J Comput Aided Mol Des (2009) 23:253–259
DOI 10.1007/s10822-008-9255-y

ChemGPS-NP$_{Web}$: chemical space navigation online

Josefin Rosén · Anders Lövgren · Thierry Kogej · Sorel Muresan · Johan Gottfries · Anders Backlund

Received: 9 September 2008 / Accepted: 22 November 2008 / Published online: 10 December 2008
© Springer Science+Business Media B.V. 2008

Abstract Internet has become a central source for information, tools, and services facilitating the work for medicinal chemists and drug discoverers worldwide. In this paper we introduce a web-based public tool, ChemGPS-NP$_{Web}$ (http://chemgps.bmc.uu.se), for comprehensive chemical space navigation and exploration in terms of global mapping onto a consistent, eight dimensional map over structure derived physico-chemical characteristics. ChemGPS-NP$_{Web}$ can assist in compound selection and prioritization; property description and interpretation; cluster analysis and neighbourhood mapping; as well as comparison and characterization of large compound datasets. By using ChemGPS-NP$_{Web}$, researchers can analyze and compare chemical libraries in a consistent manner. In this study it is demonstrated how ChemGPS-NP$_{Web}$ can assist in interpreting results from two large datasets tested for activity in biological assays for pyruvate kinase and *Bcl-2* family related protein interactions, respectively. Furthermore, a more than 30-year-old suggestion of "chemical similarity" between the natural pigments betalains and muscaflavins is tested.

Keywords ChemGPS-NP · Natural products · Chemical space · Internet tool · Drug discovery · Biologically active compounds

J. Rosén (✉) · A. Backlund
Division of Pharmacognosy, Department of Medicinal Chemistry, Uppsala University, BMC Box 574, 751 23 Uppsala, Sweden
e-mail: josefin.rosen@fkog.uu.se

A. Lövgren
IT-/Computing Department, BMC Box 570, 751 23 Uppsala, Sweden

T. Kogej · S. Muresan
DECS Global Compound Sciences, AstraZeneca R&D, 431 83 Mölndal, Sweden

J. Gottfries
Pharmnovo Inc., Sahlgrenska Science Park, 413 46 Gothenburg, Sweden

Introduction

Internet technology offers an exceptional framework to develop public scientific applications. The World Wide Web has become a central source for information, education, tools, and services to support medicinal chemists and drug discoverers and, over the last years, a growing number of web-based tools for data analysis in chemistry have been made available [1, 2]. In this paper we introduce a web-based public tool ChemGPS-NP$_{Web}$ (http://chemgps.bmc.uu.se), for comprehensive chemical space navigation and exploration. ChemGPS-NP [3, 4] is a principal component analysis (PCA) [5] based global chemical positioning system [6] tuned for exploration of biologically relevant chemical space, i.e. those areas of chemical space most likely to enclose biologically active compounds. This is achieved in terms of global mapping onto a consistent, eight-dimensional (8D) map based on structure-derived physico-chemical characteristics for a reference set of compounds. The first four dimensions of the ChemGPS-NP map, accounting for 77% of data variance, can be interpreted in such a way that the first dimension (principal component one, PC1) represents size, shape and polarizability; PC2 corresponds to aromatic and conjugation-related properties; PC3 describes lipophilicity, polarity, and H-bond capacity; and PC4 expresses flexibility and rigidity. Any compound with a known chemical structure

can be positioned onto this map using interpolation in terms of PCA score prediction. From the resulting projections properties of the compounds can be compared and easily interpreted together with trends and clusters.

