Thèse présentée pour obtenir le grade de Docteur de

l'Université Louis Pasteur Strasbourg

Faculté de Pharmacie

Ecole Doctorale des Sciences
de la Vie et de la Santé de Strasbourg

Discipline : Sciences Pharmaceutiques

Conception rationnelle de chimiothèques focalisées

Soutenue publiquement le 21 octobre 2005

par Mireille Krier

Membres du Jury

Dr. Nicolas Baurin	Examinateur
Prof. Ronan Bureau	Rapporteur externe
Prof. Marcel Hibert	Examinateur
Prof. Thierry Langer	Rapporteur externe
Prof. Alexandre Varnek	Rapporteur interne
Dr. Didier Rognan	Directeur de Thèse

A Marc

A ceux qui me sont chers

Table of Contents

Résumé français circonstancié

Le processus de découverte et développement d'un nouveau agent thérapeutique exige un engagement et un investissement à long terme (7 à 15 ans). Seul un nombre restreint des nombreux projets lancés réussit à produire une « nouvelle entité chimique » (NCE, 30-50 par an). En effet, sur les centaines de milliers de composés testés, au plus 1% est identifié comme « touche ». Des analyses plus approfondies permettent de sélectionner parmi ces touches les structures dites « tête de file » qui continueront vers la phase de développement et cliniques. Jusqu'à ce qu'un NCE soit admis sur le marché par la Food and Drug Administration (FDA) des États-Unis (ou son homologue européen ou japonais), les raisons d'échec peuvent être diverses (Kubinyi, 2003). La qualité initiale des chimiothèques aide à éviter ces impasses. En effet, le problème de la conception de chimiothèques a joué un rôle important dans beaucoup d'approches essayant accélérer le processus de conception d'un médicament. La tendance actuelle privilégie la conception rationnelle des ligands par rapport à la découverte fortuite ou le criblage systématique. Face à l'impossibilité de l'étude complète de l'espace chimique contre l'espace des cibles, l'organisation des molécules de faible poids moléculaire et des protéines en des classes de châssis moléculaires et des familles de protéines, respectivement, offre une certaine abstraction qui laisse établir des rapports de chemogénomique. En effet, la pratique de la pharmacochimie et celle du criblage biologique ont révélées que les composés agissant sur une famille donnée de protéines sont rarement éloignés dans l'espace chimique (Lipinski & Hopkins, 2004). Cependant, l'évaluation du potentiel « candidat-médicament » d'une molécule ainsi que sa catégorisation dans les classes de châssis ont été traditionnellement réalisées par inspection visuelle par les pharmacochimistes. Ceci a habituellement conduit à des contradictions dans la

conception de chimiothèques ainsi que dans l'acquisition de composés (Lajiness *et al.*, 2004) et accentue le besoin de développer des méthodes informatiques robustes qui peuvent aider le chimiste.

Le défi de la "simplexité" a été récemment formulé par Compain (Compain, 2003) dans le contexte de la combinaison du simple et de la complexité dans la synthèse organique. Cette idée réinterprétée pour l'informatique combinatoire moléculaire, notre stratégie consistera dans la construction de molécules virtuelles qui acquièrent leur complexité structurale par des combinaisons d'entités moléculaires simples. Les travaux fondamentaux de Bemis et de Murcko (Bemis & Murcko, 1996; Bemis & Murcko, 1999) concernant l'analyse sousstructurale des médicaments admis sur le marché ont fortement inspirés ce travail. Afin de concilier les avancées de la chimie combinatoire et de la conception rationnelle basée sur la structure protéinique, le défi scientifique abordé par ce sujet de thèse est de créer le maximum de diversité en générant un minimum de molécules constituant une chimiothèque. Nous avons choisi de partir d'une structure de Markush (Markush, 1924) ayant le rôle de châssis moléculaire (scaffold) et de lui greffer à chaque position de substitution un espaceur (linker), puis un groupement fonctionnel(functional group); nous avons intitulé cette approche SLF (Figure I).

C'est le chimiste organicien qui propose cette structure ou le sélectionne parmi une chimiothèque de châssis moléculaires dont il attend une activité. Le nombre totale de molécules générées s'obtient par :

$$(L \times F)^S$$

où L = nombre d'espaceurs
F = nombre de groupements fonctionnels
S = nombre de positions de substitution sur la structure de Markush

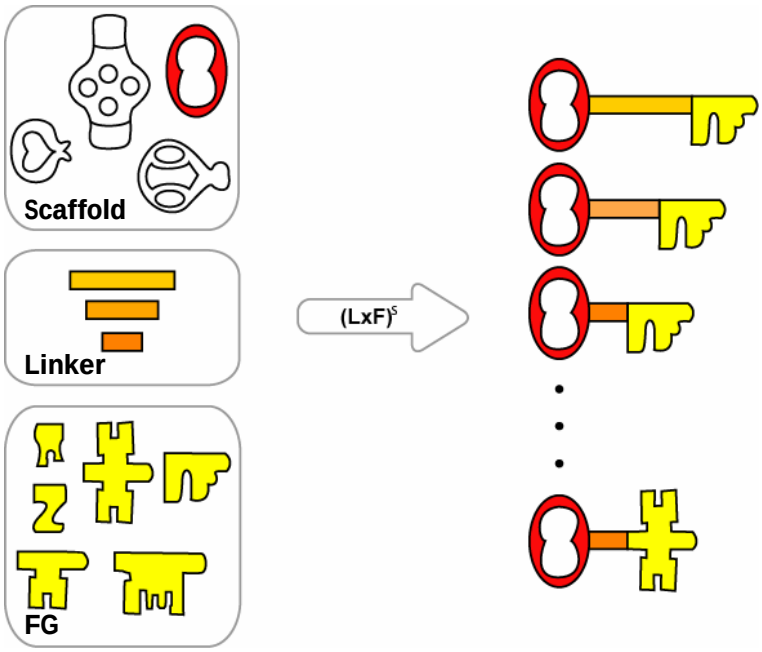


Figure I. Illustration de SLF

Le châssis moléculaire adéquat est supposé orienter ses substituents dans des positions optimales à l'interaction avec la protéine, mais aussi interagir lui-même avec la protéine. Son nombre de points de substitution (S) peut théoriquement varier de l'unité à l'infini, mais en général S varie pour la majeure partie des châssis moléculaires entre 1 et 10.

Le rôle de l'espaceur est de moduler la distance d'interaction entre le « châssis/groupement fonctionnel » et la protéine cible. Dans un premier tour de criblage, trois à quatre espaceurs sont choisis parmi les séries polyméthyléniques ouvertes (-[CH_2]-). Les groupements fonctionnels sélectionnés (entre huit et dix fragments) représentent les propriétés pharmacophoriques (charge négative/positive, accepteur/donneur d'hydrogène, lipophile/aromatique). Ils servent à sonder le site actif si peu d'informations structurales sont disponibles.

Dans un premier temps, il fallait trouver la représentation adéquate pour coder les structures de Markush et les fragments moléculaires (espaceurs et groupements fonctionnels). Plusieurs critères ont guidé notre choix :

1. Diminuer au mieux les pertes d'informations lors de conversions en différents formats moléculaires.
2. Garder la taille du fichier de sortie contenant les structures moléculaires raisonnable (=> mémoire morte).
3. Empêcher une erreur de mémoire vive insuffisante lors de l'exécution du programme.

Le SMILES (Weininger, 1988; Weininger *et al.*, 1989), un format « 0D » (zéro dimension), a été adopté. L'information d'une molécule (atomes, liaisons, topologie) est représentée par une ligne de caractères alphanumériques. Le caractère « * » (étoile) marque la position de substitution sur les fragments moléculaires.

L'application SLF_LibMaker (Scaffold-Linker-Functional Group Library Maker) implémentant cette méthode de génération de molécules est écrite en langage C++ et, outre les bibliothèques standards du C++, se base sur la bibliothèque OEChem développée par OpenEye Scientific Software("OEChem", 2004). Les combinaisons espaceurs / groupement fonctionnels sont d'abord générées et stockées dans un fichier temporaire. Puis le nombre de points de substitution sur le châssis moléculaire est déterminé pour énumérer de façon systématique toutes les combinaisons possibles entre châssis moléculaire et espaceurs / groupement fonctionnels.

L'application garde un caractère générique, car elle peut générer des chimiothèques de taille importante (de l'ordre du milliard de molécules) qui peuvent servir au criblage virtuel à haut débit. Cependant notre objectif est de générer une chimiothèque d'une taille raisonnable (une centaine de structures) pour la synthèse et l'évaluation biologique.

En cas d'absence d'idée concrète sur le châssis moléculaire, le chimiste a la possibilité de consulter une chimiothèque de châssis moléculaire pour planifier la conception des prochaines molécules à synthétiser. Soucieux de la réalité physique des molécules, la stratégie de constitution de la chimiothèque de châssis moléculaires (« scaffoldthèque ») a été basée sur une classification de chimiothèques commerciales qui de plus a conduit au développement de nouvelles métriques de la diversité moléculaire (Figure II).

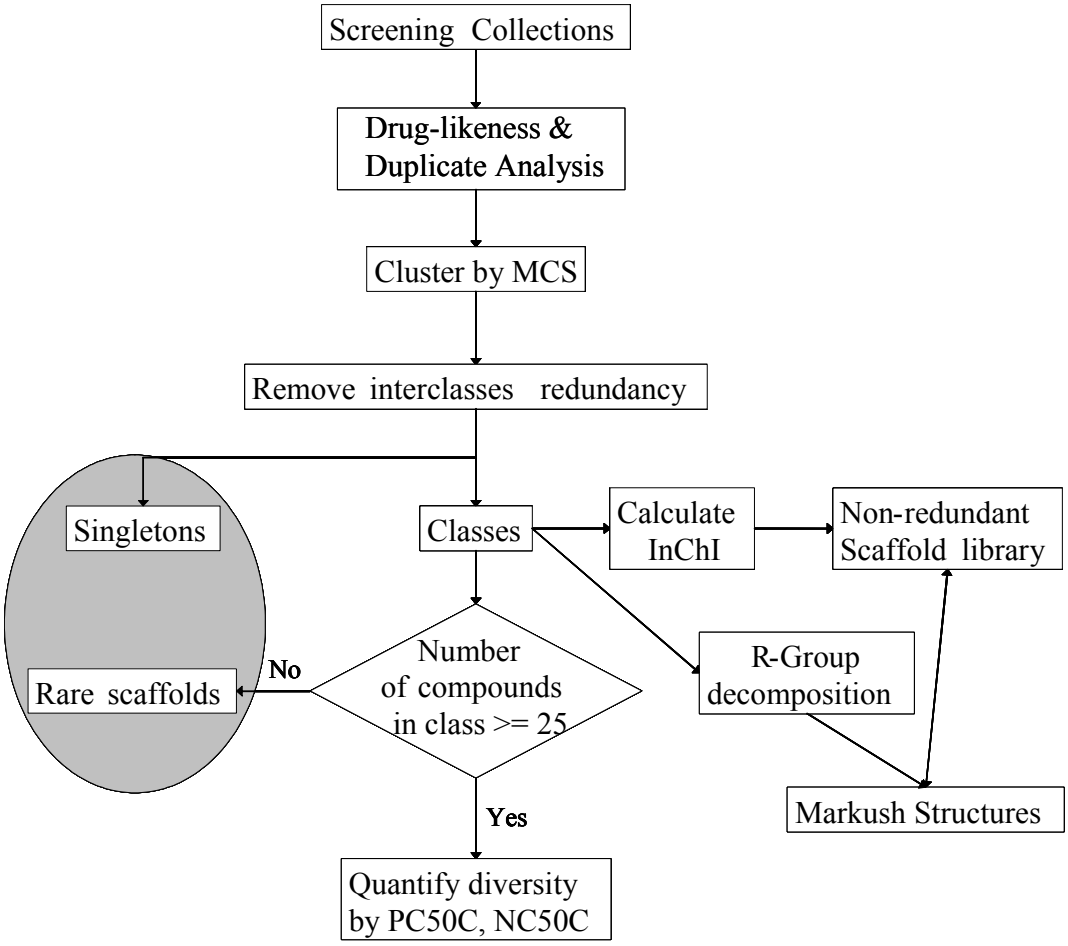


Figure II. Organigramme du traitement des chimiothèques commerciales et de la constitution de la scaffoldthèque

Ces collections de criblage de criblage constituent une source importante de molécules potentiellement bioactives (environ 2 millions de molécules disponibles dans un délai d'un mois). L'analyse du potentiel médicament (« drug-likeness ») (Baurin *et al.*, 2004; Charifson & Walters, 2002) et de la diversité moléculaire (Bradley, 2002) de ces

collections de criblage est importante dans la sélection des molécules à tester. Dans notre cas, la procédure de préparation de chaque chimiothèque passe par une étape de filtrage, puis par l'élimination des structures redondantes (Figure II). Les molécules qui ont été classées correspondent à celles ayant passées les filtres, définis comme un certain intervalle de propriétés physico-chimiques et l'absence de certains groupements fonctionnels jugés trop réactifs. L'analyse de diversité a été alors effectuée sur les classes comportant au moins 25 composés. Ce nombre de molécules analogues comportant le même châssis moléculaire est nécessaire du point de vue du pharmacochimiste (Nilakantan & Nunn, 2003) pour garantir une exploration minimale lors d'un criblage. Un autre avantage réside dans la possibilité d'établir une relation structure-activité préliminaire. Les classes ne remplissant pas ce critère de taille minimale ont été conservées dans une chimiothèque de châssis rares. Les descripteurs choisis lors de l'analyse de diversité doivent quantifier la diversité moléculaire intrinsèque d'une chimiothèque. Le nombre (NC50C) et le pourcentage (PC50C) de classes auxquelles appartiennent 50% des composés classés sont extrapolés de la courbe de distribution cumulative des pourcentages des composés appartenant à une classe. Nous avons retrouvé de manière qualitative la nature de la chimiothèque en reportant la taille de chimiothèque en fonction des deux métriques de diversité (PC50C, NC50C) (Figure III). En effet, les collections issues de chimie combinatoire sont de taille importante mais peu diversifiées en châssis moléculaires. Diverses collections de taille intermédiaire présentent également une faible diversité. Seul un petit nombre de collections, qui sont connues pour être optimisées par leur fournisseur, présente un bien meilleur rapport taille/diversité. On remarquera que ce sont celles qui se rapprochent également le plus de la MDDR (MDL Drug Data Report), chimiothèque de référence

pour molécules bioactives, ayant subi le même processus que les collections commerciales.

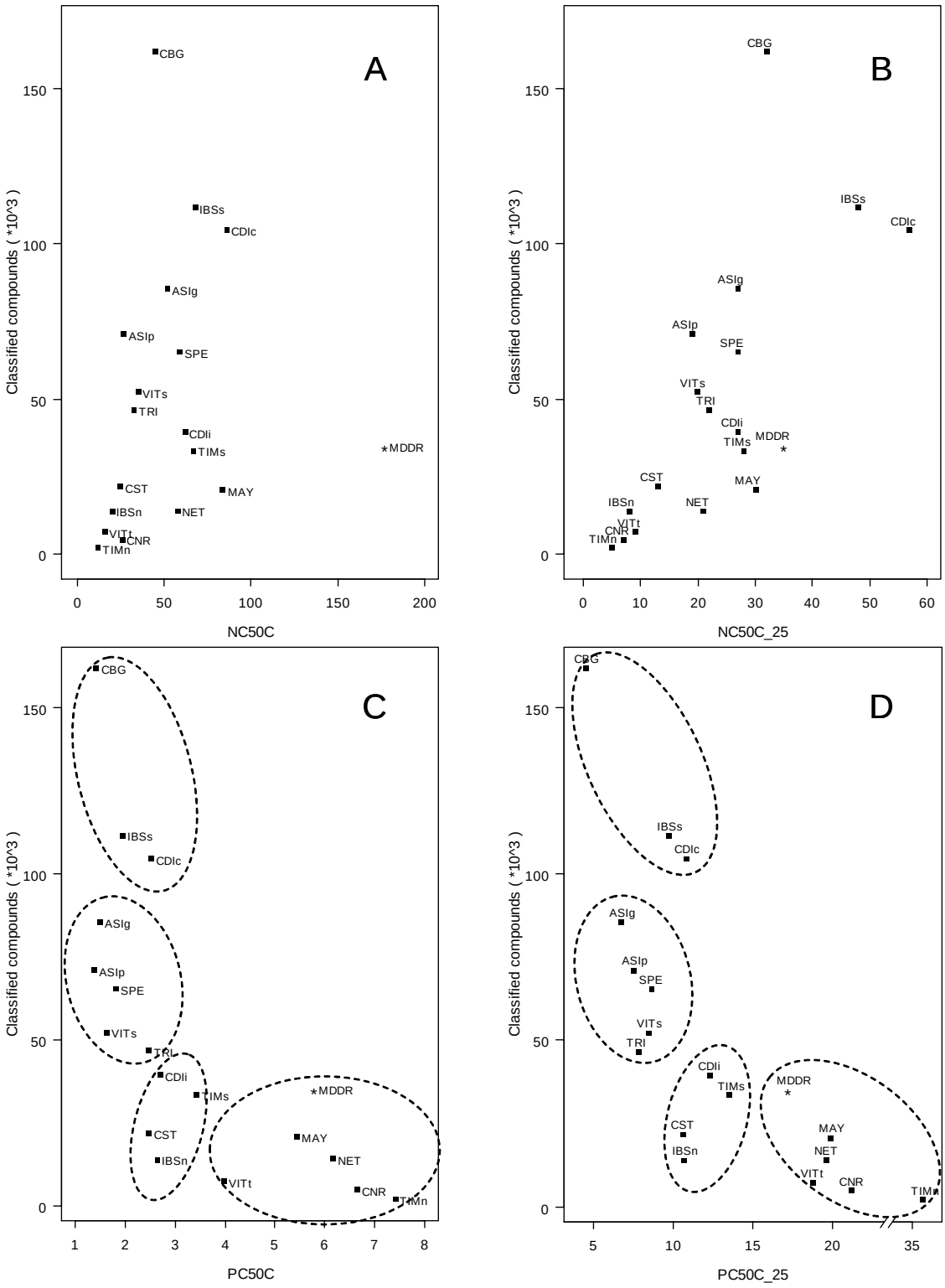


Figure III. Diversité des châssis moléculaires issues des collections de criblages disponibles chez des fournisseurs commerciaux

La méthode développée sur le concept de l'énumération complète de « scaffold, linker, functional group » a été appliquée à l'optimisation d'inhibiteurs de PDE4 (phosphodiesterase IV). Grâce aux données structurales récemment disponibles sur la PDE4 (Huai *et al.*, 2003; Lee *et al.*, 2002) et en dérivant la zardaverine (molécules spécifique de la cible), les 320 molécules de la chimiothèque ainsi générée pouvaient être évaluées à partir de leurs interactions moléculaires avec la protéine. Plus concrètement, ceci se traduit par le « docking automatisé » utilisant le logiciel FlexX (Rarey *et al.*). Ainsi neuf molécules ont été choisies prioritairement pour la synthèse chimique et pour des mesures d'affinité consécutives. Cinq des neufs composés se sont avérés être de inhibiteurs plus puissants in vitro que la zardaverine. Le composé ayant montré la plus haute affinité (CI_{50} = 0,88 nM) (Figure IV) a permis un gain en affinité de trois unités logarithmiques par rapport à la référence interne utilisée (CI_{50} = 2 µM).

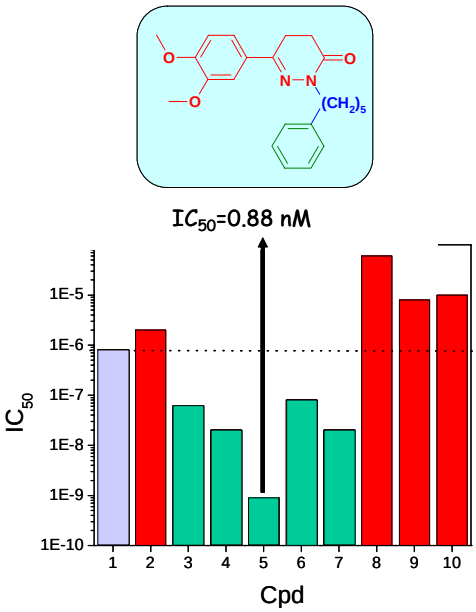


Figure IV. Distribution de l'affinité des molécules sélectionnées par rapport à la référence (en 1).

De plus, une sous-poche dans le site actif de la PDE4 a été identifiée, présentant une variation en acides aminés par rapport aux autres PDE. Ceci peut fournir le point de

départ à la conception de structures surmontant les problèmes de sélectivité connus jusqu'à présent chez les inhibiteurs en phase cliniques.

Au cours d'autres projets impliquant la génération d'une chimiothèque virtuelle, le logiciel a été ajusté au besoin de l'utilisateur. Afin de faciliter l'analyse « post-criblage », une nomenclature unique rendant compte du nom initial des fragments a été mis en place. Un autre point à ajuster est le cas des châssis moléculaires adoptant une conformation difficilement prévisible, car les outils usuels de générations de coordonnées trois-dimensionnelles s'avèrent insuffisants. Par conséquent, le logiciel va évoluer vers l'extension de l'approche en trois dimensions, permettant ainsi de prendre en charge une conformation de châssis moléculaire prédéfinie.

Ce travail de thèse nous a permis de développer une méthodologie qui applique les principes d'explorations topologiques d'une cible thérapeutique de façon systématique à la conception des ligands, ainsi permettant d'accélérer le processus de la touche au candidat médicament. De plus, des chimiothèques des fragments élémentaires d'intérêt thérapeutique ont été constituées : d'un part, celles d'espaceurs et de groupement fonctionnels sur des critères empiriques et d'autre part, celle de châssis moléculaires extraits à partir de collections de criblage commerciales. La scaffoldthèque peut en outre servir à la rationalisation de l'acquisition de composés ; c'est-à-dire (i) on comble l'espace chimique « intéressant » et (ii) empêche d'acquérir des composés ayant des propriétés indésirables. De plus, la constitution de la scaffoldthèque a mené au développement de nouvelles métriques de la diversité moléculaire.

Bibliographie

Baurin, N., Baker, R., Richardson, C., Chen, I., Foloppe, N., Potter, A., et al. (2004). Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *Journal of Chemical Information and Computer Sciences, 44*, 643-651.

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem., 39*, 2887-2893.

Bemis, G. W., & Murcko, M. A. (1999). Properties of known drugs. 2. Side chains. *J. Med. Chem., 42*, 5095-5099.

Bradley, M. P. (2002). An overview of the diversity represented in commercially-available databases. *Journal of Computer-Aided Molecular Design, 16*, 301-309.

Charifson, P. S., & Walters, W. P. (2002). Filtering databases and chemical libraries. *Journal of Computer-Aided Molecular Design, 16*, 311-323.

Compain, P. (2003). The challenge of simplexity. The simple and the complex in organic synthesis [Le pari de la simplexité: Le simple et le complexe en synthèse organique]. *Actualité Chimique*, 129-134.

Huai, Q., Wang, H., Sun, Y., Kim, H. Y., Liu, Y., & Ke, H. (2003). Three-dimensional structures of PDE4D in complex with roliprams and implication on inhibitor selectivity. *Structure (Camb), 11*, 865-873.

Kubinyi, H. (2003). Drug research: myths, hype and reality. *Nat Rev Drug Discov, 2*, 665-668.

Lajiness, M. S., Maggiora, G. M., & Shanmugasundaram, V. (2004). Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem., 47*, 4891-4896.

Lee, M. E., Markowitz, J., Lee, J. O., & Lee, H. (2002). Crystal structure of phosphodiesterase 4D and inhibitor complex. *Febs Letters, 530*, 53-58.

Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature, 432*, 855-861.

Markush, E. A. (1924). Pyrazolone Dye and Process of Making the Salt. US Pat. 1,506,316.

Nilakantan, R., & Nunn, D. S. (2003). A fresh look at pharmaceutical screening library design. *Drug Discov. Today, 8*, 668-672.

OEChem. (Version 1.3)(2004). OpenEye Scientific Software, Inc.

Rarey, M., Claussen, C., Kramer, B., & Lengauer, T. FlexX (Version 1.12): BioSolveIT GmbH, Sankt Augustin , Germany.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences, 28*, 31-36.

Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES: 2. Algorithm for generation of unique SMILES notation. *Journal of Chemical Information and Computer Sciences, 29*, 97-101.

Acknowledgments

"*An idea belongs to nobody because nobody can have an idea out of nil;*

what a strange idea?"

Adapted and translated from the French

Confession d'un voleur, Internet: la liberté confisquée, Laurent Chelma

I would like to thank Dr. Didier Rognan, my Ph.D. advisor for accepting me as student and introducing these interesting topics, for his patience and valuable comments. I have learned so much more than my initial chemistry background. This also leads me to thank Prof. Marcel Hibert and the whole UMR 7081 (now UMR7175-LC1) that conceal at ease high scientific standards and pleasure of doing science.

I want to acknowledge Jean-Jacques Bourguignon, Martine Schmitt and their students for their collaboration on the SLF/PDE4 project. They also made the link between our group and the biologists headed by Claire Lugnier which conducted the in vitro tests.

Furthermore, I want to thank the following persons:

Vincent Vivien and his collegues from Bioreason for computing expertise and assistance.

Bruno Didier and Françoise Herth for their logistic talents.

Serge Petit and the Idéalp'Pharma collegues for financial support and throughout kindness, giving me the opportunity to realize this work.

The jury members Alexandre Varnek, Nicolas Baurin, Thierry Langer, Ronan Bureau for accepting to review this work.

Thanks also go to Sinan Karaboga and Jérôme Hert, Jörg Kurt Wegner and the chemoinformatics community.

I have also enjoyed immensely the people I worked closely with in the Bioinformatic Lab, namely Jordi Rodrigo de Losada, Guillaume Bret, Jean-Sébastien Surgand, Pascal Muller, Eui-Ki Kim, Gilles Marcou, Esther Kellenberger and Claire Schalon, but also to the former members, namely Antoine Logean, Nicodème Paul, Sandrine Longuet, Michèle Mathis, Nicolas Foata, Patchreenard Saparpakorn.

I would also like to thank Marc Michel and his family. Without their support and perseverative encouragements, this work would have never been finished. Milles merci!

To my mother, my father and my brother: ouni vill chichi, merci villmols!

1. Introduction

Long term engagement and investment are required to the process of drug research and development. Only a small number of the many initiated projects succeed in producing a new chemical entity (NCE, 30-50 per year). Meanwhile hundreds of thousands of compounds will have been screened, thus providing hits. Few turn out to be lead structures requiring additionally thousands of compounds to be synthesized. Until a NCE is approved by the US Food and Drug Administration (FDA) (or its European or Japanese regulatory organization), the reasons of failure can be manifold (Kubinyi, 2003). The initial quality of the screening libraries helps avoiding dead-ends. Indeed, the library design problem has played an important role in many approaches trying to accelerate the drug design process. There has been a shift from ligands discovered by serendipity or by systematic screening to rational design. As it is impossible to investigate the chemical space against the target space, the organization of low molecular weighted molecule and proteins in scaffold classes and families, respectively, offers a certain abstraction that permits to establish chemogenomics relationships. Indeed, the practice of medicinal chemistry and biological screening let scientists deduce that compounds acting upon a given protein family lie rarely far apart in chemical space (Lipinski & Hopkins, 2004). Nevertheless, the assessment of a compound's drug-like or lead-like nature and its categorization in compound classes has been traditionally realized through visual inspection by medicinal chemists. This usually have led to inconsistencies in library design and compound acquisition (Lajiness *et al.*, 2004) and re-emphasize on the need of robust computational methods that can assist the bench chemist.

The challenge of "simplexity" has been recently formulated by Compain (Compain, 2003) in the context of the combination of the simple and the complex in organic synthesis. Taking up this idea for molecular combinatorial informatics, our strategy will consist in building virtual molecules which acquire their structural complexity through combinations of simple building blocks. The seminal papers of Bemis and Murcko (Bemis & Murcko, 1996; Bemis & Murcko, 1999) about their substructural analysis of marketed drugs strongly influenced this work. In a first part, we will define the most salient notions relative to rational design of focused libraries. The second part will deal mainly with the realization of a 'core building block' library and its related diversity issue. Finally, the investigated strategy will be implemented and applied to a protein target for a first validation.

Bibliography

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem., 39*, 2887-2893.

Bemis, G. W., & Murcko, M. A. (1999). Properties of known drugs. 2. Side chains. *Journal of Medicinal Chemistry, 42*, 5095-5099.

Compain, P. (2003). The challenge of simplexity. The simple and the complex in organic synthesis [Le pari de la simplexité: Le simple et le complexe en synthèse organique]. *Actualité Chimique*, 129-134.

Kubinyi, H. (2003). Drug research: myths, hype and reality. *Nat Rev Drug Discov, 2*, 665-668.

Lajiness, M. S., Maggiora, G. M., & Shanmugasundaram, V. (2004). Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J Med Chem, 47*, 4891-4896.

Lipinski, C., & Hopkins, A. (2004). Navigating chemical space for biology and medicine. *Nature, 432*, 855-861.

Chapter 2

2. The role of the scaffold in the drug design process

The main goal of this present chapter is to extend the existing approaches of library design and illustrate the significant potential of this combinatorial-based strategy that supplements the rational design of a core structure by an exhaustive variation of distance-modulated pharmacophoric features. Numerous examples in the literature adopt a cognate approach.

Considering the mental model on which the medicinal chemist bases his search for new lead structures helps developing new drug design methods. A combinatorial design effort usually starts upon the choice of a scaffold. Therefore, the scaffold concept is the cornerstone in many medicinal chemist projects. Computer-assisted scaffold detection among a series of compounds has evolved with the progress made in chemical information technology, but also with the practice of medicinal chemistry.

The requirement for task automation triggered the development of terminology and languages for scaffold/fragment representation which will be addressed in a first part of this chapter.

2.1 Computer representations of chemical structures

Computer methods that record and keep track of steadily increasing amount of chemical structure information have been developed over the last six decades. Computer memory evolved from plain paper to edge-notched punch cards, magnetic tapes and finally hard disks. This implied the abstraction of a molecule into its chemical graph and the storage as strings of alpha-numeric characters or as derivatives of the adjacency matrix. The

purpose of machine-readable structure representations is to mine the molecular information and have to be suitable for the most common operations on molecules that can be highlighted as (1) storage/retrieval, (2) identity, (3) mixture relationships, (4) substructure/superstructure relationships, (5) similarity, (6) transformational relationships (tautomerism, formal charge representation) and (7) multivariate relationships (clustering, discrimination).

Chemical substructures are a special class of structures. In order to understand the representation of a chemical substructure, we also emphasize on historical evolution of computational representation of a structure. Let us review the different levels of abstraction.

2.1.1 Structure diagram

The bench chemist makes regularly use of a structure diagram by sketching the structure and reaction in his lab book or in a registry system. The conventions that govern orientation, suppression of hydrogen atom and carbon atom labels, and implicitly assumed using valence model considerations, are acquired naturally by the chemist during its training. Structural diagrams are ideograms(Garfield, 1972) which have the advantage that chemists speaking different languages will nevertheless catch the same chemical information at sight of the structure diagram.

2.1.2 Connection Table

With improvement of computer technology, the storage capacity shifted away from the punch cards to hard disks. This permitted faster retrieval. Further, computer graphics capabilities were improved and later released as the first commercial system named CROSSBOW(Hyde & Thomson, 1968) that printed a molecular graph in high-quality. Graphic display became a standard.

Figure 2.1 Use of a light pen and computer to draw a chemical structure (1976), Courtesy Chemical Abstracts Service (CAS)

Following these developments, connection tables became of interest. These are structure representations of tabular form listing atoms and bonds. Molecular Design Limited (MDL) started to develop the chemical table file (CTfile) (Dalby et al., 1992; MDL, 2003) format of which Structure-Data File (SDF) is nowadays a standard exchange format for the chemoinformatics community.

2.1.3 Line notation (LN)

In the pre-computer days, the question was to find a way to search for a molecule. The answer to this problem was provided by line notation, as graphs could not be sorted into a list. With the emergence of computers in chemical documentation, line notation were further developed because the processing of a (1D) string of characters was relatively quick and easy taking into account hardware possibilities of these days. Loschmidt had standardized the line-formula delineation in 1861, but the defined symbol set had to be simplified and hence adapted to standard typewriter keyboards (Wiswesser, 1985). A

molecule is broken down in its smallest parts called tokens in computer science. The most acquainted line notations are Wiswesser Line notation (WLN) (Wiswesser, 1982), ROSDAL (Representation Of Structure Diagram Arranged Linearly) (Barnard *et al.*, 1989), SMILES(Weininger, 1988; Weininger *et al.*, 1989) and derived Sybyl Line Notation (SLN) (Ash *et al.*, 1997) and most recently InChI (Stein *et al.*, 2005). There have been great achievements in two- and three-dimensional structure display, but the enduring value of linear descriptions is that they are approximately 110 times less costly to process and it takes about 1% of a connection table for storage.

