SIMCA®-Q

Application Note

A part of ChemGPS-NPWeb

10 June 2020

Anders Backlund, Anders Lövgren, Robert Sedzik and Josefin Rosén

Introduction

In this application note we review the design, features, and possible uses of ChemGPS-NP, as facilitated by the public web tool ChemGPS-NPWeb [1,2]. ChemGPS-NP [2-4] is a principal component analysis (PCA) based global chemical positioning system tuned for exploration of biologically relevant chemical space. Any compound with a known chemical structure can be mapped using interpolation in terms of PCA score prediction onto a consistent, eight-dimensional map. This map is based on structure-derived physio-chemical characteristics for a reference set of compounds. From the resulting projections properties of the compounds can be compared and easily interpreted.

During the first two years since its introduction, ChemGPS-NPWeb has predicted the position in chemical space for 2 651 852 molecules, originating from 2110 submissions.

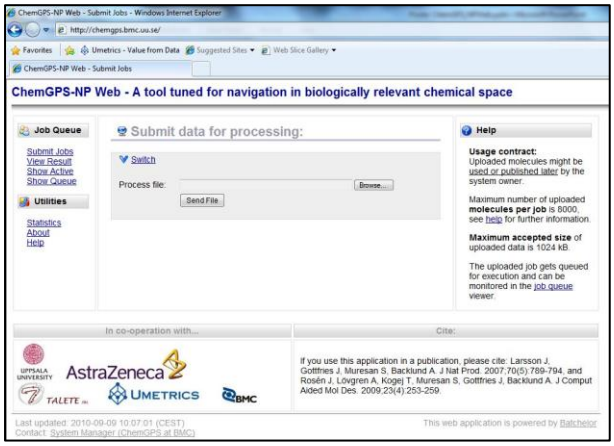

Figure 1: Screenshot from the ChemGPS-NPWeb showing the queue interface.

Web tool components

ChemGPS-NPWeb consists logically of two pieces, the Batchelor job queue manager, and the application stack. A web service layer exists with a web user interface allowing the user to submit jobs, list the queue and download results.

The application stack consists of two major parts: the DragonX [5] application and the SIMCA-QP library. Input data are SMILES strings representing the compounds to be investigated. When the workflow is initiated by the queue handler, the uploaded SMILES strings are first preprocessed. An initial Perl script removes erroneous SMILES strings, as well as information about stereochemistry and isotopes, and checks the size of the input file against the limit (at present maximum 8000 molecules per batch).

The preprocessed SMILES are then submitted to DragonX, which serves as an internal engine for calculation of the 35 molecular descriptors from which the eight principal components (PCs) representing map dimensions are extracted.

Subsequently the output from DragonX is transformed by an intermediate script and fed to SIMCA-QP, via a cgpsclt (a client) that connects to cgpsd (a server) to run SIMCA-QP. Here the relevant library functions are called, and the prediction model created in SIMCA-P+ is loaded. In ChemGPS-NPWeb SIMCA-QP performs the PCA score prediction, i.e. the actual mapping, via the library libchemgps. The server, cgpsd, subsequently returns the result, i.e. eight map-coordinates for each compound, back to cgpsclt, and the outcome of the pipeline is finally stored in the database. From the database users can monitor the status of their submitted jobs and download their result from the queue for post processing. The coordinates can then be plotted and visualized e.g. using the ChemGPS-NPViewer Java applet. Post computational statistics for ChemGPS-NPWeb are compiled based on results from each of the successive computational steps.

The apparent extra step utilizing a client/server (cgpsclt/cgpsd) was incorporated to avoid having to load the project reference set for each job. As an additional benefit it also enables predictions to be performed by one or more computers on the network. An overview of the ChemGPS-NPWeb components is presented in Figure 2.

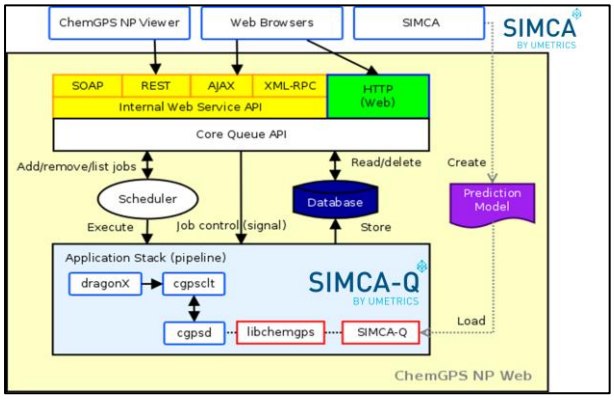

Figure 2: Overview of the ChemGPS-NPWeb components.

