

Validation Study on an Information Driven Library Design Strategy

Master of Science Thesis in the International Master's Degree Programme in Bioinformatics

Sabbath Marchend

The Author grants to Chalmers University of Technology and University of Gothenburg the non-exclusive right to publish the Work electronically and in a non-commercial purpose make it accessible on the Internet.

The Author warrants that he/she is the author to the Work, and warrants that the Work does not contain text, pictures or other material that violates copyright law.

The Author shall, when transferring the rights of the Work to a third party (for example a publisher or a company), acknowledge the third party about this agreement. If the Author has signed a copyright agreement with a third party regarding the Work, the Author warrants hereby that he/she has obtained any necessary permission from this third party to let Chalmers University of Technology and University of Gothenburg store the Work electronically and make it accessible on the Internet.

Validation Study on an Information Driven Library Design Strategy

Sabbath Marchend

© Sabbath Marchend, September 2009.

Examiner: Graham J. L. Kemp, Ph. D

Department of Computer Science and Engineering
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Cover: Illustration of combinatorial library design (Adapted from [24])

Department of Computer Science and Engineering
Göteborg, Sweden September 2009

Thesis for the Degree of Master of Science in Bioinformatics

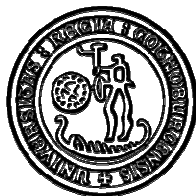
Validation Study on an Information Driven Library Design Strategy

Sabbath Marchend

Supervisor: Hongming Chen, Ph.D
Computational Chemistry, DECS Global Compound Sciences
AstraZeneca, R&D
Mölndal, Sweden

Examiner: Graham J. L. Kemp, Ph.D
Associate Professor, Department of Computer Science and Engineering
Chalmers University of Technology
Göteborg, Sweden

CHALMERS | UNIVERSITY OF GOTHENBURG



AstraZeneca 

International Master's Programme in Bioinformatics
Chalmers University of Technology and University of Gothenburg
Göteborg, Sweden
December, 2008

Acknowledgements

Thanks to God, a loving Father, for His endless love and help that I can finish this thesis in time. This thesis is one of the requirements to get a degree in International Master's Programme in Bioinformatics, Chalmers University of Technology, Sweden.

I would like to thank my supervisor, Hongming Chen Ph. D, for giving me the chance to carry out a thesis work at AstraZeneca and his endless support and encouragement throughout the completion of the thesis. Without his support, this thesis would not be completed. I also want to acknowledge Dr. Graham J. L. Kemp as my examiner for giving valuable discussions during the thesis writing. His supports inspired me a lot during the time.

I would like to thank Professor Olle Nerman as the programme director of the International Master's program in Bioinformatics and all the teachers at Chalmers University of Technology and The University of Göteborg. Thanks for the number of interesting courses and seminars.

A pleasant working environment had been occurred during the six months of thesis work at AstraZeneca. It won't be happened if not because of my friendly colleagues at Computational Chemistry, DECS Global Compound Sciences, AstraZeneca R&D Mölndal, Sweden. Special thanks goes to Niklas Blomberg, "Thanks for proofreading the thesis" and to Steffan S., Engkvist, O., and Börjesson, U., "Thanks for the scientific discussions and technical helps during the time".

For the last two years of my study, I have been blessed that I have such a lovely place to live. It happened not only because of the nice place, but also because of the warm and kind environment. Therefore, I want to thank my entire Indonesian friends in Sweden for providing a wonderful environment to live. You are like a second family for me.

Constant appreciation is given to my colleagues. Thanks for the nice environment we had during our study at CHALMERS. To my friends : Mari Chikvaidze and Santosh Dahgam, "Thanks for the discussion and friendship we had during our study".

I am more than thankful to both of my parents and all of the family members. Thanks for your given supports, love and care during my study in Sweden. Your love and kindness are like candles that give me lights in walking through the world.

Once more, I want to acknowledge all the people that contributed to this thesis. Without your supports, this thesis would not be possible. In the end, I would like to dedicate this thesis as a contribution to science and people throughout the readers.

Sabbath Marchend

The fear of the LORD is the beginning
of knowledge, but fools despise
wisdom and discipline.
Proverbs 1 : 7

For wisdom will enter your heart, and
knowledge will be pleasant to your
soul. Discretion will protect you, and
understanding will guard you.
Proverbs 2 : 10-11

The fear of the LORD teaches a man
wisdom, and humility comes before
honor.
Proverbs 15 : 33

Don't seek justifications on all things.
Wisdom don't come from a
justification of the incapable, but
rather from talking less and doing
more.
USPT

WHY HAVE YOU CHOSEN ME

*Why have You chosen me,
Out of millions Your child to be?
You know all the wrong that I have done.
Oh how could You pardon me,
Forgive my iniquities,
To save me give Jesus Your son.*

*But Lord help me be,
What You want me to be.
Your Word I will strive to obey.
My life I now give, for You I will live,
And walk by Your side all the way.*

*I am amazed to know,
That a God so great could love me so,
Is willing and wanting to bless.
His love is so wonderful,
His mercy so bountiful,
I can't understand it I confess.*

Because everything is from Him, to Him and for Him only.

Abstract

A new method is introduced for performing reagent selection for chemical library design based on topological (2D) pharmacophore fingerprints. Optimal reagent selection is achieved by optimising the Shannon entropy of the 2D pharmacophore distribution for the reagent set. The method, termed ProSAR, is therefore expected to enumerate compounds that could serve as a good starting point for deriving a structure activity relationship (SAR) in combinatorial library design. The main goal for current study is to validate this methodology by applying it on several library design examples where the active compounds were already known and comparing the performance of ProSAR libraries with random libraries and traditional diversity based libraries. The results show that ProSAR libraries generally have better pharmacophore coverage than libraries coming from other design strategies. The effectiveness of generating active compounds for the designed library is also evaluated by first doing a similarity search against GVKBio database with library compounds as query structures, then comparing the number of retrieved active compounds for different libraries. The results demonstrate that in most of cases, ProSAR libraries retrieve more active compounds than other libraries. The ProSAR strategy is further expanded to include product property profiles for aqueous solubility, hERG risk assessment etc. in the optimisation process so that the reagent pharmacophore diversity and the product property profile are optimised simultaneously via a genetic algorithm. The validation study results show that by using the ProSAR methodology, the designed libraries can achieve good pharmacophore coverage and product property profile simultaneously.

