VIRTUAL COMPUTATIONAL CHEMISTRY LABORATORY (VCC-LAB): NEW INTERNET- COMPUTER APPROACHES TO RESEARCH AND EDUCATION IN CHEMICAL PROBLEMS

*Igor V. Tetko1,2 Vsevolod Yu. Tanchuk,1 and Alexander Makarenko*3
Virtual Computational Chemistry Laboratory Project
1 - Institute of Bioorganic and Petroleum Chemistry of National Academy of Science of Ukraine, Kiev,
2 - Institute for Bioinformatics, Ingolstraedter Lanstraβe 1, D-85764, Neuherberg, Germany[1]
3 - National Technical University of Ukraine (KPI), Institute Applied System Analysis
Pobedy Avenue 37, 03056, Kiev, Ukraine[2]

Abstract *-- Exploiting electronic networks allows preparing special tools by cooperative effort of different departments and institutions. In proposed report we describe the score of our recent investigations on such issue under INTAS project. The overall objective of this research is to develop multi-platform software allowing the computational chemist to perform a comprehensive series of molecular properties calculations and data analysis on INTERNET. The proposed software is based on three-tier architecture that is becoming a widespread to provide client-server services over the world. The computational part will be written in C/C++, while the graphical interface will be done in Java. The developed software will be user friendly and will allow a simple incorporation of new modules (including proprietary software) developed by other researches. A unique feature of this system will allow such modules to run at the computers where the software was developed while the calculation results will be available worldwide.*

Index Terms -- Computational Chemistry, Remote-Access, Client-Server, Distance Education

Introduction

Internet activities have become in the last few years a major investment in information, business, communication, teaching technologies and chemistry. The WWW (World Wide Web) impact on society dramatically increases especially in the field of education and scientific research. It is clear, that Internet will become a major system for knowledge extraction and education in the new Century. A great deal of information is available for chemist in form of chemical databases such as ChemFinder, ChemExper Chemical Directory, on-line journals, conferences, etc. A number of companies have started to provide on-line demo version (e.g., Syracus corporation, Daylight, etc.) of their programs or complete services such as software to predict molecular properties including NMR spectra, logP/S/D, pKa, etc. by Advanced Chemistry Design I-labs system (http://www.acdlabs.com/ilab/), data analysis programs by SpotFire Inc. (http://www.spotfire.com/), etc. The number of companies providing on-line services is increasing very fast.

The academic scientific research has a specific place in this system by providing an access to scientific programs developed by academia. Such programs developed by professionals can become available to a general audience of the world research community. An example of successful development in this field by academia is a number of on-line software program developed by group of University of Erlangen (Prof. J. Gasteiger) including system to calculate Physicochemical Parameters of molecules (PETRA), program to convert 2D -> 3D structure of molecules CORINA and a growing number of other programs that are available for the WWW users.

A large number of available scientific programs, including molecular indices/ property calculation and data analysis programs have been developed in FORTRAN and C/C++ programming languages. Many of these programs are developed using different computer systems and are based on various input data formats. Thus it is not easy to transfer these programs between different computer platforms and laboratories even if their authors are willing to provide the source

codes of the algorithm. Therefore the developed software tools remain limited in their use to a restricted number of people, and the scientists willing to apply new algorithms very often have to re-program it again from scratch.

A big effort to deliver a consistent set of programs to calculate various parameters of molecules was done by group of Prof. R. Todeschini. Their recent "Handbook of Molecular Descriptors" includes the definitions and formulas of the most important molecular descriptors known in the field of structure-activity relationship studies. A software version for the calculation of several hundreds of molecular descriptors, called DRAGON, is available as a standalone program from their WWW site (http://www.disat.unimib.it/chm/). The groups of Dr. V. Palyulin (Moscow State University) as well as Dr. V. Tanchuk (IBPC, Kiev, Ukraine) have also developed a number of various molecular indices, including extended E-state indices, various indices of symmetry of molecules, etc. The calculation of some indices requires correct 3D structure of molecules that can be calculated by CORINA program developed. Unfortunately, these programs are not compatible since they use different input data formats and were developed for different computer platforms (Unix, MAC OS, and Windows).

The problem of incompatibility of different software developed for different computer systems can be solved using Java Native Interface (JNI). A number of useful data analysis methods (Polynomial and various back-propagation Neural Networks, Time Series Analysis methods, etc.) have been developed by our group and are currently provided on-line using this interface at http://www.lnh.unil.ch site. A similar experience in development of WWW based chemical information system has been received, where the web technology and Java tools helped to integrate various property calculation methods, statistical analysis tools and visualization programs into one compact, user friendly service [1-8].

The main idea of a new project is to integrate different methods of molecular property calculations and data analysis and make them available on-line as Virtual Computational Chemistry Laboratory (VCC-LAB) at http://vcclab.org. We are also going to provide a description of required interface that will allow other scientists to contribute their programs to VCC-LAB.