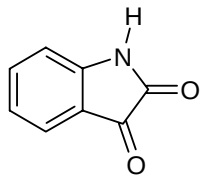
Molecule	LN type	LN code
 Isatine	WLN	T56 BMVVJ
	ROSDAL	1=2-3=4-5-6-1-9N-8-7-6,7=10O,8=11O
	SMILES	O=C(N([H])C2=CC=CC=C12)C1=O
	SLN	O=C(N(H)C[2]=CC=CC=C(@4)@2)C[4]=O
	InChI	InChI=1/C8H5NO2/c10-7-5-3-1-2-4-6(5)9-8(7)11/h1-4H,(H,9,10,11)

Table 2.1 Example molecule isatine (also known as indoledione) and its different line notations

The pioneer was Wiswesser who developed a line notation (Wiswesser, 1982) in 1949. WLN is a fragment-oriented description of a molecule, which is similar to how a chemist thinks of the molecule. It has a long list of predefined fragments, but there are no rules for canonization. Wiswesser noted in his 1950 memo file: "*The greatest difficulties in notation and nomenclature are not with the acyclic and monocyclic structures, but with the multicyclic ring structures, which seem to contain no logical beginning-to-end sequences for simple delineation*" (Wiswesser, 1985). WLN and its

evolution to advanced WLN (AWLN) were adopted for internal use by many pharmaceutical companies as tool to store, retrieve and manipulate chemical structures. It allows substructure analysis and had therefore been used to establish structure-activity relationships (Adamson & Bawden, 1975). However, WLN was not free from definitional problems.

In the late 1970s, as graphics input and display methods improved, the relevance of line notations declined (Lynch, 2002). Surprisingly, line notation had a revival with character codings aimed to be readable by a trained chemist and not such much because of its character economics. From 1985 on, ROSDAL Syntax was developed by Welford, Barnard and Lynch granted by the Beilstein Institut(Barnard et al., 1989). Six rules permit to code the organic molecule into an alpha-numeric string. Nevertheless, its use was restricted to Beilstein system. One year later, Weininger released SMILES (Simplified Molecular Input Line Entry System), but, unlike ROSDAL, atoms are encoded by their element character (Weininger, 1988). SMARTS (SMiles ARbitrary Target Specification) (Daylight, 2005) is an extension of SMILES and the equivalent of the MDL RG (ISIS query) files permitting substructural pattern search. The SMIRKS (SMIles ReaKtion Specification) (Daylight, 2005) language has been developed for encoding reaction transforms. STRAPS (Smiles TRAnsformation Pattern Specification) (Daylight, 2005) is a superset of SMIRKS and SMILES. For the latter the variation is too subtle to be noticed even by a expert. Further, the Daylight group developed languages for combinatorial chemical mixtures (CHUCKLES, CHORTLES and CHARTS) to reflect the mixture obtained by genuine combinatorial synthesis. Monomer symbols are used in CHUCKLES (Siani *et al.*, 1994) in the same way as atomic symbols are used in SMILES. An extension to the CHUCKLES language that represents regular mixtures is CHORTLES (Siani *et al.*, 1995). Finally CHARTS

(Daylight, 2005) provides a language for monomer-level patterns specified in CHUCKLES and CHORTLES much like the SMARTS language for molecular patterns specified in SMILES.

The latest line notation in the round is the IUPAC's International Chemical Identifier (InChI) (Stein et al., 2005). Like SMILES language (in fact USMILES), the InChI allows a canonical serialization of molecular structure. However, SMILES is proprietary and unlike InChI is not an open project. There have been several modifications not described in the literature to resolve canonization issues. This has led to the use of different generation algorithms and/or different implementations, and thus, different SMILES versions of the same compound can be found. On the one hand, SMILES is certainly more human-readable by an expert and can be used for substructure search and analysis. On the other hand, InChI allows with its option that detects mobile hydrogens to detect tautomeric forms and group them together. Its very first purpose is to identify a compound in a unique manner. InChI is more a concurrent for the proprietary CAS Registry Number than for SMILES which it will complement. Recently, InChI was reported to improve chemical semantic web (Coles *et al.*, 2005) and to curate, index and query 3-D structures (Prasanna *et al.*, 2005).

	WLN	ROSDAL	SMILES	InChI
Canonical			(X)	X
Normalized				X
Stereochemical	X	X	X	X
Tautomerism				X
Reaction encoding			X	
Generic atoms			X	
Widespread			X	X
Convertible back to graph	(X)	(X)	X	X

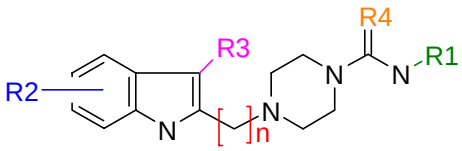
Table 2.2 Line notations and their properties The use of brackets marks that the property is in theory described, but practically flaws appeared.

2.1.4 Markush structure

Since the mid-1980s three separate groups have been developing operational systems for topological storage and retrieval of Markush structures storage and retrieval systems of chemical structure information (Berks, 2001). The origin of Markush structure goes back to Eugene Markush (Markush, 1924). Nowadays, it commonly occurs in patents for protecting compounds relating to an invention. It allows condensing the notation of all possible structures emerging from the different substituent possibilities. Abusive forms of Markush claims, the so-called “Nasties”, are very difficult to interpret and nearly impossible to index (Berks, 2001), whether for MARPAT or Markush DARC (Dubois, 2002).

Variations used in Markush structures can be of the following types (example given in Figure 2.2):

- Substituent variation; list of specific alternatives (R1).
- Positional variation; the point of attachment is variable. Example is the summarizing of ortho, meta and para substitution (R2)
- Frequency variation; multiple occurrences of groups (n)
- Homology variation; generically described groups like alkyl or aryl (R3)
- Atom variation (R4)



R1 = methyl, benzyl, naphthyl, adamantyl

R2 = F

R3 = alkyl 1-4C

R4 = O, S

n = 1, 3, 5

Figure 2.2 Example of a Markush structure (left) and its substituents (right)

The type of variability which is the most difficult to understand by computer is the positional variation, e.g. an R-group point from the center of a benzene ring.

Combinatorial chemistry adopted Markush structures and developed its own jargon (Leland et al., 1997). “Specifics” are the whole, concrete structures. Markush structures are a formal representation of all specifics of a combinatorial library. “Subgenerics” are the intermediate structures, meaning that one R-group is explicitly enumerated while the substitution point marks remain.

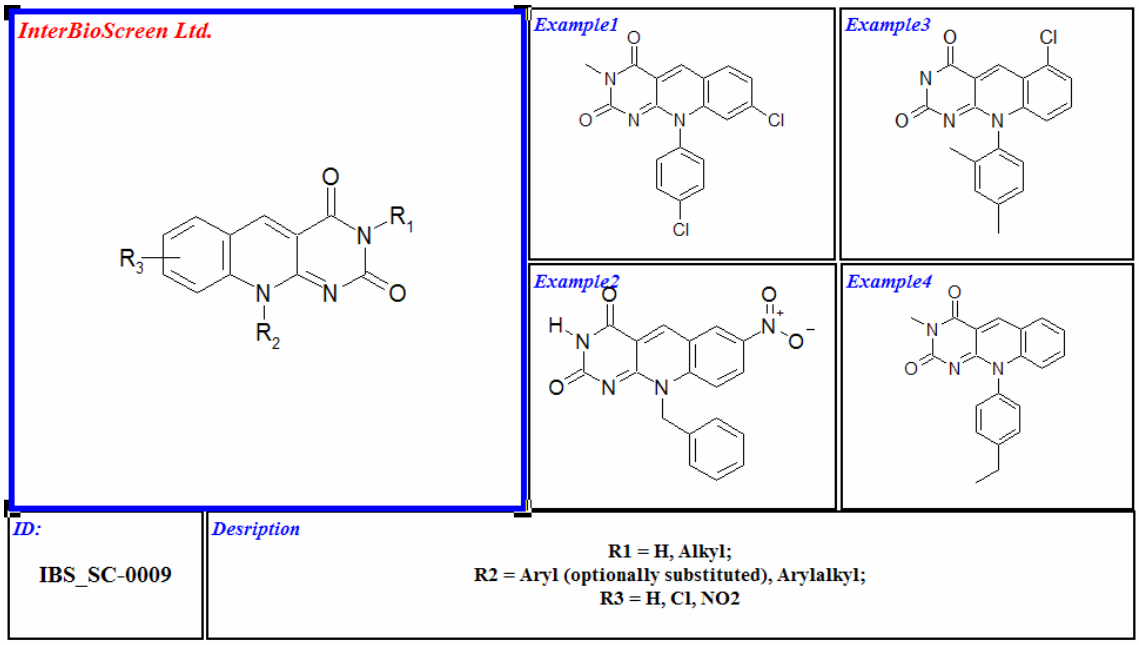


Figure 2.3 Markush Structure and specifics of the InterBioScreen Scaffold library

Markush structure handling remains cumbersome. The chemist is currently endowed with several applications to store, retrieve and manipulate specifics. Applications, permitting the elementary operations on generic structures, are only available in chemical information systems. A first case study could be the reorganization a vendor's compound collection files and its associated scaffold file like InterBioScreen (IBS) provide to their customers. As Schuffenhauer et al. point out, the Markush structure represents the scaffold as most stringent substructure query of a set of compounds (Schuffenhauer et al., 2004). However no standard have been defined which permits conversion and exchange. As there is a relationship between a generic structure and its specific (Leland et al., 1997), a given file format should be able to represent this hierarchy. From that perspective; XML-based Chemical Markup Language (CML) (Gkoutos *et al.*, 2001; Murray-Rust & Rzepa, 1999, 2001, 2003; Murray-Rust *et al.*, 2004) is the ideal candidate molecular format to represent such a hierarchy by defining a new entity in the formal specification (DTD) and could soon become the standard exchange format for molecular relational databases.

2.1.5 Fragmentation codes

Fragmentation codes were developed in order to index a molecule by specific chemical fragments. They originate with 80-column computer punch cards (Figure 2.4). For each molecule one card was used and each position on the card was allocated a structural feature. If the feature was present in the molecule, then a hole was punched at the corresponding position. Several organizations and database producers designed their own code which best represent the types of structures they dealt with. Some examples are SK&F Code (Craig & Ebert, 1969), IDC GREMAS code, Derwent's CPI code IFI Comprehensive Code, but only the latter two are survivors of the fragmentation code systems. The machine picked out all cards with given structural characteristics (Figure 2.4). This sort of search would have been almost impossible, or at least very expensive, using the traditional book-based techniques. Nowadays that sort of description is called a fingerprint (for example MACCS key fingerprint).

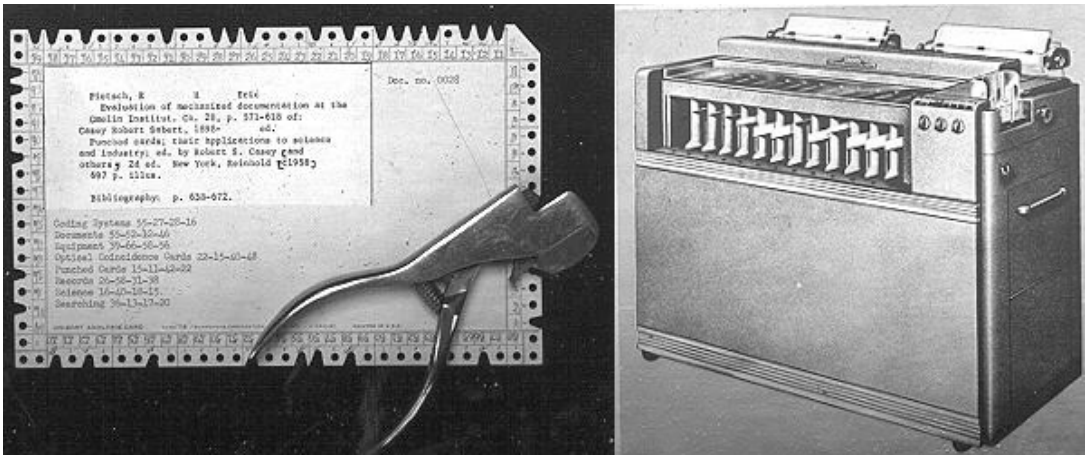


Figure 2.4 Punched cards (left) were sorted via the IBM 101 Electronic Statistical Machine (right). Courtesy from Claire K. Schultz, information scientist at Merck, Sharp & Dohme, in charge of their retrieval system.

2.2 Defining the Scaffold

2.2.1 Fragment-based drug discovery

The basic hypothesis underlying fragment-based drug discovery is that a combination of basic fragments builds every bio-active molecule.

A comprehensive review by Rees et al. (Rees *et al.*, 2004) enumerates four categories of fragment-based approaches, i.e. fragment evolution, fragment linking, fragment self-assembly and fragment optimization. The first and the last undergo structural modifications by adding supplementary fragments, but the difference consists in improvement of binding affinity for the first and of ADMET[1] for the last. Fragment linking refers to the combination of two (or more) fragments moderately affine for the same target and leading to significant gain in binding affinity. Fragment self-assembly uses mixtures (libraries) of constituents that react directly in situ the molecular target. Dynamic combinatorial libraries (Huc & Lehn, 1997; Lehn & Eliseev, 2001) and click chemistry (Kolb *et al.*, 2001; Lewis *et al.*, 2002) are associated concepts.

In a series of two papers entitled “The properties of known drugs” (Bemis & Murcko, 1996; Bemis & Murcko, 1999), the authors Bemis and Murcko conducted a substructural analysis of the 5120 marketed drugs [2] in order to infer which two-dimensional molecular shapes are prerequisite for biological activity. A similar study conducted by Cramer et al. (Cramer III *et al.*, 1974) twenty years before, had been based on a less sophisticated substructural system (SK&F fragment code) and comprised 770 structures tested for antiarthritic-immunoregulatory activity. Cramer et al.

[1] Acronym for "Absorption, Distribution, Metabolism, Excretion and Toxicity"

[2] Filtered from the Comprehensive Medicinal Chemistry (CMC) database v94.1. The CMC indexes 8757 compounds for which 8685 have 3D models (07/2005). Approximately 250 molecules, identified for the first time in the United States Approved Names (USAN) list, are added by year.

could conclude that substructural analysis is a worthwhile method to develop, since advances would be expected (i) in computer and programming technology to allow a direct computer manipulation of complete structural records and (ii) in databases acting as compendium of known drugs or bioactive molecules.

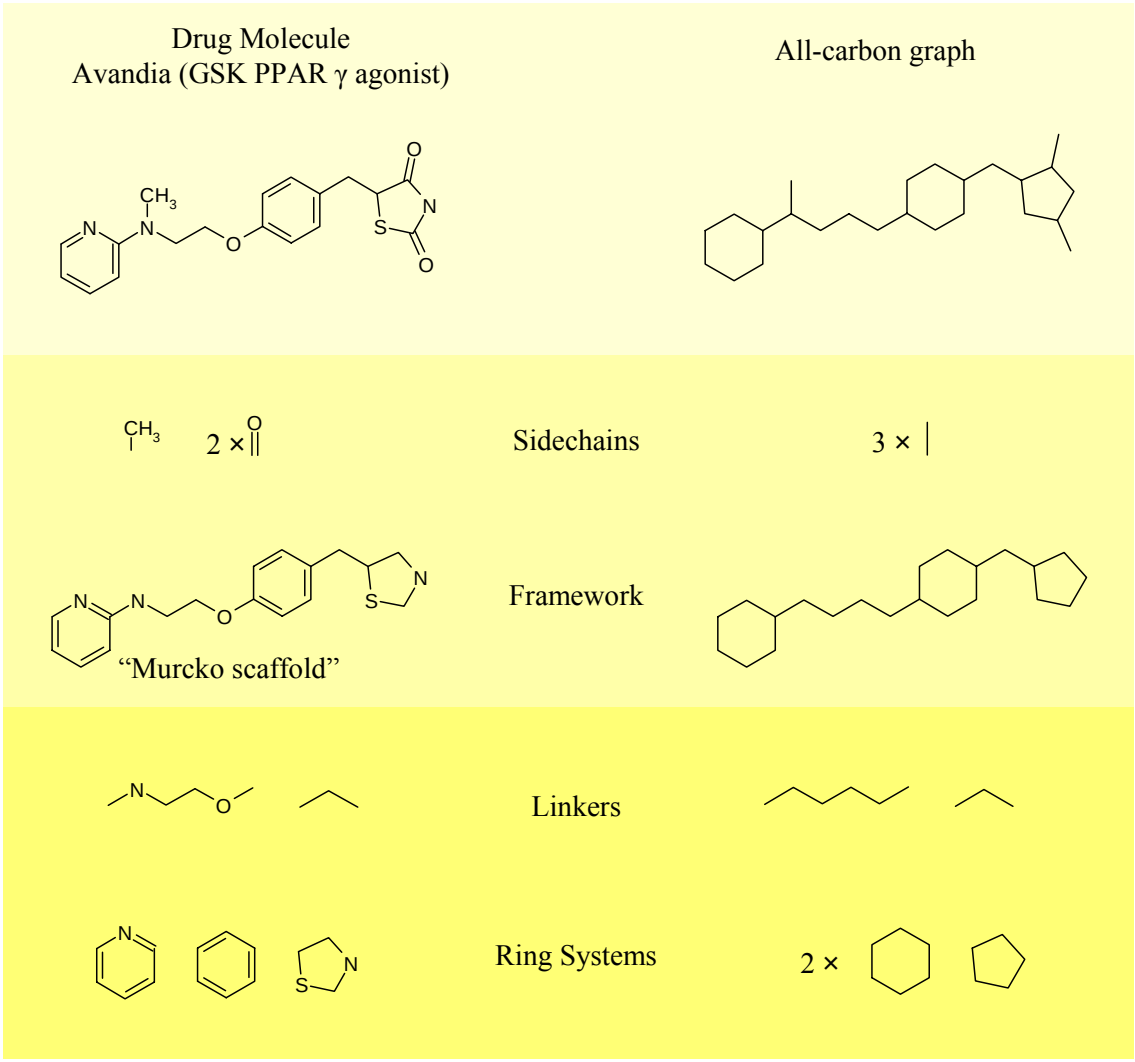


Figure 2.5 Example of molecular dissection according to Bemis and Murcko. Level 0 presents the initial drug molecule and its all-carbon graph representation. Level 1 depicts the first dissection into sidechains and framework for both atomic and all-carbon graphs. Level 2 is the result of subsequent dissection of the framework into ring systems and linkers (open valences on each end).

Back to the 90s, Bemis and Murcko (Bemis & Murcko, 1996) described a simple, but not simplistic, approach to dissect a molecule (its all-carbon graph) in four entities

(related through a hierarchical relationship). A drug molecule is composed of sidechains and a framework which are in turn composed of ring systems and linkers (Figure 2.5).

Later published papers (Schuffenhauer, 2005) refer to atomic framework as the "Murcko scaffold" and the fragmentation methodology (see below for definition and variants of both concepts scaffold and fragmentation methods) is one of most widely used method by the community. In some embodiments, carbonyl groups are considered to be part of the framework. This can be justified by the fact that the presence of the carbonyl moiety imparts significant property differences to the frameworks (Katritzky *et al.*, 2000).

A lead optimization program often generates compounds sharing the same invariant substructural moiety. In the literature, this latter moiety is named interchangeably scaffold, template, core structure, chemotype or molecular framework.

A scaffold (Bemis & Murcko, 1996; Xue & Bajorath, 1999) may also be defined as the Markush structure without R-groups. Moreover, the "R-group" is defined as a functional group or (non-ring) side chain with only one connection point to the rest of the molecule. Thus, an R-group is distinct from linkers which connect ring structures and that are part of the scaffolds. Leland et al. (Leland et al., 1997) emphasize that R-groups can be nested, i.e. an R-group can be contained in a bigger R-group. Scaffolds are associated with the computer representation of Markush structures.

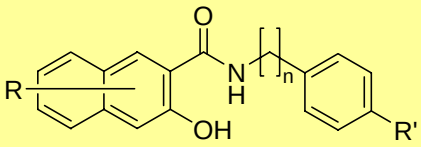
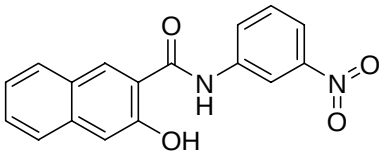
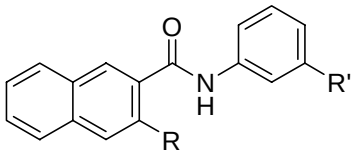
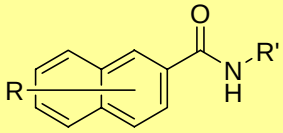
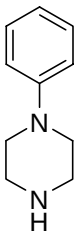
	 Biological scaffold
 Herpesvirus polymerase inhibitor	 Topological scaffold
	 Synthetic scaffold

Table 2.3 Example of compound and its respective biological, topological and synthetic scaffolds. adapted from (Xu & Johnson, 2001)

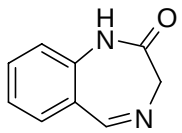
Xu (Xu, 2002) outlines the distinction between topological, biological and synthetic scaffold (Table 2.3). The "topological scaffold" is considered as the molecules depleted by its side chains (same definition as "scaffold"). The definition of "topological" implies some knowledge of the environment; thus the use of this qualification should imply explicit points on the scaffold at which R-groups are to be attached or consider atom-augmented scaffolds. The "synthetic scaffold" can also be called building block. Examples can be found at the suppliers download pages (Figure 2.3). The synthetic scaffold is somewhat smaller in terms of atom count than the others. The "biological scaffold" is thought to be the structure common to a particular compound series having

demonstrated activity (binding affinity/ADME) associated to a biological target. On one hand, the scaffold may provide the main contribution to the interaction with the biological target. On the other hand, the three-dimensional conformation of the scaffold is thought to help orientate the substitutions in space. Therefore, scaffolds can be considered as organizational units (Ramstroem & Lehn, 2001).

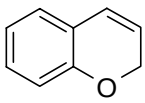
In the quest of the universal scaffold, the minimum required qualities are bioavailability and synthetic expansion. Several groups (Opatz et al., 2003; Thanh Le et al., 2003) propose sugar as platform to tailor molecular diversity. Monosaccharide-based scaffolds consist of a single cyclized poly-functionalized aldehyde or ketone unit. The six-carbon sugar D-glucose is most abundant monosaccharide in nature. Its five chiral hydroxyl functions can be diversified for example by an esterification. Carbo-hydrate like scaffolds have been used for exploration of several pharmacologically important target like SST5 and MC4 receptors, protein kinases and bacterial cell wall proteins. In this context, privileged structures are defined as scaffolds being able to bind multiple distinct protein classes (Evans et al., 1988). Recently, Horton et al. reviewed the combinatorial synthesis of some twenty bicyclic privileged structures (Figure 2.6) and prefer the term "privileged substructure" (Horton *et al.*, 2003).

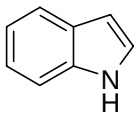


Phenyl-substituted monocycles



Fused [7-6] Ring Systems





Fused [6-6] Ring Systems Fused [5-6] Ring Systems

Figure 2.6 Representative examples of bicyclic privileged structures

Types of scaffolds can be separated mainly into linear and globular scaffolds (Ramstroem & Lehn, 2001). These topologies correspond to acyclic and ring-systems frameworks. Beilstein generic categories (Figure 2.7) provide further subdivisions based on the atom composition.

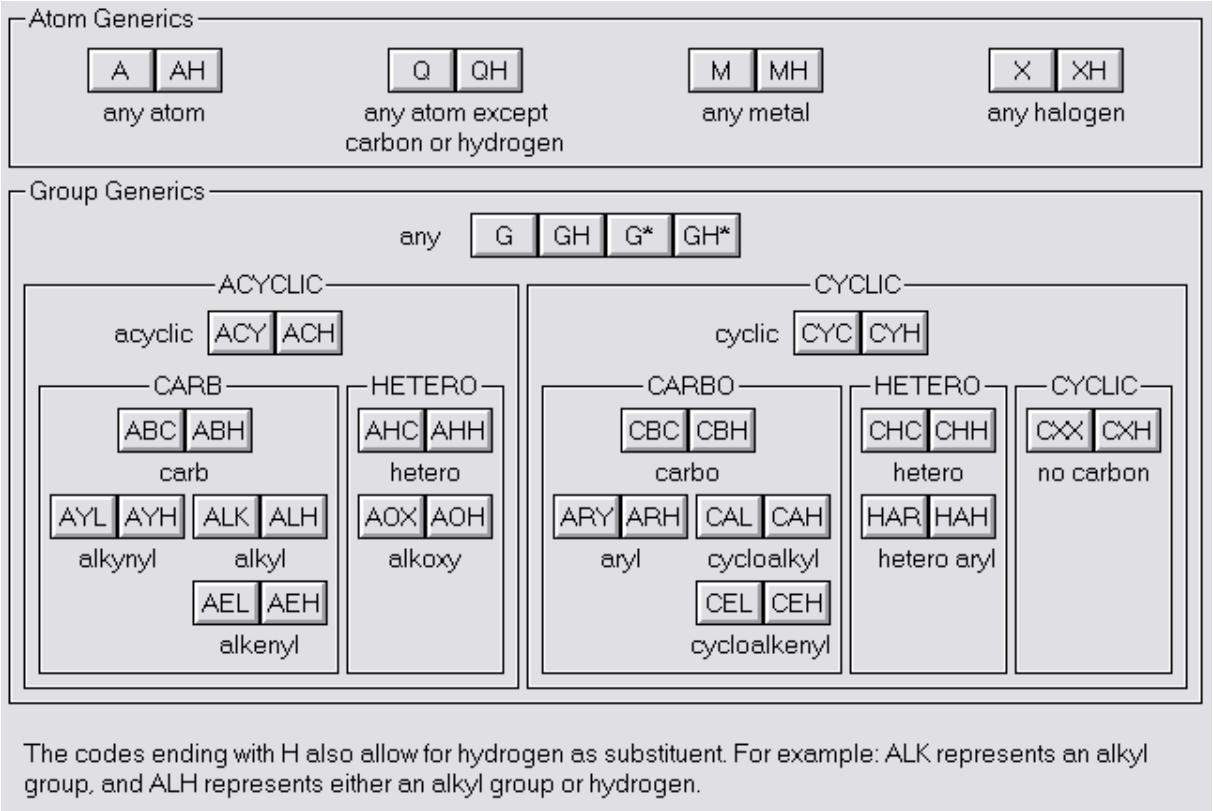


Figure 2.7 The different Beilstein Generic categories. An asterisk on the generic group (i.e. G*) allows ring closure between the group and the rest of the structure.

Acyclic frameworks account for less than ten percent of the marketed drugs (Bemis & Murcko, 1996), although a substantial number of combinatorial libraries based on acyclic structures have been produced. The main reason may be the synthetic tractability of these combinatorial libraries where a part or even the entire scaffold originates from the R-reagents during the synthesis. Brady et al. (Brady et al., 1998)

provided such an example where the urea scaffold is completely constructed by the two R-reagents. Nature's best example of a linear scaffold is the protein backbone composed of polypeptide subunits. However, a bioactive molecule coming from such libraries, carrying cycles at the extremities, will be decomposed into the ring systems and the linear part considered as linker. Another arguable case is a branching acyclic template. The question is whether is should be considered as linker (by definition, a linker has two connection points) or as a mono-atomic scaffold.

Cyclic templates, whether they are mono-, di-, poly-, fused or spiro cycles, are considered to be key features by the chemists (Lipkus, 2001; Nilakantan *et al.*, 1990). Moreover, of these, heterocylic structures dominate the chemical scaffold space. De Laet et al (De Laet *et al.*, 2000) reported that 53% of compounds contained in the Beilstein are heterocyclic. This proportion increases to 68% (at least two third) when analyzing molecules having reached at least clinical studies. Much of the structural variety of the drug-like space arises from heterocyclic rings and combinations thereof. Nevertheless, the ratio of heteroatoms to carbon atoms must lie in the recommended range between 0.1 and 1.8 (Feher & Schmidt, 2003; Zheng et al., 2005), if the application field of the structures should be pharmaceutical and not petrochemicals or explosives.

Recently, the docking problem (i.e. how to find the optimal interaction between two molecules) was associated to the scaffold concept. In a first example, Lamb et al. (Lamb et al., 2001) have developed a rapid docking method by evaluating multiple libraries against multiple biological targets during the design stage. The method involves three main stages: (i) dock the scaffold; (ii) select the best substituents at each site of diversity (scaffold plus single substituent is constructed for each site and docked); (iii) compare the resultant fully substituted molecules within and between libraries. This approach

provides a rapid way of exploring large lists of possible substituents with linear rather than combinatorial time dependence. The method has been exemplified for three libraries (one peptide, two non-peptides) docked into three different serine proteases. In this way, libraries can be designed to hit families of proteins, or conversely, selectivity issues can be explored. However, simple docking of a "bare" scaffold is not a good method of placing the scaffold into productive geometries. A site generated evaluation filter has proved to be valuable. In a second distantly related approach, Su et al (Su et al., 2001) suggest that a segregation of potential ligands into families of related molecules should increase the diversity of hits. The third example outlines scaffold-driven docking proposed by Chema et al. (Chema *et al.*, 2004). In order to identify preferred binding mode(s) of a scaffold, a large set of different ligands sharing the same scaffold is docked to the same protein target. The method is applied to members of protein kinase family as target and suggests that predicted alternative binding modes could be an aid to experimentalists.

2.2.2 MCS, mcs and MOS

We saw that the scaffold is defined as the invariant part among the compounds composing a combinatorial library. Therefore, it can be represented as a Markush structure. In many real world problems, the common invariant moieties, if not defined, have to be detected. Hattori et al. (Hattori *et al.*, 2003) for example, applied an MCS search implemented with heuristics to the KEGG/LIGAND database of metabolites. The purpose of this project was to identify biochemical meaningful substructures and establish the correspondence to the KEGG pathway map numbers. MCS algorithms are also applied by Bioreason ClassPharmer (Bioreason, 2005) and Tripos Distill (Tripos, 2005) to organize thousands of "active" compounds into meaningful groups. Statistical methods are subsequently applied to structures and related data to learn as much

information as possible. Hence, 2-3 series of compounds can be rationally selected by the medicinal chemist for follow-up studies.

The MCS problem belongs to the class of graph theoretic problems called isomorphism algorithms. Wheland (Wheland, 1949) and Mooers (Mooers, 1951) independently suggested to record chemical structural formulas as graphs on a computer for structure and substructure searches. Graphs are indeed a particularly convenient abstraction of a molecule (Figure 2.8). The atoms and bonds equal the set of vertices (nodes) and of edges (arcs), respectively (Hansen & Jurs, 1988). Lipinski recalls in his publication about the "Rule of Five" (Lipinski, 1997) that chemist's very strong skills in pattern recognition and their outstanding chemistry structural recognitions skills are likely to enhance information transfer.

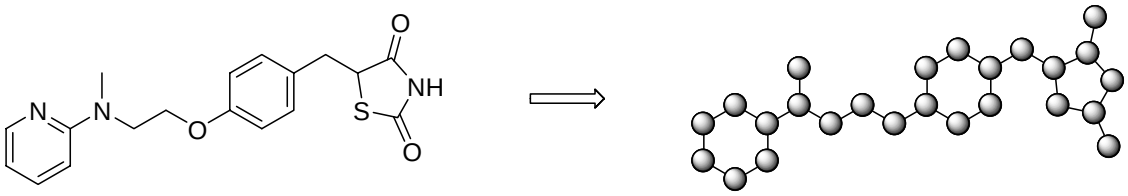


Figure 2.8 Molecule and its abstraction into graph

Several subgraph concepts have been defined in the literature (McGregor & Willett, 1981). MCS stands for maximum common substructure or subgraph. This acronym can also be employed for minimum common superstructure or supergraph, but in order to distinguish the latter from MCS, we will denote it "mcs". Their intuitive relationship is confirmed in the sense that computation of the one can be reduced to the computation of the other (Bunke *et al.*, 2000; Fernández & Valiente, 2001). MCS is often used as collective term. Sometimes MCS is used in equivalent manner than the Vleduts definition of "maximal overlapping substructure" (MOS) (Vleduts, 1977). Another definition sets the MCS equal to the largest single contiguous substructure, whereas the

substructures of MOS do not need to be connected. Yet, Raymond (Raymond *et al.*, 2002) defined the maximum common induced subgraph (MCIS) and the (disconnected) maximum common edge subgraph (MCES) as the common substructure consisting of the most atom pairs and bond pairs, respectively. Figure 2.9 depicts two PDE4 inhibitors, Rolipram and Cilomilast, with their respective MCIS and MCES highlighted in bold.