In this article we review the design, features, and proposed fields of application of ChemGPS-NP as facilitated by the public web tool ChemGPS-NP$_{Web}$. To demonstrate how interpretation of large datasets can benefit from ChemGPS-NP, two publicly available sets of data were compiled from PubChem. The first set, of close to 50,000 compounds, has been tested for activity in a pyruvate kinase assay. This set has also been evaluated by Schuffenhauer et al. [7] from a scaffold-based viewpoint, and the obtained results have different implications. The second set, of more than 190,000 compounds, has been tested in a high-throughput screening (HTS) assay to identify regulators of protein interactions between *Bim* and six *Bcl-2* family members, which have a role in apoptosis regulation. ChemGPS-NP is here used to overview the assay results, and to identify and circumscribe the volumes of chemical space in which the confirmed active compounds reside. As a central premise in medicinal chemistry stipulates that compounds with similar structures and properties in many cases have similar biological activities, it is implied that this information could assist in future selection of test compounds. By selecting compounds that in the ChemGPS-NP 8D map are positioned close to those known actives, a requirement of complex structure-based physicochemical properties will be fulfilled. The third set comprises a much smaller set of molecules, for which an assumed chemical similarity issue is discussed.

Areas of application

It has recently been demonstrated that ChemGPS-NP can be used to differentiate between different anticancer modes of action [8].

In this study we have employed ChemGPS-NP to analyse three sets of bioactive molecules; two sets extracted from the PubChem database and one in-house set, to exemplify additional uses.

The first dataset consists of 50,629 molecules tested in a pyruvate kinase assay [9]. This particular dataset is available from PubChem, and has the advantage of also been evaluated by Schuffenhauer et al. [7], using a scaffold-based approach. This is fundamentally different compared to analysis with ChemGPS-NP, which rely on PCA and score predictions of computed descriptors. Both approaches are complementary, providing different types of information. From the perspective of natural products, scaffolds hold a particular interest as it can be assumed that there is a close connection between scaffolds and biosynthetic pathways. In this particular case, the scaffold-based method evaluated the 602 active and 50,027 inactive compounds, from which eleven active substances with 2-phenyl-benzooxazole scaffolds were identified as a privileged group with high proportion of active compounds. Accessing the same dataset, including also 812 compounds that provided inconclusive results, it turned out that 587 of the active, 793 of the inconclusive, and 48,174 of the inactive had SMILES representations for which prediction scores could be immediately calculated using the ChemGPS-NP$_{Web}$-server. The additional compound representations required additional manipulations to be valid SMILES. Mapping of these data, and highlighting the privileged group mentioned above, provided us with the representations in Fig. 1.

What ChemGPS-NP tells us, is that the physicochemical properties of active, non-active and inconclusive compounds tested in this bioassay are largely overlapping—at least in the first three dimensions of ChemGPS-NP chemical property space. The privileged group discussed above all fall well within these parameters, with one exception for their compound 11 [7], the closest observation in Fig. 1. This particular compound is set aside by a much higher prediction score in PC1, primarily depending on size parameters, as compared to the others in the privileged group. This result is immediately comprehendible when comparing it to the other structures. In addition to the aberrant physico-chemical properties demonstrated for this compound, it is also from the bioassay data concluded that it is one of the least active compounds in the privileged group identified. Color coding the eleven privileged structures of this group according to their assay responses, emphasize and further narrow the volume of the potentially most interesting active compounds (Fig. 2).

The second set of SMILES was also extracted from PubChem, and has been tested in an HTS assay at University of New Mexico. One aspect of apoptosis is regulated by the *Bcl-2* protein family members, and their interactions with *Bim*. On *Bim* there is a region known as *BH3* that is required for *Bcl-2* binding and also accounts for most of *Bim* cytotoxicity. Interactions between *Bim* and members of the *Bcl-2* family can thus regulate apoptosis. Hence, for cancer therapeutic purposes it is of interest to identify compounds able to disturb or regulate this protein interaction. With this aim interactions between the *BH3* region of *Bim* and the following six *Bcl-2* family members: *Bcl-xL*, *Bcl-w*, *Bcl-B*, *Bfl-1*, *Mcl-1* and *Bcl-2*, were investigated to identify small molecule regulators of their interaction with *Bim* [10]. The primary screen was performed with 194,920 compounds of which 884 were found to be active according to the hit selection criterion (change in percentage inhibition >40%). These 884 compounds were subsequently tested in a dose response format to confirm activity and determine potency. Three compounds were found to regulate interactions