Table of Contents

ABSTRACT	i
TABLE OF CONTENTS	ii
LIST OF FIGURES	iv
LIST OF TABLES	v
CHAPTER 1: PROJECT BACKGROUND.....	1
1.1 Combinatorial Chemistry	1
1.2 Combinatorial Library Design Strategies	2
1.2.1 Focused Library Design	3
1.2.2 Targeted Library Design.....	3
1.2.3 Diversity Library Design.....	3
1.2.4 Reagent-based Library Design	4
1.2.5 Product-based Library Design.....	4
1.3 Molecular Diversity.....	5
1.3.1 Molecular Descriptor.....	5
1.3.1.1 Daylight Fingerprint	7
1.3.1.2 FOYFI Fingerprint.....	8
1.3.1.3 Pharmacophore Fingerprint	8
1.3.2 Weighting Scheme	9
1.3.3 Similarity Coefficient	9
1.3.3.1 Tanimoto Coefficient.....	10
1.4 Chemical Structure Representation	11
1.4.1 SMILES.....	11
1.4.1.1 SMILES Specification Rules	11
1.4.1.1.1 Atoms	12
1.4.1.1.2 Bonds.....	12
1.4.1.1.3 Branches	12
1.4.1.1.4 Cyclic Structures	13
1.4.1.1.5 Disconnected Structures	13
1.4.2 SMARTS.....	14
1.4.2.1 Atomic Primitives.....	14
1.4.2.2 Bond Primitives	15
1.5 Structure and Activity Relationship	15
1.6 Aims and Objectives for the Project.....	16
CHAPTER 2: MATERIALS AND METHODS	17
2.1 ProSAR Methodology	17
2.1.1 Identification and Encoding of the Pharmacophore Fingerprint	17
2.1.2 Optimisation in Pharmacophore Space for ProSAR.....	18
2.1.3 Optimisation of the Pharmacophore Entropy and the Library Property Profile	19
2.1.4 Diversity Based Library Design Strategy	20
2.2 Library Examples	20
2.3 Reagent Preparation.....	23
2.4 Computational Procedure	24
2.4.1 ProSAR Library Design	24
2.4.2 Random Library Design	25

2.4.3	Diversity Library Design.....	25
2.4.4	Preparation of Validation Set	25
2.4.5	Similarity Search	26
CHAPTER 3: RESULTS AND DISCUSSIONS		27
3.1	Comparison of Pharmacophore Coverage for Different Library Design Strategies.....	27
3.2	Comparison of Retrieved Active Compounds for Different Library Design Strategies.....	32
3.3	Extension of ProSAR to Include Property Profile Optimisation	34
CHAPTER 4: CONCLUSIONS		36
BIBLIOGRAPHY		37
APPENDIX B		46
APPENDIX C		51
APPENDIX D		54
APPENDIX E		57

List of Figures

Figure 1.1:	Reaction scheme showing the reaction of compound A with compound B to form compound AB. This method is the traditional synthesis in which only one compound can be made at a time.	1
Figure 1.2:	Reaction scheme showing the combinatorial chemistry concept. It can be seen that from m range of analogues A's and n range of analogues B's a total of (m×n) possible compounds can be synthesized at a time.	2
Figure 1.3:	Some examples of molecular descriptors and their classification calculated from 1D, 2D and 3D molecular structure [24].	6
Figure 1.4:	Identification of pharmacophore and its fingerprint. Each feature in a molecule is identified first. After that, string of bins is generated based on it. Features that are absent and present are denoted with 0 and 1, respectively.	9
Figure 1.5:	Illustration of braches in SMILES representation. In the first row, graph representations of molecules are shown, followed by its corresponding SMILES notation in the second row (Adapted from Daylight Theory Manual [28]).	13
Figure 1.6:	Deriving SMILES notation for a cyclic structure from a cyclohexane (Taken from Daylight Theory Manual [28]).	13
Figure 1.7:	Example of representing a disconnected structure in sodium phenoxide (Taken from Daylight Theory Manual [28]).	13
Figure 2.1:	Identification and encoding of a reagent pharmacophore fingerprint.	18
Figure 2.2:	Combinatorial library example for 1D_Lib1.	20
Figure 2.3:	Reaction scheme for 1D_Lib1.	21
Figure 2.4:	Combinatorial library example for 2D_Lib1.	21
Figure 2.5:	Reaction scheme for 2D_Lib1.	21
Figure 2.6:	Combinatorial library example for 2D_Lib2.	22
Figure 2.7:	Reaction scheme for 2D_Lib2.	22
Figure 2.8:	Combinatorial library example for 2D_Lib3.	22
Figure 2.9:	Reaction scheme for 2D_Lib3.	22
Figure 2.10:	Combinatorial library example for 2D_Lib4.	23
Figure 2.11:	Reaction scheme for 2D_Lib4.	23
Figure 2.12:	Combinatorial library example taking into accounts both pharmacophore entropy and property profile.	23
Figure 3.1:	Comparison of pharmacophore coverage in 1D_Lib1 library example taken from Appendix A and Appendix B. (a) 5-pharmacophore types fingerprint (b) 6-pharmacophore types fingerprint.	28
Figure 3.2:	Comparison of pharmacophore (5 pharmacophore type) coverage of designed libraries in 2D_Lib1 example on (a) R1 reagents (b) R2 reagents.	31
Figure 3.3:	Comparison of pharmacophore (5 pharmacophore type) coverage in 2D heat maps for 2D_Lib1 examples. (a) PSAR library compared with validation set. (b) Random library compared with validation set. (c) Diversity library compared with validation set.	31
Figure 3.4:	Similarity search examples. The molecules on left hand side are retrieved active compounds and the molecules on right hand side are query molecules.	33