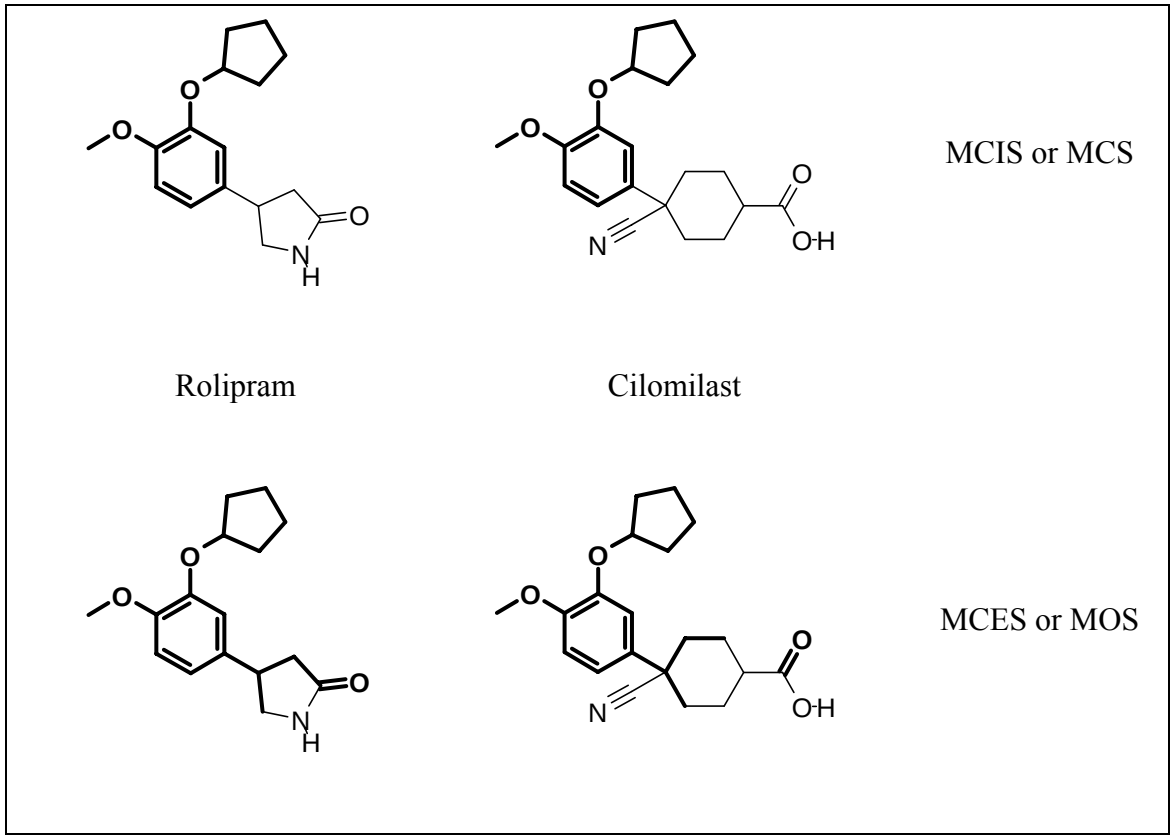


Figure 2.9 Different common subgraphs of two PDE4 inhibitors

Different algorithms to uncover the MCS have been developed. The search for the MCS is a NP-complete problem[3], which implies that all exact algorithms have a worst-case time complexity very likely being exponential to the number of vertices in the graph. The MCS problem is reduced to the maximal clique (Kann, 2000) problem, but remains a NP-complete problem. Thus, the introduction of heuristics is required. Search space

[3] NP stands for "nondeterministic polynomial time" (Lopez-Ortiz, 2000).

may be pruned by removing redundant matching units for example (Raymond et al., 2002). Often a threshold size (heavy atom count) has to be defined to ensure that the user is not overwhelmed by the huge output consisting primarily of small common substructures with little structural significance.

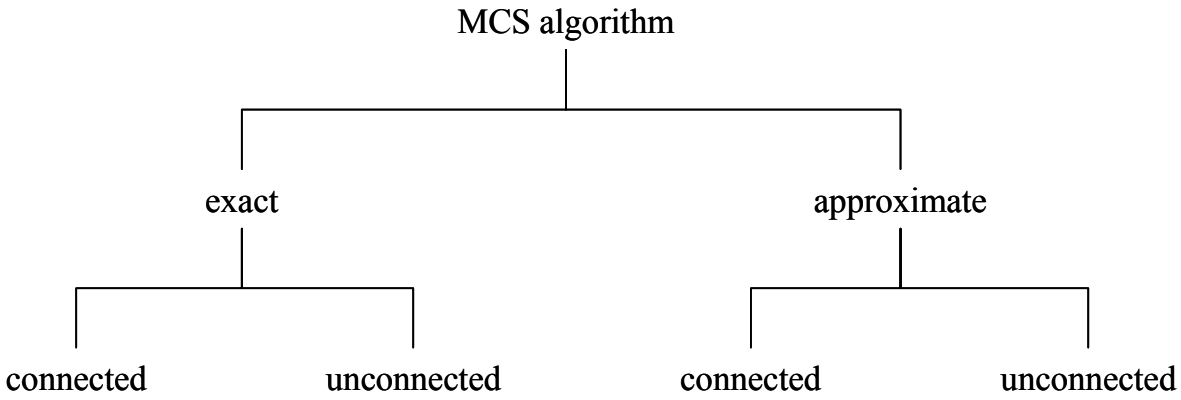


Figure 2.10 Classification of MCS algoritms, adapted from Raymond (Raymond, 2002).

A number of maximum clique detection algorithms have been published in the literature; among those can be cited Bron-Kerbosch, Crandell-Smith and Rascal algorithms. The fastest algorithm at present has been reported to be Rascal (Stahl *et al.*, 2005). These algorithms can be categorized (Figure 2.10). Nevertheless, due to vague or ambiguous descriptions, categorization appears to be very difficult to realize.

2.2.3 Fragmentation methods

In order to conduct substructure analysis or make use of one of the fragment-based approaches, existing molecules (preferred are drugs or drug-like) have to be fragmented. We describe selected approaches of partitioning a molecular graph.

The most obvious way fragmentation of a molecule is breaking it down into its elementary parts, the atoms and bonds. Solov'ev et al. generated several different ensembles of subgraphs (atom/bond sequences and "augmented atoms") from which upon statistical criteria the optimal molecular fragment is selected (Solovév *et al.*, 2000).

As described before, Murcko and Bemis provided a pregnant method of fragmentation for their substructural analysis (Bemis & Murcko, 1996). Xue and Bajorath used a kindred method to isolate scaffold and R-groups of from large compound collections (Xue & Bajorath, 1999). Another way of fragmentation is used by docking methods. Molecules are cut into fragments by splitting at rotatable bonds. This is also how the medicinal chemist proceeds when deconstructing a hit.

In the RECAP method (Retrosynthetic Combinatorial Analysis Procedure) (Lewell *et al.*, 1998), molecules are fragmented in silico based on chemical knowledge. Eleven chemical bond types have been defined to cleave the molecule in fragments (Table 2.4). These bond types are derived from common chemical reactions. The cleavage of ring bonds is prohibited in order to preserve ring motifs. Thus the rules are restricted to cleave acyclic bonds. Fragments are defined to belong to two categories: terminal monomer (one connection point) and core template (two or more connection points). In order to avoid "trivial" fragments, possible terminal fragments either consisting of hydrogen, methyl, ethyl, propyl and butyl (and in some embodiments also phenyl) is not cleaved off. Later, Lewell at al. created a drug ring database by deconstructing the molecules at single and olefinic non-cyclic bonds (Lewell et al., 2003). Katritzky et al. propose an outline of an expert system that permits to define the principal template of a molecule and defines simple rules helping to decide whether two compounds belong to the same template or not (Katritzky et al., 2000).

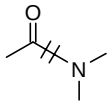
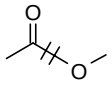
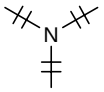
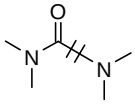
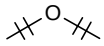
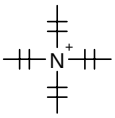
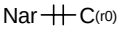
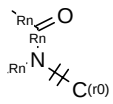
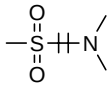
Bond Type	Cleavage rule	SMIRKS
Amide		[O:3]=[C!$(C([#7])(=O)[!#1!#6]):2]-[#7!$([#7][!#1!#6]):1] >>[O:3]=[C:2].[#7:1]
Ester		[#6!$([#6](O)~[!#1!#6])][O:2][C:1]=O >>[C:1]=O.[#6][O:2]
Amine		[#6:2]-[N!$(N[#6]=[!#6])!$(N~[!#1!#6])!X4:1] >>[N:1].[#6:2]
Urea		N[C:1]([N:2])=O >>N[C:1]=O.[N:2]
Ether		[#6]-[O!$(O[#6]~[!#1!#6]):1]-[#6:2] >>[#6:2].[O:1]-[#6]
Olefin		[C:1]=[C:1] >>[C:1].[C:1]
Quaternary nitrogen		[#6:1]-[N$(N([#6])([#6])([#6])[#6])!$(NC=[!#6]):2] >>[#6:1].[N:2]
Aromatic nitrogen - Aliphatic carbon		[n:1]-[#6!$([#6]=[!#6]):2] >>[n:1].[#6:2]
Lactam nitrogen - Aliphatic carbon		[C:3](=[O:4])@-[N:1]!@-[#6!$([#6]=[!#6]):2] >>[C:3](=[O:4])[N:1].[#6:2]
Aromatic carbon - Aromatic carbon		[c:1]-[c:1] >>[c:1].[c:1]
Sulphonamide		[#7:1][S:2](=O)=O >>[#7:1].[S:2](=O)=O

Table 2.4 Eleven RECAP bond cleavage types with corresponding SMIRKS

2.3 On the diversity and complexity in molecular library design

Diversity, like its antonym similarity, is in the eye of the beholder. Indeed, there is no consensus about the manner to quantify molecular diversity (Martin, 2001; Monev, 2004). We are electing to restrict ourselves to considerations on molecular diversity that deal with the design of a screening set. This set may stem from a collection of individual compounds or a combinatorial library. Similarity property principle states that structural similarity engenders similar properties. Conversely, diverse properties are represented by diverse structures. A 'cherry-picking' approach is adequate for collection of individual compounds. Before continuing with the review of available methods, let us first present some aspects of the quantification of molecular diversity.

A currently standard approach to assess the diversity of chemical structure databases consists in the characterization of the Tanimoto indices distribution. From a conceptual point of view, two objects are compared by their characteristics to measure their difference or overlap. In order to be comparable, the same measure composed of coefficient and descriptors must be applied to all object/molecule pairs. Holliday et al. investigated on the characteristics of 14 standard similarity coefficients (Holliday *et al.*, 2003) and were particularly interested by the tendency of the Tanimoto (alias Jaccard) coefficient to select small compounds in dissimilarity selection. It should be noted that the coefficient has to be appropriate to a specific goal. Monev provides an indicative guide which distinguishes between three types (a representative is cited in brackets): similarity coefficient (Tanimoto), dissimilarity coefficient (Euclidean distance) and composite coefficient (Tversky) (Monev, 2004). The other issue is the dependency of the measure on the chosen characteristics. For a molecule, these characteristics are interchangeably called properties, descriptors, features, attributes, etc. These have to be

independent of each other and discriminating. Several research groups have investigating descriptors selection methods appropriate for a given problem. For example, Hert et al described the use topological descriptors in ligand-based virtual screening (Hert et al., 2004). Meanwhile, the literature (Livingstone, 2000) counts a plethora of descriptors (Dragon software (Talete, 2005) calculates currently 1664), mostly developed for establishing quantitative structure-activity relationship (QSAR). Common used categories are constitutional, topological, electrostatic, geometrical, quantum chemical and statistical mechanical descriptors. An unbiased criterion for measuring dissimilarity between molecules should be established, independent of the fingerprints or descriptors used for similarity searching. Jenkins et al suggest to cluster in scaffold classes (Jenkins *et al.*, 2004) using molecular equivalence indices (Meqi) developed by Xu and Johnson (Xu & Johnson, 2001; Xu & Johnson, 2002). The Meqi scaffold is defined as a"recognizable structural feature" (in fact a pseudo-graph) shared by an exhaustive subset of molecules. These latter are reduced to their topology and their memberships do not shift, if more compounds are added to the dataset, as in other clustering methods. Graph reduction also opened up the possibility for scaffold hopping (Barnard *et al.*, 1989; Gillet *et al.*, 2003).

Beyond mere random subset selection, techniques borrowed from machine-learning literature like clustering, cell-based, dissimilarity-based and optimization-based methods have applied to the chemical context (Willett, 2000). Regardless the method, a similarity function for pairs of molecules is needed. The simple idea behind the application of clustering (Downs & Barnard, 2002) and cell-based algorithms (Xue *et al.*, 2004) to a compound collection is that once clusters are formed, one selects a limited number of compounds from each cluster to be tested. Thus, the subset is expected to be representative and diverse. If a compound has been confirmed as hit, its

cluster members are assayed subsequently. Dissimilarity-based compound selection (Snarey *et al.*, 1997) comprises several maximum-dissimilarity and sphere-exclusion methods. Whereas the maximum-dissimilarity algorithm aims at minimizing the sum of all pair similarities in the subset, in a sphere-exclusion algorithm the largest subset is selected in which the pair similarity between any pair of molecules does not exceed a given threshold. The latter has also the problem that final subset size can not be set the a priori, unless one wants to risk uneven sampling. Examples of the optimization-based methods have been implemented by simulated annealing (SA) (Agrafiotis, 1997; Brown *et al.*, 2000; Zheng *et al.*, 1999) and genetic algorithms (GA) (Gillet *et al.*, 1999).

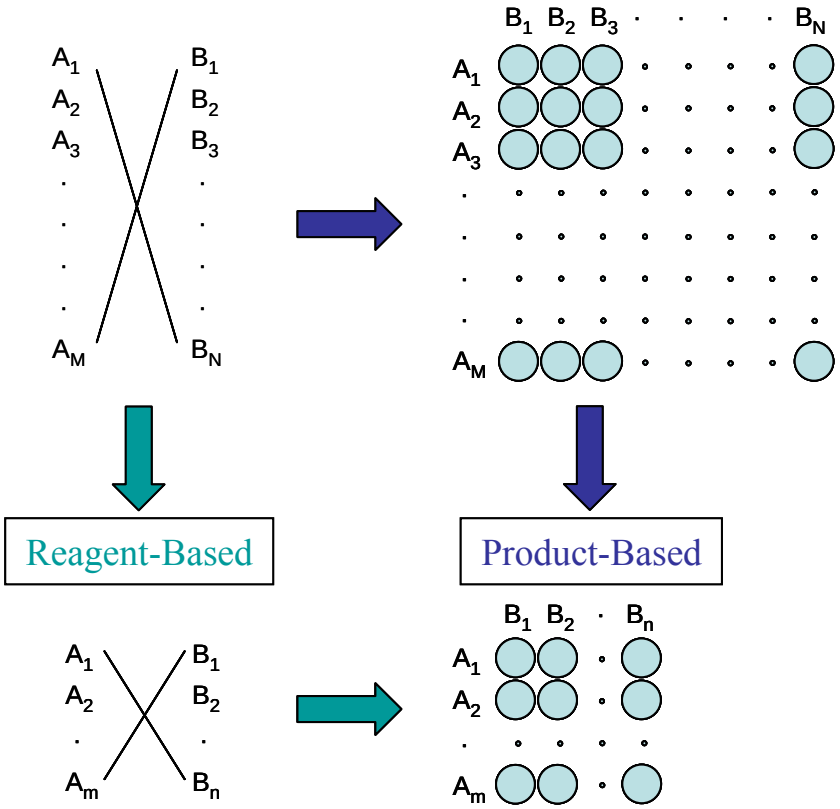


Figure 2.11 Combinatorial reagent- and product-based design

Moreover, these methods can be used for 'product-based design' of combinatorial libraries (Figure 2.11). Let us take the example of a combinatorial synthesis describing a two-component reaction with M reagents of type A and N reagents of type B, yielding

MN products of type AB. Product-based design involves selecting the mn products from the fully enumerated set of MN possible products. A reagent-based design procedure (Figure 2.11) (Gillet *et al.*, 2002), conversely, involves selecting a diverse m-member subset from all of the M available reagents of type A and a similar n-member subset from all of the N available reagents of type B (Martin et al., 1995). Whether reagent-based selection is less (Brown et al., 2000) or more (Sheridan *et al.*, 2000) demanding of computational resources may depend on its implementation, but it may result in libraries that are less diverse than those resulting from product-based approaches (Gillet *et al.*, 1997).

The objective of a library determines whether one strives for maximal diversity or a just diverse enough library (Martin, 2001). Three types of libraries in the context of screening are currently distinguished: (i) large, exploratory libraries, (ii) medium-sized, targeted libraries and small focused libraries. In the case of minimal information about the biological target, largest libraries screened nowadays contain between 10^5 and 10^6 compounds (Posner05). A decade ago, large combinatorial libraries had several drawbacks, like lack of diversity and drug-like qualities, as well as deconvolution, isolation and identification problems from compound mixtures. A high number of compounds tested did not guarantee a high number of actives.

A step further in the drug design process, more biological activity data ought to be known. A preliminary selectivity profile might be established with the help of targeted libraries. These are sets of individual pure compounds biased towards protein families, mostly G-protein coupled receptors (GPCR), kinases, nuclear receptors, proteases and ion channels. The scaffolds underlying those compounds stem from the "privileged substructures". Cell-based algorithms are often applied to the design of targeted libraries in order to ensure that the designed library covers all members of the protein class (Xue

et al., 2004). Nilakantan addresses well the question by asking whether it is better to start from 10 000 mutually dissimilar compounds or 100 analogs each of 100 different scaffolds (Nilakantan & Nunn, 2003). The author has a preference for the latter approach and argues that hit identification may be followed by a preliminary SAR. A hypothesis was formulated that small multiple-scaffold libraries are superior to large single-scaffold libraries in terms of their potential to hit a broad panel of biological targets (Sauer & Schwarz, 2003). This hypothesis was verified through computation of three-dimensional shape descriptors (normalized ratios of principal moments of inertia). Focused libraries are designed for one protein target and its primary goal is the optimization of lead structure, i.e. to bring the affinity to the low nanomolar range. The scaffold will orient the ligands in a uniform way and an increase in average affinity of these ligands to the target protein is expected. However, the difference between focused and targeted libraries concepts seems to be a question of semantics. Indeed, it should be noted that both concepts are used in an interchangeable manner and any categorization attempt will be artificial. The size of such libraries varies from 10^4 down to 10^2 compounds, the latter named mini-libraries (Nilakantan & Nunn, 2003). Hence, diversity is inversely proportional to the knowledge about the target.

With the advent of the chemical genetics concept, small organic molecules can also act as a chemical tool to help target discovery (Stockwell, 2000). Shedden et al. coined the term "supertargeted chemical library" (Shedden *et al.*, 2003) and define it as a large collection of compounds designed to localize to a specific organelle or subcompartiment. An application case was the design of a fluorescent library based on a stryryl scaffold targeting the organelle. In this example, scaffold-based library was used as molecular tool with optical properties.

Diversification of a molecular library can be achieved by incorporating not only various scaffolds, but also with various degrees of complexity. Recently, Schuffenhauer et al. reported that some diversity selection algorithms are biased towards molecules with lower complexity (Schuffenhauer et al., 2004). However, the more complex a molecule (in terms of structural features), the higher its biological activity (Hann *et al.*, 2001). Natural products as source of scaffold have been suggested to present a much higher level of sophistication (Ortholand & Ganesan, 2004). Traditional combinatorial chemistry fails to provide (among other properties) complex rings systems and enough stereocenters (Feher & Schmidt, 2003). With diversity-oriented synthesis (DOS) approach (Burke *et al.*, 2004; Burke & Schreiber, 2004; Tempest & Armstrong, 1997), it is now possible to synthesize molecules with diverse skeletons right from the start. Natural product-like cyclic architectures are predominant in compounds synthesized by DOS, although obtained by short, efficient synthetics routes.

There is general tendency to focus on smaller scaffold-based libraries and design them joining up chemoinformatics, medicinal and combinatorial chemistry approaches.

2.4 Enumerative molecular library design

Virtual library design and enumeration is currently regarded as an important capability in drug discovery. Many pharmaceutical companies have developed programs to address this need (Leach *et al.*, 1999; Yasri *et al.*, 2004). Possible objectives for virtual combinatorial library generation are:

- Enumeration of a library for registration
- Properties investigation of library
- Structure-based evaluation, diversity analysis of large libraries.

The basic enumeration strategy is illustrated by the Figure 2.12. Full connection tables (CT) are obtained by assembling the generic structures according the given instructions. A subsequent deployment step permits to generate 2D or 3D coordinates. A concatenated name is associated to the CT.

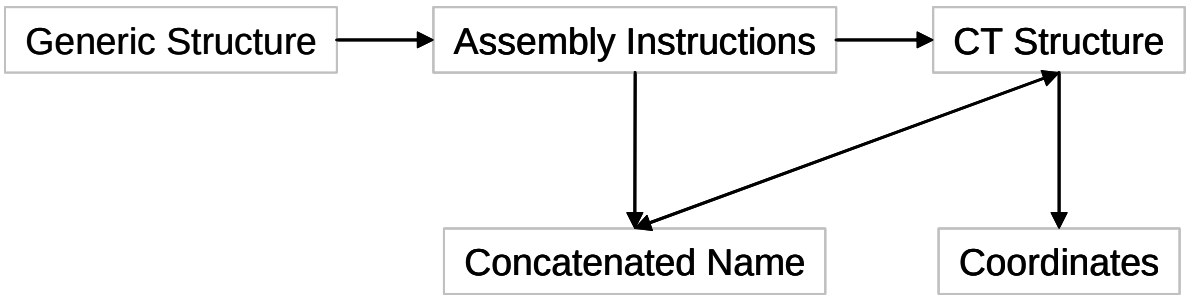


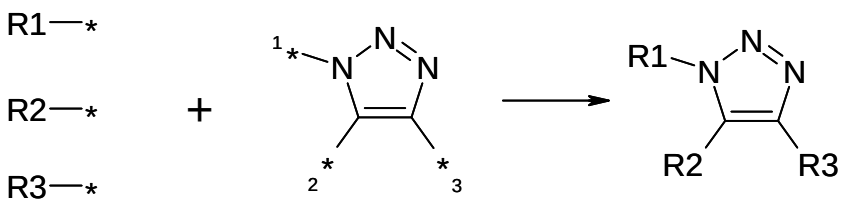
Figure 2.12 Basic enumeration strategy, adapted from Leland (Leland et al., 1997)

The following sections are technical discussions about the possibility to enumerate in a quick manner fragments to specifics.

Two types of enumeration are distinguished: complete, so-called “bulk” enumeration and constrained enumeration (Leland et al., 1997), the latter resulting in a subset of structures of the former. The enumeration can be constrained by a list of disallowed bond and angles (Rotstein & Murcko, 1993) or a set of selectivity rules imposing

property ranges For example, Nilakantan submitted a method for structure generation by random combination of known fragments (Nilakantan *et al.*, 1991). Subsequent selection removes the gross of chemically unstable or very unlikely structures. This resulted in a at least 99 % rejected structures. The selection of the reagents can be handled by subset selection algorithms (previous section) or the user interactively.

The generic structures are either joined by fragment marking (also known as the Markush approach) or reaction transform (Agrafiotis *et al.*, 2002). In the Markush structure approach, the fragments are marked by a special character to define the substitution points (Figure 2.13). The following examples illustrate the strategy. SMILIB (Schuller *et al.*, 2003) was developed to perform virtual reactions of building blocks and linkers with scaffolds using SMILES notation. The “Any” (“[A]”) atom symbol is replaced by a "%N" (N is an integer above 9), it becomes a bond symbol. Concatenation of the incomplete ring closure characters (“%”) containing strings using the dot disconnect character results in unconventional, but nevertheless valid SMILES. Serving as input for 2D or 3D coordinates generators, a complete molecule is created. The same result is obtained using an OpenEye(OpenEye, 2005) SMILES extension for external attachment points, an integer value following the “&” character corresponding to the atoms map index. "&1" is identical to "R1" and to "[*:1]".

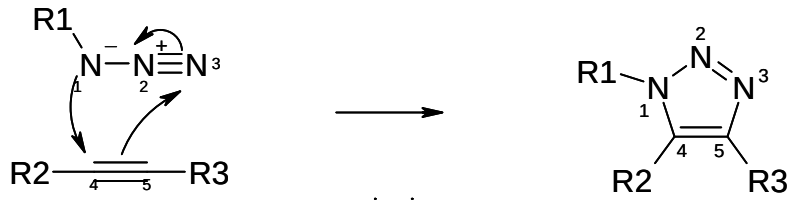


N1([*:1])N=NC([*:2])=C1[*:3]

Figure 2.13 Fragment marking of 1,2,3-triazole core and R-groups

Zauhar et al. implemented with their program ALMS (Automated Ligand-binding with Multiple Substitutions) (Zauhar *et al.*, 2003), written in the SYBYL programming language (SPL), a strategy to assemble hit fragments to a selected framework in a 3D conformation. This latter was prepared by considering each ring hydrogen of the framework and each nonring hydrogen of a hit fragment as a possible attachment point, marked by a dummy atom. A fragment was attached to the framework by removing the dummy atoms on both the hit and the target inhibitor site and replacing these with a single bond linking the inhibitor and the fragment. The orientation of the newly attached fragment was then optimized using FlexiDock, the genetic-algorithm-based optimizer included with the SYBYL modeling package.

The "reaction transform" approach proceeds by mapping the atoms of the implicated reagents. Thus atoms that change during the reaction are mapped and identified on the reactant and product side. Atom maps can convey the reaction mechanism (Figure 2.14). Reaction sequence, as implemented in ChemAxon Reactor (ChemAxon, 2005), allows in addition to process reaction sequentially. Highly adapted formats are the MDL RXN or RDF, as well as Daylight SMIRKS.



[N-:1]-[N+:2]#[N-:3].[C:4]#[C:5]>>[N:1]1[N:2]=[N:3][C:5]=[C:4]1

Figure 2.14 Huisgen 1,3-dipolar cycloaddition of azides and acetylenes by SMIRKS to give 1,2,3-triazoles

Lobanov et al. report about the construction of virtual combinatorial libraries (Lobanov & Agrafiotis, 2002) by implementing a reaction scripting language (RSL). The

definition of each combinatorial reaction equals a named Tcl (Tool command language) and is composed of three blocks; definition of reagents and the products, then assembly instructions and lastly, a statement triggering the mappings of the reactive patterns onto the reagents, compiling the assembly instructions and storing the virtual library into a file. Substructural patterns are encoded in SMARTS, not in SMIRKS. The latter are restricted in reaction functionalities, i.e. bond queries are not available and if the bond order or the connectivity changes, atomic expressions can not contain queries.

Recently, Wolber at al. proposed CombiGen (Wolber & Langer, 2000). Three phases: parameterization, compound generation and filtering. In the parameterization phase, the program has as additional feature the possibility to evaluate and prioritize parts of the fragments according to their reactivity potential. Before the compound generation process starts, the set of rules, defining assembly, is extracted from the user-defined fragment pool and interactivity permits to adapt those rules if necessary. A sequencer class generates arrays of integer upon the user-defined constraints which represents fragment index sequences. In the filtering phase, the constraints like range of molecular weight, number of atoms and chemical features are evaluated in order to retain only corresponding molecules. Selected fragments are then combined to constitute the virtual combinatorial library.

MOLGEN-COMB (Gugisch *et al.*, 2000) let the user choose or define interactively the core and building blocks. The symmetry group of a molecule is taken into account during the combinatorial generation of a non-redundant compound library.

De Novo design techniques dispose all of a structure generation module like those described herein. For a more informative introduction to the field of De novo design, detailed texts are available (Schneider & Fechner, 2005).

Virtual combinatorial chemistry, as well as its real world congener, is shifting away from the mix-and-split era, where fragments were added together just for the sake of creating new combinations and thus increase the number of molecules.

2.5 Conclusion

This chapter has presented some salient concepts that are of interest for computer-assisted library design. Of particular interest to this work is the MCS problem as it pertains to scaffold detection. Considering the literature and taking into account our objectives which are to develop and to explore a strategy for design of small-sized, yet diverse libraries and its related scaffold database, we based ourselves mainly on an "inverted" fragmentation method (Bemis & Murcko, 1996). As we saw before, this method decomposes a molecule in a hierarchical manner.

Bibliography

Adamson, G. W., & Bawden, D. (1975). A method of structure-activity correlation using Wiswesser Line Notation. *J. Chem. Inf. Comput. Sci., 15*, 215-220.

Agrafiotis, D. K. (1997). Stochastic algorithms for maximizing molecular diversity. *J. Chem. Inf. Comput. Sci., 37*, 841-851.

Agrafiotis, D. K., Lobanov, V. S., & Salemme, F. R. (2002). Combinatorial informatics in the post-genomics era. *Nat Rev Drug Discov, 1*, 337-346.

Ash, S., Cline, M. A., Homer, R. W., Hurst, T., & Smith, G. B. (1997). SYBYL Line Notation (SLN): A versatile language for chemical structure representation. *J. Chem. Inf. Comput. Sci., 37*, 71-79.

Barnard, J. M., Jochum, C. J., & Welford, S. M. (1989). *ROSDAL: A universal structure/substructure representation for PC-host communication.* In *Chemical Structure Information Systems*, W. A. Warr, Ed. pp 76-81, Washington.

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem., 39*, 2887-2893.

Bemis, G. W., & Murcko, M. A. (1999). Properties of known drugs. 2. Side chains. *J. Med. Chem., 42*, 5095-5099.

Berks, A. H. (2001). Current state of the art of Markush topological search systems. *World Patent Information, 23*, 5-13.

Bioreason. (2005). ClassPharmer Suite (Version 3.2-3.5). Santa Fe, NM, USA: Bioreason, Inc.

Brady, S. F., Stauffer, K. J., Lumma, W. C., Smith, G. M., Ramjit, H. G., Lewis, S. D., et al. (1998). Discovery and development of the novel potent orally active thrombin inhibitor N-(9-hydroxy-9-fluorenecarboxy)prolyl trans-4-aminocyclohexylmethyl amide (L-372,460): coapplication of structure-based design and rapid multiple analogue synthesis on solid support. *J. Med. Chem., 41*, 401-406.

Brown, R. D., Hassan, M., & Waldman, M. (2000). Combinatorial library design for diversity, cost efficiency, and drug-like character. *J. Mol. Graph. Model., 18*, 427-437, 537.

Bunke, H., Jiang, X., & Kandel, A. (2000). On the Minimum Common Supergraph of Two Graphs. *Computing (Vienna/New York), 65*, 13-25.

Burke, M. D., Berger, E. M., & Schreiber, S. L. (2004). A synthesis strategy yielding skeletally diverse small molecules combinatorially. *J. Am. Chem. Soc., 126*, 14095-14104.

Burke, M. D., & Schreiber, S. L. (2004). A planning strategy for diversity-oriented synthesis. *Angew. Chem. Int. Ed. Engl., 43*, 46-58.

Chema, D., Eren, D., Yayon, A., Goldblum, A., & Zaliani, A. (2004). Identifying the binding mode of a molecular scaffold. *J. Comput. Aided Mol. Des., 18*, 23-40.

ChemAxon. (2005). Reactor (Version 3.1). Budapest, Hungary: ChemAxon Ltd.

Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S., & Zhang, Y. (2005). Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem., 3*, 1832-1834.

Craig, P. N., & Ebert, H. M. (1969). Eleven years of structure retrival using the SK&F fragment codes. *J. Chem. Doc., 9*, 141-146.

Cramer III, R. D., Redl, G., & Berkoff, C. E. (1974). Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem., 17*, 533-535.

Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A., et al. (1992). Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci., 32*, 244-255.

Daylight. (2005). Daylight Theory Manual. Santa Fe, NM, USA: Daylight Chemical Information Systems, Inc.

De Laet, A., Hehenkamp, J. J. J., & Wife, R. L. (2000). Finding drug candidates in virtual and lost/emerging chemistry. *J. Heterocycl. Chem.*, 669-674.

Downs, G. M., & Barnard, J. M. (2002). Clustering methods and their uses in computational chemistry. In *Reviews in Computational Chemistry, Vol 18* (pp. 1-40).

Dubois, J.-E. (2002). *Chemical complexity and molecular topology: The DARC concepts and applications.* In *Proceedings of the 2002 Conference on the History and Heritage of Scientific and Technological Information Systems*, Philadelphia.

Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., et al. (1988). Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *J. Med. Chem., 31*, 2235-2246.

Feher, M., & Schmidt, J. M. (2003). Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci., 43*, 218-227.

Fernández, M.-L., & Valiente, G. (2001). A graph distance metric combining maximum common subgraph and minimum common supergraph. *Pattern Recognition Letters, 22*, 753-758.

Garfield, E. (1972). Are you ready for Chemical Linguistics? Chemical Semantics? Chemical Semiotics? Or, why WLN? *http://www.garfield.library.upenn.edu/essays/V1p386y1962-73.pdf*.

Gillet, V., Willett, P., & Bradshaw, J. (2003). Similarity searching using reduced graphs. *J. Chem. Inf. Comput. Sci., 43*, 338-345.

Gillet, V. J., Willett, P., & Bradshaw, J. (1997). The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *J. Chem. Inf. Comput. Sci., 37*, 731-740.

Gillet, V. J., Willett, P., Bradshaw, J., & Green, D. V. S. (1999). Selecting combinatorial libraries to optimize diversity and physical properties. *J. Chem. Inf. Comput. Sci., 39*, 169-177.

