Full text provided by www.sciencedirect.com

Protein structure similarity clustering (PSSC) and natural product structure as inspiration sources for drug development and chemical genomics

Frank J Dekker, Marcus A Koch and Herbert Waldmann

Finding small molecules that modulate protein function is of primary importance in drug development and in the emerging field of chemical genomics. To facilitate the identification of such molecules, we developed a novel strategy making use of structural conservatism found in protein domain architecture and natural product inspired compound library design. Domains and proteins identified as being structurally similar in their ligand-sensing cores are grouped in a protein structure similarity cluster (PSSC). Natural products can be considered as evolutionary pre-validated ligands for multiple proteins and therefore natural products that are known to interact with one of the PSSC member proteins are selected as guiding structures for compound library synthesis. Application of this novel strategy for compound library design provided enhanced hit rates in small compound libraries for structurally similar proteins.

Addresses
Department of Chemical Biology, Max-Planck Institute of Molecular Physiology, Otto-Hahn Str. 11, D-44227 Dortmund, Germany,
Fachbereich 3, Organic Chemistry, University of Dortmund, Otto-Hahn Str. 6, D-44227 Dortmund, Germany

Corresponding author: Waldmann, Herbert (herbert.waldmann@mpi-dortmund.mpg.de)

Current Opinion in Chemical Biology 2005, **9**:232–239

This review comes from a themed issue on
Combinatorial chemistry
Edited by Hans-Jörg Roth and Prabhat Arya

Available online 8th April 2005

1367-5931/$ – see front matter

DOI 10.1016/j.cbpa.2005.03.003

Introduction

Modern approaches addressing DNA sequence (genomics), protein structure (structural biology) and protein expression and protein interactions (proteomics) provide a wealth of data considering potential pharmaceutical targets. The pharmaceutical community faces the challenge to convert this vast amount of data into knowledge and innovations (i.e. new drugs). Combinatorial chemistry and high-throughput screening have emerged as powerful tools to generate and evaluate large compound libraries for activity against many different targets. However, initial expectations that large compound libraries should result in the discovery of many new hit and lead structures for drug development have not been fulfilled in principle. As the universe of thinkable compounds is almost infinite, the decisive question to be answered is: Where can biologically pre-validated starting points be found from which to build compound libraries?

Answers to this question may be provided by nature itself. Recent results in structural biology and bioinformatics indicate that the number of distinct protein folds is limited [1,2•,3••,4,5••]. The same structural domain is often found in many proteins in a more-or-less modified form. We proposed that protein domains or cores with similar three-dimensional structures can be clustered in so-called protein structure similarity clusters (PSSCs) and that knowledge about known ligands for members of such a cluster can be employed to guide compound library development for other members of the cluster. Thus, this approach to explore and exploit nature's structural conservatism concerning protein architectures can be used as abstracting rationale for compound library design in drug development and chemical genomics [6,7••,8].

Modern strategies in compound library design

Knowledge-based strategies and bioinformatics are gaining importance in compound library design and high-throughput screening in chemical biology and chemical genomics [9,10••,11]. Some guidelines for knowledge-based compound library design have been identified as being important; in particular, 'diversity' [12,13•], 'drug-likeness' [14–16], and 'biological relevance' [6,8,17••]. An efficient approach in designing biologically relevant compound libraries is to take natural products as inspiration sources [6,7••,8,17••,18,19••,20,21]. Natural products can be considered as biologically validated starting points in structural space, since they have to interact with different proteins in the course of their biosynthesis and when they exert their biological function (e.g. in chemical defense or communication). Another possibility to afford biologically relevant libraries is to take so-called 'privileged structures' [22] as guiding principles. Privileged structures are structural motifs found in drugs or drug-like compounds and are supposed to confer drug-like properties on compounds containing these motifs [23•].

An important knowledge-based strategy for lead finding is parallel screening of focused compound libraries for activity to clusters of related protein targets. This raises the question of how to cluster proteins to find high hit rates. Common approaches are based on similarity in amino acid sequence and/or protein function [10••]. Classification of clustered protein targets based on the structure–activity relationships of small molecules that modulate their function has been denoted as SARAH (structure–activity relationship homology) [24]. A recent example has been provided by Greenbaum *et al.* [25•] who used small molecules for enzyme family subclassification. A major limitation of the current predominantly used strategies for protein target clustering is that mostly only closely related proteins can be considered, which reduces the scope of these approaches.

