

SOMA2 – Open Source Gateway to Molecular Modelling Workflows

5th EGEE User Forum

Dr. Tapani Kinnunen

CSC – IT Center for Science Ltd., Espoo, Finland

➤ **CSC at a Glance**

- Founded in 1970 as a technical support unit for Univac 1108
- Connected Finland to Internet in 1988
- Reorganized as a company, CSC – Scientific Computing Ltd. in 1993
- All shares to the Ministry of Education of Finland in 1997
- Operates on a non-profit principle
- Facilities in Espoo, close to Otaniemi campus (of 15,000 students and 16,000 technology professionals)
- Staff 180
- Turnover 2008 19,6 million euros

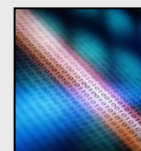
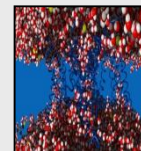
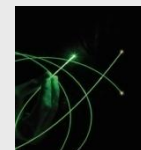


➤ **CSC's Mission**

- CSC, as part of the Finnish national research structure, develops and offers high-quality information technology services.

➤ **CSC's Services**

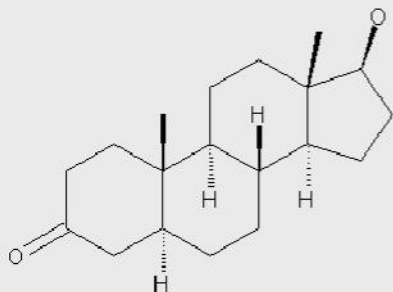
- Funet Services
- Computing Services
- Application Services
- Data Services for Science and Culture
- Information Management Services



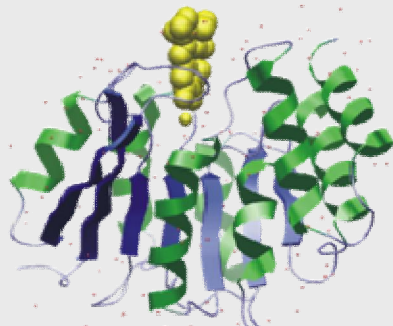
➤ **SOMA2 is a gateway for computational drug discovery and molecular modelling**

- SOMA2 is operated with WWW –browser
 - Intuitive WWW –interface provides an easy access to computational tools.
 - Offers a full scale environment from data input to result analysis.
 - System is operated with user’s own user account and access rights.
- SOMA2 makes use of scientific applications installed in the computing system
 - Uniform interface tools for applications.
 - Automatic configuration and execution of applications.
 - Different applications and tools can be integrated into application workflows.
- SOMA2 Software is open source
 - Released in May 2007 under GNU General Public License (GPL).
 - Current version: 1.3 Magnesium (3rd of September 2009).

- **SOMA2 was developed at CSC in the SOMA2 project (2002-2006)**
- Tekes (National Technology Agency of Finland) DRUG2000 program.
 - Organised and updated CSC's (the Finnish IT Center for Science) modelling program environment to meet the standards in modern computer-aided molecular design.
 - Promoted the use of computing tools in drug discovery research work in Finland.



FROM MOLECULES ...



... TO PROTEINS ...



...AND CELL-LEVEL ACTIVITIES

➤ **SOMA2 helps the users...**

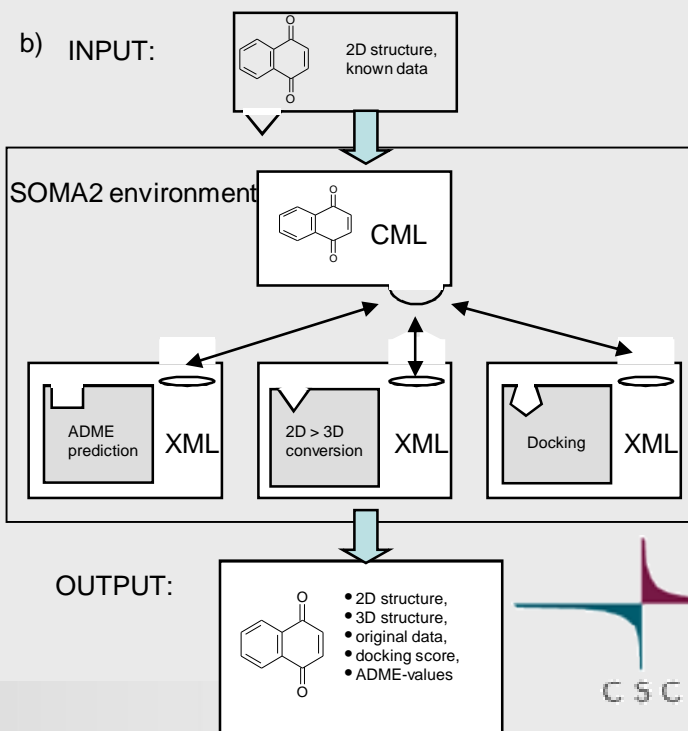
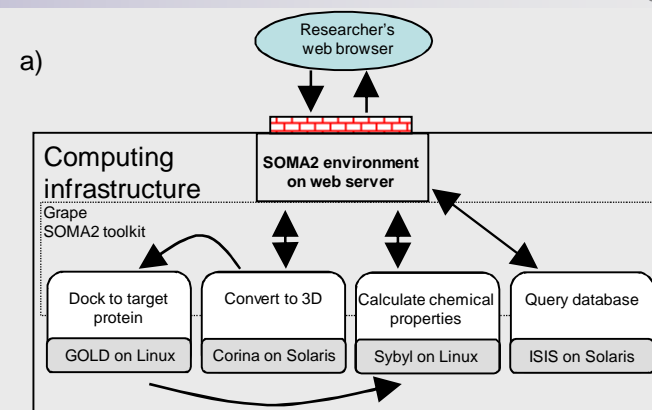
- No technical skills at all are required to use computational tools.
- Specific knowledge in Linux/UNIX systems not needed.
- Incompatible programs are integrated into seamless application workflows.
- Organisation, propagation and storing of computed data.
- Automates repeating work.
- Eliminates redundant work.
- Advanced users can benefit from automatically generated scripts.