Gillet, V. J., Willett, P., Fleming, P. J., & Green, D. V. S. (2002). Designing focused libraries using MoSELECT. *J. Mol. Graph. Model., 20*, 491-498.

Gkoutos, G. V., Murray-Rust, P., Rzepa, H. S., & Wright, M. (2001). Chemical markup, XML and the World-Wide Web. 3. Toward a signed semantic chemical web of trust. *J. Chem. Inf. Comput. Sci., 41*, 1124-1130.

Gugisch, R., Kerber, A., Laue, R., Meringer, M., & Weidinger, J. (2000). MOLGEN-COMB, a software-package for combinatorial chemistry. *Match, 41*, 189-203.

Hann, M. M., Leach, A. R., & Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci., 41*, 856-864.

Hansen, P. J., & Jurs, P. C. (1988). Chemical applications of graph theory. Part I. Fundamentals and Topological Indices. *J. Chem. Educ., 65*, 574-580.

Hattori, M., Okuno, Y., Goto, S., & Kanehisa, M. (2003). Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc., 125*, 11853-11865.

Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., et al. (2004). Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem., 2*, 3256-3266.

Holliday, J. D., Salim, N., Whittle, M., & Willett, P. (2003). Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci., 43*, 819-828.

Horton, D. A., Bourne, G. T., & Smythe, M. L. (2003). The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev., 103*, 893-930.

Huc, I., & Lehn, J. M. (1997). Virtual combinatorial libraries: dynamic generation of molecular and supramolecular diversity by self-assembly. *Proc. Natl. Acad. Sci. U. S. A., 94*, 2106-2110.

Hyde, E., & Thomson, L. D. (1968). Structure display. *J. Chem. Doc., 8*, 138-146.

Jenkins, J. L., Glick, M., & Davies, J. W. (2004). A 3D similarity method for scaffold hopping from known drugs or natural ligands to new chemotypes. *J. Med. Chem., 47*, 6144-6159.

Kann, V. (2000). Maximum clique. (08/2005), http://www.nada.kth.se/~viggo/wwwcompendium/node33.html

Katritzky, A. R., Kiely, J. S., Hebert, N., & Chassaing, C. (2000). Definition of Templates within Combinatorial Libraries. *J. Comb. Chem., 2*, 2-5.

Kolb, H. C., Finn, M. G., & Sharpless, K. B. (2001). Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angew. Chem. Int. Ed. Engl., 40*, 2004-2021.

Lamb, M. L., Burdick, K. W., Toba, S., Young, M. M., Skillman, A. G., Zou, X., et al. (2001). Design, docking, and evaluation of multiple libraries against multiple targets. *Proteins, 42*, 296-318.

Leach, A. R., Bradshaw, J., Green, D. V., Hann, M. M., & Delany, J. J., 3rd. (1999). Implementation of a system for reagent selection and library enumeration, profiling, and design. *J. Chem. Inf. Comput. Sci., 39*, 1161-1172.

Lehn, J. M., & Eliseev, A. V. (2001). Dynamic combinatorial chemistry. *Science, 291*, 2331-2332.

Leland, B. A., Christie, B. D., Nourse, J. G., Grier, D. L., Carhan, R. E., Maffett, T., et al. (1997). Managing the combinatorial explosion. *J. Chem. Inf. Comput. Sci., 37*, 62-70.

Lewell, X. Q., Jones, A. C., Bruce, C. L., Harper, G., Jones, M. M., McLay, I. M., et al. (2003). Drug rings database with web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem., 46*, 3257-3274.

Lewell, X. Q., Judd, D. B., Watson, S. P., & Hann, M. M. (1998). RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci., 38*, 511-522.

Lewis, W. G., Green, L. G., Grynszpan, F., Radic, Z., Carlier, P. R., Taylor, P., et al. (2002). Click chemistry in situ: acetylcholinesterase as a reaction vessel for the selective assembly of a femtomolar inhibitor from an array of building blocks. *Angew. Chem. Int. Ed. Engl., 41*, 1053-1057.

Lipinski, C. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev., 23*, 3-25.

Lipkus, A. H. (2001). Exploring chemical rings in a simple topological-descriptor space. *J. Chem. Inf. Comput. Sci., 41*, 430-438.

Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci., 40*, 195-209.

Lobanov, V. S., & Agrafiotis, D. K. (2002). Scalable methods for the construction and analysis of virtual combinatorial libraries. *Comb. Chem. HTS, 5*, 167-178.

Lopez-Ortiz, A. (2000). Comp.Theory FAQ. (08/2005), http://db.uwaterloo.ca/~alopez-o/comp-faq/faq.html

Lynch, M. (2002). *Introduction of Computers in Chemical Structure Information Systems.* In *Proceedings of the 2002 Conference on the History and Heritage of Scientific and Technological Information Systems*, Philadelphia.

Markush, E. A. (1924). Pyrazolone Dye and Process of Making the Salt. US Pat. 1,506,316.

Martin, E. J., Blaney, J. M., Siani, M. A., Spellmeyer, D. C., Wong, A. K., & Moos, W. H. (1995). Measuring diversity: experimental design of combinatorial libraries for drug discovery. *J. Med. Chem., 38*, 1431-1436.

Martin, Y. C. (2001). Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem., 3*, 231-250.

McGregor, J. J., & Willett, P. (1981). Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci., 21*, 137-140.

MDL. (2003). CTFile (Version October 2003). San Leandro, CA, USA: MDL Information Systems,Inc.

Monev, V. (2004). Introduction to similarity searching in chemistry. *Match, 51*, 7-38.

Mooers, C. N. (1951). Ciphering structural formulas-the Zatopleg system. *Zator Technical Bulletin*, 59.

Murray-Rust, P., & Rzepa, H. S. (1999). Chemical markup, XML and the World-Wide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci., 39*, 928-942.

Murray-Rust, P., & Rzepa, H. S. (2001). Chemical markup, XML and the World-Wide Web. 2. Information objects and the CMLDOM. *J. Chem. Inf. Comput. Sci., 41*, 1113-1123.

Murray-Rust, P., & Rzepa, H. S. (2003). Chemical markup, XML, and the World Wide Web. 4. CML schema. *J. Chem. Inf. Comput. Sci., 43*, 757-772.

Murray-Rust, P., Rzepa, H. S., Williamson, M. J., & Willighagen, E. L. (2004). Chemical markup, XML, and the World Wide Web. 5. Applications of chemical metadata in RSS aggregators. *J. Chem. Inf. Comput. Sci., 44*, 462-469.

Nilakantan, R., Bauman, N., Haraki, K. S., & Venkataraghavan, R. (1990). A ring-based chemical structural query system: Use of a novel ring-complexity heuristic. *J. Chem. Inf. Comput. Sci., 30*, 65-68.

Nilakantan, R., Bauman, N., & Venkataraghavan, R. (1991). A method for automatic generation of novel chemical structures and its potential applications to drug discovery. *J. Chem. Inf. Comput. Sci., 31*, 527-530.

Nilakantan, R., & Nunn, D. S. (2003). A fresh look at pharmaceutical screening library design. *Drug Discov. Today, 8*, 668-672.

Opatz, T., Kallus, C., Wunberg, T., Kunz, H., Schmidt, W., & Henke, S. (2003). D-glucose as a pentavalent chiral scaffold. *European Journal of Organic Chemistry*, 1527-1536.

OpenEye. (2005). OEChem (Version 1.3). Santa Fe, NM, USA: OpenEye Scientific Software, Inc.

Ortholand, J.-Y., & Ganesan, A. (2004). Natural products and combinatorial chemistry: back to the future. *Curr. Opin. Chem. Biol., 8*, 271-280.

Prasanna, M. D., Vondrasek, J., Wlodawer, A., & Bhat, T. N. (2005). Application of InChI to curate, index, and query 3-D structures. *Proteins, 60*, 1-4.

Ramstroem, O., & Lehn, J. M. (2001). Drug discovery by dynamic combinatorial libraries. *Nat. Rev. Drug. Discov., 1*, 26-36.

Raymond, J. W. (2002). *Applications of graph-based similarity in cheminformatics.* Ph.D. Thesis, University of Sheffield, Sheffield.

Raymond, J. W., Gardiner, E. J., & Willett, P. (2002). RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *Computer Journal, 45*, 631-644.

Rees, D. C., Congreve, M., Murray, C. W., & Carr, R. (2004). Fragment-based lead discovery. *Nat. Rev. Drug. Discov., 3*, 660-672.

Rotstein, S. H., & Murcko, M. A. (1993). GroupBuild: a fragment-based method for de novo drug design. *J. Med. Chem., 36*, 1700-1710.

Sauer, W. H., & Schwarz, M. K. (2003). Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci., 43*, 987-1003.

Schneider, G., & Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug. Discov., 4*, 649-663.

Schuffenhauer, A. (2005). Molecular diversity, complexity and biological activity. (08/2005), http://www.int-conf-chem-structures.org/7th_iccs_pdfs/C-3_Schuffenhauer.pdf

Schuffenhauer, A., Popov, M., Schopfer, U., Acklin, P., Stanek, J., & Jacoby, E. (2004). Molecular diversity management strategies for building and enhancement of diverse and focused lead discovery compound screening collections. *Comb. Chem. HTS, 7*, 771-781.

Schuller, A., Schneider, G., & Byvatov, E. (2003). SMILIB: Rapid assembly of combinatorial libraries in SMILES notation. *Qsar & Combinatorial Science, 22*, 719-721.

Shedden, K., Brumer, J., Chang, Y. T., & Rosania, G. R. (2003). Chemoinformatic analysis of a supertargeted combinatorial library of styryl molecules. *J. Chem. Inf. Comput. Sci., 43*, 2068-2080.

Sheridan, R. P., SanFeliciano, S. G., & Kearsley, S. K. (2000). Designing targeted libraries with genetic algorithms. *J. Mol. Graph. Model., 18*, 320-334.

Siani, M. A., Weininger, D., & Blaney, J. M. (1994). CHUCKLES: a method for representing and searching peptide and peptoid sequences on both monomer and atomic levels. *J. Chem. Inf. Comput. Sci., 34*, 588-593.

Siani, M. A., Weininger, D., James, C. A., & Blaney, J. M. (1995). CHORTLES: a method for representing oligomeric and template-based mixtures. *J. Chem. Inf. Comput. Sci., 35*, 1026-1033.

Snarey, M., Terrett, N. K., Willett, P., & Wilton, D. J. (1997). Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graph. Model., 15*, 372-385.

Solovév, V. P., Varnek, A., & Wipff, G. (2000). Modeling of ion complexation and extraction using substructural molecular fragments. *J. Chem. Inf. Comput. Sci., 40*, 847-858.

Stahl, M., Mauser, H., Tsui, M., & Taylor, N. R. (2005). A robust clustering method for chemical structures. *J. Med. Chem., 48*, 4358-4366.

Stein, S. E., Tchekhovskoi, D., & Heller, S. R. (2005). IUPAC International Chemical Identifier (InChI) (Version 1.0). http://www.iupac.org/inchi/: IUPAC and NIST.

Stockwell, B. R. (2000). Chemical Genetics: Ligand-based discovery of gene function. *Nat. Genet., 1*, 116-125.

Su, A. I., Lorber, D. M., Weston, G. S., Baase, W. A., Matthews, B. W., & Shoichet, B. K. (2001). Docking molecules by families to increase the diversity of hits in database screens: computational strategy and experimental evaluation. *Proteins, 42*, 279-293.

Talete. (2005). Dragon (Version 5). Milano: Talete SRL.

Tempest, P. A., & Armstrong, R. W. (1997). Cyclobutanedione derivatives on solid support: Toward multiple core structure libraries. *J. Am. Chem. Soc., 119*, 7607-7608.

Thanh Le, G., Abbenante, G., Becker, B., Grathwohl, M., Halliday, J., Tometzki, G., et al. (2003). Molecular diversity through sugar scaffolds. *Drug Discov. Today, 8*, 701-709.

Tripos. (2005). Distill. St.Louis, MO, USA: Tripos Inc.

Vleduts, G. E. (1977). *Development of a combined WLN/CTR multilevel approach to the algorithmical analysis of chemical reactions in view of their automatic indexing* (No. 5399). London: British Library Research and Development Department Report.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci., 28*, 31-36.

Weininger, D., Weininger, A., & Weininger, J. L. (1989). SMILES: 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci., 29*, 97-101.

Wheland, G. W. (1949). *Advanced organic chemistry*. New York: John Wiley.

Willett, P. (2000). Chemoinformatics - similarity and diversity in chemical libraries. *Curr. Opin. Biotechnol., 11*, 85-88.

Wiswesser, W. J. (1982). How the WLN began in 1949 and how it might be in 1999. *J. Chem. Inf. Comput. Sci., 22*, 88-93.

Wiswesser, W. J. (1985). Historic development of chemical notations. *J. Chem. Inf. Comput. Sci., 25*, 285-263.

Wolber, G., & Langer, T. (2000). CombiGen: A novel software package for the rapid generation of virtual combinatorial libraries. In H.-D. Höltje & W. Sippl (Eds.), *Rational Approaches to Drug Design* (pp. 390-399): Prous Science.

Xu, J. (2002). A new approach to finding natural chemical structure classes. *J. Med. Chem., 45*, 5311-5320.

Xu, Y., & Johnson, M. (2001). Algorithm for naming molecular equivalence classes represented by labeled pseudographs. *J. Chem. Inf. Comput. Sci., 41*, 181-185.

Xu, Y. J., & Johnson, M. (2002). Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci., 42*, 912-926.

Xue, L., & Bajorath, J. (1999). Distribution of molecular scaffolds and R-groups isolated from large compound databases. *J. Mol. Model., 5*, 97-102.

Xue, L., Stahura, F. L., & Bajorath, J. (2004). Cell-based partitioning. *Methods Mol. Biol., 275*, 279-290.

Yasri, A., Berthelot, D., Gijsen, H., Thielemans, T., Marichal, P., Engels, M., et al. (2004). REALISIS: a medicinal chemistry-oriented reagent selection, library design, and profiling platform. *J. Chem. Inf. Comput. Sci., 44*, 2199-2206.

Zauhar, R. J., Moyna, G., Tian, L. F., Li, Z. J., & Welsh, W. J. (2003). Shape signatures: A new approach to computer-assisted ligand- and receptor-based drug design. *J. Med. Chem., 46*, 5674-5690.

Zheng, S., Luo, X., Chen, G., Zhu, W., Shen, J., Chen, K., et al. (2005). A new rapid and effective chemistry space filter in recognizing a druglike database. *J. Chem. Inf. Model., 45*, 856-862.

Zheng, W., Cho, S. J., Waller, C. L., & Tropsha, A. (1999). Rational combinatorial library design. 3. Simulated annealing guided evaluation (SAGE) of molecular diversity: a novel computational tool for universal library design and database mining. *J. Chem. Inf. Comput. Sci., 39*, 738-746.

Chapter 3

3. Assessing the Scaffold Diversity of Screening Libraries*

*submitted to Journal of Chemical Information and Modeling for special ICCS issue; Part of this paper was the subject of a poster presentation at the 7th International Conference on Chemical Structures, 5-9 June, 2005, Noordwijkerhout, Netherlands.

3.1 Abstract

High-throughput screening nowadays requires compound libraries in which the maximal chemical diversity is reached with the minimal number of molecules. Medicinal chemists have traditionally realized assessment of diversity and subsequent compound acquisition although a recent study suggests that experts are usually inconsistent in reviewing large datasets. In order to analyze the chemical diversity of commercially available screening collections, we have developed a general workflow aimed at (1) identifying drug-like compounds, (2) cluster them by maximum common substructures (scaffolds), (3) measure the scaffold diversity encoded by each screening collection independently of its size, and finally (4) merge all common substructures in a non-redundant scaffold library that can easily be browsed by structural and topological queries. Starting from 2.4 million compounds out of 12 commercial sources, four categories of libraries could be identified: large and medium sized combinatorial libraries (low scaffold diversity), screening libraries (medium diversity, medium size) and diverse libraries (high diversity, low size). The chemical space covered by the scaffold library can be searched to prioritize scaffold-focused libraries.

3.2 Introduction

With the advent of initiatives like the CNRS “National Chemical Library”("National Chemical Library, National Center for Scientific Research (CNRS)") or the NIH Molecular Libraries Initiative (Austin *et al.*, 2004), public research rejoined the pharmaceutical industry in its effort to organize and to curate small molecular-weight molecules for the purpose of drug discovery and target de-orphanization. The drastic and steady increase of commercially available compounds beginning in the early 1990s (Webb, 2005), provided chemical information scientists the opportunity to enhance the diversity of their proprietary compound collection. The main challenge that remained over the years which is regularly revisited (Bocker *et al.*, 2005) is the way of measuring the diversity of a compound library. Very informative testimonials about this key aim were shared with the community elsewhere (Martin, 2001). Nevertheless, molecular diversity is heavily depending on descriptors, metrics and multivariate methods used to assess it. Most studies on commercially available compound libraries (Baurin, Baker *et al.*, 2004; Cummins *et al.*, 1996; Shemetulskis *et al.*, 1995) have traditionally used physicochemical and topological descriptors, summed up into a score (Sirois *et al.*, 2005) or encoded into fingerprints (McGregor & Pallai, 1997) or hash codes (Nilakantan *et al.*, 1997; Voigt *et al.*, 2001) to evaluate the uniqueness and diversity of such libraries. Although fingerprints can be quickly computed for large collections of compounds, it results in classifications of molecular libraries that are not very intuitive for medicinal chemists because a single class of compounds may contain quite different molecular scaffolds accessible by very different synthetic routes. Traditionally, medicinal chemists mining high-throughput screening (HTS) data have organized hits into homogeneous chemical series. Why not use the same partitioning method before the virtual or real screening process? Archiving compounds by scaffolds is much more

natural but computationally more demanding if calculated ad hoc. Moreover, various definitions of a scaffold (Table 2.3) are possible such as maximum common substructure (McGregor & Willett, 1981), largest rigid fragment (Su *et al.*, 2001), molecular frameworks or graph frameworks (Bemis & Murcko, 1996) with or without descriptors (e.g. topological torsions) (Nilakantan *et al.*, 1987) and fragments as generated by the RECAP method (Lewell *et al.*, 1998).

During a medicinal chemistry project, it is not uncommon that structural parts of the scaffold are redefined by either extension or reduction. If the limit of one extreme is reached by setting the full compound equal to the scaffold, how far can one reduce the compound structure to obtain a chemically meaningful scaffold?

In the present study, we classified 17 commercially-available screening collections according to graph-based maximum common substructures("ClassPharmer Suite") and joined the resulting classification into a single library of non-redundant classes. A new metric (PC50C) is proposed to assess the diversity of a screening collection, by computing the percentage of classes accounting for 50% of classified compounds. Since this metric is independent of the size of a library, it can be used to compare collections of different sizes.

3.3 Methods

The overall workflow for reading, processing and extracting molecular scaffolds out of commercial libraries is illustrated in Figure 3.1 and further detailed in the following paragraphs.

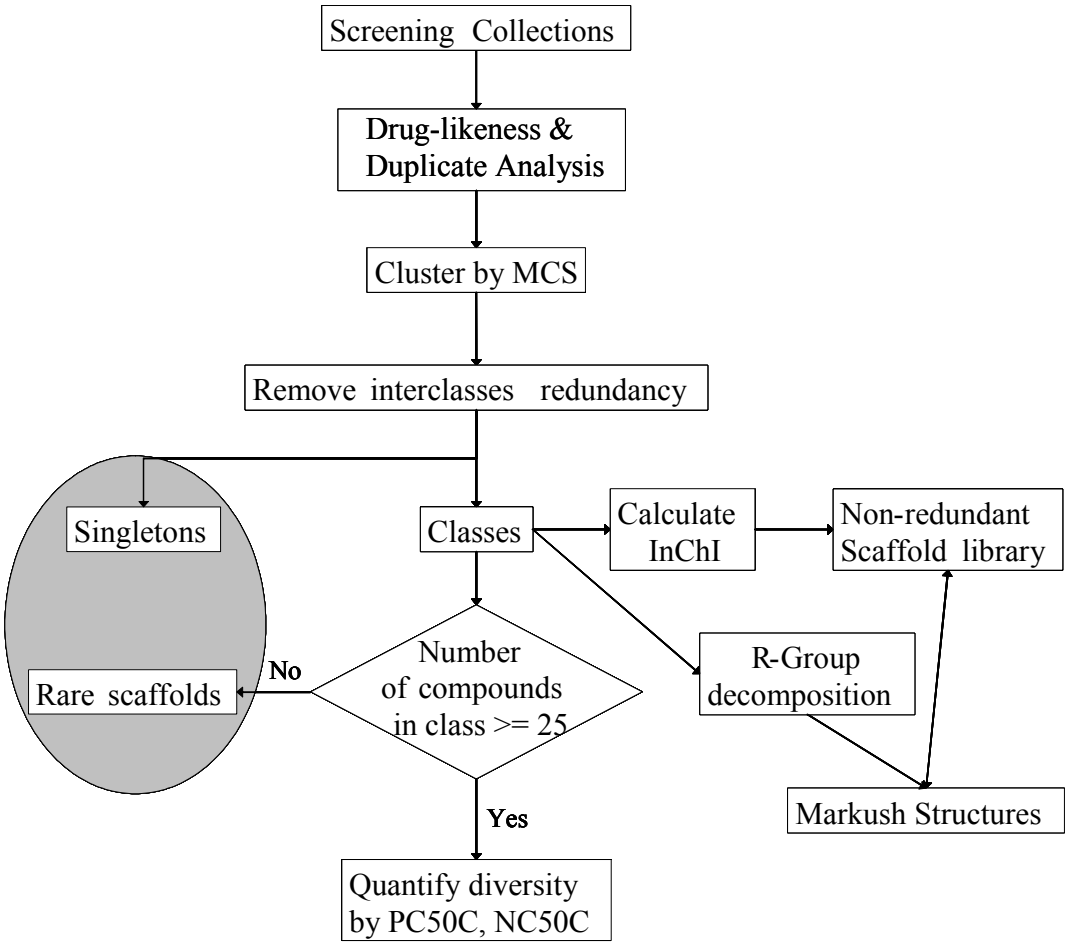


Figure 3.1 General workflow for processing screening libraries

3.3.1 Database Processing

The screening collections used in this study date from the last quarter 2003 except the MDDR for which the first 2004 release has been used. A total of 17 libraries from 12 suppliers (Table 3.1) plus the MDDR describe the commercially-available chemical space addressed by the current report.

Supplier	Collection	Code
Asinex	Gold	ASIg
	Platinum	ASIp
ChemBridge	EXPRESS-Pick + H2LS	CBG
Chemical Diversity	CombiLab	CDIc
	International Diversity	CDIi
CNRS	National Chemical Library	CNR
ChemStar		CST
InterBioScreen	Natural	IBSn
	Synthetic	IBSs
Maybridge		MAY
Bionet		NET
Specs		SPE
Timtec	Natural	TIMn
	Synthetic	TIMs
Tripos		TRI
Vitas-M	Stock	VITs
	Tulip	VITt
MDDR	2004.1	MDDR

Table 3.1 Screening collections used in the study

It covers 2 410 857 compounds easily available as powders in vials. The collections need also to have a computer-readable counterpart, delivered as SDFile on a CD-ROM or downloadable from the supplier webpage (Sirois *et al.*, 2005). The very first processing steps consisted in standardizing the structure and data headers of SD files using an in-house Perl script. Property or functional group-based filtering rules (Charifson & Walters, 2002) implemented in OpenEye's Filter ("Filter 1.0") program,

were then used to select the most suitable compounds for each library (see filtering rules in Supporting Information). In this step, counter-ions were removed and the ionization state of each compound at physiological pH was assigned. For each collection, an additional step consisted of eliminating remaining redundant compounds taking stereochemical information into account using the CLIFF program ("Cliff") (please note that CLIFF has recently been split into several separated routines).

3.3.2 Compound Classification

One of the major challenges was to obtain an organization of the screening collections into chemically meaningful classes. ClassPharmer™ Suite's proprietary clustering methodology ("ClassPharmer Suite") was adopted. In order to make a clear distinction with the common association of clustering with fingerprints in chemoinformatics, the grammatical root "class" (classes/classification) is preferred over "cluster" (clusters/clustering). But in strictly algorithmic terms, the method used herein happens to be a clustering algorithm and not a classification where one starts from predefined scaffolds (Roberts *et al.*, 2000).

Two parameters mainly influence the outcome of the classification: the homogeneity and the redundancy level. Homogeneity is related to the size (heavy atom count) of scaffold divided by the size of largest compound in the class. Redundancy describes to which extent a compound is allowed to appear in multiple classes. Hence the classes are represented by a scaffold assimilated to the maximum common substructure (MCS). The underlying algorithm is covered by trade secret but can however be approximately described as follows: given a dataset of N compounds, (i) find topologically aware (approximated) MCS for all pairs, triplets, quadruplets, …, N-1 groups of compounds; (ii) eliminate MCS that do not fulfill the user-defined homogeneity level; (iii) select the smallest number of MCS that fulfils the user-defined redundancy level while giving the

minimal number of singletons and, if the option is selected, (iv) generate subclasses with larger (exact) MCS where subsets of a class with higher homogeneity can be found. The implementation of the algorithm is preceded by a normalization process of the input structures. For the present classification, the homogeneity and redundancy were set to medium and low, respectively. Exact ring closure and exact atom match parameters were chosen to define classes. No subclasses were computed.

3.3.3 Scaffold Distribution

The non-hierarchical disjunctive algorithm which was used, allows that a compound belongs to more than one class. In order to compare the scaffold distribution of different libraries, the inter-classes redundancy of a compound was removed using a Python script based on OpenEye's OEChem1.3 library.("OEChem") This task was achieved by computing the Central Scaffold Score (CSS) of each compound/class pair and assigning the compound to the class presenting the lowest CSS, calculated by the following equation: $CSS = \frac{MW_{compound} - MW_{scaffold}}{N_R}$

where $MW_{compound}$ is the molecular weight of a compound, $MW_{scaffold}$ the molecular weight of the scaffold, and N_R the number of substitution points (R-groups).

For every screening collection, the classes were ordered by decreasing population and two metrics (NC50C, PC50C) computed. NC50C describes the number of non-redundant classes describing 50% of classified compounds. PC50C features the percentage of classes covering 50% of classified compounds. Two classifications were analyzed. In the first one, all classes of at least two unique compounds were investigated. In the second one, a threshold of 25 was assigned to the minimal size of a

class (number of unique compounds). Classes populated by less than 25 unique compounds will be referred to as 'rare scaffolds' (Figure 3.1).

3.3.4 R-group decomposition

The topology around the scaffold and generation the corresponding of SMILES strings (Weininger, 1988) were obtained by R-group decomposition. A compound subset and its corresponding scaffold were the input for finding the minimum common supergraph (Willett, 1985) under the form of a Markush structure. For each compound/scaffold pair, the substitution points were determined and a scaffold with R-groups was generated. Among this R-group labeled scaffold, isomorphs were eliminated and pair-wise substructure relationship was checked. This reduced considerably the number structures to compare. The remaining N Markush structures were then used to find the minimum common super-Markush-structure. The processing is similar to the one described by Brown and al. which identifies the hyperstructure (Brown *et al.*, 1992) and is outlined as follows:

```
SuperStructure := MarkushStructure(1)

FOR n :=2 to N DO

        BEGIN

        COMPARE(SuperStructure, MarkushStructure(n))

        UPDATE_CT()

END
```

As for the removal of inter-classes redundancy, the R-group decomposition was implemented in Python based on OpenEye's OEChem ("OEChem").

3.3.5 Setting-up a scaffold library

All classes (excluding the singletons) were assembled from the generated classifications to form a scaffold library. Computing InChI ("InChI (IUPAC International Chemical Identifier)", 2005) representations ('Mobile H Perception' option ON) for all scaffolds gave the possibility to identify tautomeric forms and group them together. All structural data were deposited in a relational database (MySQL 4.1; for database structure, see Supporting Information Scheme A). Each scaffold was annotated with molecular properties (AlogP, PSA, hydrogen bond donor and acceptor count, rotatable bond and ring systems count) and the Markush structure SMILES. The main scaffold structure table can be browsed and queried by similarity, substructure or superstructure using JChemBase ("JChemBase").

3.4 Results and discussion

3.4.1 Processing the libraries

In a first step, 17 commercially-available screening collections were processed to retain unique drug-like molecules. In addition, a prototypical collection of drug-like compounds (MDDR) was taken as reference to delimit true drug-like chemistry space. In agreement with previous reports (Baurin, Baker *et al.*, 2004; Charifson & Walters, 2002), the percentage of drug-like molecules varies from ca. 30% (ChemStar) to 60% (Asinex Platinum) (see Table 3.2). No relationships could be established between size and drug-likeness of the libraries. It should be noted that a set of very strict rules (see Supporting Information Chart A) especially regarding molecular weight (250 <MW <500) and Lipinski's rule of five violations (none) was used herein.

Code	Initial size	filtered[a]	% druglike	Unique[b]	Exclusive[c]	Classified[d]
ASIg	201 304	86 185	42.8	86 153	17 322	85 516
ASIp	120 563	71 255	59.1	71 255	69 716	70 978
CBG	709 975	327 716	49.5	181 291	72 484	161 827
CDIc	230 529	104 606	47.8	104 604	62 361	104 520
CDIi	133 085	39 859	45.9	39 831	13 571	39 401
CNR	12 670	4 978	39.3	4 806	4 571	4 770
CST	73 552	21 899	29.8	21 852	4 857	21 758
IBSn	30 749	14 196	46.2	13 936	890	13 882
IBSs	287 945	112 882	43.2	112 695	61 944	111 562
MAY	59 204	20 754	35.1	20 726	17 793	20 680
NET	38 416	14 031	36.5	14 029	13 276	13 992
SPE	172 970	65 563	37.9	65 539	20 499	65 319
TIMn	4 202	2 083	49.6	1 945	147	1 941
TIMs	95 469	33 669	35.3	33 560	7 873	33 408
TRI	84 604	46 866	55.4	46 546	44 969	46 543
VITs	134 167	52 583	39.2	52 544	8 796	52 204
VITt	21 453	7 190	33.5	7 182	3 778	7 164
MDDR	98 880	37 857	38.3	35 563	35 142	35 033

[a] Using Filter 1.0 ("Filter 1.0")

[b] Using Cliff ("Cliff") with options "-unique 1 -usestereo 1"

[c] compounds not found elsewhere by comparison of canonical SMILES with PipelinePilot 4.5. ("Pipeline Pilot")

[d] after normalization step in ClassPharmer ("ClassPharmer Suite").

Table 3.2 Library processing and classification

Considering the MDDR as a reference drug-like dataset, we can thus consider most of the screening collections investigated here to be drug-like, reflecting the effort of vendors to produce higher quality collections (Webb, 2005). Internal duplicates (compounds present several times within the same collection) ranged from none (ASIp) to 320 compounds (TRI). An exceptional high number (146 425) was found for CBG, but could be explained by the previous merge of the screening collections (EXPRESS-Pick and Hit2Lead) into a single dataset. Retrospectively, only 2 compounds would have been duplicated in CBG Express-Pick. For the MDDR, there were still 2 294 duplicates left, most of them arising from different counter ions.

An exclusivity analysis of all screening collections shows that only five of them (ASIp, CNR, MAY, NET, TRI) could be described as original as they contain more than 85% drug-like compounds not present elsewhere (Table 3.2). Significant pair-wise overlap exists between several libraries (e.g. ASIg, CBG, IBSs, CDIc, VITs; see Tables A and B in Supporting Information). However, having several commercial sources for a compound may be an advantage since it still guarantees a purchase even if the corresponding molecule is no longer available from a particular supplier.

3.4.2 What is the scaffold diversity of commercial libraries?

A first scaffold classification (Classification 1, Table 3.3) has been realized on the global set of 846 408 molecules passing the ClassPharmer normalization step. A second one (Classification 2) is a subset of the first one since it accounts for classes populated by at least 25 unique molecules. The second classification was undertaken to depict the optimization potential of each class.