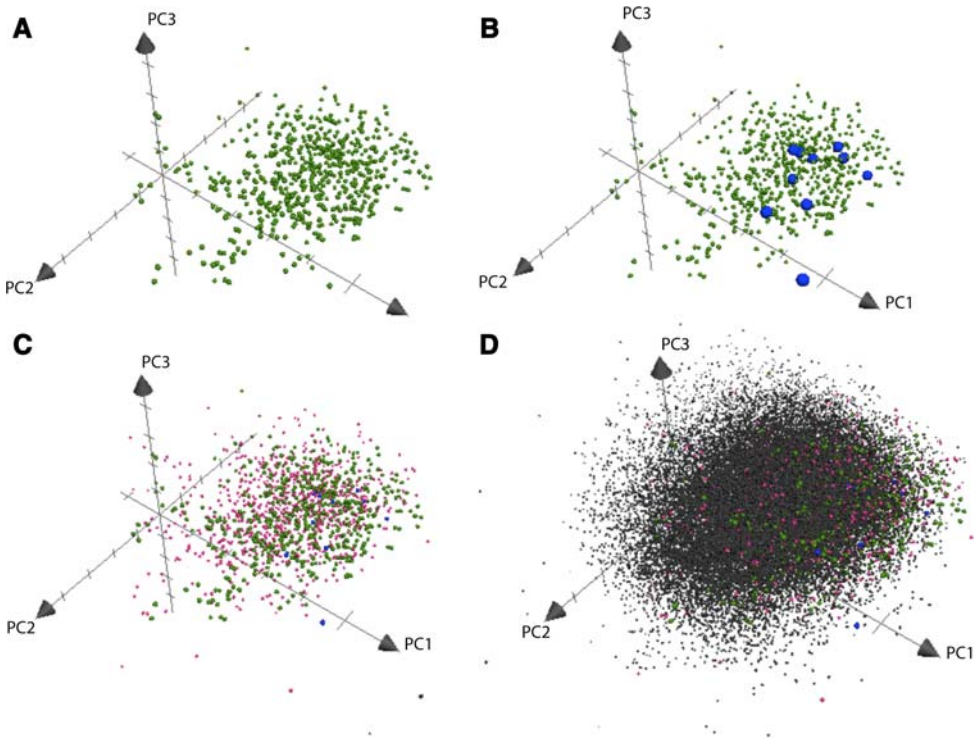

Fig. 1 Structures as SMILES downloaded from PubChem and ChemGPS-NP prediction scores calculated using the on-line tool ChemGPS-NP$_{Web}$, total processor time 16.2 s. Results plotted with Apple^{TM} system software Grapher 2.0. **a** Only confirmed active compounds (587, *green*), **b** active compounds with 2-phenyl-benzooxazoles, identified as privileged by Schuffenhauer et al. [7], highlighted (11, *blue*), **c** compounds with inconclusive results (793, *pink*), and **d** all compounds tested including non active (48,174, *gray*). Main influence in PC1 is size, in PC2 is aromaticity, and in PC3 is lipophilicity [4]

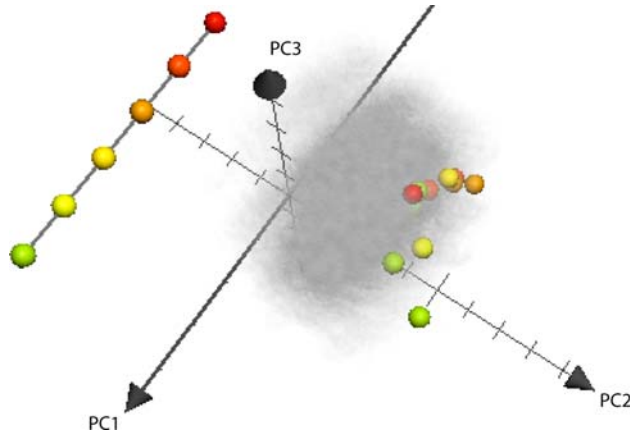

Fig. 2 ChemGPS-NP mapping of the privileged group identified by Schuffenhauer et al. [7], color coded according to legend (*green* lowest and *red* highest activity)