ChemGPS-NPViewer

ChemGPS-NPViewer is the first beta-version of a standalone multi-platform application written in Java, and still under development. It is used to visualize the eight PCs representing the eight dimensions of chemical space. Each time a compound is selected it is marked by a minute red arrow and the coordinates, the name or reference label, as well as a 2D structural representation (via a built in SMILES converter) are shown in the upper right corner (see Figure 3).

ChemGPS-NPViewer communicates with the ChemGPS-NPWeb server seamlessly to give the user full functionality from one automated graphical user interface. The plotted molecule symbols can be given specific color as well as a shape, so that it is possible to visually separate them. Groups of molecules can be selected to be visible or not in the 3D space.

Measurement of Euclidean distances [5] between molecules is another feature, as well as finding the nearest neighbors from a selected molecule to those closest in the 8D space.

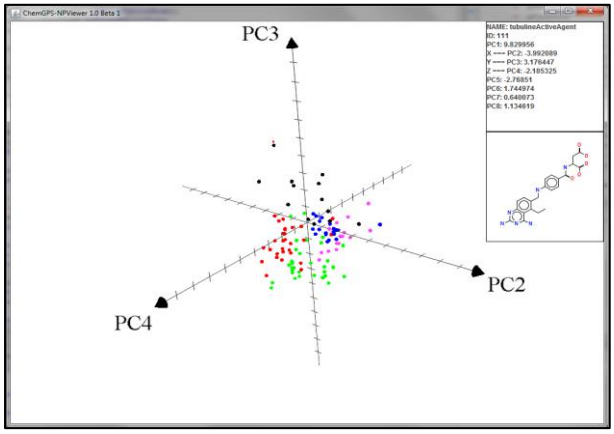

Figure 3: Screen shot from the ChemGPS-NPViewer. Anticancer agents cluster in accordance with their mode of action using ChemGPS-NPWeb. Antimetabolites are coloured green, alkylating agents red, topoisomerase I inhibitors blue, topoisomerase II inhibitors pink, and tubulin active agents black.

Potential applications of ChemGPS-NPWeb

ChemGPS-NP can be applied in several kinds of drug discovery related endeavours. Several specific examples are provided in refs [2, 6, and 7]. ChemGPS-NP can be used for cluster analyses, for evaluation of molecular similarity, and for characterization of large data sets. Furthermore ChemGPS-NP can function as a reference system by which large libraries can be compared without changing the coordinates and thus assist in prioritization and selection of suitable lead compounds in drug discovery. It is a well-known and often quoted paradigm of medicinal chemistry that compounds with similar chemical structures and properties often have similar biological activities.

Known inhibitors of a certain target can be mapped together with several available compounds. Then those situated close to the known inhibitors (neighborhood mapping) can be selected for further testing, thereby increasing the possibilities of hit-generation. This was proven successful in a recent study where a property-based similarity search, based on calculated 8D Euclidean distances from ChemGPS-NP, was used to identify natural product inspired potential leads for drug discovery [6]. ChemGPS-NP has also recently been demonstrated to be able to differentiate between different anticancer modes of action (see Figure 3) [7].

Conclusions

In this application note we describe how SIMCA-QP with its stability and flexibility forms the computational core of a publicly available web tool, ChemGPS-NPWeb. While on the one hand not being an out of the box application as SIMCA, the modular design allows for specific adaptations. In addition to SIMCA-QP a set of additional software and scripts are required to all work together to complete an entire web tool as ChemGPS-NPWeb. The industry grade performance of SIMCA-QP has proven able to handle very large datasets, and the average completion time of a job through the entire computational process for the first 2110 submissions to ChemGPS-NPWeb was a mere 27.1 seconds.

References

1. https://chemgps.bmc.uu.se (for non-commercial use only)
2. Rosén, J., Lövgren, A., Kogej, T., Muresan, S., Gottfries, J., Backlund, A. ChemGPS-NPWeb: chemical space navigation online. Journal of Computer-Aided Molecular Design 2009, 23, 253-259
3. Larsson, J.; Gottfries, J.; Muresan, S.; Backlund, A. ChemGPS-NP: tuned for navigation in biologically relevant chemical space. Journal of Natural Products 2007, 70, 789–794
4. Oprea, T. I., Gottfries, J. Chemography: the art of navigating in chemical space. Journal of Combinatorial Chemistry 2001, 3, 157–166
5. Talete srl, DragonX. Software for molecular descriptor calculations. Linux version 2007. http://www.talete.mi.it/
6. Rosén, J., Gottfries, J., Muresan, S., Backlund, A., Oprea, T. I. Novel chemical space exploration via natural products. Journal of Medicinal Chemistry 2009, 52, 1953-1962
7. Rosén, J., Rickardson, L., Backlund, A., Gullbo, J., Bohlin, L., Larsson, R., Gottfries J. ChemGPS-NP Mapping of Chemical Compounds for Prediction of Anticancer Mode of Action QSAR & Combinatorial Science 2009, 28, 436-446Text