	Classification 1[a]					Classification 2[b]		
Code	#Classes	#Singl[c]	%Red[d]	NC50C[e]	PC50C[f]	#Classes	NC50C	PC50C
ASIg	3 491	5 476	7.25	52	1.49	400	27	6.75
ASIp	1 968	2 907	9.27	27	1.37	252	19	7.54
CBG	3 199	5 269	15.79	45	1.41	709	32	4.51
CDIc	3 430	5 171	6.29	86	2.51	528	57	10.80
CDIi	2 306	3 447	7.99	62	2.69	219	27	12.33
CNR	391	662	2.74	26	6.65	33	7	21.21
CST	1 011	1 719	9.19	25	2.47	123	13	10.57
IBSn	757	1 188	2.02	20	2.64	75	8	10.67
IBSs	3 490	5 370	5.25	68	1.95	492	48	9.76
MAY	1 544	2 501	12.59	84	5.44	151	30	19.87
NET	941	1 230	5.72	58	6.16	107	21	19.63
SPE	3 261	4 971	8.11	59	1.81	313	27	8.63
TIMn	162	316	1.29	12	7.41	14	5	35.71
TIMs	1 956	3 445	7.23	67	3.43	207	28	13.53
TRI	1 341	2 041	11.55	33	2.46	282	22	7.80
VITs	2 153	3 134	8.85	35	1.63	237	20	8.44
VITt	402	513	6.59	16	3.98	48	9	18.75
MDDR	3 058	4 620	8.51	177	5.79	203	35	17.24

[a] Class defined as containing at least 2 unique compounds

[b] Class defined as containing at least 25 unique compounds

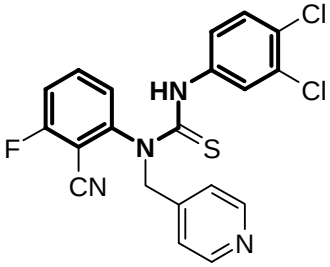
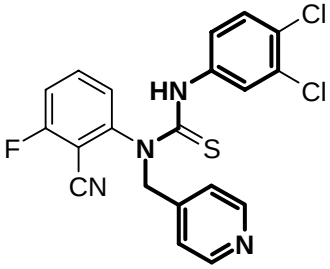
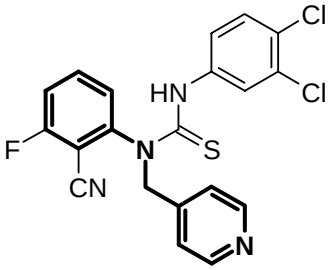
[c] Number of singletons

[d] Percentage of inter-classes redundancy (percentage of compounds present in multiple classes)
[e] Number of classes accounting for 50% of classified compounds
[f] Percentage of classes accounting for 50% of classified compounds

Table 3.3 Classification results

Hence, a class described by a low number of compounds might be of lower interest for a medicinal chemist due to a possible lack of synthetic tractability or insufficient statistics if the library has to be assayed experimentally. On account of a possible overemphasis on combinatorial libraries, the minimal size was set to a lower value than that of 100 compounds advocated by Nilakantan for ring-scaffold focused libraries (Nilakantan & Nunn, 2003).

Using our classification method (see Methods), there are generally 10-30 times less classes (scaffolds) than molecules (Table 3.3). Classification 1 afforded a total of 34 961 classes and 53 980 singletons. Interestingly, the number of singletons always exceeds that of classes for all libraries. Considering the homogeneity of the input libraries which was set to medium prior to the classification, most singletons do not describe unique scaffolds but rather compounds which failed to pass the homogeneity threshold level (i.e. the number of heavy atoms in the scaffold is too small in comparison to the overall size of the largest molecule in the class). Classification 2 (only classes populated by at least 25 compounds) led to a smaller set of 4 390 classes. Since a single compound may be classified in different classes for a single library, there is a significant level of redundancy across the classes generated by ClassPharmer (about 10% on average, Table 3.3) which biases relationships between the number of classes and the number of compounds within a library. To get unbiased relationships and a clearer comparison of the scaffold diversity of input libraries, the redundancy was removed by a simple strategy aimed at selecting the most central scaffold for a compound appearing in multiple classes (Table 3.4). It is important to point out that class redundancy among different libraries has not been considered at this stage.

Compound[a]	MW(scaffold)	N_R[b]	CSS[c]
	198	6	33
	218	4	55
	184	3	61

[a] The compound exemplified here (CD 05668, Maybridge) has a molecular weight of 413 and three proposed scaffolds highlighted in bold.

[b] NR: Number of R-groups

[c] **CSS (Central Scaffold Score) =** $\frac{MW_{compound} - MW_{scaffold}}{N_R}$

Table 3.4 Example of inter-classes redundancy

Two metrics (NC50C, PC50C) have been developed to measure and compare the scaffold diversity of screening collections. The first one (NC50C) is a simple measure of the number of classes accounting for 50% of classified compound for a particular collection. The NC50C descriptor has been derived from a first plot (Figure 3.2) describing the density (percentage of classified compounds) of each class which was then transformed into a cumulative plot (Figure 3.3) allowing to interpolate the number of classes required to describe 50% of classified compounds.

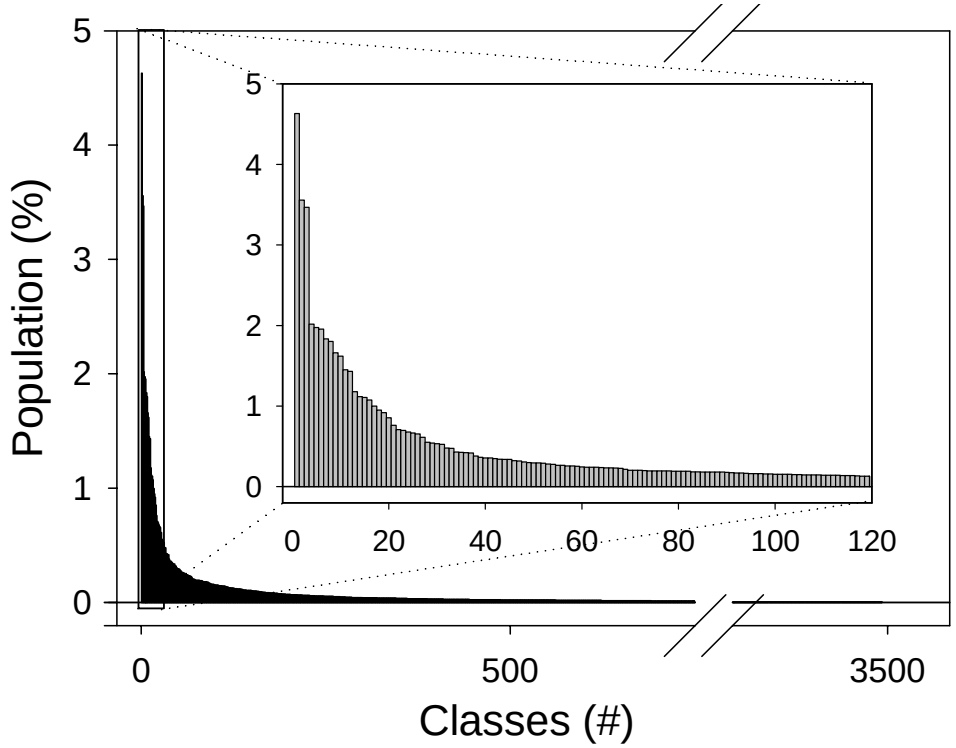


Figure 3.2 Density of ClassPharmer classes (ASIg collection) featuring the percentage of classified compounds for all classes. A zoom on the most populated classes is boxed within the graph.

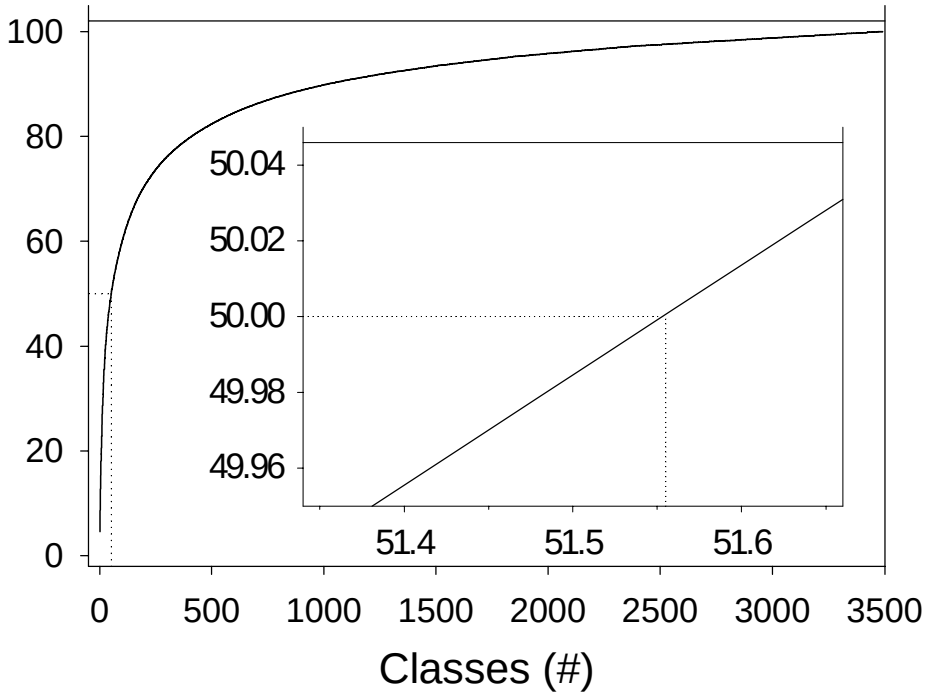


Figure 3.3 Interpolating the NC50C value by plotting the number of classes versus the cumulative percentage of classified compounds (ASIg collection). A zoom around the NC50C value is boxed within the graph.

The NC50C descriptor can be regarded as the absolute scaffold diversity of the collection. As expected, larger collections have higher NC50C values (Figure 3.4 A), except for four collections which either present a quite large panel of classes with respect to their size (MAY, NET and especially MDDR) or a low number of classes (CBG). Discarding these four libraries, a significant correlation could be found between size (number of classified drug-like compounds) and NC50C ($r = 0.70$, $n=14$, $p=0.002$). Compared to the absolute scaffold diversity for classes containing at least 25 compounds (figure 3.4 B), all collections shift to lower NC50C values with the reference MDDR (Table 3.3) performing the most notable shift to the left, thus joining commercially-available collections. For classification 2, a significant correlation is also observed between size and NC50C for all collections ($r = 0.75$, $n=16$, $p =0.002$).

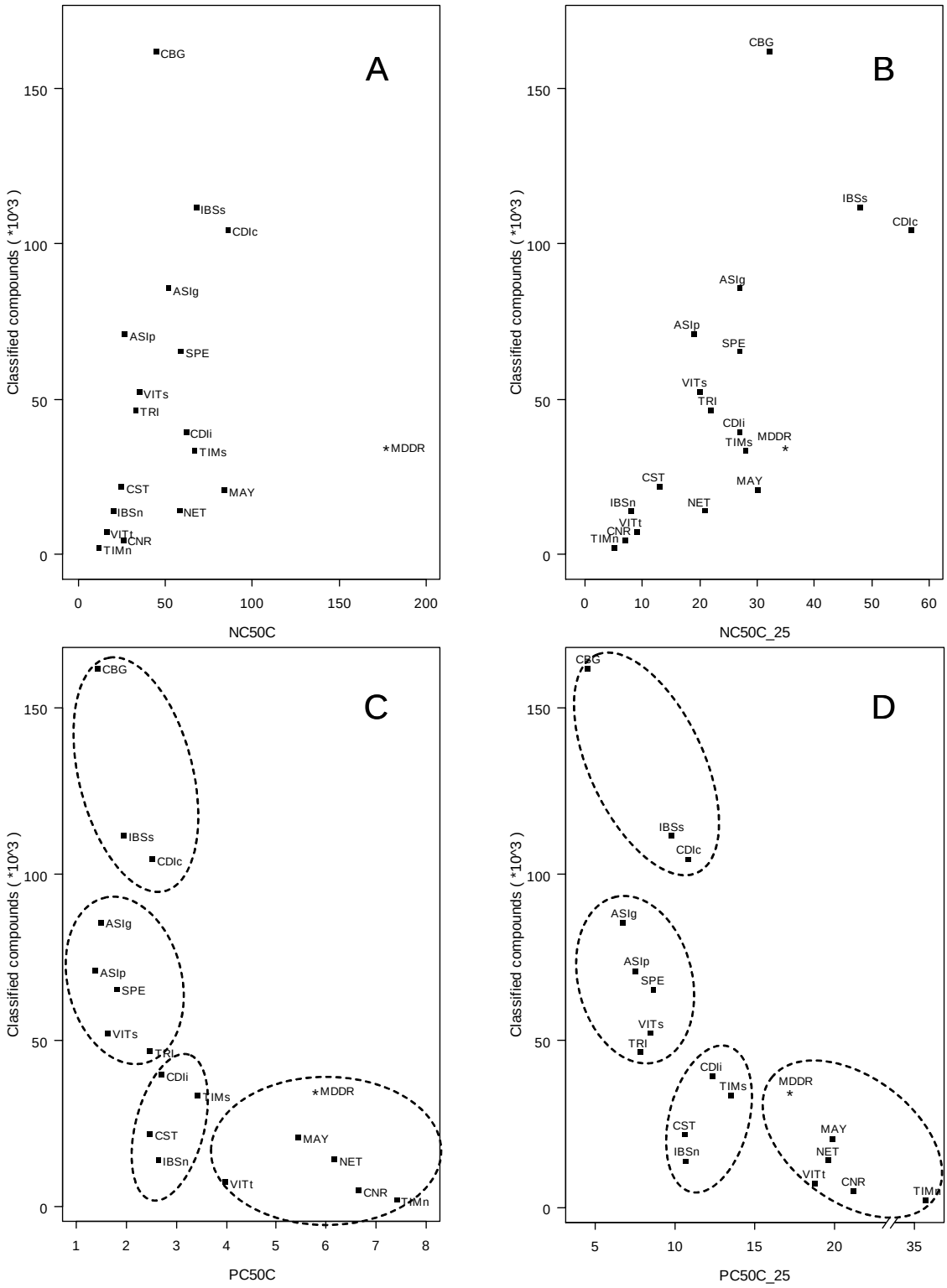


Figure 3.4 Scaffold diversity of screening collections

Since the first metric is dependent on the size of each collection, it cannot be used to compare the intrinsic scaffold diversity. We therefore computed a second descriptor (PC50C) estimating the percentage of classes accounting for 50% of classified

compounds (Table 3.3). It presents the advantage to be independent of the size of the library and therefore more suitable for a comparative analysis (Figure 3.4 C). Strikingly, plotting the PC50C versus the size of each collection allows segregating the herein analyzed 18 collections into four categories (Figure 3.4 C). A first category (CBG, IBSs, CDIc), in agreement with a previous report (Xue & Bajorath, 1999), regroups large combinatorial libraries for which a very tiny percentage of scaffold (less than 3 %) have been overrepresented.

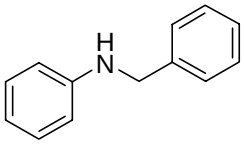
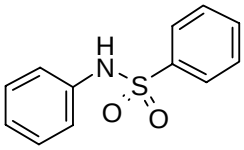
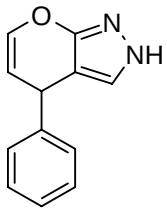
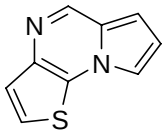
Category	Scaffold	Identifier	Suppliers	Uniques compounds
Large combinatorial Libraries		SBI_5287	15	22 988
Medium combinatorial Libraries		SBI_3167	10	8 592
Screening Libraries		SBI_2894	1	322
Diverse Libraries		SBI_22704	1	106

Table 3.5 Example of characteristic scaffolds for the four categories of screening collections

Corresponding scaffolds are usually very simple (e.g. N-benzylaniline, quinoline, Table 3.5), account for over 10,000 unique compounds and are available at a majority of suppliers. Typical scaffolds from the first category of libraries (e.g. N-phenylbenzenesulfonamide, Table 3.5) are also found in the second category which also regroups combinatorial libraries but of lower size (ASIg, ASIp, SPE, TRI, VITs).

In a third group are found libraries of smaller size (<50,000 drug-like unique compounds) with more original and less populated scaffolds (CDIi, CST, IBSn, TIMs). Last, a fourth category of libraries (MAY, NET, TIMn, VITt) was identified nearby the reference MDDR dataset (Figure 3.4 C). The latter libraries are really diverse in terms of scaffold architecture and generally present a larger choice of proprietary low-populated scaffolds. These libraries are either collections of compounds from various origins (CNR, MDDR), from natural sources (TIMn, VITt) or have been synthesized by the supplier itself with the purpose of optimizing diversity versus size (NET, MAY). For example, the French National Chemical Library (CNR) ("National Chemical Library, National Center for Scientific Research (CNRS)") is a repository of compounds collected at 22 academic laboratories, each of them with a different medicinal chemistry history. Likewise, collections labeled “natural products” (TIMn, VITt) are in fact synthetic compound libraries that are based on structures found in nature (Feher & Schmidt, 2003). Acknowledging the high scaffold diversity found in natural products, it is therefore logical to group them into the fourth category of diverse libraries. Interestingly, looking at the scaffold diversity of the same libraries considering only those scaffolds populated by at least 25 compounds leads to identical clusters with a simple shift of PC50C towards higher values (Figure 3.4 D). Simple rules based on the size (number of classified drug-like and unique compounds) and on PC50C values (all classes, classes with more than 25 compounds) of 18 collections are provided (Table 3.6) as a guide to classify libraries not investigated herein.

Category	Libraries[a]	Size[b]	PC50C[c]	PC50C_25[d]
Large combinatorial Libraries	CBG, CDIc, IbSs	>100 K	< 3	<13
Medium combinatorial Libraries	ASIg, ASIp, SPE, VITs	50-100 K	< 3	< 13
Screening Libraries	CDIi, CST, IBSn, TIMs	<50K	< 4	10-15
Diverse Libraries	CNR, MAY, MDDR, NET, TIMn, VITt	< 50K	> 4	> 15

[a] Libraries are indexed as shown in Table 1

[b] Number of drug-like and unique compounds passing the ClassPharmer normalization step

[c] PC50C value derived from all classes of a library

[d] PC50C value derived from classes populated by at least 25 representatives.

Table 3.6 Classification of collections according to their size and relative scaffold diversity (PC50C)

3.4.3 Setting-up a library (SBI) of non-redundant classes

In order to set-up a single dataset for registering all commercially-available scaffolds, all classes (except those arising from the reference MDDR database) depicted by the previous analysis were merged into a single library. Redundancy of the scaffolds was removed by working with InChI codes which enable the detection of duplicates and of tautomers. The resulting SBI collection contains a total of 21 393 unique classes out of which a surprisingly high number (16 583) are exclusively found at one supplier (Tables 3.7 and 3.8).

Library	Classes[a]	Exclusive Classes[b]
ASIg	3 485	1 240 (36%)
ASIp	1 964	1 729 (88%)
CBG	3 194	1 213 (38%)
CDIc	3 425	2 325 (68%)
CDIi	2 306	974 (42%)
CNR	391	299 (76%)
CST	1 010	307 (30%)
IBSn	756	504 (67%)
IBSs	3 484	1 845 (57%)
MAY	1 543	1 052 (68%)
NET	941	722 (77%)
SPE	3 260	1 504 (46%)
TIMn	162	48 (30%)
TIMs	1 954	700 (36%)
TRI	1 338	1 098 (82%)
VITs	2 149	759 (35%)
VITt	402	264 (66%)

[a] non-redundant classes by comparison of INChI codes (Mobile H Perception' option on). For duplicate classes, a single copy has been conserved corresponding to the first encountered library sorted by alphabetical order.

[b] classes not found elsewhere (by comparison of INChI codes)

Table 3.7 Distribution of classes for the SBI scaffold library

	InChI		InChI & at least 25 compds.	
# of DBs n	# of scaffolds in at least n DBs	# of scaffolds in exactly n DBs	# of scaffolds in at least n DBs	# of scaffolds in exactly n DBs
1	21 393	16 583	2 498	921
2	4 810	2 532	1 577	431
3	2 278	1 009	1 146	106
4	1 269	501	853	251
5	768	300	602	194
6	468	179	408	133
7	289	97	275	84
8	192	71	191	70
9	121	39	121	39
10	82	25	82	25
11	57	20	57	20
12	37	19	37	19
13	18	3	18	3
14	15	6	15	6
15	9	7	9	7
16	2	1	2	1
17	1	1	1	1

Table 3.8 Number of scaffolds which are at least/exactly in n screening collections (DBs)

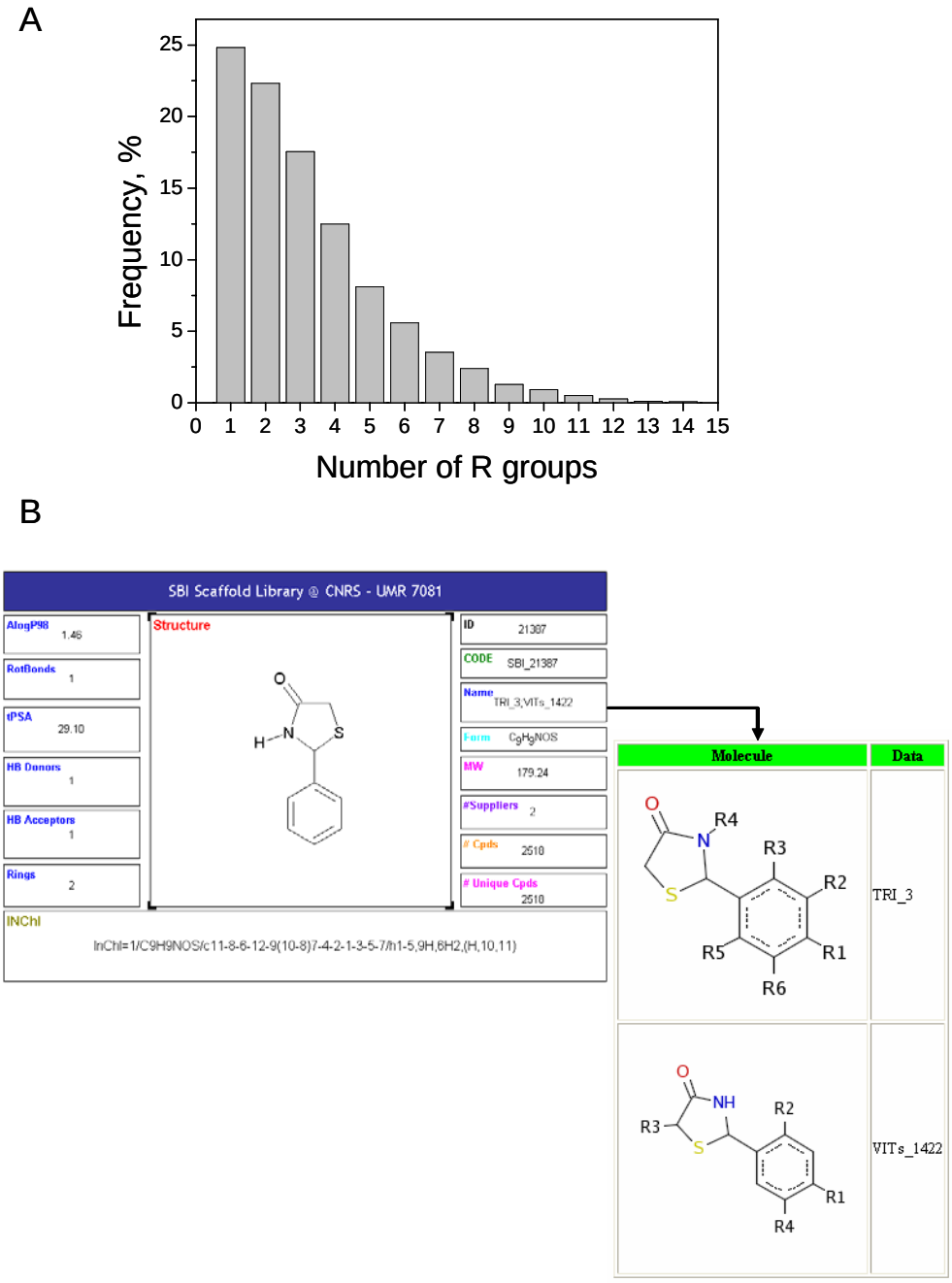


Figure 3.5 The SBI scaffold library. A) Distribution of the number of R-groups for each scaffold, B) Browsing the library. For each scaffold, molecular descriptors (AlogP98, Number of rotatable bonds, topological polar surface area, number of H-bond donors and acceptors, number of rings, molecular weight, number of unique compounds), vendor information (identity and number of suppliers) and a unique SBI code enables an easy navigation in the chemistry space covered by commercial scaffolds. Selecting a particular scaffold (e.g. 2-phenylthiazolidin-4-one) returns the corresponding classes indexed by commercial sources (TRI_3, VITs_1422; see a list of index in Table 1) and the related Markush structures

Compounds contained in the classification represent 811 375 compounds out of which 556 107 have a unique InChI representation. A more restrictive dataset of 2 498 classes comprises scaffolds with a density of at least 25 compounds. Of these, 921 classes have only one supplier as source. 329 (1.5%) scaffolds have been discarded when the compounds contained in a class are checked for uniqueness by InChI. An R-group decomposition of all classes into Markush structures indicates a distribution of substituents following a mono-exponential decay (Figure 3.5 A). 75% of the stored scaffolds offer at least two substituents and thus real diversity. The scaffold library can be easily browsed by substructure, physicochemical properties or suppliers of the corresponding compounds (Figure 3.5 B). A unique code for each scaffold refers to the individual suppliers and the corresponding Markush structures thereby enabling the comparison of commercial sources for a particular scaffold (Figure 3.5 B).

A molecular complexity of the SBI scaffold library was investigated as described by Selzer et al. (Selzer *et al.*, 2005) by computing circular FCFP_4 fingerprints and extracting FCFP_4 sizes and densities (Figure 3.6). FCP4_density calculated for all scaffolds of the SBI library indicate that a large majority of scaffolds are complex enough (FCFP_4 density > 1) to ensure biological activity. A putative application of the SBI library could then be the selection of low molecular-weight fragments for NMR screening (Baurin, Aboul-Ela *et al.*, 2004; Schuffenhauer *et al.*, 2005; Zartler & Shapiro, 2005). Due to their small size, the scaffolds selected herein present a relatively high average self-similarity (average Tanimoto coefficient of 0.74 using FCFP_4 fingerprints). Customizing a fragment library out of the SBI dataset would therefore require the selection of the "least-substituted" compounds for a subset of dissimilar molecular scaffolds.

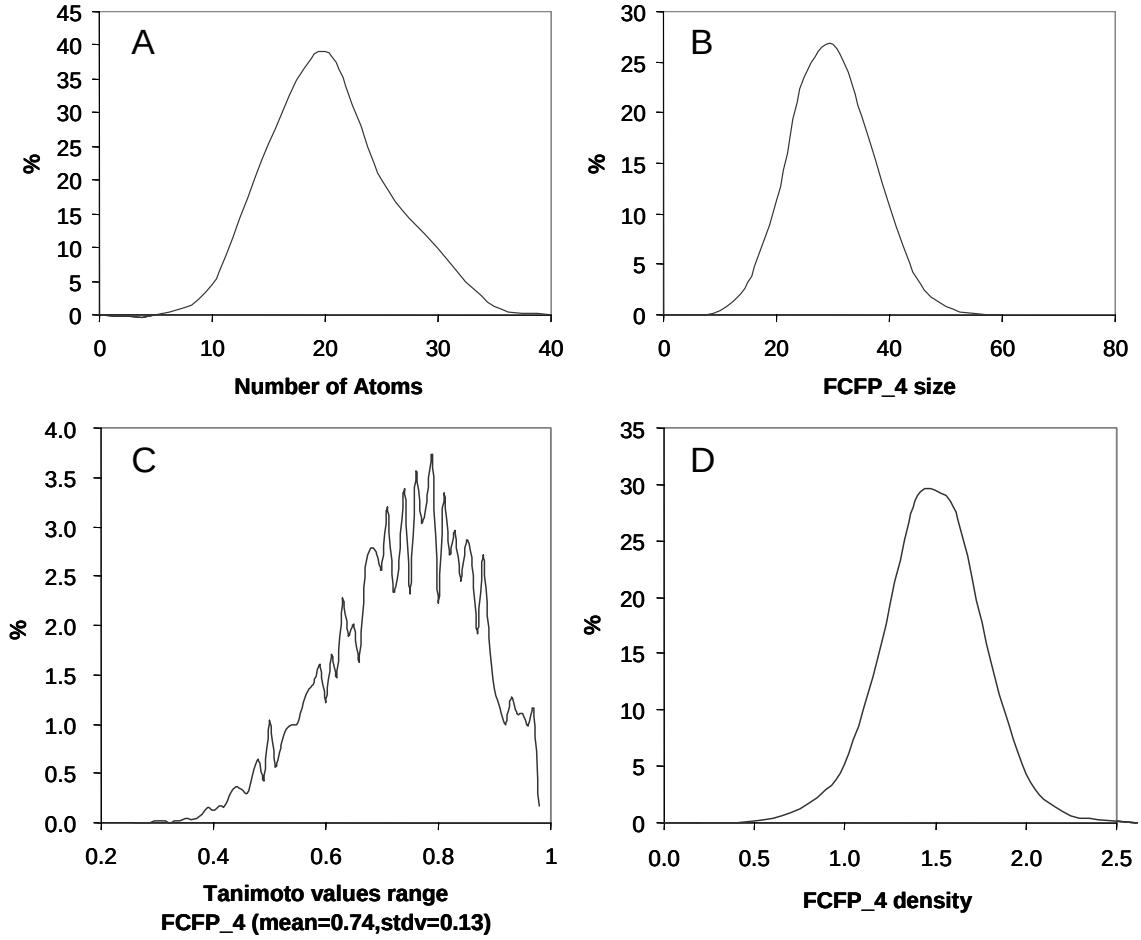


Figure 3.6 Analyzing the molecular complexity of the scaffold library. A) number of heavy atoms; B) FCFP_4 size: number of bits set in the SciTegic functional connectivity fingerprints("Pipeline Pilot") using a fragment diameter up to four bonds; C) Self-similarity plot using FCFP_4 fingerprints and Tanimoto coefficient; D) FCFP_4 density: FCFP_4 size / number of heavy atoms.

It should be noted that several scaffold-based libraries have already been reported in the past. Agrafiotis et al. (Agrafiotis *et al.*, 2002) described a probe library based on. 50 representative scaffolds comprising 300 000 drug-like compounds dedicated to primary screening. Another design of a scaffold-library was recently reported by Card et al. (Card *et al.*, 2005) where 275 555 compounds (starting with 1 994 133 molecules from 17 vendors, then filtered by MW range) have been clustered according to their constituent fragments (segmented at rotatable bonds) and similar compounds were grouped (Tanimoto index > 0.85). This resulted in 20 360 small molecular-weight

fragments covering approximately 80 % of the scaffold component space. Our library presents the advantage to cover most commercially-available compounds and to archive scaffolds as a medicinal chemist would do by intuition, thus enabling an easy navigation in scaffold space and the selection of the most relevant compounds according to simple user-defined queries.

3.4.4 On the use of ClassPharmer scaffolds

There are both advantages and drawbacks in utilizing ClassPharmer for computing molecular scaffolds out of large libraries. A first true advantage is the ad-hoc detection of MCS which enables a classification of all compounds of the library. Alternative strategies based on the storage of pre-computed chemical features (Roberts *et al.*, 2000) do not guarantee this exhaustiveness. Secondly, the fuzziness of ring closure and atom match definitions can be customized depending on purpose. We here chose exact definitions of the latter parameters to ensure a chemically unique definition of each scaffold. Although tolerating non exact atom matches would enable taking into account bioisosterism in the scaffold definition and thus significantly decrease the number of scaffolds, fuzzy ring closure is clearly not suited for archiving scaffolds as it would allow the definition of substructures (e.g. 3 carbon atoms of a phenyl ring) as classes. Thirdly, ClassPharmer MCS describe not only the minimum common substructure but also its chemical environment which enables a classification mirroring mostly the intuition of a medicinal chemist. Hence, many scaffolds already identified by vendors within their collections (De Laet *et al.*, 2000) can be recovered in our SBI scaffold library. Lastly, importing compounds from a new collection into an existing classification allows the quick evaluation of the scaffold overlap of both collections.

A clear drawback of our approach is its low speed. Using a standard PC with 1GB RAM, only collections with less than 150 K compounds can be classified within 48 cpu hours.

The regular upgrade of the scaffold library is thus considerably penalized. Meanwhile alternative classification approaches (Stahl & Mauser, 2005; Wilkens *et al.*, 2005) have been developed and might be considered under the conditions that (i) it also produces chemically meaningful classes and (ii) a significant increase in performance can be observed for the same initial (huge) library size.

A limitation, for the purpose of scaffold archiving, is the redundancy observed in the clustering (e.g. a particular compound is often found in multiple classes). Although class redundancy is not necessarily a problem in mining HTS data as it exactly reflects the point of view of a medicinal chemist, it was a real hurdle in our study to quantify the population covered by each class. To overcome this problem, we developed a very simple approach which selects the most 'central scaffold' in the structure of each ligand. It should be stated that our protocol still generates a significant number of singletons. Because of the overall low speed of the classification procedure, we have not considered merging all singletons and reclassifying this subset to populate existing classes or to generate additional clusters. Likewise, reclustering singletons by similarity to existing cluster substructures (Stahl & Mauser, 2005) is another interesting alternative to reduce the number of singletons.

3.5 Conclusions

The molecular diversity of 17 commercially-available screening collections covering 2.4 million compounds was evaluated by computing graph-based maximum common substructures for each library. Two metrics (NC50C, PC50C) were developed to facilitate the comparison of libraries of various sizes. The herein analyzed commercial collections could be grouped into four categories depending on their size and PC50C value (percentage of scaffolds accounting for 50% of classified compounds). Our

classification reflects the history of each collection and the way it had been compiled (combinatorial libraries, screening collections, medicinal chemistry libraries). Merging all classes led to a library of non-redundant scaffolds that can easily be browsed for different purposes like (i) defining a scaffold-focused library (Krier *et al.*, 2005) starting from an existing hit and thus quickly generate structure-activity relationships, (ii) defining a general purpose library where a few copies of user-selected diverse scaffolds are cherry picked (Card *et al.*, 2005), (iii) setting-up a collection of small molecular weight fragments for structural biology screening (Zartler & Shapiro, 2005) (X-ray, NMR) by selecting the least substituted compound(s) for user-defined classes.

