

## DSM Science & Technology Awards 2006

Name	Marcus Koch
University	University of Dortmund (D)
Department	Max Planck-Institute of Molecular Physiology, Department Chemical Biology
PhD Supervisor	Prof. Dr. H. Waldmann

# Protein and Natural Product Structure as Guiding Principles for Compound Library Design

- Summary of Ph.D. Thesis -

Dr. Marcus A. Koch

In the Ph.D. thesis "Protein- und Naturstoffstruktur als Leitprinzipien für die Entwicklung von Verbindungsbibliotheken", structure-based principles of the "world of proteins" and the "world of natural products" were used in order to develop guidelines and concepts for the targeted design of compound libraries. The guidelines were applied in concrete ligand design situations, i.e. chemical libraries were assembled and tested in biochemical and cell-based assays for validation of the underlying concepts.

The first compound library design concept developed within this Ph.D. thesis, "Protein Structure Similarity Clustering" (PSSC), is based on a purely structure-driven view on proteins. A bioinformatics procedure was being developed in order to identify structural similarities in proteins (see Fig. 1) using web-based databases (Dali/FSSP, CE). [1]

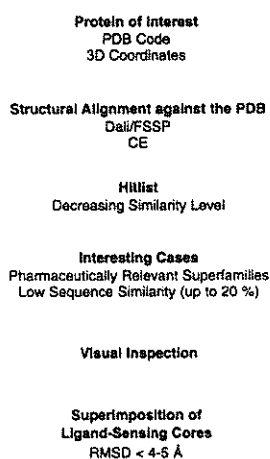


Fig. 1: Database search strategy and procedure developed for the identification of protein structure similarity clusters. Database searches, for example, in the Dali/FSSP (<http://www.ebi.ac.uk/dali>) and the Combinatorial Extension (CE) (<http://cl.sdsc.edu/ce.html>) databases, using the 3D coordinates of a query protein could provide insights into their structural neighborhood. For compound library development, the ligand-sensing cores of the proteins are of paramount importance. Therefore, it is imperative that these relevant parts of the protein domains share structural similarity.

This bioinformatics search algorithm was based on the insight that the majority of all proteins is modularly built from a limited set of approximately 1,000 structural domains.

Searches were performed using the Protein Data Bank (PDB) code of the protein domain of interest. Alternatively, 3D coordinates of a query protein domain structure may be used. The employed databases are based on exhaustive all-against-all 3D structure comparisons of protein structures currently included in the PDB based on their C<sup>α</sup> traces. The output of such a database search is a hit list in which protein domains are ranked with decreasing similarity level (3D and sequence similarity). From this hit list proteins belonging to pharmaceutically relevant protein families or super-families with low sequence similarity (sequence identities up to 20%) which usually cannot be grouped according to structural similarity using common sequence-based comparison algorithms were chosen and visually inspected. For a comparative analysis of protein domains with respect to the development of small molecule protein ligands, the relevant domain subsets, i.e. the catalytic or ligand-sensing cores, were identified and defined. This was of paramount importance whenever protein domains of significantly different sizes were compared. It was investigated whether the smaller domain in the set of protein domains to be compared was part of the larger domain with respect to principal structural elements. If this was the case it was investigated whether these common structural elements describe the catalytic or ligand-sensing cores of the larger domain. In the end, the ligand-sensing cores had to be similar and superimposed. This led to the formation of so-called protein structure similarity clusters in which domain core structures were clustered based exclusively on structural similarity considerations irrespective of any evolutionary or functional relationships. The scaffolds of known ligands of member proteins in such a similarity cluster were then used as biologically pre-validated guiding structures in chemical space for the development of ligands for the other member proteins.

The validity of the PSSC approach was verified in a concrete compound library design project. Thus, performing the above delineated database search strategy, using the core of the catalytic domain of Cdc25A phosphatase as search motif, a protein structure similarity cluster was formed containing Cdc25A, the dismantled catalytic core of acetylcholinesterase (AChE) and the ligand-sensing cores of 11 $\beta$ -hydroxysteroid dehydrogenase (11 $\beta$ HSD) type 1 and type 2 (see Figure 2A). As at the time of investigation no crystal structures of both 11 $\beta$ HSD isoenzymes were available, homology models had to be constructed and used as basis for the structural comparison. A known naturally occurring inhibitor of Cdc25A, dysidiolide (**1**, Figure 2B), was used as leitmotif for the synthesis of a 147 compound library. This compound collection was subjected to biochemical investigation for possible