➤ **...As well as the service providers**

- Knowledge transfer and documentation in machine readable form.
- Steer the usage of the computing system.
- Heterogeneous computing system can be made invisible to the users.
- Centralise the maintenance of scientific programs.
- Automate repeating support routines.
- SOMA2 suits for both small and large computing infrastructures.

➤ Basic technical concept

- WWW –interface for configuring a scientific program is based on XML –description of the program.
- Different machine architectures are hidden.
- Automatic generation of program and platform specific configuration files.
- CML (Chemical Markup Language, <http://cml.sourceforge.net>) is used as internal data format (data transferred in XML).
- Unique computational workflows.



➤ **Modular components of SOMA2**

- A. WWW interface
 - User authentication, input of molecular data, building the program configurations, performing database queries, creating a workflow and analysing the results.
 - Tools: Perl, JavaScript, HTML, CSS.
- B. Workflow manager program Grape
 - Execution, logistics and monitoring of program execution (2D XML graph).
 - Tools: Java.
- C. SOMA2 capsules
 - eXtended Markup Language (XML) description for attaching a scientific program to be used via SOMA2.
 - Templates of program configuration files, command scripts for executing programs, batch queue system scripts and program output parsers.
 - Tools: XML, shell scripts.
- D. Toolkit of helper applications
 - Programs for molecule format conversions, building the execution files from the templates and managing the internal data.
 - Tools: Perl, shell scripts.

- **Modular program integration with generic configuration interface generation**
 - All information needed in integrating and executing a program is in SOMA2 capsule.
 - Program configuration interface generated from description that is based on XML schema (“template”).
 - Programming skills are not required to produce SOMA2 capsule for a program.
 - Programs are easily added to be used via the SOMA2 –environment without a need to change SOMA2 program code itself.
 - Expert user knowledge of a program can be saved in SOMA2 capsule.

- **Security**
 - System is operated with user’s own user account and access rights.
 - Data is not accessible to the other users.

- **Flexibility**
 - Almost any molecular modelling program can be attached to be used via the SOMA2 –system.
 - Only condition is that a program can be operated from the command line or through API.
 - Programs can be executed interactively or via a batch system.

➤ **SOMA2 is open source**

- Initially open source released in May 2007.
- SOMA2 source code is licensed under GNU General Public License (GPL).
- All interested parties can install SOMA2 to their computing environment and make local applications easily available to the users.
- Downloads available from SOMA2 WWW –pages: <http://www.csc.fi/soma>.
- SOMA2 demo installation with limited features available at: <http://soma2demo.csc.fi>

➤ **Distribution contains example SOMA2 capsules**

- Can be used as examples in creating own capsules
 - obenergy (Open Babel single point energy calculator, <http://openbabel.sourceforge.net>).
 - obgen (Open Babel 3D coordinate generator, <http://openbabel.sourceforge.net>).
 - obprop (Open Babel molecular property calculator, <http://openbabel.sourceforge.net>).
 - identity / identity_batch (SOMA2 test capsule).

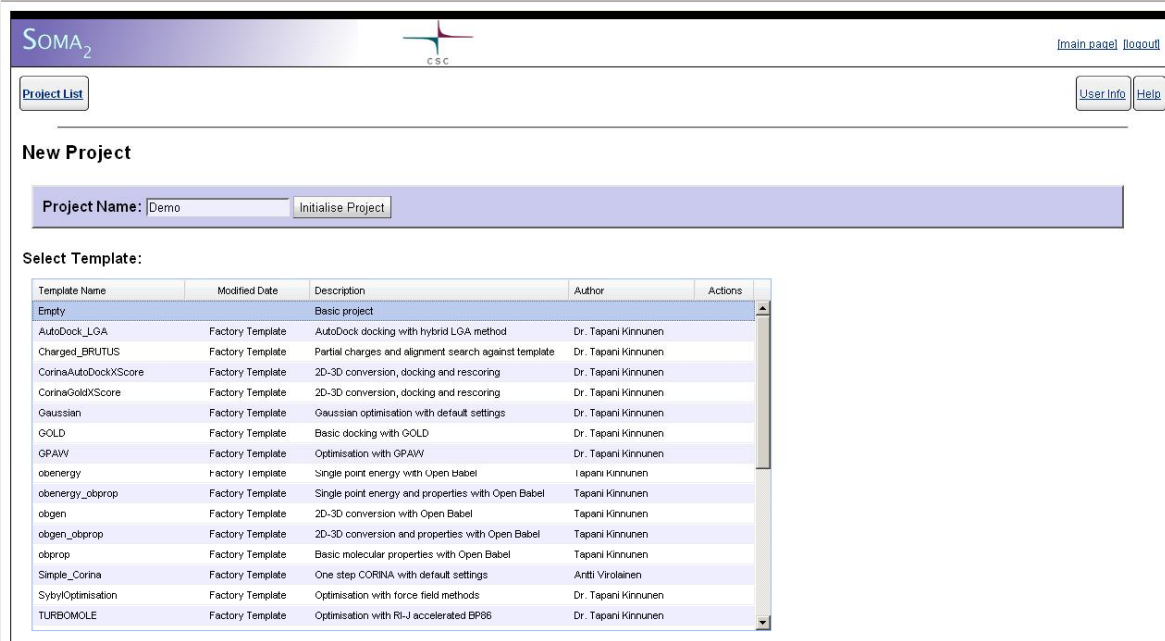
➤ **SOMA2 capsules can be discussed in the development forum**