List of Tables

Table 1.1:	SMARTS Atomic Primitives (Taken from Daylight Theory Manual [28])	14
Table 1.2:	SMARTS Bond Primitives (Taken from Daylight Theory Manual [28])	15
Table 2.1:	Reagents used in all library examples	24
Table 2.2:	The number of retrieved active compounds from GVKBio database	25
Table 3.1:	Shannon entropy values (SE) of selected reagents from all library examples.....	29
Table 3.2:	The number of retrieved active compounds from GVKBio database	32
Table 3.3:	Results for the GA optimized libraries using 5 pharmacophore types	34
Table 3.4:	Results for the GA optimized libraries using 6 pharmacophore types	35

Chapter 1

Project Background

1.1 Combinatorial Chemistry

In the past, the approach of making compounds was the traditional synthetic approach where compounds were made one by one by reacting a particular reaction at a time. This approach is time consuming and inefficient since only one compound can be made from a particular reaction. For example, compound A which reacted with compound B will give a product AB as shown in Figure 1.1. However, during the early 1990s this paradigm of chemical synthesis had been changed completely [1-3]. It changed from the old fashioned traditional synthesis to a modern method where a huge amount of compounds can be made at the same time. This new methodology is known as the combinatorial chemistry.

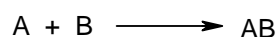


Figure 1.1: Reaction scheme showing the reaction of compound A with compound B to form compound AB. This method is the traditional synthesis in which only one compound can be made at a time.

Combinatorial chemistry had been developed in industry in early 1990s, but its origin can be traced back as early as 1960s when a research group from Rockefeller University [4-7] started the investigation of solid-stated synthesis of peptides. During their investigation, it was realized that synthesizing peptides were difficult and time consuming. So they proposed an approach of synthesizing peptides in a more efficient way by having peptides assembly attached to a solid support. Their method was then developed further by Geysen *et al.* [8] in which arrays of peptides was synthesized on pin-shaped solid supports in the 1980s. Since then the idea of combinatorial chemistry has become popular.

The concept of combinatorial chemistry is the parallel synthesis of compounds from a large range of analog reagents under the same reaction conditions and the same reaction vessels [3]. This can be described as follow. For two sets range of analog reagents $[A_1, A_2, A_3, \dots, A_m]$ and $[B_1, B_2, B_3, \dots, B_n]$, the total combination when both analog reagents are reacted will be the products of $m \times n$ compounds. By performing all these combination at the same time, it is possible to speed up the process of compounds production since numerous compounds can be made in a time. Figure 1.2 shows the illustration of combinatorial chemistry concept.

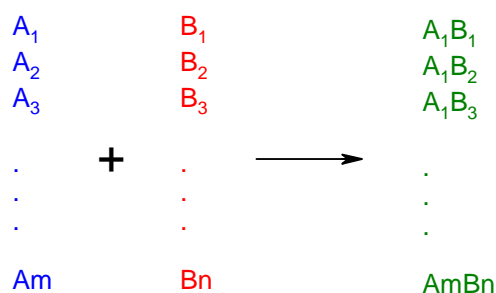


Figure 1.2: Reaction scheme showing the combinatorial chemistry concept. It can be seen that from m range of analogues A's and n range of analogues B's a total of $(m \times n)$ possible compounds can be synthesized at a time.

The key advantage of combinatorial chemistry is the massive production of compounds. When it is combined with robotics and high throughput screening (HTS) technology, hundreds or even thousands of compounds can be made and screened in a relatively short time [9]. This may lead to a rapid discovery of bioactive compounds and an increased chance of identifying lead compounds in the drug discovery process. This powerful method of combinatorial chemistry has dramatically changed drug discovery paradigm of pharmaceutical industries [10-13].

1.2 Combinatorial Library Design Strategies

Combinatorial library design plays an important role in the discovery of new drugs. It has become the common approach to progress hits into potential lead compounds after active compounds are identified in the early phase. By applying combinatorial library design, hundreds of compounds can be synthesized to explore the chemical space around the scaffold of the identified lead compound. For a specific chemical scaffold, millions products could be easily enumerated. But to physically synthesize and purify all possible compounds from such reaction is almost impossible since the number of compounds that chemist can afford to process is very limited [14]. Therefore, how to select a subset of reagents to make a smaller

scale of library is very important. This library sub-setting method is called the combinatorial library design strategy. Library design strategies can be classified into focused library design, targeted library design and diversity library design [15]. It can also be divided into reagent-based library design and product-based library design based on the way compounds are being selected [16].

1.2.1 Focused Library Design

A focused library is built from a lead molecule and aim for a particular target [15]. This library is designed to occupy certain chemical space around the target of interest. It has a large amount of information regarding the molecular design such as core structure and pharmacophoric elements and can be constructed in two different ways depending what is known about the target. For example, when the three-dimensional structure of a target is known, a library can be constructed to screen out molecules that can not fit into the active site of a particular target. On the other hand, if an active compound is known, a library could be enumerated to contain molecules that are similar to the known actives or a library could be designed to contain molecules that are predicted to be active and have a SAR [17].

1.2.2 Targeted Library Design

A targeted library is designed for finding lead compounds against a specific target class. This kind of library design is a broader version of a focused library in a sense that it comprises compounds which are supposed to be active against several proteins/receptors which belong to same target class. Some common target classes are G-protein-coupled-receptors (GPCR), kinases, ion-channels, proteases, etc. For these targets classes, the library is designed by taking account on information like important components of the structural space, privilege fragments or pharmacophoric elements of the chosen target family from all available literatures [15].