This project is based on fruitful collaboration that was established in the previous INTAS 95-0060 and INTAS-OPEN 97-168 grants. These projects developed several innovative approaches, such as new Associative Neural Networks algorithms, lipophilicity (logP) and aqueous solubility (logS) calculation programs. These programs will be further elaborated and made available at our Internet site.

Scientific Description

The developed software is based on the extension of the existing three-tier architecture successfully used for on-line calculation of general data analysis programs, such as Artificial Neural Networks and Polynomial Neural Networks, programs to calculate properties of molecules and a number of programs for data analysis in Neurophysiology. The general layout of the developed software includes three main parts, namely **Client Applet**, **Super Server** and **Calculations Servers**.

- The first important part of the project will be to extend functionality of this software for applications in Computer Chemistry.
- The second part will be to include modules for calculation of molecular properties and data analysis program to integrate them in VCC-LAB
- The third part will be to test the different sets of indices for several databases.

Client Applet is the only part of the program that is visible to the user. The user can upload the input data files, specify data analysis parameters and submit their task for calculation. The calculated results can be saved on local disk. This part of work is already functional for the data analysis and calculation of logP/S. The design of convenient interface to perform calculation of various molecular indices will be the purpose of some investigations.

An upload and saving of files in current version is done using HTML-based interface. This is not very convenient in some cases since the files should be sent back and forth through Internet. While preserving this possibility for the first trial user, we will also extend the functionality of Applet Clients using Signed Applet. This extension will be based on Java 1.2 platform and will allow user to read and write files directly from his local disk. This will provide a more convenient interface for the user. The use of Signed Applets, however, will be allowed only for registered users (the registration will be free for Universities and participants) and will require installation of Java 1.2 plug-in. The instructions how to do this for different

computer platforms will be provided on the project web pages.

The user can also visualize the uploaded molecules using JME structure editor of Dr. P. Ertl or can use this editor to create the molecules directly within a web page. This interface is already available for molecular property calculation programs. However, a current version visualizes molecules that were only created in JME or saved in internal JME format. We will extend the JME editor to depict molecules in other formats too.

Super Server. The main purpose of the Super Server is to collect the requests from the client, to deliver the tasks to the Calculation Servers and to deliver the calculated results back to the client. The current version is already functional and it provides basic service of this kind.

A considerable effort will be devoted to create a macro language that will allow an easy configuration of complex calculations. For example, let us imagine that user would like to calculate a lipophilicity of a molecule from its SMILES code and this property requires some symmetry indices of molecules that should be calculated using 3D structure of molecules. Thus, a calculation of such property will include 1) conversion of SMILES to 3D Structure using CORINA program (server of University of Erlangen), 2) calculation of 3D indices (server of University of Milano) and 3) calculation of the lipophilicity (servers of Institute of Bioorganic & Petroleum Chemistry and Institute of Bioinformatics).

A use of macro language will provide an easy configuration of the routing of the tasks and will allow extension of this software for new properties that can be programmed by the other users. It is also possible that all these programs will be available from the same server. However, it will make no difference to the Super Server to organize an integration of these programs.

Calculation Servers will provide conversions of data files, calculation of the properties of molecules and data analysis. This server can be running on a remote computer. The Internet connection by Java allows it to establish connection with Super Server, to receive task, to perform calculations and to send results back to the Super Server. The Calculation Servers can be running everywhere in the world provided the server computer should have (fast) Internet connections. The available libraries accessible by Java Native Interface determine tasks that can be calculated by the Calculation Server. Since the Java language is supported by virtually all computer systems, such interface makes it possible to use the VCC-LAB the programs developed and running at different computer systems. The current version of Calculation Server integrates Java and C/C++ programs only by means of re-defined FILE system. This requires to change and to recompile the original software and can be applied only to C/C++ or converted FORTRAN programs. A new version will also include a possibility to do such integration through standard input/output streams, thus leaving the developed software intact. This will allow including also software developed in other languages. A detailed description of the interface will be prepared for the users who would like to contribute to the development of the VCC-LAB.

Registering of new modules. A user who would like to contribute a new property calculation method or index will be required to provide an information with a short description of the property, macro command for Super Server, e-mail address, name and telephone number as well as original literature reference for the method. This information will be saved in a local database and can be available from the help menu for each data property. The user can also decide to provide source code (and thus this property will be calculated at servers of the project participants) or IP number of a computer on which this property will be calculated. The registration of new methods will be possible on-line using WWW interface. This system will be used at first by the project participants and, after careful testing, will be made publicly available for other users.

Programming of calculation modules. The project participants will develop a number of the property analysis modules. This will include an integration of topological indices calculation program DRAGON, extended E-state indices and various topological indices. Since some indices were developed by several groups (e.g., topological indices of Balaban, Kier, etc. were developed) a part of these tasks will be to verify and compare different programs and to correct possible programming errors. As result of these tasks about 1,200 indices will be available for the users. All these indices will be classified on groups according to rules proposed in book of Prof. R. Todeschini. A use of Java interface developed will allow user to select and calculate groups of parameters or single parameters.