between *Bim* and Mcl-1, six compounds between *Bim* and *Bcl-B*, 18 compounds between *Bim* and *Bcl-w*, 97 compounds between *Bim* and *Bfl-1*, six compounds between *Bim* and *Bcl-xL*, and nine compounds regulated the interactions between *Bim* and *Bcl-2*. The majority of the 194,920 compounds did not affect the protein interactions. In Fig. 3 the 194,830 inactive compounds are shown in blue, while the compounds active against any of the targets are shown in black. By predicting them in ChemGPS-NP it is possible to locate those volumes of chemical space where the active compounds reside. It is obvious from the plot that the inactive compounds span a much larger area. The more extreme volumes of space spanned is with no exceptions populated by inactive compounds. Based on the well known argument in medicinal chemistry that compounds with similar structures and properties often have similar biological activities [11], this information can be used when selecting test compounds for future screenings. Compound libraries could be positioned with ChemGPS-NP, and those compounds residing outside the 'active area' could be excluded for the benefit of compounds located closer to the active compounds. To demonstrate this we selected a subset of 42 confirmed active compounds using D-optimal onion design [12–14]. D-optimal designs select the most extreme points of the candidate set and give a minimal set of selected compounds with maximum diversity. D-optimal onion designs divide the set into a number of selected layers where one separate D-optimal design is made in each layer. Thereby it samples more evenly through the region and provides a better representation of the data set. The subset of 42 active compounds was positioned in ChemGPS-NP and the 'active volume' was defined. Subsequently the remainder of this compound set was positioned in ChemGPS-NP. After eliminating those compounds positioned outside of the 'active volume', only 52% of the compound set remained. This set included 91% of the active compounds. Accordingly, if this were a set aimed for future screening in this assay, 52% of the compounds could have been left out after initial ChemGPS-NP positioning, while still finding 91% of the actives, saving both time and resources.

The third dataset plotted in Fig. 4 is an in-house compiled set of betalains. The betalains are found in nine of the

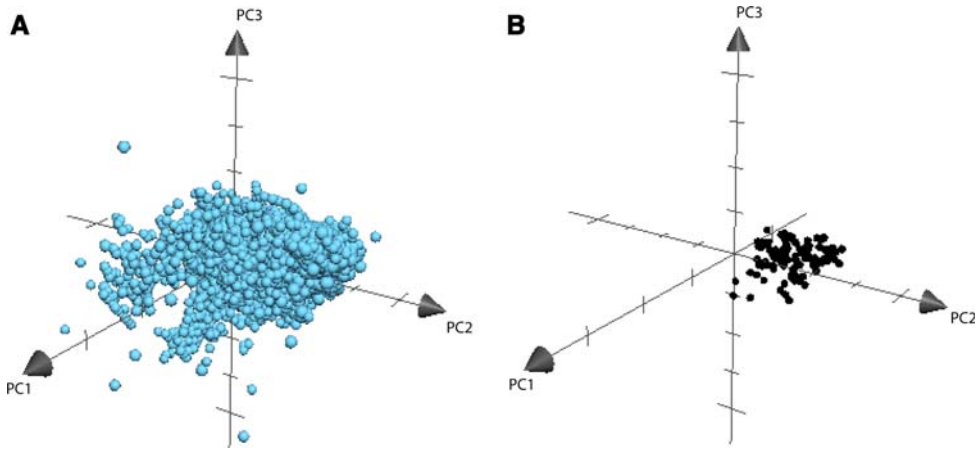

Fig. 3 **a** ChemGPS-NP mapping (three-first dimensions, PC1–PC3) of all compounds tested in the *Bcl-2* assay at University of New Mexico **b** confirmed actives. All active compounds are localized to one sector of chemical space, forming a cluster with distinct *borderlines*