- 32 SOMA2 capsules have been made for 14 different scientific programs at CSC.

- **2D-Property (Sybyl module)**
 - Molecular properties that are based on the 2D structure.
- **3D-Property (Sybyl module)**
 - Molecular properties that are based on the 3D structure.
- **CORINA**
 - 2D – 3D coordinate conversion or multiple ring conformation generation.
- **ROTATE**
 - Rotamer generation.
- **AutoDock**
 - Ligand docking and scoring.
- **GOLD**
 - Ligand docking and scoring.
- **Overlay**
 - Flexible molecular alignment search tool.
- **BRUTUS**
 - Rigid molecular alignment search tool.
- **Volsurf (Sybyl module)**
 - Calculation on molecular descriptors and molecular response values.
- **Tanimoto similarity (Sybyl module)**
 - Calculation of Tanimoto similarity index against template.
- **Sybyl**
 - Calculations based of force field methods. Charges, energies and optimisation.
- **X-Score**
 - Rescoring of docked ligands with several scoring functions.
- **Gaussian 09**
 - Versatile quantum chemistry software package.
- **TURBOMOLE**
 - Versatile quantum chemistry software package.
- **GPAW**
 - Versatile DFT software package.

➤ Model workflows

- User can choose a predefined workflow for specific task.
- Predefined workflow can still be freely modified.
- Possibility to save own workflows as a template.