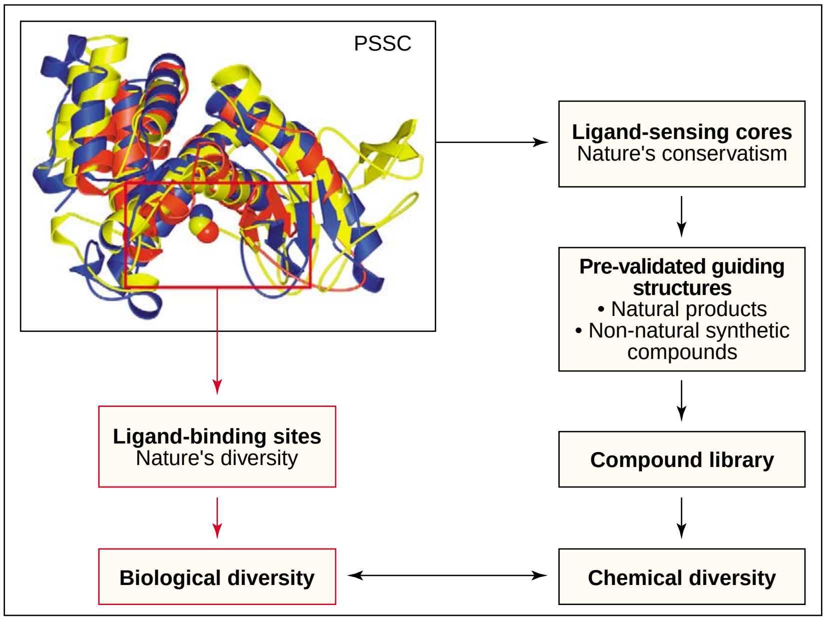
In this article we describe how clustering of protein targets exclusively based on structure similarity considerations provides opportunities to find novel classes of small-molecule modulators of protein function.

PSSC as a guiding principle for compound library development

Proteins can be regarded as modularly built macromolecules whose building blocks are so-called domains, which are evolutionary and genetically mobile units [26,27,28•,29]. Recent results in bioinformatics indicate that the number of distinct domain fold types as structural motifs is fairly limited. Current estimates of the total number of protein domain folds range from 1000 to 10 000 depending on the models and approximations applied [2•,3••,4,5••]. It seems certain that a great majority of protein domains can be attributed to about 1000 most commonly observed folds [2•,3••]. Ligand binding or catalytic sites are the most relevant parts of protein domains from the point of view of development of small molecule ligands. There is evidence that the structural conservatism of nature is not only restricted to protein folds, but also often applies to the topological location of the functional sites in proteins [30,31•]. However, correlations between protein functions and proteins folds remain a matter of debate. The functional sites of proteins with similar architectures show significant diversity due to differences in the amino acid sequences that form the binding sites, which is crucial to maintain specificity in functional behavior. On the other hand, some conservatism can be observed in functional sites of structurally similar proteins, such as the similar topological location of catalytic residues in enzyme active sites. This is true for the example described later in this review showing similarity of Cdc25A phosphatase, acetylcholinesterase and 11β-hydroxysteroid dehydrogenases.

These findings led to the development of a novel strategy, which exploits nature's structural conservatism concerning protein architecture, for the identification of small-molecule modulators of protein function (Figure 1). We conceived protein structure similarity clustering (PSSC) as a guiding principle for the selection of biologically pre-validated starting points for compound library synthesis [6,7••,8]. In this concept, the structures of natural or non-natural products that bind to one member protein of the PSSC are taken as guiding structures for compound library synthesis. Proteins that share structural similarity

Figure 1



Nature's conservatism and diversity: opportunities for the directed development of biologically relevant compound libraries. See Figure 3 for further details on the three catalytic sites compared here. Reproduced from [52] with permission. Copyright 2005, Elsevier.