eleven families of the plant order Caryophyllales, where they constitute red, violet, and yellow pigments [15]. Despite their small number and apparent structural homogeneity, the betalains span a comparably large sector of the chemical property space, in particular along PC1. Betalains have several applications in food industry [16, 17], have also recently been the object of several pharmaceutical studies [18–20], and their biosynthesis has been thoroughly investigated [21–23]. The two main biosynthetic groups betacyanins (red and violet pigments) and betaxanthins (yellow pigments) are clearly separated by their physicochemical properties. The betacyanins are mapped further out in the positive direction of PC1 (Fig. 4), indicating larger size and higher polarizability while the betaxanthins are positioned further out in the positive direction of PC4 (not shown), indicating that they are more flexible then the betacyanins. These features are also evident from the sample structures shown in Fig. 5. The muscaflavins, identified as yellow pigments in e.g. the common toadstool *Amanita muscaria* have, on biosynthetic grounds, been suggested as chemical relatives of the betalains, as they share a common route to the enzyme DOPA-dioxygenase [23, 24]. In this study two muscaflavins were included, and they clustered together with the other yellow pigments, the betaxanthins. This confirms a long-standing suggestion that the muscaflavins show a 'chemical similarity' to the betalains, although originating from very distantly related organisms and partly different biosynthesis routes.

Technical details

General

ChemGPS-NP$_{Web}$ includes a number of different programs and libraries that interact with each other according to the

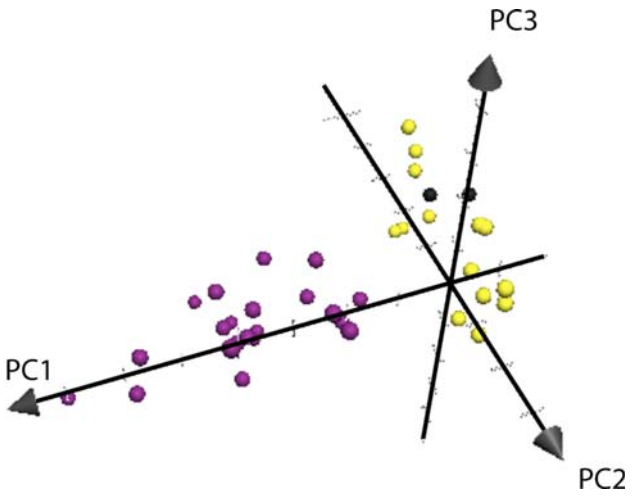

Fig. 4 The obtained coordinates for the betalains plotted in the first three dimensions of ChemGPS-NP. Betacyanins are shown in *violet*, betaxanthins in *yellow*, and the muscaflavins in *black*. The two yellow pigments muscaflavins and betaxanthins cluster together and are clearly separated from the betacyanins which, according to the model, are less flexible and larger

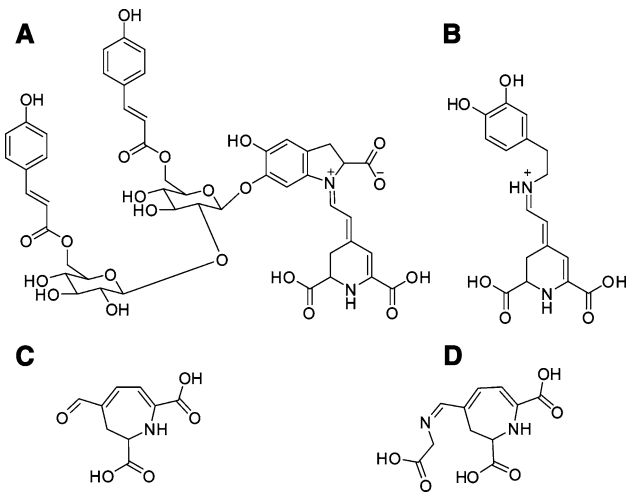

Fig. 5 Sample structure formulas for: **a** a betacyanin (betanidin 6-*O*-(6′,6″-di-*O*-*E*-4-coumaroyl)-*β*-sophoroside); **b** a betaxanthin (miraxanthin V); and **c**, **d** the two fungi-derived compounds muscaflavin and hygroaurin

traditional UNIX-model, where each part performs a well defined task and together they produce the desired output.

The system includes three main elements: DragonX [25] for calculation of molecular descriptors; Simca-QP [26] for multivariate predictions; and Batchelor [27] the web interface and batch queue manager that allows jobs with long run times to be submitted to the web server and scheduled for later execution by its batch queue. The programs exchange information with the web interface by storing information on the file system, which acts as the database.

