Similarity in Chemical and Protein Space: Finding novel starting points for library design

Summary of the Ph.D. thesis of Stefan Wetzel

Chemical genomics, i.e. the use of small molecule modulators of protein function to study the underlying biological processes, lies at the heart of chemical biology. The work presented herein aimed at the development and application of computational approaches for compound library design. The methods that were developed build on the structural complementarity of small molecule and protein space to map and explore biologically relevant parts of chemical space as well as the corresponding protein targets.

1 Cartography of and Navigation in Chemical Space

1.1 Scaffold Tree

The scaffold tree is a hierarchical classification of chemical structures based on chemically meaningful scaffolds, i.e. all rings and connecting linker chains, as well as all double bonds directly attached to these. Iterative deconstruction by one ring at a time guided by a set of rules developed in collaboration with Novartis generates a branch of scaffolds rooted in the one ring scaffold. In this hierarchy, the smaller scaffold is termed the “child” scaffold and the larger scaffold the “parent”. Classification of many structures results in a tree-like scaffold diagram. Scaffolds populating the same branch may also share other properties, i.e. bioactivity. The term “brachiation” describes movement from larger scaffolds towards smaller scaffolds while keeping similar biochemical activity. Scaffolds that do not represent molecules in the dataset are incorporated into the scaffold tree and termed “virtual scaffolds”.

1.2 Scaffold Hunter

Scaffold trees chart chemical space in a chemically meaningful and intuitive way. To enhance their applicability an interactive scaffold tree browsing program named “Scaffold Hunter” was developed within a joint student project with the Chair of Algorithm Engineering at the Technical University of Dortmund.

Scaffold Hunter automatically generates a visual representation of the scaffold tree based on the data read from a database. It enables navigation in the scaffold tree including filtering, zooming, colour shading according to properties and bookmarking. The scaffold tree database can be easily generated with a second tool, the “Scaffold Tree Generator” written by Dr. Steffen Renner from any SD file, a standard open file format for molecular structures. Both programs are available free of charge under an open source license via www.scaffoldhunter.com.

1.3 Finding and filling gaps in chemical space

As described above, virtual scaffolds represent gaps in the chemical space that can be exploited by identification of promising virtual scaffolds by their proximity to scaffolds representing potent compounds. An initial retrospective study provided proof of concept comparing virtual scaffolds from PubChem data to known active compounds in literature. In a prospective study the pyruvate kinase screen data stored in PubChem was analyzed with Scaffold Hunter. Out of the 65 promising virtual scaffold identified, small focused libraries based on four of these scaffolds were acquired and tested as modulators of pyruvate kinase activity. The biochemical screen yielded eight confirmed hits that had not been described as modulators of any protein before according to a SciFinder search.

1.4 Exploring Natural Products: the γ-pyrones

One natural application of Scaffold Hunter is the comparison of different sets of compounds by merging of their scaffold trees. This may also be used to annotate one of the sets with the properties of the other, i.e. protein target information. In a first model case, the natural product chemical space was annotated with target information from the WOMBAT database. The branch of the γ-pyrones showed promising biochemical activities and was synthetically accessible. A library of higher γ-pyrones (2- to 4-ring scaffolds) was compiled and screened against the monoamine oxidases (MAO) subtype A and B, the signal transducers and activators of transcription (STAT) proteins and acid sphingomyelinase, which were selected as targets because activity was described for one or two scaffold types of the γ-pyrone branch. All together, the screens yielded a significant number of selective hits for all three protein families with remarkable selectivity – also on the iso-enzyme level.

1.5 Outlook

Future extensions of Scaffold Hunter are described, e.g. several mechanisms to enable the generation of scaffold trees according to customized set of rules, also invoking biology-guided scaffold trees as developed by S. Renner. The possible extension of the scaffold tree for the generation of chemically meaningful natural product fragments is described. Such fragments possibly facilitate the exploration of nature's diversity with a reasonable synthetic effort.

2 Exploration of Proteomic Space – Protein Structure Similarity Clustering (PSSC)

PSSC builds on the concept that proteins with structurally similar binding sites bind similar ligands. Hence, the structural similarity can be explored to cluster protein targets more likely to be addressed by certain for small compound structural motifs.

2.1 Method Development – Automated PSSC

The newly devised PSSC process centres on the structural alignment of "ligand-sensing cores", spherical cut-outs around the binding site, instead of full proteins, which drastically reduces false positive results. Definition of criteria for the size of ligand-sensing cores was based on an evaluation of catalytic sites annotated in the Catalytic Site Atlas (CSA). A self-written computer program

automatically extracted ligand-sensing core structures optimized for subsequent structure comparison from PDB files using the binding site information from the CSA. Structural alignments were computed with DaliLite. Results are comparable with those of earlier analyses since the FSSP database was also compiled with DaliLite. This alignment data is then clustered by an adapted implementation of the OptiSim clustering algorithm written in Java. Comparison with the SCOP database, the “gold standard” in structural similarity, yielded a large number of clusters that were also classified as similar by SCOP but also 33 clusters where SCOP predicts only less than half of the cluster members. For experimental validation, two proteins from a PSSC cluster, pyruvate kinase (PK) and dihydropteroate synthetase (DHPS) were chosen and a library of 740 sulfanilamides, a known class of DHPS inhibitors, was screened for PK inhibition. The screen, however, did not yield any hit compounds, which may be due to the assay conditions of the screening system used but could also imply that PSSC does not work in this case. A final conclusion about the applicability of PSSC is not possible based on these results.

For large scale PSSC analyses, development of a faster structure comparison method based on protein fingerprints was begun. Initial cross-validation showed the new fingerprint based similarity in fair agreement with the Dali data for the 15,000-membered core set identifying further potential for optimisation.

2.2 Method development – PSSC with dynamic protein structures

One inherent drawback of the PSSC method is the demand for a structure of the protein of interest. Although PSSC has been shown to work with homology models induced fit might still be problematic. Together with B.D. Charette from Prof. Berkowitz‘s group in Lincoln, Nebraska it was successfully shown that molecular dynamics could facilitate the transition from an apo protein to a ligand bound structure simulating the induced fit needed for subsequent PSSC analysis.

3 Miscellaneous projects

During the doctoral work present herein, several other projects were supported as well. This work includes NMR-restrained conformational analysis of macrocycles using force fields as well as docking of inhibitor structures to gain insights into experimentally observed selectivities. Further work included the evaluation, statistical analysis and quality control of biochemical screening data and subsequent optimization of the experimental screening protocol.