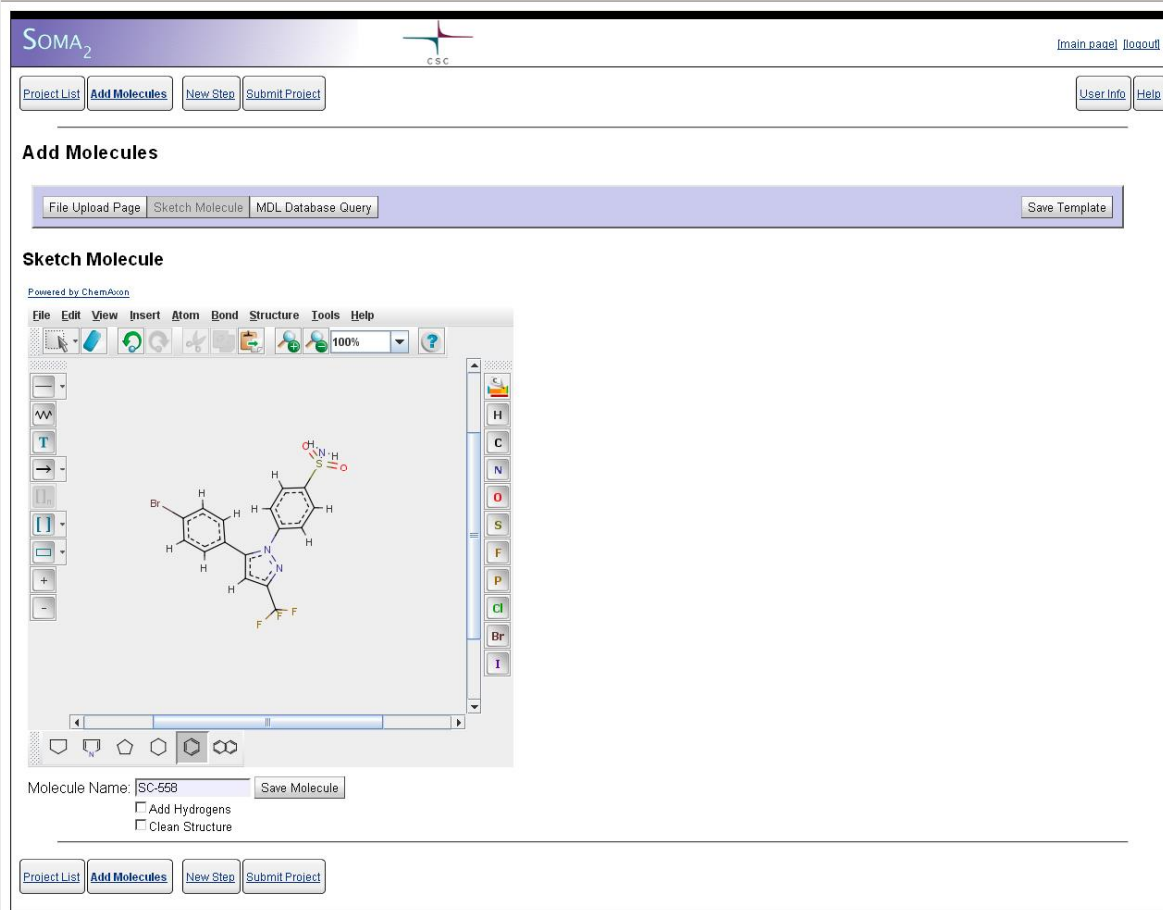


The screenshot displays the SOMA₂ web application interface. At the top, there is a navigation bar with the SOMA₂ logo, the CSC logo, and links for 'main page' and 'logout'. Below the navigation bar, there are buttons for 'Project List', 'User Info', and 'Help'. The main content area is titled 'New Project' and features a 'Project Name' input field containing the text 'Demo' and an 'Initialise Project' button. Below this, the 'Select Template' section contains a table listing various predefined workflows.

Template Name	Modified Date	Description	Author	Actions
Empty		Basic project		
AutoDock_LGA	Factory Template	AutoDock docking with hybrid LGA method	Dr. Tapani Kinnunen	
Charged_BRUTUS	Factory Template	Partial charges and alignment search against template	Dr. Tapani Kinnunen	
CorinaAutoDockXScore	Factory Template	2D-3D conversion, docking and rescoring	Dr. Tapani Kinnunen	
CorinaGoldXScore	Factory Template	2D-3D conversion, docking and rescoring	Dr. Tapani Kinnunen	
Gaussian	Factory Template	Gaussian optimisation with default settings	Dr. Tapani Kinnunen	
GOLD	Factory Template	Basic docking with GOLD	Dr. Tapani Kinnunen	
GPAVV	Factory Template	Optimisation with GPAVV	Dr. Tapani Kinnunen	
obenergy	Factory Template	Single point energy with Open Babel	Tapani Kinnunen	
obenergy_obprop	Factory Template	Single point energy and properties with Open Babel	Tapani Kinnunen	
obgen	Factory Template	2D-3D conversion with Open Babel	Tapani Kinnunen	
obgen_obprop	Factory Template	2D-3D conversion and properties with Open Babel	Tapani Kinnunen	
obprop	Factory Template	Basic molecular properties with Open Babel	Tapani Kinnunen	
Simple_Corina	Factory Template	One step CORINA with default settings	Antti Virolainen	
SybylOptimisation	Factory Template	Optimisation with force field methods	Dr. Tapani Kinnunen	
TURBOMOLE	Factory Template	Optimisation with RI-J accelerated BP86	Dr. Tapani Kinnunen	

➤ Input molecules

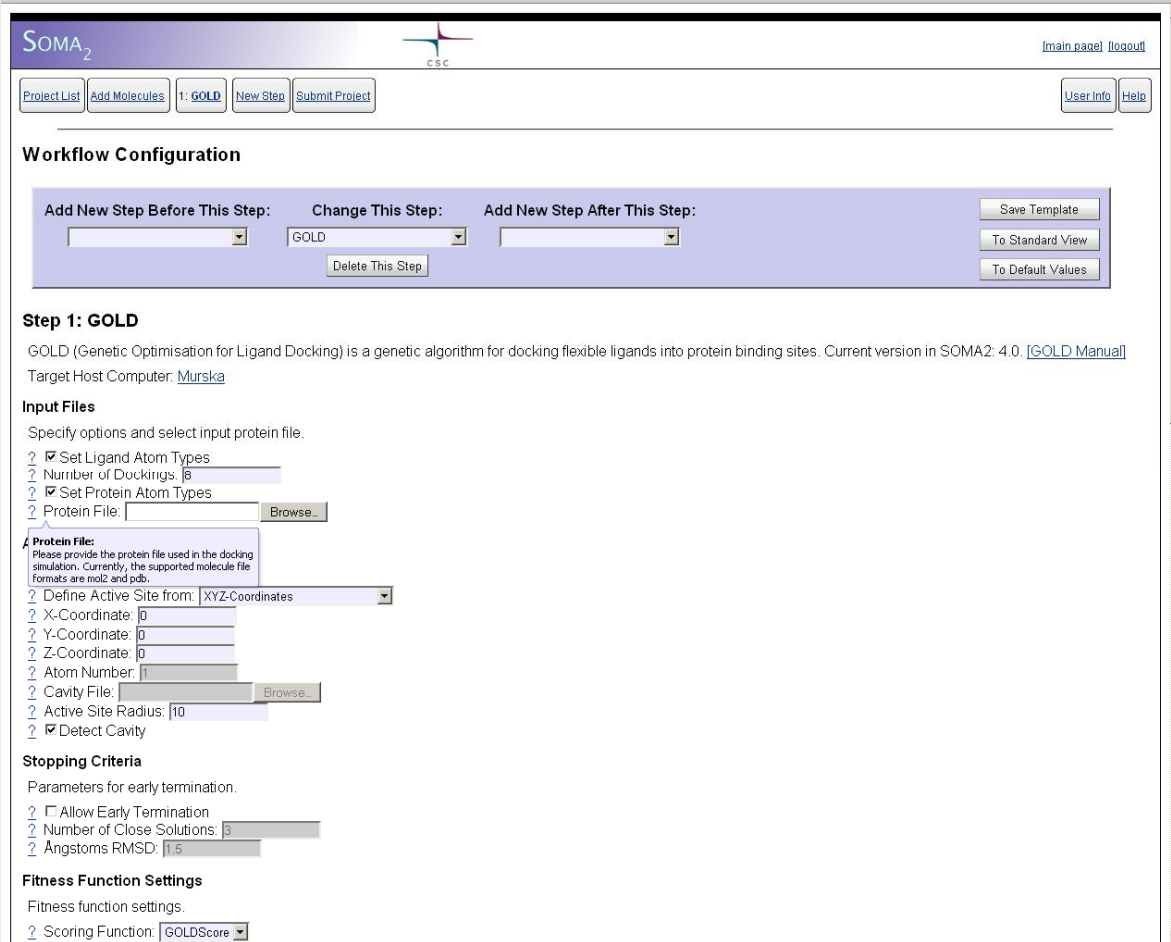
- Upload files from local computer.
- Sketch molecules within the user interface.