Work-flow

When the work-flow is initiated by the queue handler, the uploaded SMILES-strings [28] are first preprocessed. An initial Perl script removes erroneous SMILES, as well as information about stereochemistry and isotopes. Subsequently, the number of rows in the file is checked against the maximum allowed (at present 8,000). The preprocessed SMILES are then submitted to DragonX [25], which serves as an internal engine for calculation of the molecular descriptors from which the eight principal components are extracted [4]. ChemGPS-NP_{Web} uses 40 molecular descriptors from DragonX. Six of these are subsequently summarized into an additional descriptor (n_amid) by a Perl script. In total the initial model as well as the final score predictions are based on 35 molecular descriptors. The complete list of molecular descriptors used is published by the authors elsewhere [4]. The Perl script then organizes and prepares the obtained data and the resulting matrix is used as input data to cgpsclt (a client) that connects to cgpsd (a server) to run Simca-QP. The latter is a calculation engine for multivariate predictions. In ChemGPS-NP_{Web} 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 coordinates for each compound) back to cgpsclt, which stores them in the database. If cgpsd (the server) is not available the prediction will instead be performed locally by cgpsstd (standalone program). Users can monitor the status of their submitted jobs (pending, running or finished) and later download the result from the queue. The coordinates can then be plotted using any preferred software. Here we have used Grapher 2.0 distributed together with MacOS X. Post computational statistics are compiled based on results from each of the successive computational steps.

Figure 6 describes how the different computing elements interact with each other.

The extra step with 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.

All elements (DragonX, Simca-QP, and cgpsd) are multithreaded, which becomes more and more important to take advantage of the increasing number of cores (CPUs) in server hardware.

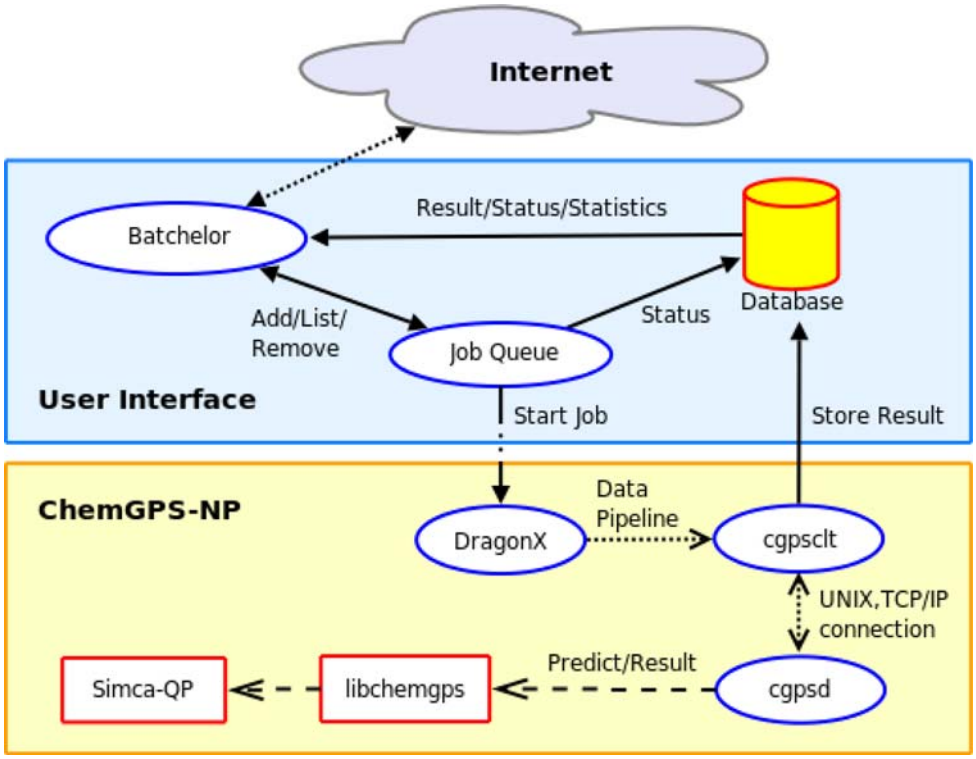

Fig. 6 Flowchart describing the interaction between the different elements of ChemGPS-NP_{Web} application