1.2.3 Diversity Library Design

A diversity library (or usually known as primary screening library) is a compound collection used for screening to find new leads or novel scaffolds. Diversity library design is suitable to be used when almost nothing about the target is known or when no information is available about what kind of molecules will interact with what target [15]. This library design could produce a library having large structural diversity. On the other hand, some physico-chemical property filters could be applied to filter out all molecules which don't satisfy with

properties of being a “good drug” such as the properties suggested by Lipinski’s “rule of 5” [18].

1.2.4 Reagent-based Library Design

Reagent-based library design is a way of sub-setting library directly from the reagent pool itself rather than from the products constructed by those reagents. For example, for a reaction involving aldehydes and amines, if there are 2000 molecules available for each type of reagents, in total, there will be 4×10^6 of possible products (2000×2000) enumerated from this combinatorial synthesis. Given the aim of library design is to build a set of 10^4 products. In the reagent-based design, sub-setting is independently done to the reagent pools. So, for this case, the procedure is to select 100 aldehyde reagents from 2000 available aldehydes and then repeating the same procedure for the amines. By doing this, 200 reagents (100 for aldehydes and 100 for amines) are picked out from the total of 4000 reagents (2000 from aldehydes and 2000 from amines); and obviously it is quicker than directly selecting 10^4 subset out of 4×10^6 products.

Reagent-based design has been used by chemist for many years. Its practical usage and efficiency have been the reasons why this method is favored compared to product-based design. Nevertheless, there is a limitation to this method. The natural properties of the products enumerated are sometimes not manifested even after they have been considered at the reagents level [16, 19]. Sometimes the properties of the product can be very different from the properties of their corresponding building blocks. For example, in constructing a library we may take into considerations all the “drug-likeness” properties on a building block of a molecule; however, the resulting library might not contain products which have the expected “drug-like” properties. However in the other hand, it is also possible to have a building block which is considered to be unsuitable when taken individually, but when it is combined with other reagents, it may result in products that have the “drug-like” properties.

1.2.5 Product-based Library Design

If a reagent-based library design generates a library by selecting molecules directly from the reagent sets, a product-based library design does the opposite way. It generates a library from the products rather than from the reagents. A product-based library design involves more complicated optimisation procedure compared to reagent-based library design. In product-based library design, a library is constructed first from all available reagents and then the optimal combinatorial subset is selected from the product pool. In this way, the

combinatorial characteristic of the sub-library is retained through combinatorial boundary, whereas diversity evaluations, focusing on other criteria are performed on the product side.

Gillet *et al.* [16, 20] investigated the diversity of products obtained from three different approaches. They compared the diversity of libraries from reagent-based design, product-based design and random design. It was shown that among these three different approaches, product-based library design gave higher diversity than the reagent-based design and both of these libraries are more diverse than the random library design. However, despite the fact that product-based library design gives the highest diversity, there lies an additional problem. The subset of product molecules chosen as the most diverse set probably can not be made from a simple combinatorial reaction of the reagents and will involve more reagents. Adding constraints on reagent matrix can overcome this problem.

1.3 Molecular Diversity

The main goal of combinatorial library design is to select a subset of compounds from a given library that is diverse as possible. A lot of interest has been shown in molecular diversity analysis by using different descriptors [21, 22]. The rationale lies on the assumption that maximizing diversity would result in a maximal coverage of bioactivity space, hence, maximizing the chance of finding new lead compounds. Diversity selection usually uses the similarity property concept, which means structurally similar compounds are likely to have similar properties [23]. This concept assumes that compounds which are structurally similar to a known biologically active compound are likely to exhibit the same activity.

There are three main components that are usually involved in measuring molecular similarity or diversity. They are the structural descriptors, which are used to describe the molecules; the weighting scheme, which is used to differentiate more important characteristics from less important characteristics of molecule; and the similarity coefficient, which is used to quantify the degree of similarity between pairs of molecules. Some introductions are given on these three components as follow.

1.3.1 Molecular Descriptor

The choice of descriptor is important because diversity or similarity measurement depends largely on it, since descriptor is needed to compare and assess molecules quantitatively. Before the comparison can start, the molecules needed to be converted first into other representation that could be easily measured. This could be achieved by converting the molecules into numbers, array of numbers or a vector of bit strings. Molecular descriptors

can be categorized into two broad classes: the whole molecular entity and the molecular parts. The first category is the whole molecular entity where a descriptor represents some physical property of the whole molecule. Some descriptors that can fall into this category are the clogP (a measurement of lipophilicity), CMR (calculated molar refractivity, a measure of size), several type of topological indices (for example those representing molecular shape of molecular connectivity) and functional group counts. This type of descriptor is usually in the form of one-dimension (1D) representation. The benefits of using this type of descriptor are the easiness of calculation and the additional properties information related to small molecule-macromolecule interaction. However, one disadvantage that could happen is that information regarding specific molecular features such as atom type, bond type, and connectivity is lost.

The other category of descriptor is called molecular parts. It separates molecules into parts that are considered to be structurally important and describes those parts numerically. In this way, atom types, bond types, and connectivity information can all be encoded easily. This category can be further classified into two-dimension (2D) descriptor and three-dimension (3D) descriptor which encode two and three dimensional properties of a molecule, respectively. An example of several descriptors used which are based on molecular structure is shown in Figure 1.3.

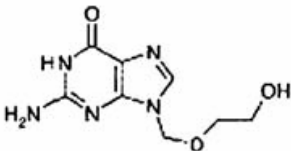

Typical Representation		Typical Descriptors
1D	<chem>C8H10N5O3</chem>	Molecular weight Atom counts
2D		Fragment counts Topological indices Connectivity
3D		Molecular surface Molecular volume Interaction energies

Figure 1.3: Some examples of molecular descriptors and their classification calculated from 1D, 2D and 3D molecular structure [24].