Bibliography

Agrafiotis, D. K., Lobanov, V. S., & Salemme, F. R. (2002). Combinatorial informatics in the post-genomics ERA. *Nat. Rev. Drug. Discov., 1*, 337-346.

Austin, C. P., Brady, L. S., Insel, T. R., & Collins, F. S. (2004). NIH Molecular Libraries Initiative. *Science, 306*, 1138-1139.

Baurin, N., Aboul-Ela, F., Barril, X., Davis, B., Drysdale, M., Dymock, B., et al. (2004). Design and characterization of libraries of molecular fragments for use in NMR screening against protein targets. *J. Chem. Inf. Comput. Sci., 44*, 2157-2166.

Baurin, N., Baker, R., Richardson, C., Chen, I., Foloppe, N., Potter, A., et al. (2004). Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci., 44*, 643-651.

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem., 39*, 2887-2893.

Bocker, A., Derksen, S., Schmidt, E., Teckentrup, A., & Schneider, G. (2005). A hierarchical clustering approach for large compound libraries. *J. Chem. Inf. Model., 45*, 807-815.

Brown, R. D., Downs, G. M., Willett, P., & Cook, A. P. F. (1992). A hyperstructure model for chemical structure handling: Generation and atom-by-atom searching of hyperstructures. *J. Chem. Inf. Comput. Sci., 32*, 522-531.

Card, G. L., Blasdel, L., England, B. P., Zhang, C., Suzuki, Y., Gillette, S., et al. (2005). A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. *Nat. Biotechnol., 23*, 201-207.

Charifson, P. S., & Walters, W. P. (2002). Filtering databases and chemical libraries. *J. Comput. Aided Mol. Des., 16*, 311-323.

ClassPharmer Suite. (Version 3.2-3.5): Bioreason, Inc., 3900 Paseo del Sol, Santa Fe, NM 87507, USA.

Cliff. (Version 1.23): Molecular Networks GmbH, D-91052 Erlangen, Germany.

Cummins, D. J., Andrews, C. W., Bentley, J. A., & Cory, M. (1996). Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. *J. Chem. Inf. Comput. Sci., 36*, 750-763.

De Laet, A., Hehenkamp, J. J. J., & Wife, R. L. (2000). Finding drug candidates in virtual and lost/emerging chemistry. *J. Heterocycl. Chem.*, 669-674.

Feher, M., & Schmidt, J. M. (2003). Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. *J. Chem. Inf. Comput. Sci., 43*, 218-227.

Filter 1.0. OpenEye Scientific Software, Inc., Santa Fe, NM, USA.

InChI (IUPAC International Chemical Identifier). (Version 1.0)(2005). IUPAC.

JChemBase. ChemAxon Ltd., Budapest, 1037 Hungary.

Krier, M., Araujo-Junior, J. X., Schmitt, M., Duranton, J., Justiano-Basaran, H., Lugnier, C., et al. (2005). Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. *J. Med. Chem., 48*, 3816-3822.

Lewell, X. Q., Judd, D. B., Watson, S. P., & Hann, M. M. (1998). RECAP--retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. *J. Chem. Inf. Comput. Sci., 38*, 511-522.

Martin, Y. C. (2001). Diverse viewpoints on computational aspects of molecular diversity. *J. Comb. Chem., 3*, 231-250.

McGregor, J. J., & Willett, P. (1981). Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci., 21*, 137-140.

McGregor, M. J., & Pallai, P. V. (1997). Clustering of large databases of compounds: Using the MDL "keys" as structural descriptors. *J. Chem. Inf. Comput. Sci., 37*, 443-448.

National Chemical Library, National Center for Scientific Research (CNRS).), http://chimiotheque.ujf-grenoble.fr/induk.html

Nilakantan, R., Bauman, N., Dixon, J. S., & Venkataraghavan, R. (1987). Topological torsions: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci., 27*, 82-85.

Nilakantan, R., Bauman, N., & Haraki, K. S. (1997). Database diversity assessment: new ideas, concepts, and tools. *J. Comput. Aided Mol. Des., 11*, 447-452.

Nilakantan, R., & Nunn, D. S. (2003). A fresh look at pharmaceutical screening library design. *Drug Discov. Today, 8*, 668-672.

OEChem. (Version 1.3): OpenEye Scientific Software, Inc., Santa Fe, NM, USA.

Pipeline Pilot. (Version 4.2): SciTegic Inc., San Diego, CA 92123-1365, USA.

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., & Blower Jr., P. E. (2000). LeadScope : Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci., 40*, 1302-1314.

Schuffenhauer, A., Ruedisser, S., Marzinzik, A. L., Jahnke, W., Blommers, M., Selzer, P., et al. (2005). Library design for fragment based screening. *Curr Top Med Chem, 5*, 751-762.

Selzer, P., Roth, H. J., Ertl, P., & Schuffenhauer, A. (2005). Complex molecules: do they add value? *Curr. Opin. Chem. Biol., 9*, 310-316.

Shemetulskis, N. E., Dunbar, J. B., Jr., Dunbar, B. W., Moreland, D. W., & Humblet, C. (1995). Enhancing the diversity of a corporate database using chemical database clustering and analysis. *J. Comput. Aided Mol. Des., 9*, 407-416.

Sirois, S., Hatzakis, G., Wei, D., Du, Q., & Chou, K. C. (2005). Assessment of chemical libraries for their druggability. *Comput. Biol. Chem., 29*, 55-67.

Stahl, M., & Mauser, H. (2005). Database clustering with a combination of fingerprint and maximum common substructure methods. *J. Chem. Inf. Model., 45*, 542-548.

Su, A. I., Lorber, D. M., Weston, G. S., Baase, W. A., Matthews, B. W., & Shoichet, B. K. (2001). Docking molecules by families to increase the diversity of hits in database screens: computational strategy and experimental evaluation. *Proteins, 42*, 279-293.

Voigt, J. H., Bienfait, B., Wang, S., & Nicklaus, M. C. (2001). Comparison of the NCI open database with seven large chemical structural databases. *J. Chem. Inf. Comput. Sci., 41*, 702-712.

Webb, T. R. (2005). Current directions in the evolution of compound libraries. *Curr. Opin. Drug Discov. Devel., 8*, 303-308.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci., 28*, 31-36.

Wilkens, S. J., Janes, J., & Su, A. I. (2005). HierS: hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem., 48*, 3182-3193.

Willett, P. (1985). An algorithm for chemical superstructure searching. *J. Chem. Inf. Comput. Sci., 25*, 114-116.

Xue, L., & Bajorath, J. (1999). Distribution of molecular scaffolds and R-groups isolated from large compound databases. *Journal of Molecular Modeling, 5*, 97-102.

Zartler, E. R., & Shapiro, M. J. (2005). Fragonomics: fragment-based drug discovery. *Curr. Opin. Chem. Biol., 9*, 366-370.

Chapter 4

4. Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold : Application to the structure-based optimization of a phosphodiesterase 4 inhibitor*

*J. Med. Chem. 45, 3816-3822, 2005; Part of this paper was the subject of a poster presentation at the 3rd Joint Conference on Chemoinformatics, 21-23 April, 2004, Sheffield, UK.

4.1 Scope and critical evaluation of the project

(Not included in the publication)

The goal of this project was to validate the Scaffold-Linker-Functional Group Approach implemented during this thesis in a C++ program. Together with the bench chemists and biologists at the Faculty of Pharmacy, a common protein target, the phosphodiesterase IV (PDE4), was adopted in order to be our application case. Phosphodiesterase inhibitors have been developed for more than 20 years, but an X-ray structure was only obtained in the 2001. Two years later, a 3D-structure of a PDE4 co-crystallized with zardaverine has been resolved and as the chemist have been attracted to optimize this structure in order to develop a more potent inhibitor, it seemed to be the ideal test case for our SLF approach.

Phosphodiesterase IV is a member of the 3'5'-cyclic nucleotide phosphodiesterases, which is composed of 21 human PDE genes coding for11 family members and over 60 isoforms. A typical signature of the catalytic domain H-D-[LIVMFY]-x-H-x-[AG]-x(2)-[NQ]-x-[LIVMFY] of this family have been deposited in PROSITE under the entry name PDEASE_1 (Accession number PS00126). PDEs are soluble proteins that functions as intracellular second messenger. The function of PDEs is to maintain cellular levels of cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP) which mediate various biological responses from a great number of extracellular stimuli. PDEs (E.C. 3.1.4.17) catalyze the hydrolysis of cAMP or cGMP to the corresponding nucleoside 5' monophosphates. We refer the reader to the recently published review by Lugnier (Lugnier, 2005) to enhance their knowledge about the properties, tissue and subcellular distributions of cyclic nucleotide phosphodiesterase superfamily.

PDE ligands are widely used in a variety of clinical applications like anti-inflammatory agents, anti-asthmatics, vasodilators, smooth muscle relaxants, cardiotonic agents, antithrombotics, antidepressants, and improving cognitive functions. Manallack et al. (Manallack *et al.*, 2005) summarized a large number of compounds that bind to PDEs. Table 4.1 focuses on the PDE ligands co-crystallized and deposited in the Protein Data Bank.

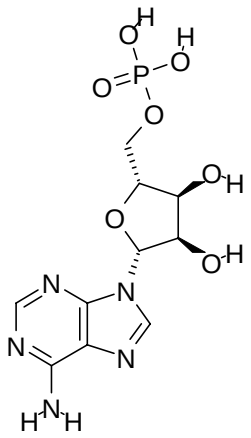
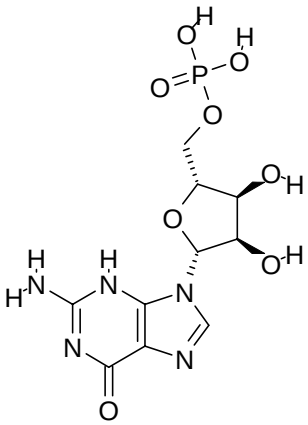
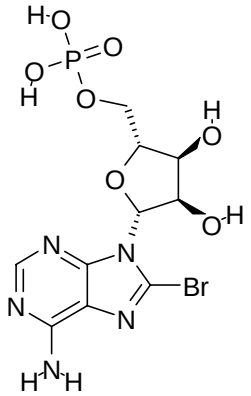
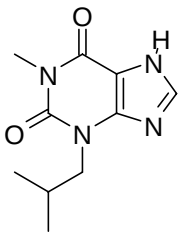
Ligand	Name	Isoform	PDB code (resolution, Å)
	AMP	4B 4B 4D 4D	1TB5 (2.15) 1ROR (2.0) 1PTW (2.4) 1TB7 (1.63)
	GMP	5A	1T9S (2.0)
	8-BromoAMP	4B	1RO9 (2.13)
	IBMX	3B 4D 5A 7A 9A	1SOJ (2.9) 1ZKN (2.1) 1RKP (2.05) 1ZKL (1.67) 1TBM (2.23)

Table 4.1 Co-crystallized ligands with proteins of the PDE family

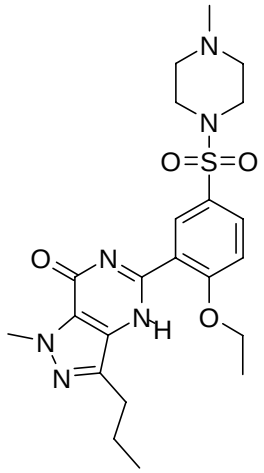
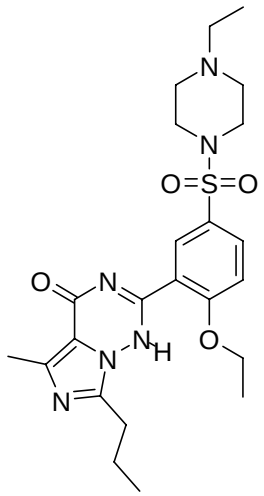
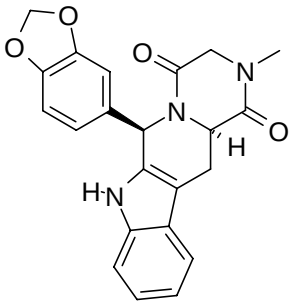
Ligand	Name	Isoform	PDB code (resolution, Å)
	Sildenafil	4B 5A 5A	1XOS (2.28) 1TBF (1.3) 1UDT (2.3)
	Vardenafil	5A 5A 4B	1UHO (2.5) 1XP0 (1.79) 1XOT (2.34)
	Tadalafil	5A 5A	1UDU (2.83) 1XOZ (1.37)

Table 4.1 (continued) Co-crystallized ligands with proteins of the PDE family

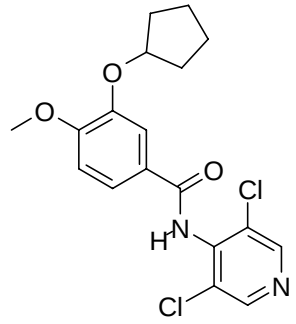
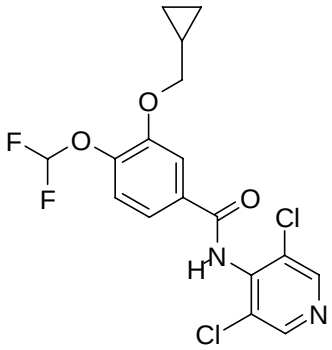
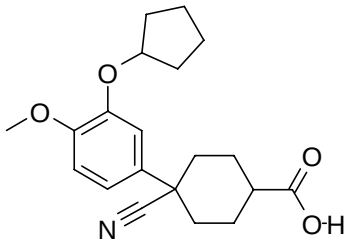
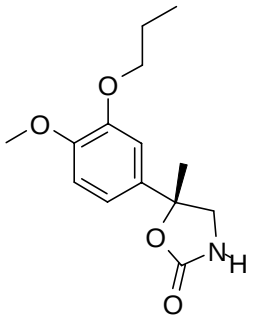
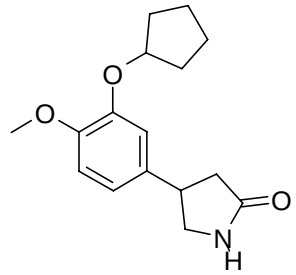
Ligand	Name	Isoform	PDB code (resolution, Å)
	Piclamilast	4D 4B	1XON (1.72) 1XM4 (2.31)
	Roflumilast	4B 4D	1XMU (2.3) 1XOQ (1.83)
	Cilomilast	4B 4D	1XLX (2.19) 1XOM (1.55)
	(R)-mesopram	4B	1XM6 (1.92)
	(R,S)-rolipram	4B 4D 4B 4D	1XN0 (2.31) 1OYN (2.0) 1RO6 (2.0) 1TBB (1.6)

Table 4.1 (continued) Co-crystallized ligands with proteins of the PDE family

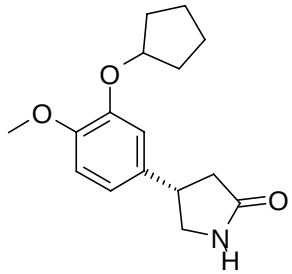
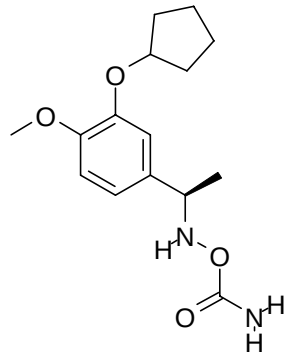
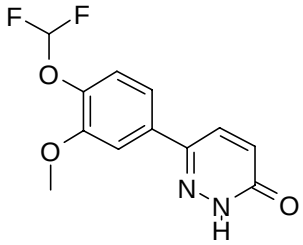
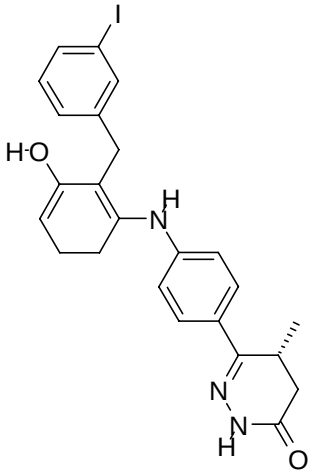
Ligand	Name	Isoform	PDB code (resolution, Å)
	(R)-rolipram	4B 4D	1XMY (2.40) 1Q9M (2.3)
	Filaminast	4B	1XLZ (2.06)
	Zardaverine	4D 4D	1MKD (2.9) 1XOR (1.54)
	Merck-1	3B	1SO2 (2.4)

Table 4.1 (continued) Co-crystallized ligands with proteins of the PDE family

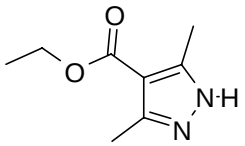
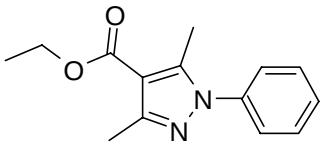
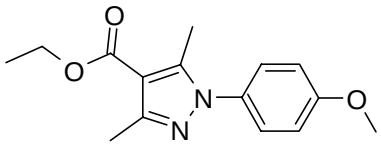
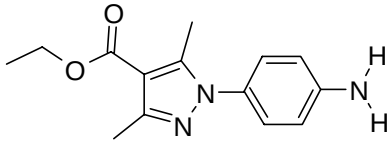
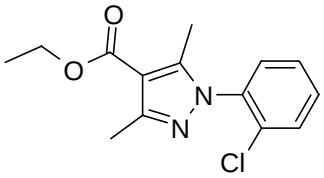
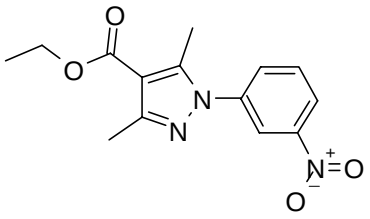
Ligand	Name	Isoform	PDB code (resolution, Å)
	Plexxikon2	4D	1Y2B (1.4)
	Plexxikon8	4D	1Y2C (1.67)
	Plexxikon17	4D	1Y2D (1.7)
	Plexxikon19	4D	1Y2E (2.1)
	Plexxikon20	4B	1Y2H (2.55)
	Plexxikon21	4B 4D	1Y2J (2.55) 1Y2K (1.36)

Table 4.1 (continued) Co-crystallized ligands with proteins of the PDE family

Synthetic PDE ligands such as Theophylline and IBMX were first inhibitors described in literature. They have been used as non-specific inhibitors and serve mainly as pharmacological tool for characterization. Manallack et al. (Manallack *et al.*, 2005) suggests that most of the PDE inhibitors adopt a planar conformation in order to mimic the planarity of the purine structure. The ability to interact with the active site in more than one orientation contributes to the non-selectivity. The authors therefore propose to

take reduction of multiple binding orientations into account when attempting to increase selectivity of new inhibitors.

Rolipram has been the lead structure targeting PDE4 for a series of bicyclic compounds containing a catechol moiety. Side effects like emesis stopped their development in clinical phase. The clinical candidates that are the furthest advanced are cilomilast (Compton et al., 2001) and roflumilast (Timmer et al., 2002). These next generation inihibitors promise to be less emetic, but "*unimpressive results from a phase III study of roflumilast do not appear to provide strong enough long-term efficacy data to warrant approval by the FDA*" (http://www.pharmaceutical-business-review.com, Feature of 6 July 2005).

Although IBMX is a non-specific inhibitor for PDE1 to PDE5 and both PDE8 and PDE9 are IBMX insensitive, this structure is a valuable core template. Heizmann and Eberle (Heizmann & Eberle, 1997) showed that xanthines are synthetically tractable scaffold for combinatorial chemistry. Their paper described a five-step solid-phase synthesis of xanthine derivatives consisting of alkylations, a nucleophilic displacement reaction at a heterocycle and a ring closure reaction by condensation of a nitroso function with an activated methylene group. Hence, a small-molecule combinatorial compound library being highly diverse was set up.

Recently, Card et al. (Card *et al.*, 2005) described PDE4 co-crystallized ligands on the basis of a ethyl 3,5-dimethyl-1H-pyrazole-4-carboxylate scaffold increasing the activity from 60/82 μM (IC_{50} PDE4B/PDE4D) to 0.033/0.021 μM (IC_{50} PDE4B/PDE4D, PDB code 1Y2J/1Y2K).

Some additional difficulties like water-mediated binding and metal coordination made the virtual screening (under the form of docking) more challenging. As experience in

our laboratory showed before, the active site properties induce the choice for the appropriate docking program/scoring function. FlexX performed in the most accurate way (Table 4.4). The PDE4 active site can be subdivided in several pockets: (i) the catalytic site with the two metal centers, (ii) the hydrophobic clamp with its two subpockets and its superfamily invariant glutamine. The later permits directional interactions. The absence of metal centers contributed to the failure of the other combinations. Gold has the constraint that an atom parameterized as a metal ion has to be coordinated to at least two protein atoms or water molecules so that Gold can determine the correct coordination geometry. But alignment of the different three-dimensional protein structures along alpha carbons didn't reveal any structural (invariant) water molecule for the second metal ion. Surflex is building up a negative mold of the active site by placing voxels (contraction for "volumetric pixel") complementary in property to the protein amino acids. This pseudo-molecule, also called the protomol, is composed of molecular fragments being of nature C=O (probe-a), CH4 (probe-s), NH (probe-d). The three ways of protomol generation were investigated: (1) ligand-based, (2) residue-based and (3) neither of them (in other words the user makes no indication of putative active site and ligand). Nevertheless, Surflex did not succeed in placing the reference ligand accurately.

This published test case demonstrated the successful application of our approach. Indeed potency was increased by a factor of almost 10^3 by evaluating computationally only 320 compounds build by the approach and testing 9 of these. Five showed an improved potency over the reference compound zardaverine.

Another application case for SLF approach was the generation of a combinatorial library based on the scaffold of secretory phospholipase A_2 inhibitors (unpublished data, Muller et al.). This project is the logical sequel to the in silico guided target identification by a molecular scaffold. Briefly describing this precedent work, a collection of 2,150 active sites from the Protein Data Bank was screened by high-throughput docking to identify putative targets for five representative molecules of a combinatorial library sharing a triazepanedione scaffold. Five targets were prioritized for experimental evaluation by computing enrichment rates of individual protein entries among the top 2% scoring targets. Out of the five proposed proteins, phospholipase A2 had shown to be a true target for a panel of triazepanediones which exhibited micromolar affinities toward several isoforms of the latter enzyme.

The virtual combinatorial library was based on a cyclic urea, the triazepanedione scaffold. It was created in two steps. First, R2-substituted 16 scaffolds were created. Secondly, these scaffolds were substituted independently at the three positions R3, R4 and R5 by 3 linkers and 10 functional groups. Hence 1440 (16 × 3× 10 × 3) have been rapidly assembled by this way (Table 4.2).

Docking studies were performed using GOLD (v2.0) for the binding predictions of the 1440 structures. As protein structure model served the X-ray PLA_2 isoform hGX (PDB code 1LE6). 48 structures scored above 61 (GoldScore) and are mainly substituted on R5 position and carry a benzyl group on R2 position.

Four molecules were subsequently synthesized and for two, a preliminary activity was determined on isoforms V and X.

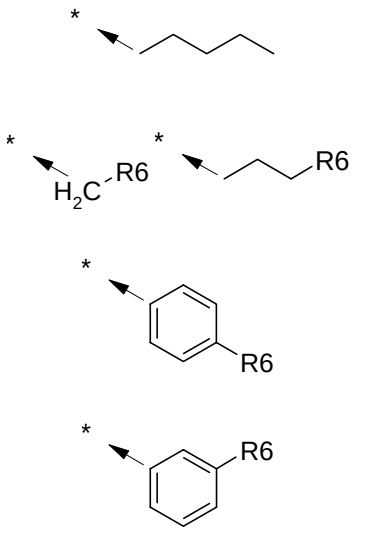
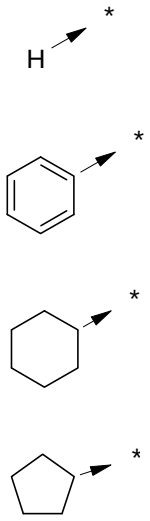
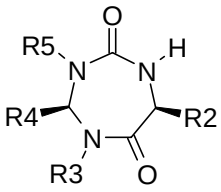
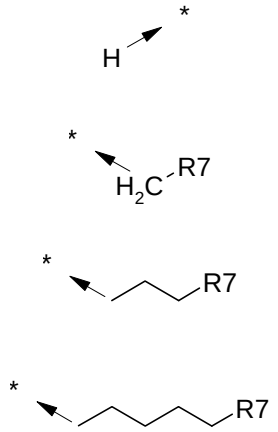
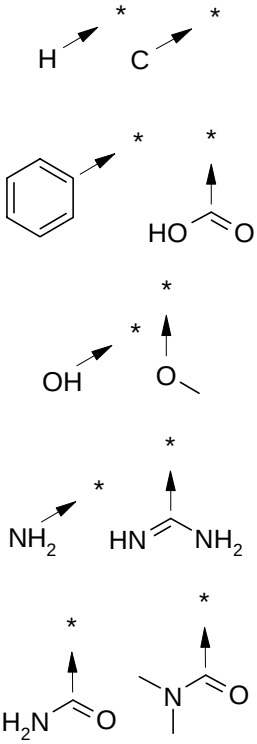
Scaffold	Linker	Functional Group
	R2	R6
		
	R3 = R4 = R5 If R3 < > H then R4 = R5 = H If R4 < > H then R3 = R5 = H If R5 < > H then R3 = R4 = H	R7
		

Table 4.2 Scaffold, linkers and functional groups used for the combinatorial enumeration of a PLA_2-focused library

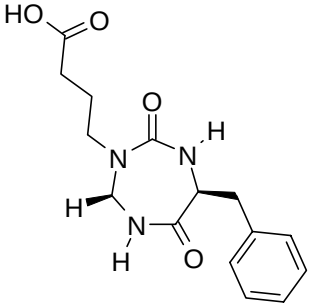
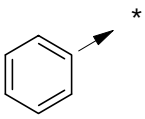
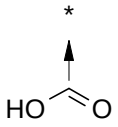
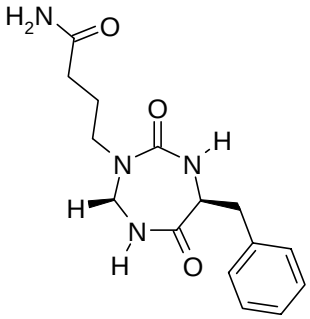
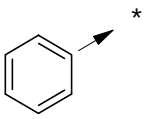
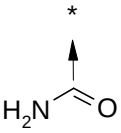
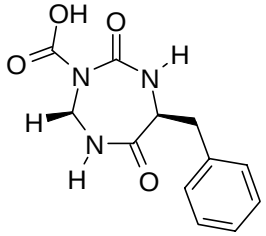
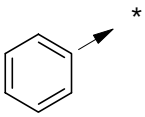
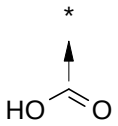
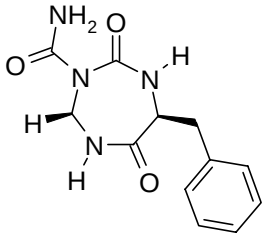
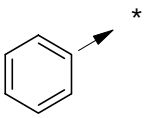
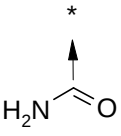
Compound	R2 L	R2 FG	R5 L	R5 FG	IC_{50} (μM) PLA_2 hGV	IC_{50} (μM) PLA_2 hGX
	1		3		41	52
	1		3		40	77
	1		1		N/A	N/A
	1		1		N/A	N/A

Table 4.3 PLA_2 hGV and PLA_2 hGX inhibition of compounds

During other structure-based design projects involving the SLF approach, the generation of three-dimensional coordinates by external application led to conformations of the core structure which were too distant (in terms of root-mean square deviation) from initial co-crystallized scaffold conformation. Therefore, the SLF approach should be extended to a topographic mode taking a predefined scaffold conformation into account.

4.2 Abstract

Combinatorial chemistry and library design have been reconciled by applying simple medicinal chemistry concepts to virtual library design. The herein reported "Scaffold-Linker-Functional Group" (SLF) approach has the aim to maximize diversity while minimizing the size of a scaffold-focused library with the aid of simple molecular variations in order to identify critical pharmacophoric elements. Straightforward rules define the way of assembling three building blocks: a conserved scaffold, a variable linker and a variable functional group. By carefully selecting a limited number of functional groups not overlapping in pharmacophoric space, the size of the library is kept to a few hundred. As an application of the SLF approach, a small-sized combinatorial library (320 compounds) was derived from the scaffold of the known phosphodiesterase 4 inhibitor zardaverine. The most interesting analogs were further prioritized for synthesis and enzyme inhibition by FlexX docking to the X-ray structure of the enzyme, leading to a 900-fold increased affinity within 9 synthesized compounds and a single screening round.

4.3 Introduction

With combinatorial chemistry as a tool, infinite variations on a core template are theoretically possible (Alper, 1994). However, in the drug discovery process, it is desirable to gain the maximum of information (Bradley *et al.*, 2003) out of a minimum of experiments. For the medicinal chemist, this means the optimization of a screening library, i.e. a minimal size with a maximal chemical diversity. Hence, focused libraries designed around a selected scaffold can only span a wide range of physicochemical and structural properties, when the decorations are diverse enough. To date, a lot of combinatorial structure generation tools have been developed and most of them are generating rather large virtual libraries (Beavers & Chen, 2002; De Julian-Ortiz, 2001). However, awareness that the highest possible number of compounds do not automatically increase the hit rate and the fact that most of the generated molecules are synthetically not easily available make computational chemists apply a second algorithm to select a representative subset. Indeed, to stay in the order of magnitude of a hundred compounds (Nilakantan & Nunn, 2003), virtual combinatorial libraries are assessed by different techniques like Monte Carlo calculations (Langer & Wolber, 2004), genetic algorithms (Wright *et al.*, 2003), artificial neutral network or simply statistical sampling with user-defined property ranges (Weaver, 2004).

In contrast to the latest, we propose an alternative approach to optimize size versus diversity that relies on the combinatorial assembly of user-selected building blocks: a scaffold, a linker and a functional group (Figure 4.1). Thus, each enumerated molecule can be considered as a chemical tool to probe the protein active site. Similar approaches published recently are implemented in COREGEN (Aronov & Bemis, 2004) and SMILIB (Schuller *et al.*, 2003). Based on homology and molecular diversity concepts, combining a limited number of linkers and functional groups (cations, anions, hydrogen

bond acceptor-donor systems, aromatics/lipophilics) easily affords small-sized polyfunctionalized compounds (Bemis & Murcko, 1996; De Laet *et al.*, 2000).

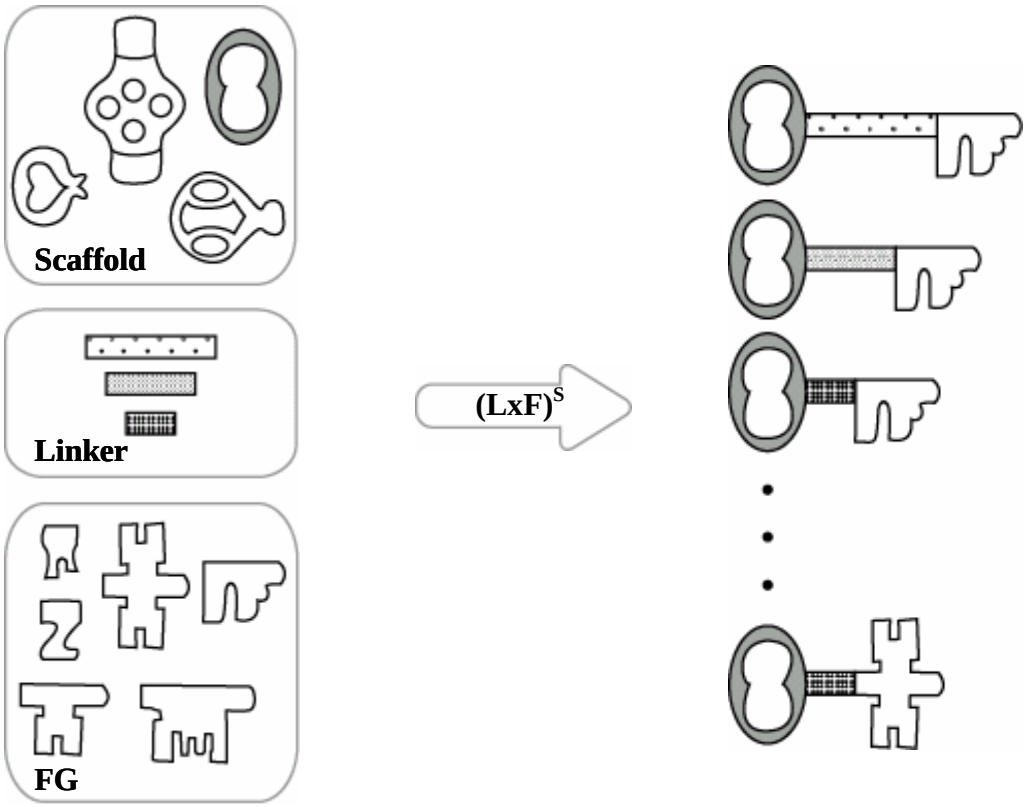


Figure 4.1 Schematic representation of the three types of molecular fragments and the assembly rule for the complete library enumeration

Indeed, chemists succeeded over the years to apply combinatorial synthesis strategies to simple rings and chains to form small organic molecules (Janvier *et al.*, 2002; Kuznetsov *et al.*, 2004) and not stay limited to peptides and oligonucleotides polymers. Thus, the virtual combinatorial library has to be designed in order to have its physical counterpart and to guarantee that all compounds are synthesizable. We herewith present the combinatorial assembly method encoded in the SLF_Libmaker program and its coupling to the structure-based prioritization of the most interesting compounds applied to the optimization of a known phosphodiesterase 4 (PDE4) inhibitor.