inhibition of Cdc25A, AChE, or 11 $\beta$ HSD1/2. Compounds displaying IC<sub>50</sub> values  $\leq 10 \mu\text{M}$  were considered as hits. Of the 147 compounds investigated, 42 qualified as hits in the Cdc25A assay. The most potent compound (**2**) had an IC<sub>50</sub> value of 350 nM, which is significantly lower than the reported IC<sub>50</sub> value for dysidiolide (9.4  $\mu\text{M}$ ). Three compounds inhibited AChE with IC<sub>50</sub> values of 1.3–4.5  $\mu\text{M}$ . The collection contained three 11 $\beta$ HSD1 inhibitors with IC<sub>50</sub> values of 7.8–10  $\mu\text{M}$  and four 11 $\beta$ HSD2 inhibitors with IC<sub>50</sub> values of 2.4–6.7  $\mu\text{M}$ . Thus, the hit rates for the enzymes identified as being similar to Cdc25A are  $\sim 2\text{--}3\%$ , which is a very acceptable value for an initial screen aimed exclusively at identifying hit classes. Gratifyingly, even at this small library size, the hits indicated a pronounced degree of selectivity for individual enzymes and also for the isoenzymes 11 $\beta$ HSD1 and 11 $\beta$ HSD2. Most remarkably, the  $\alpha,\beta$ -unsaturated lactone **3** inhibited only the therapeutically relevant 11 $\beta$ HSD1 but not or only very weakly the other enzymes investigated.

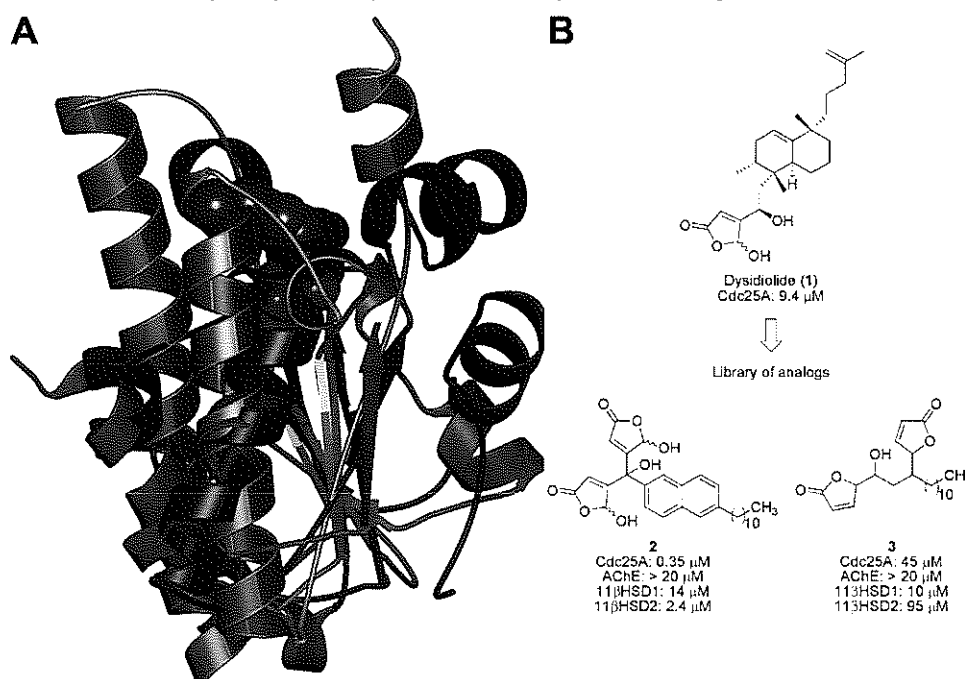


Fig. 2: Protein Structure Similarity Cluster consisting of the ligand-sensing cores of Cdc25A, AChE and 11 HSD1. A) The superimposition of the cores of Cdc25A (red), 11 -hydroxysteroid dehydrogenase 1 (green) and acetylcholinesterase (blue) shows that the key catalytic residues, Cys430 (Cdc25A), Tyr183 (11 HSD1) and Ser200 (AChE), shown in Corey–Pauling–Koltun representation, are located similarly. (B) Analogs of the naturally occurring Cdc25A inhibitor dysidiolide (**1**) profiled against the PSSC member proteins Cdc25A, AChE, 11 HSD1 and 11 HSD2 (IC<sub>50</sub> values are given).