The screenshot displays the SOMA₂ web interface. At the top, there is a navigation bar with buttons for 'Project List', 'Add Molecules', 'New Step', and 'Submit Project'. On the right side of this bar are 'User Info' and 'Help' buttons. Below the navigation bar is a section titled 'Add Molecules' with tabs for 'File Upload Page', 'Sketch Molecule', and 'MDL Database Query', and a 'Save Template' button on the right. The main area is titled 'Sketch Molecule' and is powered by ChemAxon. It features a standard software menu (File, Edit, View, Insert, Atom, Bond, Structure, Tools, Help) and a toolbar with various drawing tools. The central workspace shows a chemical structure of a complex molecule with a brominated benzene ring, a pyrazole ring, and a sulfonamide group. To the right of the workspace is a vertical toolbar with buttons for elements: H, C, N, O, S, F, P, Cl, Br, and I. Below the workspace, there is a 'Molecule Name' field containing 'SC-568', a 'Save Molecule' button, and two checkboxes: 'Add Hydrogens' and 'Clean Structure'. At the bottom of the interface, there is another set of navigation buttons: 'Project List', 'Add Molecules', 'New Step', and 'Submit Project'.

➤ Program configuration

- Easy configuration of programs with interactive web form.
- Useful help texts, reasonable default values, thresholds and requirements.
- Interactive parameter validation on web form.
- SOMA2 capsule includes configuration file templates for running a program.



SOMA₂  [Main page](#) [Logout](#)

[Project List](#) [Add Molecules](#) 1: GOLD [New Step](#) [Submit Project](#) [User Info](#) [Help](#)

Workflow Configuration

Add New Step Before This Step: Change This Step: GOLD Add New Step After This Step:

[Delete This Step](#) [Save Template](#) [To Standard View](#) [To Default Values](#)

Step 1: GOLD

GOLD (Genetic Optimisation for Ligand Docking) is a genetic algorithm for docking flexible ligands into protein binding sites. Current version in SOMA2: 4.0. [\[GOLD Manual\]](#)
Target Host Computer: [Murska](#)

Input Files

Specify options and select input protein file.

Set Ligand Atom Types
Number of Dockings:
 Set Protein Atom Types
Protein File: [Browse...](#)

Protein File:
Please provide the protein file used in the docking simulation. Currently, the supported molecule file formats are mol2 and pdb.

Define Active Site from: XYZ-Coordinates
X-Coordinate:
Y-Coordinate:
Z-Coordinate:
Atom Number:
Cavity File: [Browse...](#)
Active Site Radius:
 Detect Cavity

Stopping Criteria

Parameters for early termination.

Allow Early Termination
Number of Close Solutions:
Angstroms RMSD:

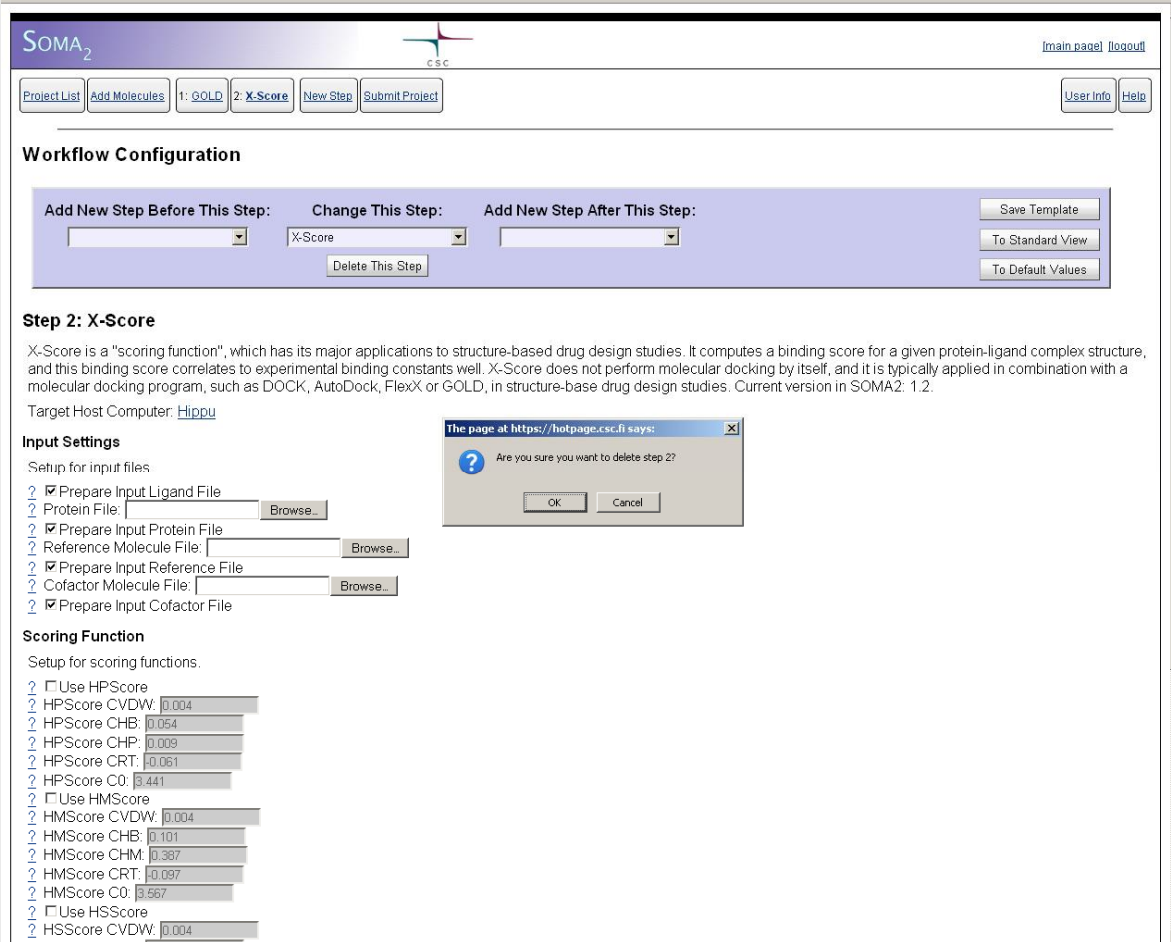
Fitness Function Settings

Fitness function settings.

Scoring Function:

➤ Workflow management

- Free navigation between steps.
- Insert, change and delete operation supported.
- Validation of the user constructed workflow.



The screenshot displays the SOMA₂ web interface. At the top, there is a navigation bar with buttons for 'Project List', 'Add Molecules', '1: GOLD', '2: X-Score', 'New Step', and 'Submit Project'. On the right, there are 'User Info' and 'Help' buttons. The main content area is titled 'Workflow Configuration' and includes three dropdown menus for 'Add New Step Before This Step:', 'Change This Step:', and 'Add New Step After This Step:'. The 'Change This Step:' dropdown is currently set to 'X-Score'. Below these are buttons for 'Delete This Step', 'Save Template', 'To Standard View', and 'To Default Values'.

Step 2: X-Score

X-Score is a "scoring function", which has its major applications to structure-based drug design studies. It computes a binding score for a given protein-ligand complex structure, and this binding score correlates to experimental binding constants well. X-Score does not perform molecular docking by itself, and it is typically applied in combination with a molecular docking program, such as DOCK, AutoDock, FlexX or GOLD, in structure-base drug design studies. Current version in SOMA2: 1.2.

Target Host Computer: [Hippu](http://hippu)

Input Settings

Setup for input files.

- Prepare Input Ligand File
- Protein File:
- Prepare Input Protein File
- Reference Molecule File:
- Prepare Input Reference File
- Cofactor Molecule File:
- Prepare Input Cofactor File

Scoring Function


Setup for scoring functions.

- Use HPScore
- HPScore CVDW:
- HPScore CHB:
- HPScore CHP:
- HPScore CRT:
- HPScore C0:
- Use HMScore
- HMScore CVDW:
- HMScore CHB:
- HMScore CHM:
- HMScore CRT:
- HMScore C0:
- Use HSScore
- HSScore CVDW:
- HSScore CHB:

A dialog box is overlaid on the interface, titled 'The page at https://hotpage.csc.fi says:'. It contains a question mark icon and the text 'Are you sure you want to delete step 2?'. There are 'OK' and 'Cancel' buttons at the bottom.

➤ Result view

- Exportable spreadsheet like result view.
- Tools for sorting and filtering data.
- Save molecular data in different formats.