A fingerprint is a representation of a molecule in an abstractive way regarding its structural features. A fingerprint can be used to quantify the similarity of two molecules on one hand and also be used to eliminate molecules which are not similar to query structure on the other hand. There are several commonly used fingerprint methods, these include the MACCS structural fragment keys (MDL software) [25] and Unity [26] or Daylight fingerprint [27]. MACCS keys have been used as molecular descriptor for substructure searching. It describes molecules by assigning numbers in bit strings such as zero and one to represent the absence and presence of particular fragments of molecules. However, it is only able to describe small substructures but not the large scale ones.

The Unity and Daylight fingerprint are quite different from MACCS structural keys although they were built on the same principle concept of structural keys. The fingerprint was built to overcome the lack of generality of the MACCS structural keys. Instead of using the presence or the absence of fragments from a predefined fragment library, it uses the presence or the absence of a set of two to seven atom patterns in the molecules as a template to create a bit string. This makes the generated patterns dynamic and various depending on the individual molecule. The advantages of this fingerprinting method are avoiding missing features in the predefined structural keys and providing a more general and precise description of individual molecule since it is based on all possible atomic patterns (from two to seven atoms) of the molecules. However, the dimensionality of the constructed fingerprint could raise a problem when it is too high, especially when dealing with highly complex molecules.

1.3.1.1 Daylight Fingerprint

The Daylight fingerprint of a molecule is generated by first examining the molecule and generating the following patterns: a pattern for each atom, a pattern for two neighboring atoms with one bond length, a pattern for three neighboring atoms with two bonds length, a pattern of four neighboring atoms with three bonds length and continuing up to eight atoms with seven bonds length [28]. For example, the molecule AB=CD would generate the following patterns:

0-bond paths:	A	B	C
1-bond paths:	AB	B=C	CD
2-bond paths:	AB=C	B=CD	
3-bond paths:	AB=CD		

Then each pattern is used as a seed to generate pseudo random number (hashed) and give output as a set of bits (typically 4 or 5 bits per pattern). The bits for each pattern are combined to form fingerprint. These fingerprints are rearranged to have a fixed size (the size is large enough to represent any normal small molecule) and then folded by using logical OR

to increase the information density. So each bit in a Daylight fingerprint doesn't correspond to a particular substructure like a MACCS structural key does. In general a Daylight fingerprint has much higher bit density and hence gives greater discriminating power

1.3.1.2 FOYFI Fingerprint

The FOYFI fingerprint is an AstraZeneca in-house developed fingerprint. The FOYFI [29] fingerprint is generated in a very similar way to the Daylight fingerprint. The methodology is implemented as follows: All possible paths in a molecule are first enumerated up to a predefined length (number of bonds). Atom information (e.g. type, charge, hybridization etc.) is recorded and recursively updated via integer hashes. After that, the resulting large integer is iteratively divided by fingerprint length, and the remainder is used to set the corresponding bit. Usually several bits are set for a given path, and consequently there is no direct correspondence between a specific bit and an atom or substructure.

1.3.1.3 Pharmacophore Fingerprint

One of the descriptors commonly used is the pharmacophore fingerprint [30, 31]. This type of descriptor is built on the basis of several interactions observed in molecules such as hydrogen bonding, ionic charge interaction and hydrophobic interaction. A pharmacophore is a functional group in a small molecule or ligand which makes interaction with a specific target receptor. This interaction is usually thought to be responsible for a pharmacological action of the corresponding molecule. Pharmacophore fingerprinting is a technique of encoding molecules into string of bins that represents the molecular interactions in terms of chemical features. These features are relative and arbitrary in a sense that no strict rule is available for this, thus enabling a flexible choice of features depending on the need. In general, these features can be consisting of hydrogen bond donor (HD), hydrogen bond acceptor (HA), positive charge center (POS), negative charge center (NEG) and lipophilic groups (LIP). This kind of pharmacophore employs 5 different types of chemical feature that can be expanded further into 6 by separating the lipophilic groups (LIP) into two distinct types which are the aliphatic and aromatic lipophilic groups. Other expansion of pharmacophore features is also possible, depending the kind of interaction observed or required in a particular molecule. Figure 1.4 shows the pharmacophore concept used as descriptor.

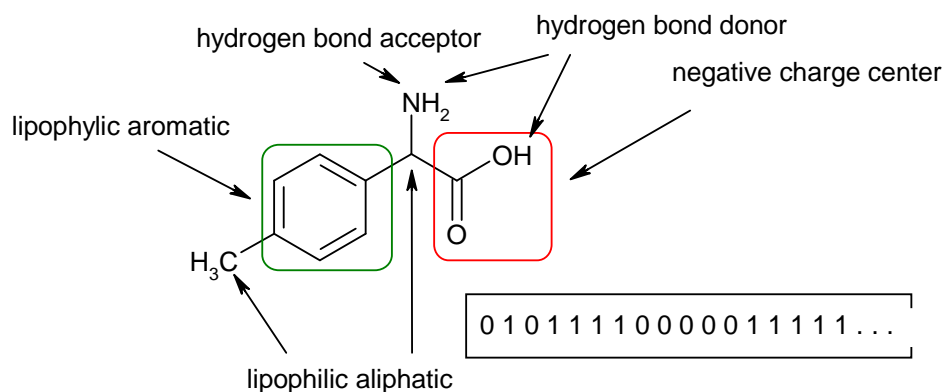


Figure 1.4: Identification of pharmacophore and its fingerprint. Each feature in a molecule is identified first. After that, string of bins is generated based on it. Features that are absent and present are denoted with 0 and 1, respectively.

1.3.2 Weighting Scheme

The second main component in determining similarity of molecules is the weighting scheme. This component assigns the degree of importance of the various characteristics of the descriptor. There are some publications [32, 33] that reported how the weighting scheme affects the utility of molecular measure. However, molecular descriptors and similarity coefficients are far more common in the literature than weighting scheme.