The queue manager and web interface

The queue manager (Batchelor) web interface makes it possible to upload data (SMILES) and to download results from the ChemGPS-NP$_{Web}$ runs. The job queue can be filtered and sorted according to different criteria. Uploaded data and corresponding results are personal and can only be reached from the same computer as the job was initiated from. The information presented to the user is in part obtained from the database (results and statistics), and in part directly from the job queue (job status).

System information

Currently, the entire process runs on one single computer, a 64 bit 2 × Quad Core Xeon operating at 1.6 GHz with 4 GB RAM, using GNU/Linux as operating system.

Discussion

The drug discovery process is today held back by increasing costs and high attrition-rates, with an overall decrease in the number of annually registered new chemical entities. Considering the immensity of chemical space, which is estimated to exceed 10^{60} possible compounds when only small carbon-based compounds are considered [29], it is clear that the process of compound selection and prioritization is essential. An efficient selection process will increase the quality of a hit or a lead and considerably speed-up the hit and lead identification process. ChemGPS-NP provides a framework for making compound selection more efficient, thereby increasing the probability of hit generation in the search for novel bioactive molecules. One feature is the function as a reference system by which large libraries can be compared without changing the co-ordinates as novel compounds are handled via interpolation and therefore avoids extrapolations. This becomes possible since the PCA property space is well covered in all directions by relevant structures in the reference set. Through its carefully selected reference set of compounds and molecular descriptors, ChemGPS-NP is tuned for exploration of the regions of chemical space most likely to enclose compounds with biologically relevant functions and activities. In this context it provides compound property description, clustering overview and property interpretation via the PCA loading vectors, and has also recently been demonstrated to be able to differentiate between different anticancer modes of action [8].

The benefits of ChemGPS-NP$_{Web}$ are, in one way, comparable to the possibilities opened in molecular biology by rigorous application of the BLAST algorithms [30]. These allow, for example through web-interfaces, the research community to easily compare sections of nucleotide or amino-acid sequences for homology searching, identifying genes, or preparing datasets for phylogenetic analyses, all in huge datasets.

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 [11]. Known inhibitors of a certain target can be mapped by ChemGPS-NP together with a number of available compounds from which those situated close to the known inhibitors (neighbourhood mapping) can be selected, thereby increasing the possibilities of hit generation. ChemGPS-NP can handle the processing of very large data sets, which makes this approach useful for the analysis of results of HTS campaigns as illustrated in Figs. 1, 2, and 3. The possibility to color the observations on the basis of the presence of molecules with a specific type of activity makes this method particularly useful for presenting the results to colleagues in a pedagogic way. An alternative selection procedure could be performed if, for instance, only a small number of compounds from an initial large set needs to be selected for testing (e.g. because of high screening costs) against a target without prior knowledge of active compounds. The initial set can then be mapped in ChemGPS-NP and a diverse set of compounds can be selected using sampling techniques based on cluster analysis and neighbourhood mapping [31] to explore as large parts of chemical space as possible and at the same time avoiding testing too similar compounds.

In summary, we have developed a free public internet tool for chemical space navigation. ChemGPS-NP$_{Web}$ can assist in compound selection and prioritization; property description and interpretation; clustering overviews; as well as comparison and characterization of large datasets. ChemGPS-NP$_{Web}$ enables researchers worldwide to analyze and compare their chemical libraries online in a consistent manner. During the first 3 months online, more than 1.4 million compounds have been predicted and positioned in the ChemGPS-NP chemical space map.

Acknowledgments Instrumental at initial stages in implementing the ChemGPS-NP$_{Web}$ were Gustavo Gonzales-Wall and Nils-Einar Eriksson at the IT-/Computing Department at BMC. The authors are grateful for software support from UMETRICS and TALETE.