4.4 Materials and Methods

4.4.1 Virtual library construction

The three-fragment assembly rule is "scaffold, linker, functional group" (Lewell *et al.*, 2003) instead of a more common use of binary combinations of building blocks (De Julian-Ortiz *et al.*, 1999). A more detailed look at the different building blocks is given as follows. Two hypotheses about the scaffold (Lipkus, 2001) are usually cited: (i) a suitable scaffold is believed to optimally orient the attached substituents for binding and (ii) the scaffold itself interacts with the protein as an anchor. The maximum number of possible connection points equals the number of removable hydrogen atoms, but the most encountered examples of scaffolds have from one to four substituted positions.

The linker has two substitutions points. Its main role is frequency variation, i.e. to modulate the distance between the molecular scaffold and the protein active site. For a first screening round, the linkers are chosen in the acyclic polymethylenic series (Scaffold-$[CH_2]_n$-FG; FG: Functional Group). Thus, by expanding an alkyl chain, the hydrophobicity of the molecule is increased (Wermuth, 2003).

Functional groups represent basic pharmacophoric features resulting from steric, electronic, lipophilic and H-bonding properties. "H" is always the reference substituent. The other substituents are smallest possible representative fragments and will mostly mix property information, e.g. carboxylate shows anionic or H-bond acceptor behavior depending on its interaction partner.

The complete number N of enumerated molecules (Scheme 1) can be expressed as:

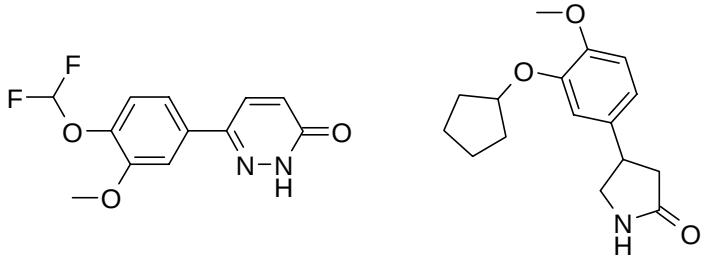
$$N = (L \times F)^S$$

where L is the number of linker fragments, F is the number of functional groups and S is the number of substitution points marked on the scaffold. The complete enumeration with a core structure showing a local symmetry in substitution points gives a lower number of unique structures according to the Pólya counting theory (van Almsick *et al.*, 2000). The complete enumeration method is implemented in SLF_LibMaker, a C++ program based on OpenEye's OEChem1.3 library ("OEChem", 2004). The molecular fragments (scaffolds, linkers, functional groups) are encoded as SMILES (Weininger, 1988). or SDF (MDL) file formats with connecting pseudo-atoms represented by "*" in both formats. Depending on the user-defined selection of scaffolds, linkers and functional groups in separate data files, the desired combinatorial library is assembled in SMILES or SD file format and converted into 3-D structures using Concord (Pearlman). In the current example, a library of 320 compounds was build from a selection of 4 scaffolds, 5 linkers and 16 functional groups, chosen in agreement with medicinal chemists. OpenEyes's Filter ("OEChem", 2004) program was finally used to ionize compounds at physiological pH. For the specific case of the benzylamino linker, it should be noticed that both neutral and ionized states were explicitly considered.

4.4.2 Automated docking

The crystal structures of the human phosphodiesterase 4D (PDE4D) catalytic domain in complex with zardaverine (Lee *et al.*, 2002) and Rolipram (Huai, Wang *et al.*, 2003) (Chart 4.1) were retrieved from the Protein Data Bank (pdb entries 1mkd and 1q9m, respectively) (Berman *et al.*, 2000). The numbering of the UniProt (Apweiler *et al.*, 2004) entry CN4D_HUMAN (Q08499) was selected as a reference. These structures were used to generate two series of input coordinates including the holo-protein, the corresponding active site and its native ligand. The protein active site was defined as the set of amino acids for which at least one atom is included in a 6.5 Å-radius sphere

surrounding any non-hydrogen atom of the bound ligand. All metal ions were assigned as Zn^{2+} ions, although their real nature is still a matter of debate (Huai, Colicelli *et al.*, 2003; Huai, Wang *et al.*, 2003; Xu *et al.*, 2000), and included in the binding site. All water molecules were removed, except the one supposed to be a hydroxide ion[4] that is thought to be a bridging element between the metal ions and the pyridazinone moiety of zardaverine. Atomic types and protonation states of protein atoms were manually checked. Hydrogen atoms were finally added by using the BIOPOLYMER module of SYBYL package ("SYBYL", 2003).



zardaverine (**1**) rolipram

Chart 4.1 Structure of a two PDE4 inhibitors

In order to determine which docking tool was the most appropriate in the current context, FlexX1.12 (Rarey *et al.*, 1996), Gold2.1 (Verdonk *et al.*, 2003) and Surflex1.1(Jain, 2003) programs were used as previously described (Kellenberger *et al.*, 2004), to reproduce the enzyme-bound pose of zardaverine. Docking was considered successful, when the best-scored pose was found within 2.0 Å root-mean square deviation (rmsd) from the X-ray pose. Cross-docking of zardaverine to the 1q9m entry and of rolipram to the 1mkd structure was then achieved in order to select the best set of holoprotein coordinates for both inhibitors.

Full database docking was realized using the 1q9m coordinates and FlexX as described above. The final hitlist was prioritized (i) by FlexX-score; (ii) by analysis of binding

modes achieved by a nearest-neighbor clustering of FlexX poses based on the Cartesian coordinates of the common dimethoxyphenyl substructure; and (iii) visual inspection of all compounds.

4.4.3 Synthesis.

The synthesis of compounds **2-10** and structurally-related molecules will be described elsewhere (M.S and J-J.B, manuscript in preparation)

4.4.4. PDE4 inhibition

PDE4 was isolated from the media layer of bovine aorta by anion exchange chromatography as previously described (Lugnier & Komas, 1993; Lugnier *et al.*, 1986) and its activity was measured at a concentration of 1 µM cAMP by radioenzymatic assay (Keravis *et al.*, 1980). To prevent the interaction of contaminating PDE3 in the assay of isolated PDE4, studies were always carried out in the presence of 100 µM cGMP. New compounds were dissolved in DMSO or ethanol with a final concentration (1%) which did not significantly affect PDE activity. The inhibition study on PDE4 activity included six concentrations of the drug. The IC_{50} values were calculated by nonlinear regression using the Prism Software (GraphPad Software, Inc., San Diego, CA 92130 USA)

4.5 Results

4.5.1 Selection of the most appropriate docking tool

Predicting the best possible docking/scoring strategy from the simple knowledge of a protein binding site is still very difficult (Bissantz *et al.*, 2000). Therefore, three accurate docking engines (Kellenberger *et al.*, 2004) (FlexX, Gold, Surflex) in combination with four scoring functions (FlexXscore, Goldscore, Chemscore, Surflex) were selected for a preliminary study aimed at determining which X-ray structure

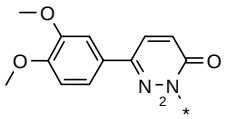
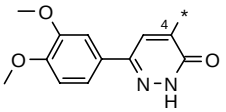
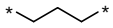
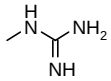
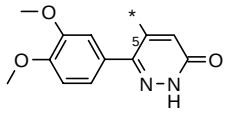
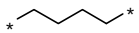
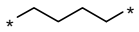
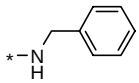
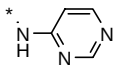
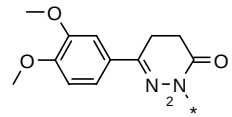
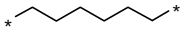
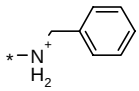
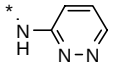
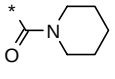
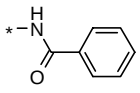
(1q9m, 1mkd) is the most suitable and which docking strategy recovers the X-ray pose of zardaverine. Out of the three docking tools tested herein, FlexX was the only program able to predict with a reliable accuracy (below 2.0 Å rmsd) the X-ray pose of zardaverine, whatever the scoring function used and the protein coordinates (Table 4.4). FlexX was then selected for further docking the zardaverine-focused library, using the original FlexX score for primary sorting the virtual hits and the 1q9m coordinates of the holoprotein.

Docking/Scoring Method	1mkd		1q9m	
Docking/Scoring Method	best score[a]	best rmsd[b]	best score	best rmsd
FlexX/FlexScore	0.66	0.66 (1)	1.02	0.51 (7)
Gold/GoldScore	8.14	6.44 (29)	10.04	7.46 (10)
Gold/ChemScore	6.73	1.40 (7)	1.39	1.27 (3)
Surflex	7.05	1.49 (3)	5.84	5.56 (4)

[a] best-scored solution

[b] solution with the lowest rmsd from the X-ray pose. The ranking of the corresponding pose is indicated in commas.

Table 4.4 Docking of zardaverine to two 1mkd and 1q9m coordinates of human PDE4.

Scaffold	Linker	Functional Group
		  
		   
	 	   
		   

a "*" indicates a connecting pseudo-atom used for the combinatorial assembly of scaffolds, linkers and fragments.

Table 4.5 zardaverine-derived scaffolds, linkers and functional groups used for the combinatorial enumeration of a PDE4-focused library

4.5.2 Setup and docking of a PDE4 focused library

The molecular fragments (Table 4.5) were assembled according to the rules described above. Four scaffolds with a single substitution point were derived from zardaverine, replacing the difluoromethoxy by a methoxy group. Moreover the unsaturated pyridazinone moiety was topologically explored at position N2, C4 and C5, whereas the dihydropyridazinone was substituted only at position N2 (Table 4.5). The linkers were chosen to be three odd-numbered (C_1, C_3, C_5) and two even-numbered polymethylene chains (C_4, C_6). The functional groups were finally selected for their pharmacophoric properties and for their synthetic feasibility. 320 structures were altogether generated to be part of the virtual library that was docked against the PDE4 target. FlexX scores

range from –42.9 to –13.5 kJ/mol, zardaverine being scored at -22.3 kJ/mol. For most of the structures, a single binding mode of the dimethoxyphenyl substructure, very close to that observed for zardaverine was found (Figure 4.2).

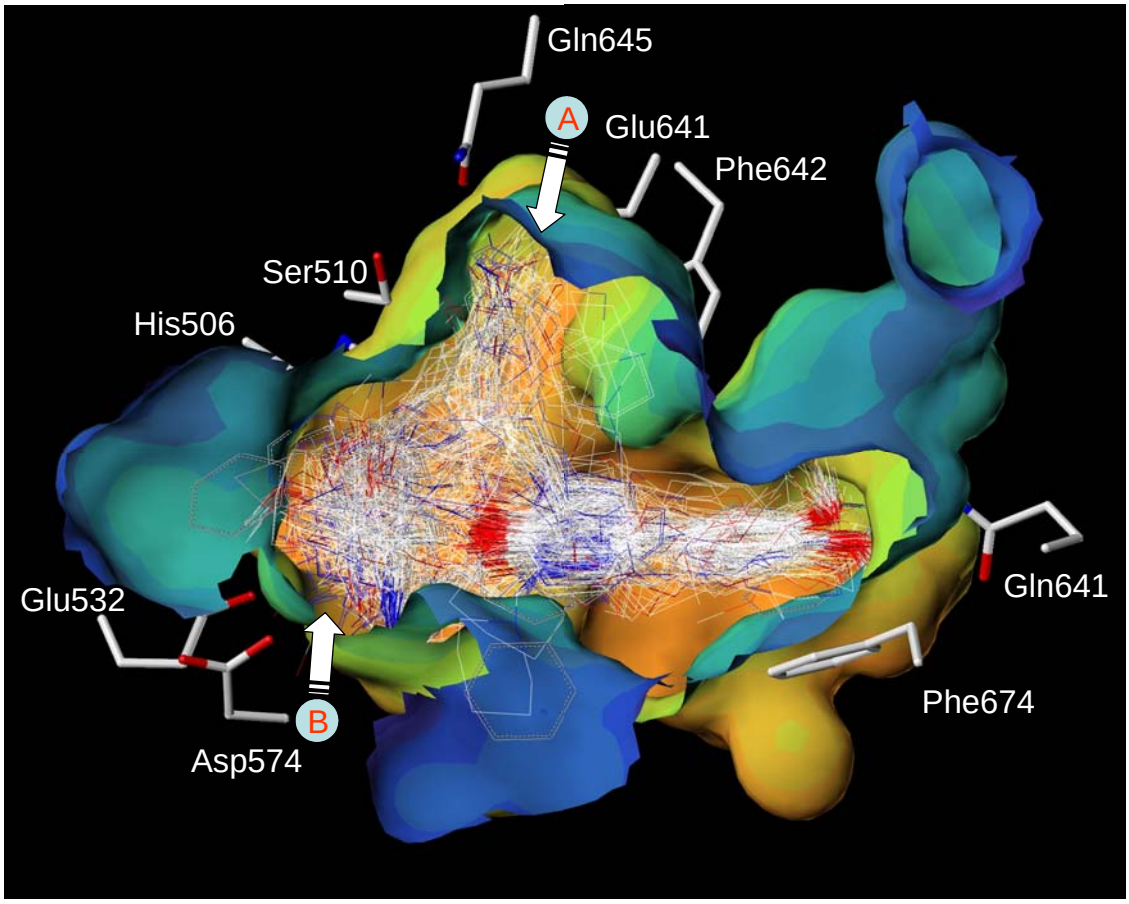
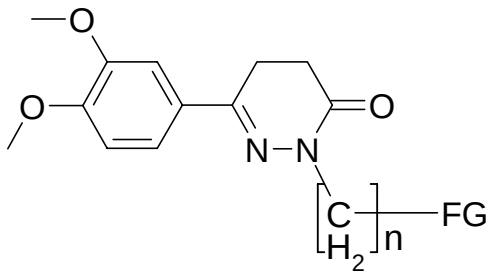


Figure 4.2 Docking of a scaffold-based library of 320 compounds into the X-ray structure of the human PDE4 catalytic domain. The best-ranked pose of each compound is displayed as a color-coded wireframe in the active site of PDE4D represented as a MOLCAD("SYBYL", 2003) solid surface color-coded by cavity depth (blue → yellow: accessible→ buried surfaces). Important side chains are displayed as capped sticks and labeled at the Cα atom. Subsites A and B are indicated by white arrows.

Other energetically-favored binding modes were not discovered through visual inspection of all poses. Browsing the top-ranked pose of all compounds suggest that two additional pockets (named A and B in Figure 4.2) could be targeted by numerous compounds. Hitlist prioritization was then achieved by selecting any compound whose Flex score was lower than -15 kJ/mol and for which the rmsd of the dimethoxyphenyl

substructure from that of zardaverine in its X-ray pose was lower than 1.0 Å. Nine compounds exploring additional pockets A and B unoccupied by zardaverine were finally selected, synthesized and tested (Tables 4.6 and 4.7). Seven out of these nine compounds were N2-substituted dihydropyridazinones exploring two additional pockets of PDE4 not investigated by either zardaverine or rolipram. A first hydrophobic channel (His462, His506, Phe642; site A) topped by polar side chains (Glu641, Gln645) favors a phenyl ring 1 to 6 carbon atoms from the N2-pyridazine ring (Figure 4.3 B). A second negatively-charged subsite around Glu532 and Asp574 (site B) favors basic amines (primary amine or amidine) 6 carbon atoms from the N2 pyridazine ring (Figure 4.3 C).



Compound	Functional Group (FG)	n	FlexX score[a]	IC_{50}, nM[b]
1 (Zardaverine)			-22.31	800
2	H	0	-20.84	2000
3	Ph	1	-19.45	60
4	Ph	3	-17.17	20
5	Ph	5	-15.63	0.9
6	Ph	6	-16.85	80
7	NH_2	6	-26.84	20
8	$NHC(NH)NH_2$	6	-20.74	60000

[a] FlexX score, in kJ/mol

[b] The IC_{50} was calculated by non linear regression and represents the mean value of three independent determinations. The experimental error is about 15%.

Table 4.6 PDE4 inhibition of compounds **1**-**8**

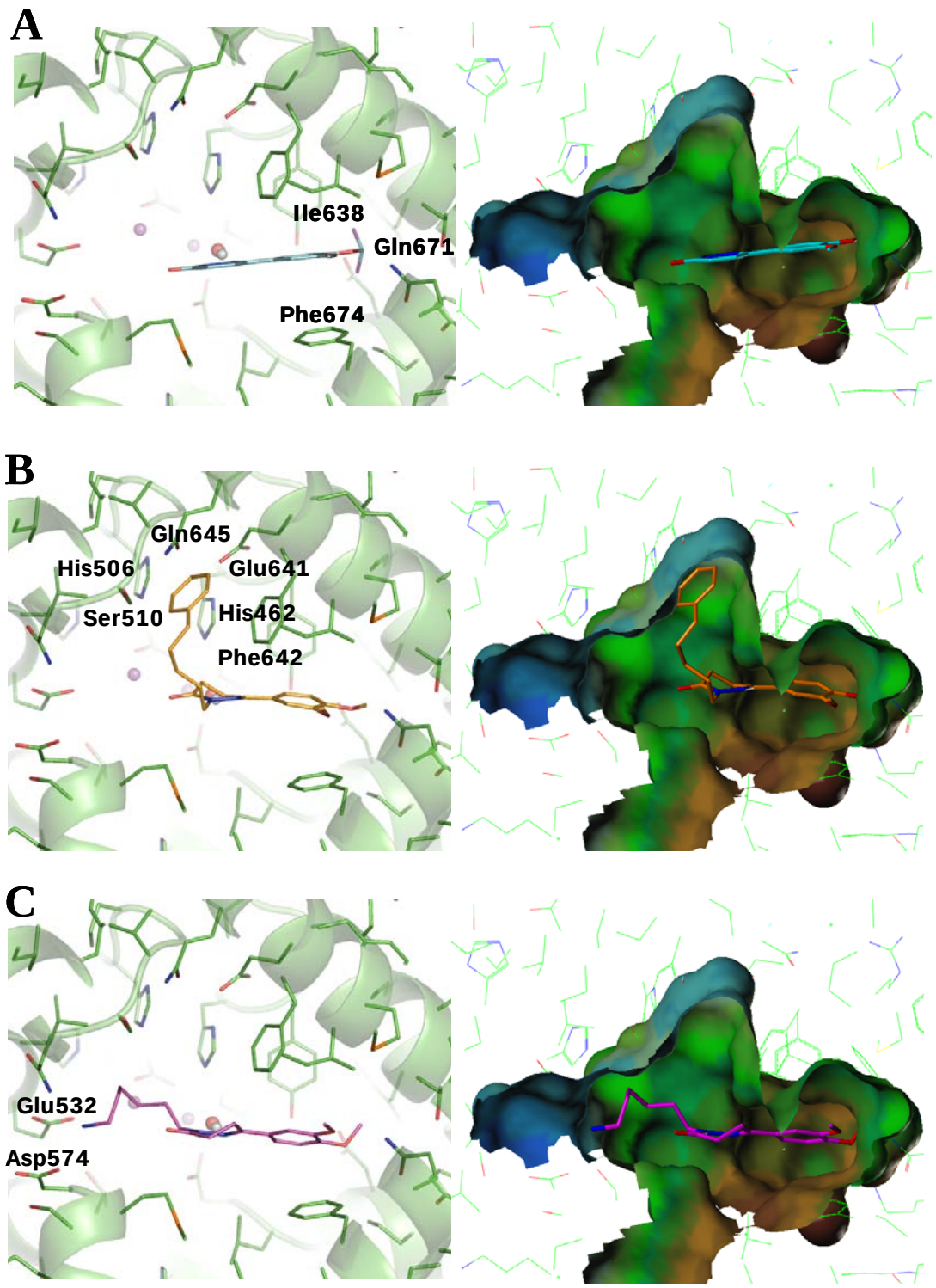
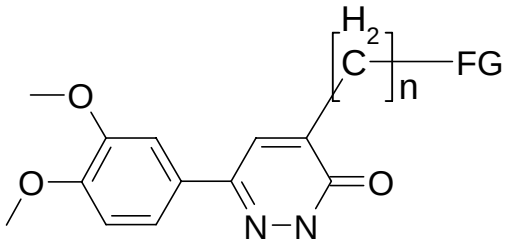


Figure 4.3 Close up to the human PDE4 inhibitor binding site filled with zardaverine **1** (A, X-ray pose), compound **5** (B, FlexX best-scored pose) and compound **7** (C, FlexX best-scored pose). Left panels represent the inhibitor (zardaverine, cyan; compound **5**, orange; compound **7**, magenta) bound to PDE4 (green). Zn^{2+} and OH^- ions are displayed by balls. The molecular surface of the binding site was rendered using the SYBYL implementation of MOLCAD("SYBYL", 2003) and color-coded by hydrophobicity (brown → blue: hydrophobic → hydrophilic). The view was prepared with PyMol version 0.95 (http://www.pymol.org)



Compound	Functional Group (FG)	n	FlexX score[a]	IC_{50}, nM[b]
9	Ph	1	-22.08	8,000
10	Ph	3	-20.40	>10,000

[a] FlexX score, in kJ/mol

b The IC50 was calculated by non linear regression and represents the mean value of three independent determinations. The experimental error is about 15%.

Table4.7 PDE4 inhibition of compounds **9-10**

The last two selected compounds (Table 4.7) belong to the series of 4-substituted pyridazinones, with again a phenyl ring connected via one or three carbon atoms to the C4 pyridazine atom, and predicted to interact with the above mentioned hydrophobic channel after 180° rotation of the dihedral angle linking the dimethoxyphenyl moiety to the pyridazine ring.

4.5.3 PDE4 inhibitory potency

Out of the nine synthesized compounds, five exhibit a stronger *in vitro* inhibition of bovine smooth muscle PDE4 than zardaverine **1.** Considering the very high sequence conservation among PDE4s in mammals, we can expect very similar results with the human PDE4 that has been modeled in the current study. A significant enhancement of enzymatic inhibition is observed by adding a phenyl ring at various distance (1 to 6 carbon atoms) from the N2-pyridazine atom, the best inhibitor being compound **5,** bearing a phenylpentyl substituent and exhibiting a subnanomolar IC_{50} (Table 4.6). A potent inhibitor (compound **7**, IC_{50} = 20 nM) combining a hexyl linker and a primary

amine functional group was also discovered (Table 4.6). Surprisingly, the corresponding amidine **8** was found much less potent. Last, the two 4-substituted pyridazinones **9** and **10** (Table 4.7) were much less active than the corresponding 2-substituted dihydropyridazinones **3** and **4** and show only micromolar affinities for the PDE4.

4.6 Discussion

We herewith present a simple and straightforward method to design small combinatorial libraries while optimizing size versus diversity. A key advantage of the SLF approach is that diversity is encoded in a simple pharmacophoric space. The method relies on three building blocks that are all under the control of the user: an invariable scaffold, a variable linker, a variable functional group. The linker has the simple role of varying the distance between the core of the molecule (the scaffold) and a few functional groups carefully selected to cover all possible intermolecular interactions. Therefore, additional interactions may be gained either locally or at a remote site within a single round of library design.

The concept of enumerating combinatorial libraries by assembling building elements (scaffolds, ring systems, linkers, building blocks) has been recently described in several methods (Aronov & Bemis, 2004; Schuller *et al.*, 2003). COREGEN (Aronov & Bemis, 2004) is a fragment-based design method for assembling linkers and rings frequently occurring in known kinase inhibitors. By decomposing a molecule into R ring-building blocks and L linkers, a combinatorial library of R * L * P compounds (P being the number of positions that can be derivatized) is generated. SMILIB (Schuller *et al.*, 2003) assembles scaffolds, linkers and functional groups in product-space. However, the latter do not necessarily describe the pharmacophoric space well. Thus the combinatorial assembly is unrestricted and generates very large libraries unless subset selection

according to user-defined queries (e.g. drug-likeness) is performed to reduce the size of the library. A basic difference with the above-mentioned methods is that our approach does not fully optimize the starting lead in a single round, but ensures at each design step, a significant affinity gain by an incremental optimization of both the linker and the functional group. Once a linker-functional group combination has been identified in the first design round, both building blocks may be optimized in a second round to fine tune the best possible combination by exploring the local chemical space around the selected building blocks.

The SLF approach has been applied to the structure-based optimization of a known micromolar PDE4 inhibitor, zardaverine. By carefully selecting, in agreement with medicinal chemists, a limited number of linkers and functional groups, a zardaverine-focused library of 320 compounds has been enumerated and docked to the X-ray structure of PDE4 according to settings previously known to reproduce the X-ray pose of zardaverine in the enzyme. The catechol substructure in both zardaverine and rolipram, two known PDE4 inhibitors, are positioned in the most hydrophobic sub-pocket of the active site between Ile638 and Phe674 (Huai, Wang *et al.*, 2003; Lee *et al.*, 2002). Both ether oxygen atoms are involved in bifurcated hydrogen bonds to the side-chain of Gln671 (Figure 4.3 A). The same binding mode is observed for the very large majority of compounds in the virtual library (Figure 4.2) indicating that unrestrained FlexX docking is able to properly locate most of these ligands in the protein active site. The advantage of the herein proposed library design is exemplified by a series of zardaverine-based compounds for which additional interactions with remote pockets have been disclosed (Figures 4.2, 4.3). N2-substituted pyridazines **3**-**6** (Table 4.6) interact with hydrophobic pocket A (His462, His506, Phe642) through a phenyl functional group that can be linked to the pyridazine core by polymethylene spacers of

various lengths (1 to 6 carbon atoms; Figure 4.3 B). Compound **7** discloses another remote polar subsite B (Glu532, Asp574) through a primary amine 6 carbon atoms from the N2 pyridazine atom. The folded conformation of the spacer however suggests that shorter polymethylene spacers (e.g. butyl) may be appropriate as well.

In the current study, the design effort has only been focused towards potency for a given PDE. However, the lack of selectivity of most PDE inhibitors towards other PDE isoforms (Bischoff, 2004) and genes probably account for observed side effects such as emesis and arrhythmia, which dramatically restricts the clinical development of PDE4 inhibitors as anti-inflammatory compounds(Lipworth, 2005). Thus, it may be valuable to identify specific 3-D features in the selected PDE target to direct the design of potent and selective inhibitors. A systematic survey of the amino acid sequence of 21 human PDEs in the UniProt database (Apweiler *et al.*, 2004) and a subsequent multiple alignment indicates that the acidic subsite B is fully conserved in all PDEs (see Supporting Information by (Manallack *et al.*, 2005)). We therefore anticipate that the additional interactions gained by compound **7** will not affect its selectivity profile with respect to zardaverine. Conversely, hydrophobic subsite A targeted by compounds **3-6** shows some degree of variation among PDEs, especially at Ser510 and Cys660, as exemplified by the recently-described crystal structure of phosphodiesterase 4B in complex with (S)- and (R)-rolipram (Xu *et al.*, 2004). The present data can thus be used to try and design potent and selective PDE inhibitors by simultaneously targeting the two remote subsites and directing the interaction with variable residues of the hydrophobic pocket A.

4.7 Conclusion

The SLF method allows medicinal chemists to use their knowledge in an iterative "design-synthesize-screen-analyze" process. There are still certain shortcomings to the current implementation of the method and many further refinements are possible. For example, a user-defined 3-D conformation (e.g. X-ray conformation) of a scaffold could be selected as a rigid body to avoid incorrect conformer generation from a simple 1D representation (i.e. complex ring systems). Furthermore, the automatic detection of symmetry centers and /or axes would avoid the enumeration of duplicates and spare an additional post-processing step. Last, a scaffold library designed from commercially-available screening collections (Baurin *et al.*, 2004) will soon enable the choice of multiple scaffolds fulfilling similarity/diversity-based queries. The SLF method can be used for a fast lead optimization consisting of the systematic search of remote subpockets in the neighborhood of a given scaffold by optimizing both the length of the necessary linker and the nature of the terminal functional groups.

Bibliography

Alper, J. (1994). Drug discovery on the assembly line. *Science, 264*, 1399-1401.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res, 32 Database issue*, D115-119.

Aronov, A. M., & Bemis, G. W. (2004). A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins, 57*, 36-50.

Baurin, N., Baker, R., Richardson, C., Chen, I., Foloppe, N., Potter, A., et al. (2004). Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J Chem Inf Comput Sci, 44*, 643-651.

Beavers, M. P., & Chen, X. (2002). Structure-based combinatorial library design: methodologies and applications. *J. Mol. Graph. Model., 20*, 463-468.

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem., 39*, 2887-2893.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*, 235-242.

Bischoff, E. (2004). Potency, selectivity, and consequences of nonselectivity of PDE inhibition. *Int J Impot Res, 16 Suppl 1*, S11-14.

Bissantz, C., Folkers, G., & Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem, 43*, 4759-4767.

Bradley, E. K., Miller, J. L., Saiah, E., & Grootenhuis, P. D. J. (2003). Informative library design as an efficient strategy to identify and optimize leads: Application to cyclin-dependent kinase 2 antagonists. *Journal of Medicinal Chemistry*, 4360-4364.

Card, G. L., Blasdel, L., England, B. P., Zhang, C., Suzuki, Y., Gillette, S., et al. (2005). A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. *Nat. Biotechnol., 23*, 201-207.

De Julian-Ortiz, J. V. (2001). Virtual Darwinian drug design: QSAR inverse problem, virtual combinatorial chemistry, and computational screening. *Combinatorial Chemistry and High Throughput Screening*, 295-310.

De Julian-Ortiz, J. V., Galvez, J., Munoz-Collado, C., Garcia-Domenech, R., & Gimeno-Cardona, C. (1999). Virtual combinatorial syntheses and computational screening of new potential anti-herpes compounds. *Journal of Medicinal Chemistry*, 3308-3314.

De Laet, A., Hehenkamp, J. J. J., & Wife, R. L. (2000). Finding drug candidates in virtual and lost/emerging chemistry. *J. Heterocycl. Chem.*, 669-674.

Galvez, J., Garcia-Domenech, R., de Julian-Ortiz, J. V., & Soler, R. (1995). Topological approach to drug design. *J. Chem. Inf. Comput. Sci., 35*, 272-284.

Heizmann, G., & Eberle, A. N. (1997). Xanthines as a scaffold for molecular diversity. *Mol. Divers., 2*, 171-174.

Huai, Q., Colicelli, J., & Ke, H. (2003). The crystal structure of AMP-bound PDE4 suggests a mechanism for phosphodiesterase catalysis. *Biochemistry, 42*, 13220-13226.

Huai, Q., Wang, H., Sun, Y., Kim, H. Y., Liu, Y., & Ke, H. (2003). Three-dimensional structures of PDE4D in complex with roliprams and implication on inhibitor selectivity. *Structure (Camb), 11*, 865-873.

Jain, A. N. (2003). Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J Med Chem, 46*, 499-511.

Janvier, P., Sun, X., Bienayme, H., & Zhu, J. (2002). Ammonium chloride-promoted four-component synthesis of pyrrolo[3,4-b] pyridin-5-one. *J. Am. Chem. Soc.*, 2560-2567.

Kellenberger, E., Rodrigo, J., Muller, P., & Rognan, D. (2004). Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins, 57*, 225-242.

Keravis, T. M., Wells, J. N., & Hardman, J. G. (1980). Cyclic nucleotide phosphodiesterase activities from pig coronary arteries. Lack of interconvertibility of major forms. *Biochim Biophys Acta, 613*, 116-129.

Kuznetsov, V., Gorohovsky, S., Levy, A., Meir, S., Shkoulev, V., Menashe, N., et al. (2004). Approaches for introducing high molecular diversity in scaffolds: Fast parallel synthesis of highly substituted 1H-quinolin-4-one libraries. *Mol. Divers.*, 437-448.

Langer, T., & Wolber, G. (2004). Virtual combinatorial chemistry and in silico screening: Efficient tools for lead structure discovery? *Pure Appl. Chem., 76*, 991-996.

Lee, M. E., Markowitz, J., Lee, J. O., & Lee, H. (2002). Crystal structure of phosphodiesterase 4D and inhibitor complex. *FEBS Lett., 530*, 53-58.