This was the first example to demonstrate that a clustering of target proteins exclusively driven by structural considerations in conjunction with natural product-inspired compound collections can serve as a valuable guiding concept for the focused development of protein

ligands. This is all the more true as the identified structural compound classes have not yet been described as inhibitors of the respective enzymes.

The second compound library design concept developed within this Ph.D. thesis comprises a complementary view on the chemical structural space of natural products. Natural products are synthesized by enzymes and were selected during the course of evolution to fulfill a certain function that is most often mediated through interactions with proteins. Thus, natural products inherently represent "privileged" structural motifs. Their structural scaffolds represent the biologically relevant and pre-validated fractions of chemical structure space explored by nature so far. Thus, in this Ph.D. thesis, the natural product scaffolds were investigated. As a result of these investigations a structure-based classification of natural products was being developed that can serve as navigator through the chemical space of natural products. Primary and secondary cheminformatics processing of the CRC "Dictionary of Natural Products", which lists 190,939 records, led to the extraction of the natural product scaffolds which were then classified hierarchically with increasing complexity in a tree-like fashion. Basis for the classification was the generation of a structure-based genealogy for each natural product scaffold. These genealogies were subsequently correlated with one another. Thus, every natural product scaffold was traced back to a mono cycle as basic "taxon", leading to a kind of "phylogenetic tree" in which the natural product scaffolds are arranged according to their structural interrelationship with increasing number of rings in their scaffolds. Each node in the tree diagram constitutes a distinct scaffold from which further arborization may lead to more complex scaffolds.

This scaffold-based „Structural Classification of Natural Products“ (SCONP) was annotated on each node, i.e. scaffold, with the individual natural products bearing the scaffold and the biological source organism and biological effects as far as known. [2]

The applicability of SCONP as hypotheses and idea generator in the context of the design of compound libraries was experimentally validated.

Using SCONP in conjunction with PSSC, structurally simplified, potent and highly selective inhibitors of 11 $\beta$ HSD1 could be developed using the complex pentacyclic natural product and non-selective 11 $\beta$ HSD inhibitor glycyrrhetic acid (GA, 4, Figure 3) as starting structure. The scaffold of GA was first structurally assigned to a scaffold node of the SCONP tree. Subsequent brachiation in the direction of reduced complexity led to a subset of two- and three-ring systems that occur most frequently in nature according to the statistical analyses performed on the basis of the SCONP tree. From this subset, the 1,2,3,4,4a,5,6,7-

octahydronaphthalene scaffold also occurring in dysidiolide (**1**, Figure 3) was chosen as the precise scaffold for the generation of a focused compound library using PSSC as second, independent criterion. A collection of 162 compounds bearing this scaffold was biochemically evaluated with respect to inhibition of 11 $\beta$ HSD type 1 and type 2. Twenty-eight compounds selectively inhibited 11 $\beta$ HSD1, four of them in the nanomolar range. To demonstrate cellular activity of the previously undescribed inhibitor class, translocation and transactivation assays were performed for compound **5**, one of the most potent and selective 11 $\beta$ HSD1 inhibitors ( $IC_{50} = 0.35 \mu M$ ).

It has to be stressed that finally the complementary worlds of natural products and proteins were brought together through the synergistic use of SCONP and PSSC for the design of compound libraries.