Data analysis methods. The calculated indices could be easily analyzed on WWW using previously developed Back propagation and Polynomial Neural Network methods. We will also add to these methods Multiple Linear Regression Analysis and Partial Least Squares (both of them are already programmed) as well as method of Continuum Regression.

Application of the developed software. The existing software will be used to test the information content of the programmed indexes using large databases of structure-property including logP (ca. 13,000 compound), logS, melting point (ca. 15,000 compounds) and enthalpy of fusion (ca. 2,300 compounds) using several different methods of data analysis developed in as well as methods already available on Internet. All these databases are coded using SMILES codes and can be easily used to calculate molecular properties by the developed programs. The databases are already available for analysis.

This study will evaluate how the different sets of indices reflect various aspects (features) of molecular structure necessary for the modeling of the analyzed properties of organic compounds. This study will also allow us to propose novel indices in order to feel the "gaps" in the sets of existing indices.

The result of this study will be new powerful programs for prediction of the molecular properties of chemical compounds. It should be noted that many groups have already a considerable experience in developing molecular property prediction software, including lipophilicity, aqueous solubility, melting point, enthalpy of fusion, drug transport properties and substituent parameters.

Summary

This project will develop integrated multi-platform software and will allow the computational chemist to perform a comprehensive series of molecular properties calculation and data analysis on Internet. The developed software will be user friendly and will provide on-line calculation of ca. 1,200 different topological indices and six data analysis methods. All modules can be used separately or as one functional system.

The developed software will allow a simple incorporation of new modules (including proprietary software) developed by other researchers. A unique feature of this software will allow these modules to run at the computers where the software was developed while the calculation results will be available worldwide. This will provide a wide dissemination of the project results and will be important to speed up developments in the drug design and computational chemistry. The developed on-line software can be also used in Universities to teach the modern approaches in computational chemistry.

The developed indices will be evaluated using four large databases of molecular parameters, including lipophilicity, aqueous solubility, melting point and fusion enthalpy. This will help us to identify "gaps" in the sets of existing indices and to propose new indices. The analysis of these databases will also provide new methods for prediction of these important physico-chemical parameters.

This is only very short description of presumable possibilities of proposed approach. More detailed description will be posed in coming publications. Also we welcome including of new modules into the system as the exploitation in educational and scientific practice. It is planned that the developed software will be available at http://vcclab.org in the autumn of 2002. The proposed approach and structure may be useful as prototype for another virtual computer educational tools.

Acknowledgement

The authors would like to bring deep gratitude to Prof. Johann Gasteiger, Prof. Roberto Todeschini, Dr. Peter Ertl, Dr. David Livingstone, Dr. Vladimir Palyulin, Dr. Vsevolod Tanchuk, Dr. Tetyana Aksyonova and many others our collaborators for close cooperation in implementation of the proposed approach. This study was supported with Virtual Computational Chemistry Laboratory grant INTAS 00-0363.

References

[1] Tetko, I.V.; Tanchuk, V.Yu,; Kasheva, T.N.; Villa, A.E.P. Internet Software for Calculation of Lipophilicity and Aqueous Solubility of Chemical Compounds, *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 246-252.

[2] Tetko, I.V.; Aksenova, T.I.; Volkovich, V.V.; Kasheva, T.N.; Filipov, D.V.; Villa, A.E.P.; Welsh, W.J.; Livingstone, D.J. Polynomial Neural Network for Linear and Non-linear Model Selection in Quantitative-Structure Activity Relationship Studies on WWW SAR and QSAR in Environmental Research. **2000**, 11(3/4), 263-280.

[3] Voigt K., Gasteiger J., Brüggemann R. Comparative Evaluation of Chemical and Environmental Online and CD-ROM Databases. *J. Chem. Inf. Comput. Sci.*, **2000**, 40, 44-49.

[4] Making the Computer Understand Chemistry J. Gasteiger *Internet J. Chemistry*, 1, 33, (**1998**) URL: http://www.ijc.com/articles/1998v1/33/.

[5] Todeschini, R. & Consonni, V. (**2000**). Handbook of Molecular Descriptors. WILEY-VCH, Mannheim (DE).

[6] Ertl, P. QSAR Analysis through the World Wide Web Chimia (special issue Chemistry and the Internet), 52, 673-677 (**1998**)

[7] Palyulin V.A., Radchenko E.V., Zefirov N.S. Molecular Field Topology Analysis (MFTA) method

in QSAR Studues of Organic Compounds. *J. Chem. Inf. Comput. Sci.*, **2000**, 33, 659-667.

[8] Makarenko A. Geometrical Approach to the Measure of Individual Object Complexity. Proc. 15th European Meeting on Cybernetics and Systems Research, Vienna, Austria, 25-28 April, **2000**. Vol.1. pp.25-30.