SOMA₂  [Main page](#) [Logout](#)

[Project List](#) [Project Overview](#) [1: GOLD View Config](#) [2: X-Score View Config](#) [Results](#) [User Info](#) [Help](#)

Results: Project "Demo"

Format Selection: sdf [Save Selected](#) [Start Project With Selected](#) [Export Data](#) [Save Template](#)

Result Data:

Molecule Name	GOLD: Fitness	X-Score: HPScore	X-Score: MW	X-Score: LogP	GOLD: Docking Protein	Step Number	GOLD: S(hb_ext)	GOLD: S(hb_int)	GOLD: S(int)
ER-34122.1.1	-14.11	7.97	507.8	3.91	4cox-A.pdb	2	0.00	0.00	
ER-34122.2.1	10.74	7.80	507.8	3.91	4cox-A.pdb	2	0.00	0.00	
FR-123826.1.1	56.99	6.58	341.3	2.66	4cox-A.pdb	2	0.00	0.00	
FR-123826.2.1	55.77	6.58	341.3	2.66	4cox-A.pdb	2	0.00	0.00	
FR-188582.1.1	55.43	6.57	332.7	3.40	4cox-A.pdb	2	0.34	0.00	
FR-188582.2.1	58.28	6.74	332.7	3.40	4cox-A.pdb	2	0.00	0.00	
Ribuprofen.1.1	44.81	6.28	206.1	3.64	4cox-A.pdb	2	0.29	0.00	
Ribuprofen.2.1	47.64	6.34	206.1	3.64	4cox-A.pdb	2	0.54	0.00	
Sibuprofen.1.1	39.54	6.29	206.1	3.64	4cox-A.pdb	2	0.00	0.00	
Sibuprofen.2.1	40.24	6.24	206.1	3.64	4cox-A.pdb	2	0.00	0.00	
SC-558.1.1	62.50	6.84	446.2	3.32	4cox-A.pdb	2	0.00	0.00	
SC-558.2.1	59.77	6.81	446.2	3.32	4cox-A.pdb	2	0.00	0.00	
SC-58125.1.1	53.81	6.87	384.3	3.86	4cox-A.pdb	2	2.00	0.00	
SC-58125.2.1	54.69	6.89	384.3	3.86	4cox-A.pdb	2	3.97	0.00	
Tepezalim.1.1	57.15	6.96	385.7	2.63	4cox-A.pdb	2	0.00	0.00	
Tepezalim.2.1	42.63	7.10	385.7	2.63	4cox-A.pdb	2	0.00	0.00	
Vioxx.1.1	54.81	6.67	314.2	3.22	4cox-A.pdb	2	0.00	0.00	
Vioxx.2.1	52.58	6.46	314.2	3.22	4cox-A.pdb	2	0.00	0.00	

Filter Tools: Hide Show

Filter Rules: X-Score: LogP > 2


[Filter](#) [Reset View](#) [\[Project Overview\]](#)

Columns:

- GOLD: Fitness
- X-Score: HPScore -6.54
- X-Score: MW -4.19
- X-Score: LogP -5.81
- GOLD: Docking Protein -4.93
- Step Number -5.40
- GOLD: S(hb_ext) -5.19
- GOLD: S(hb_int) -3.73
- GOLD: S(int) -4.13
- GOLD: S(vdw_ext) -6.35
- SOMA2 Original Name -10.01
- X-Score: Average -9.62
- X-Score: Docking Protein -9.35
- X-Score: Formula -6.12
- X-Score: HMScore
- X-Score: HSScore

➤ Result details

- Visualisation of the result molecules.
- Summary of computed properties.

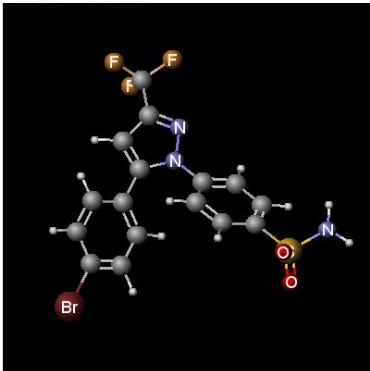
SOMA₂  [main page](#) [logout](#)

[Project List](#) [Project Overview](#) [1: GOLD View Config](#) [2: X-Score View Config](#) [Results](#) [User Info](#) [Help](#)

Result Details: "SC-558.1.1"

Parent: [SC-558.1](#)

Select Viewer:
 Viewer Sketcher



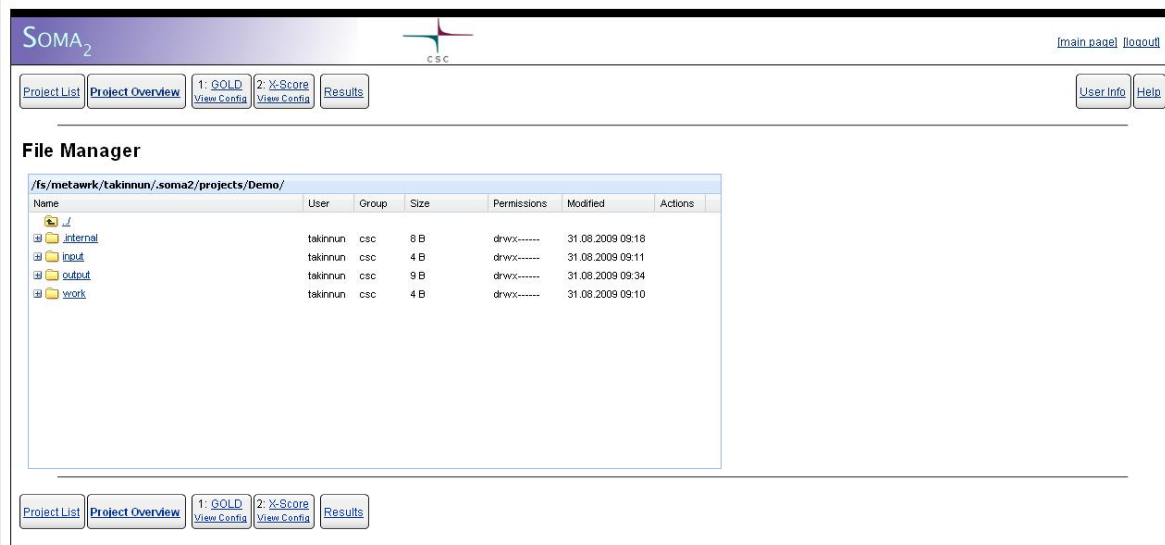
Powered by ChemAxon

[\[Results\]](#)