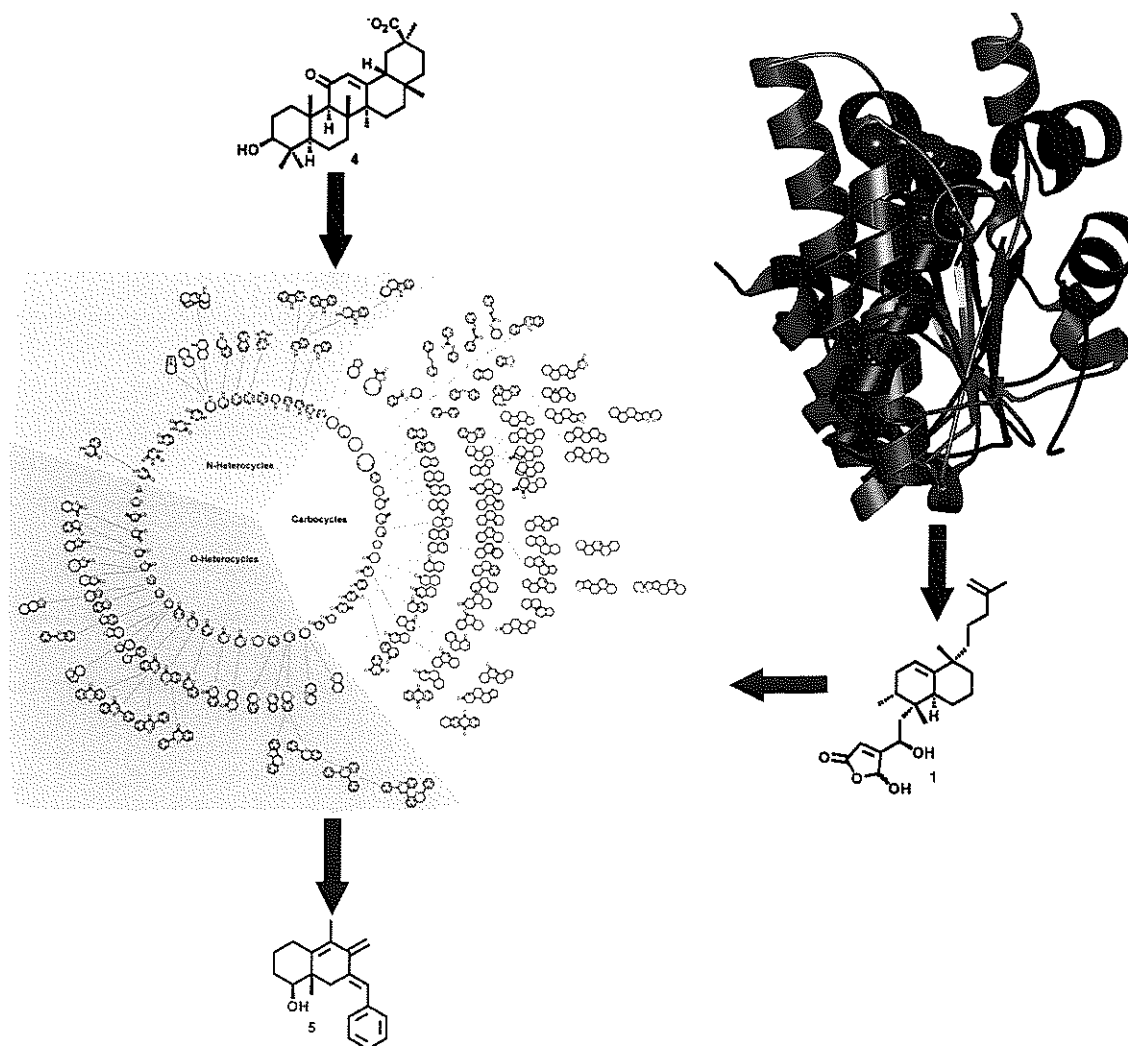


Fig. 3: Synergistic application of SCONP and PSSC in the development of potent and selective 11 $\beta$ HSD1 inhibitors. The octahydronaphthalene scaffold was chosen as starting point for the development of a focused combinatorial library. It is present in both natural products, dysidiolide (**1**) and glycyrrhetic acid (**4**). The PSSC analysis was used as a second, independent criterion for the simplification of the complex scaffold of glycyrrhetic acid towards the octahydronaphthalene

core guided by the SCONP tree. The compound collection yielded 28 selective  $11\beta$ HSD1 inhibitors. Compound 5 was one of the most potent and selective  $11\beta$ HSD1 inhibitors, and proved to be active in biological cells.

The concepts developed within this Ph.D. thesis may serve as conceptually new principles guiding the development of compound libraries in particular for medicinal chemistry research. However, beyond this, PSSC and SCONP may open up new opportunities for research in the currently developing field of 'chemical genomics'. In a general sense, 'chemical genomics' may be defined as the genomic response to chemical compounds, i.e. chemistry is employed to probe a biological system. A more focused, workable definition appears to be the identification of small molecule lead-like compounds for a member of a gene family product and the subsequent use of these compounds to elucidate the function of other (disease-associated) members of the gene family. Currently in this approach, the gene family products are predominantly classified on the basis of sequence similarities and function, i.e. into kinases, phosphatases, proteases, etc. (see Figure 4).