Lewell, X. Q., Jones, A. C., Bruce, C. L., Harper, G., Jones, M. M., McLay, I. M., et al. (2003). Drug rings database with Web interface. A tool for identifying alternative chemical rings in lead discovery programs. *J. Med. Chem.*, 3257-3274.

Lipkus, A. H. (2001). Exploring chemical rings in a simple topological-descriptor space. *J. Chem. Inf. Comput. Sci.*, 430-438.

Lipworth, B. J. (2005). Phosphodiesterase-4 inhibitors for asthma and chronic obstructive pulmonary disease. *Lancet, 365*, 167-175.

Lugnier, C. (2005). Cyclic nucleotide phosphodiesterase (PDE) superfamily: A new target for the development of specific therapeutic agents. *Pharmacol. Ther.*

Lugnier, C., & Komas, N. (1993). Modulation of vascular cyclic nucleotide phosphodiesterases by cyclic GMP: role in vasodilatation. *Eur Heart J, 14 Suppl I*, 141-148.

Lugnier, C., Schoeffter, P., Le Bec, A., Strouthou, E., & Stoclet, J. C. (1986). Selective inhibition of cyclic nucleotide phosphodiesterases of human, bovine and rat aorta. *Biochem Pharmacol, 35*, 1743-1751.

Manallack, D. T., Hughes, R. A., & Thompson, P. E. (2005). The next generation of phosphodiesterase inhibitors: structural clues to ligand and substrate selectivity of phosphodiesterases. *J. Med. Chem., 48*, 3449-3462.

MDL, E. San Leandro, CA 94577.

Nilakantan, R., & Nunn, D. S. (2003). A fresh look at pharmaceutical screening library design. *Drug Discovery Today, 8*, 668-672.

OEChem. (Version 1.3)(2004). OpenEye Scientific Software, Inc., Santa Fe, NM, USA.

Pearlman, R. S. "Concord User's Manual" (Version v4.08): Tripos,Inc.: 1699 South Hanley Road, St. Louis, MO 63144-2319.

Rarey, M., Kramer, B., Lengauer, T., & Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology, 261*, 470-489.

Schuller, A., Schneider, G., & Byvatov, E. (2003). SMILIB: Rapid assembly of combinatorial libraries in SMILES notation. *Qsar & Combinatorial Science, 22*, 719-721.

SYBYL. (Version 6.91)(2003). St-Louis, MO: TRIPOS, Assoc., Inc.

van Almsick, M., Dolhaine, H., & Honig, H. (2000). Efficient algorithms to enumerate isomers and diamutamers with more than one type of substituent. *J. Chem. Inf. Comput. Sci., 40*, 956-966.

Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., & Taylor, R. D. (2003). Improved protein-ligand docking using GOLD. *Proteins, 52*, 609-623.

Weaver, D. C. (2004). Applying data mining techniques to library design, lead generation and lead optimization. *Curr. Opin. Chem. Biol.*, 264-270.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci., 28*, 31-36.

Wermuth, C. G. (Ed.). (2003). *The Practice of Medicinal Chemistry* (Second ed.): Academic Press.

Wright, T., Gillet, V. J., Green, D. V. S., & Pickett, S. D. (2003). Optimizing the size and configuration of combinatorial libraries. *J. Chem. Inf. Comput. Sci., 43*, 381-390.

Xu, R. X., Hassell, A. M., Vanderwall, D., Lambert, M. H., Holmes, W. D., Luther, M. A., et al. (2000). Atomic structure of PDE4: insights into phosphodiesterase mechanism and specificity. *Science, 288*, 1822-1825.

Xu, R. X., Rocque, W. J., Lambert, M. H., Vanderwall, D. E., Luther, M. A., & Nolte, R. T. (2004). Crystal structures of the catalytic domain of phosphodiesterase 4B complexed with AMP, 8-Br-AMP, and rolipram. *J. Mol. Biol., 337*, 355-365.

Chapter 5

5. Conclusion

During this thesis work, we extended existing approaches of rational library design by a combinatorial-based strategy that combines a core structure by an exhaustive variation of distance-modulated pharmacophoric features. This strategy originates from the practice of medicinal chemistry and conceals it with combinatorial chemistry principles. Thereby, we achieve maximum diversity while keeping the size of the library low.

At first, the assembling rules were investigated and defined to be the inverse of the hierarchical fragmentation method described by Bemis and Murcko (Bemis & Murcko, 1996; Bemis & Murcko, 1999) during their substructural analysis of 5120 marketed drugs. We dubbed it the SLF approach (Figure 5.1) because it is based on three types of molecular building blocks: the invariant scaffold (S), the distance-modulating linker (L) and variable functional group (F).

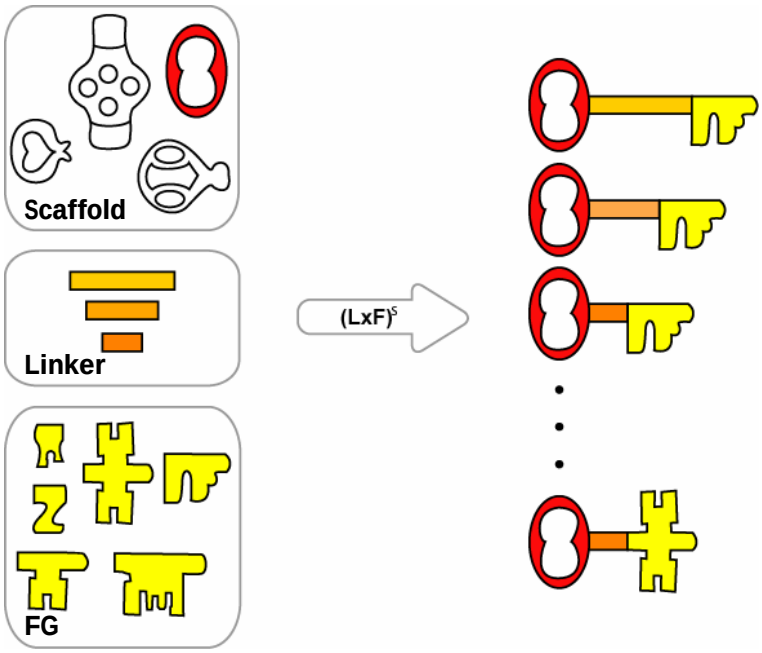


Figure 5.1 SLF illustration: a metaphor to Emil Fisher's lock –and-key concept

The implementation of this assembly rule is kept generic, i.e. a huge number of compounds could be generated, but we deliberately focused on small-sized libraries.

The careful selection of the fragments was, beneath the way of assembling the fragments, the most critical step. Hence, we created three databases for the different building blocks types. The functional groups were chosen to be simple non-overlapping representational fragments of the pharmacophoric space. The linker library is mainly composed of alkyl chains to fulfill the role as distance-modulator. For a second-round screening, ethers, amide or benzene have been foreseen. For the scaffold, the user (i.e. the medicinal chemist) has the possibility to bring in his scaffold or to browse through a scaffold library for selection. This scaffold library has been extracted from supplier screening collections. This has the advantage that the scaffolds are synthetically tractable. Starting initially from 2.4 million compounds provided by 17 commercially available screening collections, we ended up with a scaffold library of 21 393 non-redundant scaffolds computed by MCS detection algorithms. While constructing the scaffold library, we developed a work flow including two novel diversity metrics in order to evaluate the diversity of the different commercially available screening collections. These two metrics, named NC50C and PC50C, represent the number and the percentage of scaffolds accounting for 50% of classified compounds, respectively. The latter permitted to partition these collections in four categories corresponding to their origin. Moreover, in addition to dispose of a scaffold library required for the SLF approach, a cherry-picked probe library or a fragment library can be constructed.

In order to demonstrate the strategy, we generated 320 molecules out of four scaffolds derived from zardaverine, a known PDE4 inhibitor, with all possible combinations of five linkers and sixteen functional groups. Subsequently, these molecules have been docked into to X-ray structure of PDE4 and ranked by predicted binding probability.

Nine compounds had been selected for subsequent synthesis and assaying. Of these, five compounds had a higher affinity then the reference compound. Within a single-screening round, the binding affinity had been increased by 900-fold.

Further application cases will be implemented to confirm the ability of the SLF approach to accelerate the drug design process.

Bibliography

Bemis, G. W., & Murcko, M. A. (1996). The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem., 39*, 2887-2893.

Bemis, G. W., & Murcko, M. A. (1999). Properties of known drugs. 2. Side chains. *Journal of Medicinal Chemistry, 42*, 5095-5099.

Annexes

Annex 1

Supporting Information for:

Chapter 3 Assessing the Scaffold Diversity of Screening Libraries

Contents:

Chart A. filtering rules in OpenEye's Filter program (filter.txt configuration file)

```
MIN_MOLWT                   250       "Minimum molecular weight"
MAX_MOLWT                   500       "Maximum molecular weight"

MIN_SSSR_RINGS        0                 "Minumum number of SSSR rings"
MAX_SSSR_RINGS        7                 "Maximum number of SSSR rings"

MAX_RING_SIZE         7                 "Maximum size of any SSSR ring"

MIN_CARBONS           5                 "Minimum number of carbons"
MIN_HETEROATOMS       2                 "Minimum number of heteroatoms"
MIN_Het_C_Ratio       0.10          "Minimum heteroatom to carbon ratio"
MAX_Het_C_Ratio       1.0               "Maximum heteroatom to carbon ratio"

#count ring degrees of freedom = (#BondsInRing) - 4 - (RigidBondsInRing) -
(BondsSharedWithOtherRings)
#must be >= 0, from JCAMD 14:251-265,2000.
ADJUST_ROT_FOR_RING_TRUE                  "BOOLEAN for weather to estimate
degrees of freedom in rings"
#ADJUST_ROT_FOR_RING_FALSE

MIN_ROT_BONDS         0                 "Minimum number of rotatable bonds"
MAX_ROT_BONDS         15                "Maximum number of rotatable bonds"

MIN_RIGID_BONDS       0                 "Minimum number of rigid bonds"
MAX_RIGID_BONDS       50                "Maximum number of rigid bonds"

MIN_HBOND_DONORS      0                 "Minimum number of hydrogen-bond
donors"
MAX_HBOND_DONORS      5                 "Minimum number of hydrogen-bond
donors"
```

MIN_HBOND_ACCEPTORS 0 "Minimum number of hydrogen-bond acceptors"
MAX_HBOND_ACCEPTORS 10 "Minimum number of hydrogen-bond acceptors"

MIN_COUNT_FORMAL_CRG 0 "Minimum number formal charges"
MAX_COUNT_FORMAL_CRG 3 "Maximum number of formal charges"
MIN_SUM_FORMAL_CRG -2 "Minimum sum of formal charges"
MAX_SUM_FORMAL_CRG 2 "Maximum sum of formal charges"

MIN_XLOGP -2.0 "Minimum XLogP"
MAX_XLOGP 6.0 "Maximum XLogP"

MIN_2D_PSA 0.0 "Minimum 2-Dimensional (SMILES) Polar Surface Area"
MAX_2D_PSA 140.0 "Maximum 2-Dimensional (SMILES) Polar Surface Area"

ALLOWED_ELEMENTS H,C,N,O,F,S,P,Cl,Br,I
ELIMINATE_METALS
Sc,Ti,V,Cr,Mn,Fe,Co,Ni,Cu,Zn,Y,Zr,Nb,Mo,Tc,Ru,Rh,Pd,Ag,Cd

#acceptable molecules must have <= instances of each of the patterns below

#specific, undesirable functional groups
RULE 0 Carbazides
RULE 0 Acid_anhydrides
RULE 0 Pentafluorophenyl_esters
RULE 0 Paranitrophenyl_esters
RULE 0 HOBT_esters

RULE 0 Triflates
RULE 0 Lawesson_s_reagent
RULE 0 Phosphoramides
RULE 0 Aromatic_azides
RULE 0 Beta_carbonyl_quart_nitrogen
RULE 0 Acylhydrazide
RULE 0 Quarternary_C_Cl_I_P_or_S
RULE 0 Phosphoranes
RULE 0 Chloramidines
RULE 0 Nitroso
RULE 0 P_S_Halides
RULE 0 Carbodiimide
RULE 0 Isonitrile
RULE 0 Triacyloxime
RULE 0 Cyanohydrins
RULE 0 Acyl_cyanides
RULE 0 Sulfonyl_cyanides
RULE 0 Cyanophosphonates
RULE 0 Azocyanamides
RULE 0 Azoalkanals
RULE 0 Polyenes
RULE 0 Saponin_derivatives
RULE 0 Cytochalasin_derivatives
RULE 0 Cycloheximide_derivatives
RULE 0 Monensin_derivatives
RULE 0 Cyanidin_derivatives
RULE 0 Squalestatin_derivatives

#functional groups which often eliminate compounds from consideration
RULE 0 acid_halide
RULE 0 aldehyde
RULE 0 alkyl_halide
RULE 0 anhydride
RULE 0 azide

```
RULE   0     azo
RULE   0     di_peptide
RULE   0     long_aliphatic_chain              //(>7 atoms)
RULE   0   michael_acceptor
RULE   0     beta_halo_carbonyl
RULE   0   nitro
RULE   0     peroxide
RULE   0     phosphonic_acid
RULE   0     phosphonic_ester
RULE   0     phosphoric_acid
RULE   0     phosphoric_ester
RULE   0     sulfonic_acid
RULE   0     sulfonic_ester
RULE   0   triphenyl_phosphene
RULE   0     epoxide
RULE   0     hetero_hetero
RULE   0     sulfonyl_halide
RULE   0     halopyrimidine
RULE   0     perhalo_ketone
RULE   0     methyl_ketone
RULE   0     aziridine
RULE   0     imine
RULE   0     oxalyl
```

#the dye group includes a set of patterns which describe all cpds with colors in their names from the ACD98.2

```
RULE   0     dye
```

#functional groups which are allowed, but may not be wanted in high quantities
#common functional groups

```
RULE   6     alcohol
```

RULE 8 alkene
RULE 4 amide
RULE 4 amino_acid
RULE 4 amine
RULE 4 primary_amine
RULE 4 secondary_amine
RULE 4 tertiary_amine
RULE 4 carboxylic_acid
RULE 6 halide
RULE 1 iodine
RULE 4 ketone
RULE 4 phenol

#other functional groups

RULE 4 alkyne
RULE 4 aniline
RULE 4 aryl_halide
RULE 4 carbamate
RULE 4 ester
RULE 4 ether
RULE 4 hydrazine
RULE 4 hydrazone
RULE 4 hydroxylamine
RULE 4 nitrile
RULE 4 sulfide
RULE 4 sulfone
RULE 4 sulfoxide
RULE 4 thiourea
RULE 4 thioamide
RULE 4 thiol
RULE 4 urea

Chart B. Queries to analyze the overlap of the screening collections

In order to evaluate inter-supplier classes overlap, we have several possibilities to query the scaffold database under mysql.
The following query helped establish Table A.

```
SELECT count(*) FROM cpdsinclass as c1,cpdsinclass as c2
WHERE c1.class_id rlike 'DB1' AND c2.class_id rlike 'DB2'
AND c1.inchi=c2.inchi
AND c2.class_id <> c1.class_id;
```

Table B is derived from Table A.
Another approach for evaluation is obtained by the following query leading to Tables C and D

```
SELECT count(*) FROM supplierscaf as s1, supplierscaf as s2
WHERE s1.inchi=s2.inchi AND s1.class_id<>s2.class_id
AND s1.class_id rlike 'DB1' AND s2.class_id rlike 'DB2' ;
```

Scheme A. Database scheme with the central table "structures" containing the essential information relative to 21,393 scaffolds. The four tables in the third column contain the re-classification of the scaffold library with different parameters.

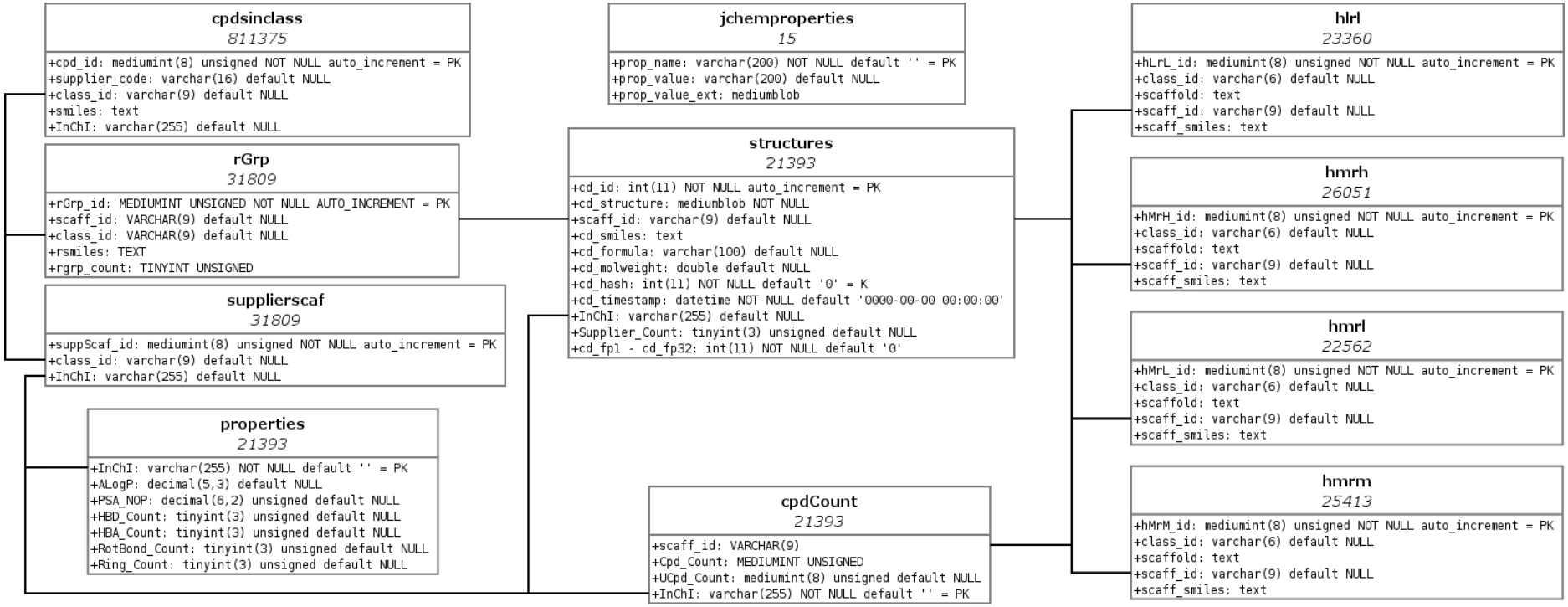


Table A. Number of classified compounds overlapping pair-wise

	ASIg	ASIp	CBG	CDIc	CDIi	CNR	CST	IBSn	IBSs	MAY	NET	SPE	TIMn	TIMs	TRI	VITs	VITt
ASIg	0	0	39 490	17 042	8 800	39	7 385	1 274	18 829	953	211	14 830	593	8 178	371	13 826	1 053
ASIp	0	0	11	237	25	0	8	2	45	31	27	916	0	27	32	69	3
CBG	39 490	11	57	21 452	14 132	62	9 884	2 259	24 504	1 156	240	23 125	642	11 189	980	28 243	1 285
CDIc	17 042	237	21 452	17	49	17	4 284	984	14 109	562	124	9 833	466	7 831	109	7 396	588
CDIi	8 800	25	14 132	49	2	24	2 592	672		329	111	4 949	129	5 016	218	4 388	262
CNR	39	0	62	17	24	0	25	16	23	34	11	16	5	29	1	13	3
CST	7 385	8	9 884	4 284	2 592	25	1	208	3 070	264	48	5 161	116	3 586	68	5 969	251
IBSn	1 274	2	2 259	984	672	16	208	11	29	93	10	401	197	428	4	298	130
IBSs	18 829	45	24 504	14 109		23	3 070	29	46	638	132	8 774	444	4 811	145	11 111	798
MAY	953	31	1 156	562	329	34	264	93	638	10	160	404	56	532	183	434	43
NET	211	27	240	124	111	11	48	10	132	160	0	83	7	82	24	87	4
SPE	14 830	916	23 125	9 833	4 949	16	5 161	401	8 774	404	83	0	319	8 923	204	9 575	550
TIMn	593	0	642	466	129	5	116	197	444	56	7	319	0	1 210	0	225	36
TIMs	8 178	27	11 189	7 831	5 016	29	3 586	428	4 811	532	82	8 923	1 210	10	136	6 003	344
TRI	371	32	980	109	218	1	68	4	145	183	24	204	0	136	0	340	0
VITs	13 826	69	28 243	7 396	4 388	13	5 969	298	11 111	434	87	9 575	225	6 003	340	14	1 599
VITt	1 053	3	1 285	588	262	3	251	130	798	43	4	550	36	344	0	1 599	0

Table B. Percentage of overlap of classified compounds of a database A with database B. For example, 4.6% of the Maybridge compounds can also be found in the Asinex Gold collection and conversely 1.1% of the Asinex Gold classified compounds are also among the Maybridge compounds.

	ASIg	ASIp	CBG	CDIc	CDIi	CNR	CST	IBSn	IBSs	MAY	NET	SPE	TIMn	TIMs	TRI	VITs	VITt
ASIg	0.0	0.0	24.4	16.3	22.3	0.8	33.9	9.2	16.9	4.6	1.5	22.7	30.6	24.5	0.8	26.5	14.7
ASIp	0.0	0.0	0.0	0.2	0.1	0.0	0.0	0.0	0.0	0.1	0.2	1.4	0.0	0.1	0.1	0.1	0.0
CBG	46.2	0.0	0.0	20.5	35.9	1.3	45.4	16.3	22.0	5.6	1.7	35.4	33.1	33.5	2.1	54.1	17.9
CDIc	19.9	0.3	13.3	0.0	0.1	0.4	19.7	7.1	12.6	2.7	0.9	15.1	24.0	23.4	0.2	14.2	8.2
CDIi	10.3	0.0	8.7	0.0	0.0	0.5	11.9	4.8	0.0	1.6	0.8	7.6	6.6	15.0	0.5	8.4	3.7
CNR	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.1	0.0	0.2	0.1	0.0	0.3	0.1	0.0	0.0	0.0
CST	8.6	0.0	6.1	4.1	6.6	0.5	0.0	1.5	2.8	1.3	0.3	7.9	6.0	10.7	0.1	11.4	3.5
IBSn	1.5	0.0	1.4	0.9	1.7	0.3	1.0	0.1	0.0	0.4	0.1	0.6	10.1	1.3	0.0	0.6	1.8
IBSs	22.0	0.1	15.1	13.5	0.0	0.5	14.1	0.2	0.0	3.1	0.9	13.4	22.9	14.4	0.3	21.3	11.1
MAY	1.1	0.0	0.7	0.5	0.8	0.7	1.2	0.7	0.6	0.0	1.1	0.6	2.9	1.6	0.4	0.8	0.6
NET	0.2	0.0	0.1	0.1	0.3	0.2	0.2	0.1	0.1	0.8	0.0	0.1	0.4	0.2	0.1	0.2	0.1
SPE	17.3	1.3	14.3	9.4	12.6	0.3	23.7	2.9	7.9	2.0	0.6	0.0	16.4	26.7	0.4	18.3	7.7
TIMn	0.7	0.0	0.4	0.4	0.3	0.1	0.5	1.4	0.4	0.3	0.1	0.5	0.0	3.6	0.0	0.4	0.5
TIMs	9.6	0.0	6.9	7.5	12.7	0.6	16.5	3.1	4.3	2.6	0.6	13.7	62.3	0.0	0.3	11.5	4.8
TRI	0.4	0.0	0.6	0.1	0.6	0.0	0.3	0.0	0.1	0.9	0.2	0.3	0.0	0.4	0.0	0.7	0.0
VITs	16.2	0.1	17.5	7.1	11.1	0.3	27.4	2.1	10.0	2.1	0.6	14.7	11.6	18.0	0.7	0.0	22.3
VITt	1.2	0.0	0.8	0.6	0.7	0.1	1.2	0.9	0.7	0.2	0.0	0.8	1.9	1.0	0.0	3.1	0.0

Table C. Number of overlapping classes by pair-wise comparison

	ASIg	ASIp	CBG	CDIc	CDIi	CNR	CST	IBSn	IBSs	MAY	NET	SPE	TIMn	TIMs	TRI	VITs	VITt
ASIg	6	138	1 046	543	644	45	383	110	844	239	93	836	42	583	111	726	63
ASIp	138	4	110	85	66	23	60	15	81	78	43	110	8	84	33	103	20
CBG	1 046	110	5	446	611	47	368	103	671	221	110	892	34	534	110	694	55
CDIc	543	85	446	6	146	32	201	68	458	132	64	432	30	330	50	326	48
CDIi	644	66	611	146	3	46	264	93	477	177	76	518	28	482	107	351	46
CNR	45	23	47	32	46	0	38	23	45	56	35	49	6	45	22	37	19
CST	383	60	368	201	264	38	1	54	278	143	60	320	18	295	61	344	45
IBSn	110	15	103	68	93	23	54	1	87	36	19	66	20	71	13	56	28
IBSs	844	81	671	458	477	45	278	87	6	178	82	584	29	417	71	511	52
MAY	239	78	221	132	177	56	143	36	178	1	114	228	19	191	92	174	33
NET	93	43	110	64	76	35	60	19	82	114	0	105	7	85	34	67	21
SPE	836	110	892	432	518	49	320	66	584	228	105	1	30	579	99	565	55
TIMn	42	8	34	30	28	6	18	20	29	19	7	30	0	80	7	24	7
TIMs	583	84	534	330	482	45	295	71	417	191	85	579	80	2	84	400	51
TRI	111	33	110	50	107	22	61	13	71	92	34	99	7	84	0	96	18
VITs	726	103	694	326	351	37	344	56	511	174	67	565	24	400	96	4	68
VITt	63	20	55	48	46	19	45	28	52	33	21	55	7	51	18	68	0

Table D. Percentage of of overlapping classes by pair-wise comparison of a database A with database B. For example, 15.5% of the Maybridge classes can also be found in the Asinex Gold collection and conversely 6.8% of the Asinex Gold classes are also among the Maybridge classes.

	ASIg	ASIp	CBG	CDIc	CDIi	CNR	CST	IBSn	IBSs	MAY	NET	SPE	TIMn	TIMs	TRI	VITs	VITt
ASIg	0.3	7.0	32.7	15.8	27.9	11.5	37.9	14.5	24.2	15.5	9.9	25.6	25.9	29.8	8.3	33.7	15.7
ASIp	4.0	0.2	3.4	2.5	2.9	5.9	5.9	2.0	2.3	5.1	4.6	3.4	4.9	4.3	2.5	4.8	5.0
CBG	30.0	5.6	0.2	13.0	26.5	12.0	36.4	13.6	19.2	14.3	11.7	27.4	21.0	27.3	8.2	32.2	13.7
CDIc	15.6	4.3	13.9	0.2	6.3	8.2	19.9	9.0	13.1	8.5	6.8	13.2	18.5	16.9	3.7	15.1	11.9
CDIi	18.4	3.4	19.1	4.3	0.1	11.8	26.1	12.3	13.7	11.5	8.1	15.9	17.3	24.6	8.0	16.3	11.4
CNR	1.3	1.2	1.5	0.9	2.0	0.0	3.8	3.0	1.3	3.6	3.7	1.5	3.7	2.3	1.6	1.7	4.7
CST	11.0	3.0	11.5	5.9	11.4	9.7	0.1	7.1	8.0	9.3	6.4	9.8	11.1	15.1	4.5	16.0	11.2
IBSn	3.2	0.8	3.2	2.0	4.0	5.9	5.3	0.1	2.5	2.3	2.0	2.0	12.3	3.6	1.0	2.6	7.0
IBSs	24.2	4.1	21.0	13.4	20.7	11.5	27.5	11.5	0.2	11.5	8.7	17.9	17.9	21.3	5.3	23.7	12.9
MAY	6.8	4.0	6.9	3.8	7.7	14.3	14.1	4.8	5.1	0.1	12.1	7.0	11.7	9.8	6.9	8.1	8.2
NET	2.7	2.2	3.4	1.9	3.3	9.0	5.9	2.5	2.3	7.4	0.0	3.2	4.3	4.3	2.5	3.1	5.2
SPE	23.9	5.6	27.9	12.6	22.5	12.5	31.7	8.7	16.7	14.8	11.2	0.0	18.5	29.6	7.4	26.2	13.7
TIMn	1.2	0.4	1.1	0.9	1.2	1.5	1.8	2.6	0.8	1.2	0.7	0.9	0.0	4.1	0.5	1.1	1.7
TIMs	16.7	4.3	16.7	9.6	20.9	11.5	29.2	9.4	11.9	12.4	9.0	17.8	49.4	0.1	6.3	18.6	12.7
TRI	3.2	1.7	3.4	1.5	4.6	5.6	6.0	1.7	2.0	6.0	3.6	3.0	4.3	4.3	0.0	4.5	4.5
VITs	20.8	5.2	21.7	9.5	15.2	9.5	34.0	7.4	14.6	11.3	7.1	17.3	14.8	20.4	7.2	0.2	16.9
VITt	1.8	1.0	1.7	1.4	2.0	4.9	4.5	3.7	1.5	2.1	2.2	1.7	4.3	2.6	1.3	3.2	0.0

Annex 2

SLF implementation

The reformulation of the algorithm in C++ uses the OpenEye OEChem Toolkit library.

As the implementation depends on the OpenEye OEChem Toolkit (i.e. linkable libraries), the portability depends on the latter and on the disposability of C++ compilers. In the laboratory, the program is operational under SGI/IRIX, PC/LINUX and PC/Windows with gcc as compiler.

The program is launched in a command-line manner in a console window with the following options:

Mandatory

-scaf	<SCAFFOLD-FILE>	Name of the file containing the scaffold(s)
-link	<LINKER-FILE>	Name of the file containing the linkers
-func	<FUNCTIONAL-FILE>	Name of the file containing the functional groups

Optional

-out	<OUTPUT-FILE>	Name of the output file, different molecular file formats possible
-help	HELP	Displays this guidelines

Example :

SLF_LibMaker –scaf Scaffold.smi –link AlkylLinker135.smi –func FG.smi –out Specifics.smi

Annex 3

Nature Reviews Drug Discovery. Comment on the paper by

Krier, M., Araujo-Junior, J. X., Schmitt, M., Duranton, J., Justiano-Basaran, H., Lugnier, C., et al. (2005). Design of small-sized libraries by combinatorial assembly of linkers and functional groups to a given scaffold: application to the structure-based optimization of a phosphodiesterase 4 inhibitor. J. Med. Chem., 48, 3816-3822.

Kirkpatrick, P. (2005). Medicinal chemistry - Best of both worlds? Nature Reviews Drug Discovery, 4, 540-540.

***Nature Reviews Drug Discovery*, 2005, Vol 4, Page 540-540**

Abstract

The main goal of this present work was to develop a new approach making the compromise between a minimum size and maximal molecular diversity for a compound library. This key issue in medicinal chemistry was tackled with computer assisted library design enumerating systematically molecules which acquire their structural complexity through combinations of three types of building blocks: scaffolds (S), linkers (L) and functional groups (F). A scaffold library was created from commercially available screening collections. In order to analyze the chemical diversity of these libraries, a general workflow was developed aimed at (1) identifying drug-like compounds, (2) cluster them by common substructures (scaffolds) and (3) measure the scaffold diversity encoded by each screening collection independently of its size. The combinatorial scaffold-based library concept (SLF) was illustrated by the lead optimization of a known PDE4 inhibitor, zardaverine. A library of 320 molecules was evaluated by virtual screening techniques which selected 9 compounds for synthesis and biological assay. This led to a 900 fold increase of affinity in comparison to zardaverine in one screening cycle.

Résumé

Le but principal de ce travail de thèse a été de développer une nouvelle approche cherchant le compromis entre une taille minimale et une diversité moléculaire maximale pour une chimiothèque. Cette problématique clef de la pharmacochimie a été traitée à l'aide d'une conception assistée par ordinateur énumérant systématiquement des molécules qui acquièrent leur complexité structurale par des combinaisons de trois types d'entités moléculaires : les châssis moléculaires (S), les espaceurs (L) et les groupements fonctionnels (F). Une chimiothèque de châssis moléculaires a été crée à partir de collections de criblage disponibles chez des fournisseurs. Afin d'analyser la diversité de ces collections, une chaîne de travail a été développée dans le but (1) d'identifier des composés à potentiel médicamenteux, (2) de les regrouper par sous-structures communes (châssis) et (3) de mesurer la diversité en châssis contenus dans chaque collections de criblage indépendamment de leur taille. Le concept de chimiothèque combinatoire basée sur châssis (SLF) a été illustré par l'optimisation d'une tête de série d'un inhibiteur connu de la PDE, la zardaverine. Une chimiothèque de 320 molécules a été évaluée par des techniques de criblage virtuel sélectionnant ainsi 9 composés pour la synthèse et le test biologique. Ceci a mené à une augmentation de l'affinité par 900 fois par rapport à la zardaverine en un seul cycle de criblage.

Mots-clefs

Chimiothèque, Diversité moléculaire, Phosphodiesterase, Chemoinformatique