1.3.3 Similarity Coefficient

For comparing similarity between a pair of molecules, a numerical measure is needed. This measure (known as similarity coefficient) will provides a quantitative assessment of the degree of structural similarity between a pair of molecules.

Generally, similarity coefficients can be differentiated into two main groups. The first group is the measure of distance and the second group is the measure of direct similarity. These two groups are complementary to each other. In the measure of distance, the value of 0 would be given to molecules that are identical while in the direct similarity, a maximum value is given to identical ones. These values can be any value ranging from 0 to anything, but for convenience that range could be normalized to a coefficient ranging from 0 to 1.

Similarity coefficients are usually calculated in the following way. Consider the molecules to be compared had been encoded using a descriptor. This descriptor converts the molecules into a set of numbers representing their structural attributes. Let A be a molecule which can be described by means of a vector X_A of n attributes (Equation 1.1) such that

$$X_A = \{x_{1A}, x_{2A}, x_{3A}, \dots, x_{jA}, \dots, x_{nA}\} \quad (1.1)$$

where X_{jA} is the value of the j^{th} attributes of molecule A. The values of these attributes may be real numbers over any range or discrete values such as binary, indicating the absent or present of some particular features of the molecule. If binary numbers were used, an entry of 0 in the attributes would mean the absence of the particular feature on the molecules while an entry of 1 would mean the opposite. Once the molecules have been represented into vectors, similarity coefficients that quantify similarity between molecules are calculated by measuring distance of those respective vectors.

There are several ways to measure the similarity or distance coefficient of molecular descriptor vectors, including Euclidean distance, Hamming distance [34], Soergel distance [35] and Tanimoto coefficient [35]. Among those, the two most popular measures are the Euclidean distance and the Tanimoto coefficient. However, Tanimoto coefficient is more popular than Euclidean distance for comparing two arbitrary molecules. Euclidean distance is better only for “relative” distance comparison (i.e., the distance of molecules to the same target) but for “absolute” comparison (comparison between two independent pairs of molecules), Tanimoto coefficient performs better.

1.3.3.1 Tanimoto Coefficient

Tanimoto coefficient (also known as Jaccard coefficient) is computed as the number of attributes shared by two objects divided by the total number of their attributes. For the case of comparing molecule A and B, Tanimoto coefficient can be calculated as follow: Let “a” be the number of present features in molecule A; “b” be the number of present features in molecule B; and “c” be the number of present features in both molecule A and B. The Tanimoto coefficient is expressed as:

$$Tanimoto_{A,B} = \frac{c}{(a + b - c)} \quad (1.2)$$

Tanimoto coefficient is a straightforward calculation of similarity measurement, as it is adjusted to account the number of attributes that might be in common relative to the number attributes that are in common. As seen from Equation 1.2, Tanimoto coefficient of 1 indicates identical molecule, while Tanimoto coefficient of 0 indicates that two molecules have nothing in common.

1.4 Chemical Structure Representation

Several methods have been introduced to represent the structure of a molecule. These methods can be divided into two main groups consisting coordinate-based and graph-based. The coordinate-based is known as the connection table while the graph-based is known as the line notations.

Coordinate-based representation has several characteristics in the way it is built. It requires the x, y, z coordinate of the molecule along with the formal charges and the bond information between each atom. This information is put together into one or more tables. After that, the tables are constructed uniquely so they can provide sufficient information for a molecule to be characterized. Some examples of coordinate-based representation are the MDL “MOL” File (or called as SD file) which is developed by MDL Inc, MOL2 file format which is developed by Tripos Inc., MAE file format which is developed by Schrödinger and other formats.

Graph-based representation describes molecules based on atoms and their bonds connection. It is constructed to cover all information regarding atoms and bonds in a molecule. It is constructed by first converting a molecule into its atom type, and then adding the connection and branching information between bonded atoms. By doing this, a string of characters is assigned to denote a molecule. The most commonly used graph-based representation is the Simplified Molecular Input Line Entry System (SMILES) [28] which was developed by Daylight, Inc.

1.4.1 SMILES

SMILES is a notation used for representing a molecule or a chemical reaction. It is a graph-based representation which can provide a collection of information covering atom elements, atom connectivities and bond types in a molecule along with other information like chirality, ionization state etc. Compared to a coordinate-based representation, SMILES notation is more compact and efficient. A typical SMILES string can take up to 70% less space than a coordinate-based representation. However, the information stored in a SMILES string can be as much as the coordinate based representation except for 3D coordinates.

1.4.1.1 SMILES Specification Rules

SMILES notation can be considered as a notation of chemical language since it has rules with several vocabularies and grammars. Rules in SMILES are designed to regulate an interpretation of a graphical representation of a molecule in a standard way. There are several

terms used in SMILES. SMILES which denotes information about chirality of a molecule is called “isomeric SMILES” while those that are not are called “generic SMILES”. SMILES representation of a molecule can exist in several notations. However, they are all valid as long as they follow the rules in SMILES. SMILES notation which is unique to a specific molecule is called “unique SMILES”.

The rules in SMILES usually regulate the atoms and bonds specification of a molecule. However, there are also rules for branches, ring closures, and disconnections. Other typical rules of SMILES are the way of how it should be written. For example, SMILES notation should consist of a series of characters without spaces. The next one, hydrogen atoms may be omitted or included and aromatic structures may be specified directly or in Kekulé form. There are five generic SMILES encoding rules, these are described briefly here.

1.4.1.1.1 Atoms

There are two different ways of representing atoms in a SMILES notation. First, atoms that are categorized as metal elements should be written in square brackets [], for example [Ag] for elemental silver and [K] for elemental potassium. Atoms that are non metal elements such as B, C, N, O, P, S, F, Cl, Br, and I can be written without brackets as long as they are in the lowest normal valence. If not, they should be written inside the brackets, and any hydrogens attached to them should also be included. Lower case characters should also be used for atoms that are in the aromatic ring, while non aromatic atoms are represented by capital letters. For example aromatic carbon should be written as “c” and aliphatic carbon should be written as “C” instead. Atoms that are in the formal charges should be written in the brackets with the symbols + or – followed by an optional digit, for example positive charged aliphatic N atom can be written as [N+].