References

1. Tetko IV (2005) Drug Discov Today 10:1497. doi:10.1016/S1359-6446(05)03584-1
2. Tetko IV, Gasteiger J, Todeschini R et al (2005) J Comput Aided Mol Des 19:453. doi:10.1007/s10822-005-8694-y
3. Larsson J, Gottfries J, Bohlin L et al (2005) J Nat Prod 68:985. doi:10.1021/np049655u
4. Larsson J, Gottfries J, Muresan S et al (2007) J Nat Prod 70:789. doi:10.1021/np070002y

5. Eriksson L, Johansson E, Kettaneh-Wold N et al (2006) Multi- and megavariate data analysis part i basic principles and applications, 2nd edn. Umetrics AB, Umeå
6. Oprea TI, Gottfries J (2001) J Comb Chem 3:157. doi:10.1021/cc0000388
7. Schuffenhauer A, Ertl P, Roggo S et al (2007) J Chem Inf Model 47:47. doi:10.1021/ci600338x
8. Rosén J, Rickardson L, Backlund A et al (2008) QSAR Comb Sci (submitted)
9. Inglese J, Auld DS, Jadhav A et al (2006) Proc Natl Acad Sci USA 103:11473. doi:10.1073/pnas.0604348103
10. Adams JM, Cory S (1998) Science 281:1322. doi:10.1126/science.281.5381.1322
11. Martin YC, Kofron JL, Traphagen LM (2002) J Med Chem 45:4350. doi:10.1021/jm020155c
12. de Aguiar PF, Bourguignon B, Khots MS et al (1995) Chemometr Intell Lab Syst 30:199. doi:10.1016/0169-7439(94)00076-X
13. Olsson IM, Gottfries J, Wold S (2004) J Chemometr 18:548. doi:10.1002/cem.901
14. Olsson I-M, Gottfries J, Wold S (2004) Chemometr Intell Lab Syst 73:37. doi:10.1016/j.chemolab.2004.04.001
15. Grotewold E (2006) Annu Rev Plant Biol 57:761. doi:10.1146/annurev.arplant.57.032905.105248
16. Castellar R, Obon JM, Alacid M et al (2003) J Agric Food Chem 51:2772. doi:10.1021/jf021045h
17. Kanner J, Harel S, Granit R (2001) J Agric Food Chem 49:5178. doi:10.1021/jf010456f
18. Allegra M, Tesoriere L, Livrea MA (2007) Free Radic Res 41:335. doi:10.1080/10715760601038783
19. Galati EM, Mondello MR, Lauriano ER et al (2005) Phytother Res 19:796. doi:10.1002/ptr.1741
20. Tesoriere L, Butera D, Pintaudi AM et al (2004) Am J Clin Nutr 80:391
21. Delgado-Vargas F, Jimenez AR, Paredes-Lopez O (2000) Crit Rev Food Sci Nutr 40:173. doi:10.1080/10408690091189257
22. Kobayashi N, Schmidt J, Wray V et al (2001) Phytochemistry 56:429. doi:10.1016/S0031-9422(00)00383-6
23. Strack D, Vogt T, Schliemann W (2003) Phytochemistry 62:247. doi:10.1016/S0031-9422(02)00564-2
24. Mueller LA, Hinz U, Zryd J-P (1997) Phytochemistry 44:567. doi:10.1016/S0031-9422(96)00625-5
25. Talete srl, DragonX (2008) Software for molecular descriptor calculations. Linux version-2007. http://www.talete.mi.it/. Accessed 15 November 2008
26. SIMCA-QP software Umetrics AB (2008) Umeå, Sweden. http://www.umetrics.com/. Accessed 15 November 2008
27. Batchelor (2008) http://it.bmc.uu.se/andlov/proj/batchelor/. Accessed 15 November 2008
28. Weininger D (1988) J Chem Inf Comput Sci 28:31. doi:10.1021/ci00057a005
29. Bohacek RS, McMartin C, Guida WC (1996) Med Res Rev 16:3. doi :10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6
30. Altschul SF, Gish W, Miller W et al (1990) J Mol Biol 215:403
31. Brown RD, Martin YC (1996) J Chem Inf Model 36:572. doi:10.1021/ci9501047