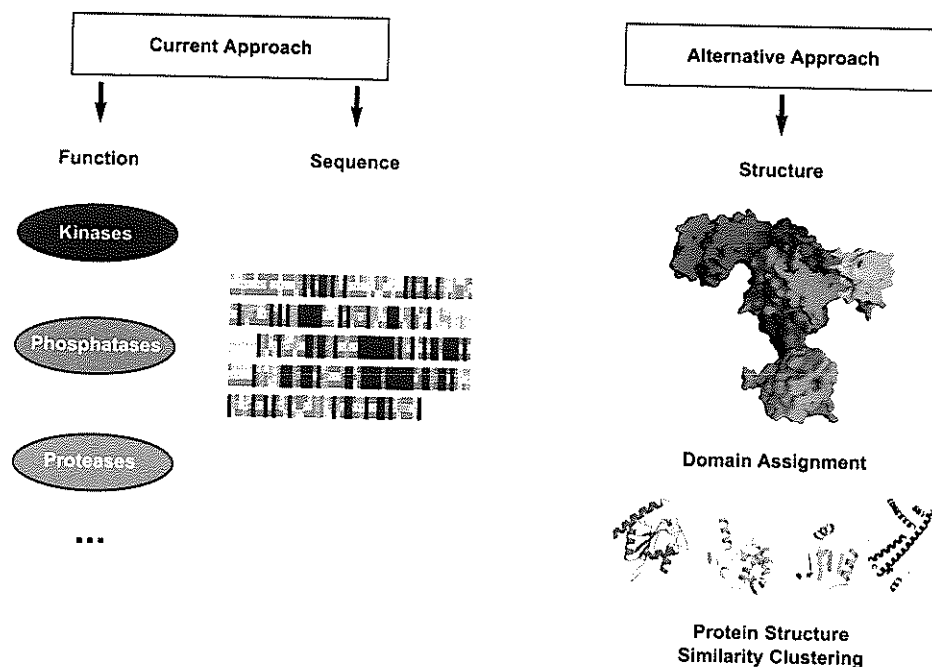


Fig. 4: Approaches for protein categorization. The currently predominating approach in chemical genomics, which is based on the clustering of target proteins according to their sequence and function, might be complemented by an alternative approach based on a purely structural view of protein domains or cores.

A protein domain core-centered approach that considers domain organization and architecture, however, may provide a new guiding principle for the combinatorial development of compounds that will pave the way to a new series of chemical proteomics and genomics experiments. A purely structure-driven concept for the grouping of potential

target proteins like PSSC could be regarded as a valuable alternative or complement to current approaches, as in principle the clustering can be performed without further knowledge about the function and the physiological context of the target protein. The consideration of protein structures strengthens the clustering perspective as protein spatial structures are evolutionarily more conserved than the determining primary sequences. Structures may be related, but primary sequences so different that, using sequence-based approaches alone, structural relatedness cannot be detected. Thus, PSSC represents a protein structure-based extension of the sequence-driven methodologies. The main disadvantage in the case of PSSC, the need for structural information, is partly relativized by the fact that generally, for the principal assignment of structurally similar structures to a cluster, structural models generated by state of the art techniques are sufficient.

The complementary mapping of the chemical structural space of natural products by SCOMP allows for navigation through these biologically relevant fractions of chemistry space explored by nature. It has been demonstrated that, at a high level of abstraction, starting from complicated natural product scaffolds simplified and structurally innovative solutions for chemical ligands with retention of biological activity can be found.

Both concepts, PSSC and SCOMP, bear certain indeterminateness because of the high abstraction level. But finally, the indeterminateness is overcome by the combinatorial approach. Combinatorial variations of the identified structural "leitmotifs" are of utmost importance in order to address the biological diversity of the binding pockets with an appropriate chemical diversity for the sake of potency and selectivity.

PSSC and SCOMP may be helpful as initial rationales in early discovery phases, when little may be known about the function and the biological context of a potential target protein. Novel structural compound classes allowing for the investigation of the target biology may be identified more efficiently. It has to be underlined that the structural frameworks then usually have to be refined in a medicinal chemistry program to optimize selectivity and to reduce unwanted activities. The identification of the right starting point for this optimization, however, already represents an important step ahead.

## References

- [1] M. A. Koch, L.-O. Wittenberg, S. Basu, D. A. Jeyaraj, E. Gourzoulidou, K. Reinecke, A. Odermatt, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 16721-16726.
- [2] M. A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzler, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 17272-17277.