1.4.1.1.2 Bonds

Bonds in SMILES notation are represented by several symbols. Character -, =, and # are used for denoting single, double and triple bonds, respectively; while character “:” is used for aromatic bond. However, normally single and aromatic bond are omitted in the SMILES notation for simplicity.

1.4.1.1.3 Branches

Branches are described in SMILES notation using parenthesis. This is done by putting the branches in a parenthesis and merging the left side of the parenthesis to the node. Figure 1.5 illustrates the application of branches in SMILES.

CCN(CC)CC	CC(C)C(=O)O	C=CC(CCC)C(C(C)C)CCC
Triethylamine	Isobutyric acid	3-propyl-4-isopropyl-1-heptene

Figure 1.5: Illustration of braches in SMILES representation. In the first row, graph representations of molecules are shown, followed by its corresponding SMILES notation in the second row (Adapted from Daylight Theory Manual [28]).

1.4.1.1.4 Cyclic Structures

Cyclic structures can be represented in SMILES notation by breaking one of the bonds in the ring and translating it into a linear notation. Once a bond is broken, a number is assigned to the beginning and ending atoms of that bond to mark the broken bond positions. By doing in this way, it is able to find out which atoms are actually connected by a ring closure just by looking and matching the number that follows an atom. Figure 1.6 shows the illustration by using cyclohexane as an example.

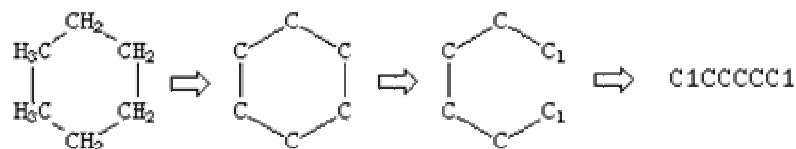


Figure 1.6: Deriving SMILES notation for a cyclic structure from a cyclohexane (Taken from Daylight Theory Manual [28]).

1.4.1.1.5 Disconnected Structures

To display a disconnected structures in a SMILES notation, a “.” (period) is needed. This period can be place anywhere on the notation as long as it is between individual structures. Regarding the placement of these individual structures, there is no order of how they should be put. An example of this can be seen in Figure 1.7.

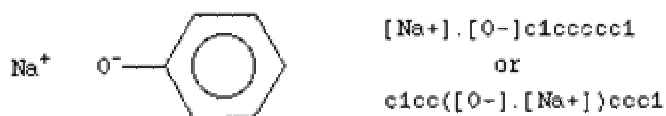


Figure 1.7: Example of representing a disconnected structure in sodium phenoxide (Taken from Daylight Theory Manual [28]).

1.4.2 SMARTS

In chemical searching, sometimes we are only interested in some particular sub-structure of a molecule rather than the whole structure. For example, we are interested in identifying all the molecules that have a phenol substructure in a database. In order to do this we need a pattern to easily identify a phenol sub-structure in a molecule and SMARTS [28] is a “language” that can help in doing this. SMARTS is an extension form of the SMILES language. This means SMILES rules are generally applied in SMARTS except for some additional symbols of atoms, bonds and logical operator. These additional rules make SMARTS notation becomes more general than a SMILES notation.

1.4.2.1 Atomic Primitives

SMARTS has additional symbols beyond the ones that are used in SMILES. These symbols describe atomic properties such as atomic symbol, charge and isotopic information. The following Table 1.1 lists all the atomic primitives used in SMARTS.

Table 1.1: SMARTS Atomic Primitives (Taken from Daylight Theory Manual [28])

Symbol	Symbol Name	Atomic Property Requirements	Default
*	Wildcard	any atom	(no default)
A	Aromatic	aromatic	(no default)
A	Aliphatic	aliphatic	(no default)
D<n>	Degree	<n> explicit connections	exactly one
H<n>	total-H-count	<n> attached hydrogens	exactly one
h<n>	implicit-H-count	<n> implicit hydrogens	at least one
R<n>	ring membership	in <n> SSSR rings	any ring atom
r<n>	ring size	in smallest SSSR ring of size <n>	any ring atom
v<n>	Valence	total bond order <n>	exactly one
X<n>	Connectivity	<n> total connections	exactly one
x<n>	ring connectivity	<n> total ring connections	at least one
-<n>	negative charge	-<n> charge	-1 charge (-- is -2, etc)
+<n>	positive charge	+<n> formal charge	+1 charge (++ is +2, etc)
#n	atomic number	atomic number <n>	(no default)
@	Chirality	anticlockwise	anticlockwise, default class
@@	Chirality	clockwise	clockwise, default class
@<c><n>	Chirality	chiral class <c> chirality <n>	(no default)
@<c><n>?	chiral or unspec	Chirality <c><n> or unspecified	(no default)
<n>	atomic mass	Explicit atomic mass	Unspecified mass

1.4.2.2 Bond Primitives

Several symbols are also introduced in SMARTS to define and characterise bonds in a substructure searching. These symbols mostly are the same as in SMILES notation. For example, the symbol -, =, # and : have the same meaning in SMARTS and SMILES. Table 1.2 lists all the bond symbols used in SMARTS with their corresponding meaning.

Table 1.2: SMARTS Bond Primitives (Taken from Daylight Theory Manual [28])

Symbol	Atomic Property Requirements
-	single bond (aliphatic)
/	directional bond "up"
\	directional bond "down"
/?	directional bond "up or unspecified"
\?	directional bond "down or unspecified"
=	double bond
#	triple bond
:	aromatic bond
~	any bond (wildcard)
@	any ring bond

1.5 Structure and Activity Relationship

Structure and activity relationship (SAR) is the approach of investigating the relationship between activity of a particular compound with its molecular structure. This approach is usually conducted by medicinal chemists once the bioactivity data from a particular compound is known. The exploration is done by altering some chemical groups in the compound and observing the biological response of that modification. Sometimes, small chemical modification can significantly alter the biological response. In contrast, large modifications (which makes it structurally different from origin) can result in having similar activity [36].

