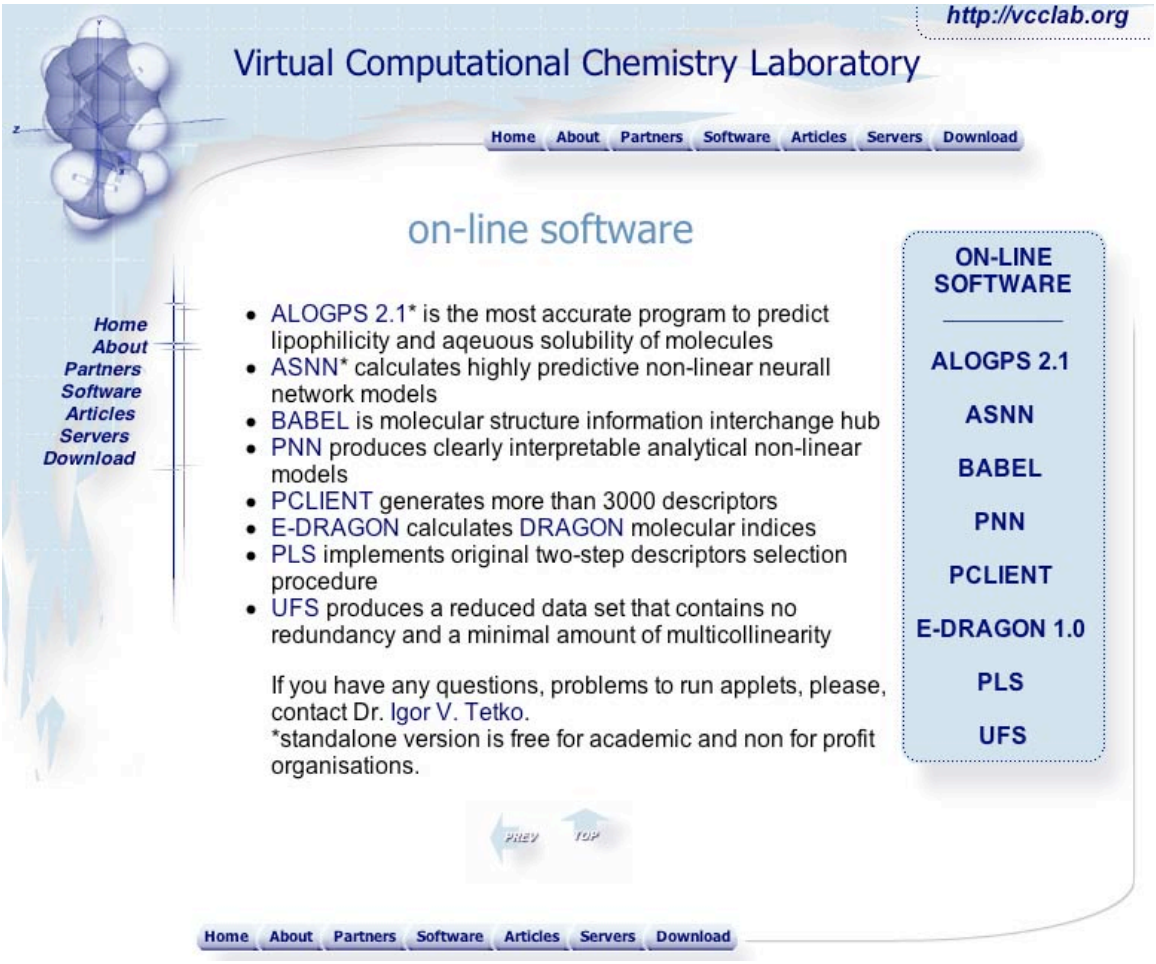

Figure 1. On-line software tools available at the Virtual Computational Chemistry Laboratory site.

ALOGPS applet calculates and compares lipophilicity and aqueous solubility of molecules using six methods including the ALOGPS 2.1,[9,26] CLOGP (http://www.daylight.com/daycgi/clogp), KOWWIN (http://www.syrres.com/esc/est_kowdemo.htm), MiLogP (http://www.molinspiration.com), IA_logP (http://www.logp.com) and XLOGP [27]. The user can draw the molecule using the JME applet of Peter Ertl or submit it in a format supported by OpenBabel (http://openbabel.sourceforge.net). Associative neural network method (ASNN)[28,29] calculate models with high prediction ability by correcting the bias of the neural network ensemble. Polynomial Neural Networks (PNN)[30] calculates analytical non-linear models between descriptors of molecules and the target activity and provides a clear interpretation of the detected relations. e-DRAGON and its extension Parameter Client (PCLIENT) calculate more than 1,600 and 3,000 descriptors per

molecule, respectively. The user can either provide optimized molecules or seamlessly convert molecules from 2D to 3D representation using the integrated CORINA program.[8] The unsupervised forward selection (UFS)[7] decreases the number of descriptors and produces a reduced data set that contains no redundancy and a minimal amount of multicollinearity.