GOLD: Docking Protein:	4cox-A.pdb
GOLD: Fitness:	62.50
GOLD: S(hb_ext):	0.00
GOLD: S(hb_int):	0.00
GOLD: S(int):	-3.73
GOLD: S(vdw_ext):	48.17
SOMA2 Molecule Name:	SC-558.1.1
SOMA2 Original Name:	SC-558
Step Number:	2
X-Score: Average:	7.13
X-Score: Docking Protein:	protein.pdb
X-Score: Formula:	C16H11N3O2SF3Br
X-Score: HMScore:	7.75
X-Score: HPScore:	6.84
X-Score: HSScore:	6.82
X-Score: LogP:	3.32
X-Score: MW:	446.2

➤ File manager

- Provides access to the file system.
- Basic file operations supported (browse, view, save).
- Access allowed only to user's own SOMA2 project directory.



The screenshot shows the SOMA₂ File Manager interface. At the top, there is a navigation bar with the SOMA₂ logo and the CSC logo. The interface includes several tabs: "Project List", "Project Overview", "1: GOLD View Config", "2: X-Score View Config", and "Results". There are also "User Info" and "Help" buttons on the right. The main content area is titled "File Manager" and displays a directory listing for the path "/fs/metawrk/takinnun/soma2/projects/Demo/". The listing is as follows:

Name	User	Group	Size	Permissions	Modified	Actions
internal	takinnun	csc	8 B	drwx-----	31.08.2009 09:18	
input	takinnun	csc	4 B	drwx-----	31.08.2009 09:11	
output	takinnun	csc	9 B	drwx-----	31.08.2009 09:34	
work	takinnun	csc	4 B	drwx-----	31.08.2009 09:10	

At the bottom of the interface, there are additional navigation buttons: "Project List", "Project Overview", "1: GOLD View Config", "2: X-Score View Config", and "Results".

➤ Future Improvements

- DCI integration
 - Instead of regular batch scheduler, make use of Grid middleware.
 - Users' X509 certificate handling needs to be added.
- UI Enhancements
 - Currently has "traditional" look and feel, incorporate more web2.0 components
- Enhancements in Data Logistics
 - Data logistics is currently based on flat files, works fine but becomes inefficient when number of molecules is very large

➤ Technical requirements

- Linux server
 - Standard GNU utilities (gsed, awk, etc.).
 - Java JDK 1.5 or later (<http://www.sun.com/java>) with additional libraries.
 - JGraphT Java library (<http://jgrapht.sourceforge.net>).
 - Apache ant build tool (<http://ant.apache.org>).
 - Perl 5.8 or later with additional libraries.
 - Perl core modules.
 - XML::Twig (<http://www.xmltwig.com>).
 - Template Toolkit (<http://www.template-toolkit.org>).

➤ Technical requirements

- Passwordless SSH connections, shared disk system, user accounts
 - Communication between server running SOMA2 and the computation platforms.
 - Excluding “localhost”.
 - The same user accounts must exist in both the server running SOMA2 and the computation platforms.
- Apache WWW server (<http://httpd.apache.org>)
 - User authentication (HTTP Basic, PAM, Cookie, etc.).
 - SSL protocol for secure communication.
 - *Altered* suEXEC module to enable CGI program execution as an authenticated user.
 - Source code and instructions are available from the SOMA2 –WWW pages, see: <http://www.csc.fi/english/pages/soma/downloads>.

➤ Client requirements

- WWW –browser (Internet Explorer, Firefox, Opera, Mozilla)
 - JavaScript support enabled.
 - Java Plug-in installed.
 - Cookie support enabled (if cookie authentication used).

➤ **Third-Party components**

- Ext JS JavaScript library (<http://www.extjs.com>)
 - All tables, popups and trees in SOMA2 user interface
- Open Babel (<http://openbabel.sourceforge.net>)
 - Molecule file format conversions and property calculation.
- JDB (<http://www.isi.edu/~johnh/SOFTWARE/JDB/index.html>)
 - ASCII data filtering tools.
- ChemAxon Marvin Java applets (<http://www.chemaxon.com>)
 - Tools for building and visualising molecular structures.

Additional information

- **CSC – the Finnish IT Center for Science:**
 - <http://www.csc.fi>
- **SOMA2 –homepage and download site:**
 - <http://www.csc.fi/soma>
- **SOMA2 Demo (limited features, no authentication required):**
 - <http://soma2demo.csc.fi>

Acknowledgements

- **Tekes (National Technology Agency of Finland):**
 - <http://www.tekes.fi>
- **ChemAxon Ltd.:**
 - <http://www.chemaxon.com>