The exploration of SAR is an important task in medicinal chemistry and drug design. SAR analysis provides a basis for chemical optimization of hits or leads and the identification of novel active compound. However, it is also reported that SAR characteristics often highly dependent on the types of molecules under investigation.

Sometimes, the molecules under investigation don't have a large coverage of chemical variation. This sometimes could lead to a difficult situation on deriving a SAR. This situation, however, could be solved if the molecules under investigation have a large coverage of chemical variations. One way to do this is by making the set of molecules to be as diverse

as possible. Since a SAR analysis is usually done by modifying the substituent (R) group of a molecule, it will be better if the variation is focused on the substituent group level rather in the whole molecular level. Therefore, in this particular problem, a diversity selection on the reagent level is more appropriate. A diverse selection of reagents will lead to a higher variation of the substituent group, thus, enabling it to be better in deriving a SAR.

1.6 Aims and Objectives for the Project

Combinatorial chemistry is now established as a powerful tool available to the medicinal chemist in the pursuit of new drug candidates. It provides a way to generate very large numbers of compounds in a relatively short period of time (compared to traditional synthesis of molecules). However this aspect of combinatorial chemistry in itself presents a problem. A balance needs to be struck between making everything possible and the constraints of economics, logistics and time. In other words, there is a strong need to do combinatorial library design to synthesize small number of compounds from the vast compound pool of those that could possibly be made. The art and science of computational library design has been reviewed extensively [37-39]. Chemical diversity [40-42] is often used as an optimisation function for library design, either on the reagent side [43, 44] or on the product side [16, 19]. Such library design strategies are often very efficient at selecting diverse compounds, but one drawback is that when the designed libraries are tested in assays, sometimes it can be hard to derive a clear structure activity relationship (SAR) from the experimental results since the selected building blocks could have little or no relationship to one another.

A new library design strategy, which called as ProSAR [45], has been developed in AstraZeneca to address this issue for designing libraries, which are helpful to derive an SAR after the screening. The goal for this project is to carry out validation study for the new methodology. ProSAR method has been applied to several literature examples to validate the ProSAR concept.

Chapter 2

Materials and Methods

2.1 ProSAR Methodology

2.1.1 Identification and Encoding of the Pharmacophore Fingerprint

A two-point pharmacophore is designed to encode the reagent pharmacophore information. The fingerprint consists of two kinds of information; one part is the pharmacophore type of the reagent and the other part is the topological distance (i.e. bond distance) between the single pharmacophore element and the attachment atom of the reagent (as shown in Figure 2.1). Five standard pharmacophore types are used here: hydrogen bond donor (HD), hydrogen bond acceptor (HA), positive charge center (POS), negative charge center (NEG) and lipophilic groups (LIP). Pharmacophores are defined via a set of SMARTS [28] patterns. In order to keep the reagent complexity low and avoid adding too long side chains to the scaffold, the bond distance between the each pharmacophore element and the attachment atom of the reagent is limited to 6 bonds. The total number of HD, HA, POS and NEG functional groups on a reagent was restricted to no more than 2 to further reduce the complexity on pharmacophore elements. The total number of unique two-point pharmacophores in a reagent is therefore 30 (5×6) and a 30-bin pharmacophore fingerprint can therefore be constructed, in which each bin refers to a specific two-point pharmacophore. The value in each bin is the frequency of the specific pharmacophore in the reagent corresponding to that bin. Figure 2.1 shows an example of such a pharmacophore fingerprint for an amine reagent. The pharmacophore fingerprint is constructed in such an order that the first 6 bins represent for HA pharmacophore element, then followed by 6 HD bins, 6 LIP bins, 6 POS bins and with 6 NEG bins at the end.

The lipophilic groups can be further divided into aromatic and nonaromatic lipophilic groups, and thus a 6-pharmacophore-type fingerprint can be obtained. In the 6-pharmacophore-type fingerprint, the pharmacophore fingerprint length will be 36 bins (6×6) and the order of pharmacophore element is kept the same as the 5 pharmacophore type fingerprint.

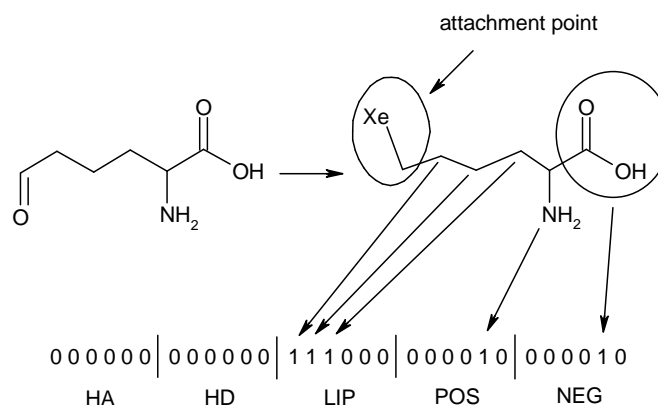


Figure 2.1: Identification and encoding of a reagent pharmacophore fingerprint.

2.1.2 Optimisation in Pharmacophore Space for ProSAR

The main goal of new library design strategy is to search for a set of reagents which cover the pharmacophore space optimally, while keeping the pharmacophore distribution as even as possible. Shannon entropy [46] was used widely to characterise the variation of descriptor space [47], so it was used here to represent the distribution of selected reagent subset based on the “pharmacophore fingerprint space”. The SE is defined as:

$$SE = -\sum_i p_i \log_2 p_i \quad (2.1)$$

where p_i is the probability of having a certain pharmacophore in the whole reagent set. p_i is calculated as:

$$p_i = \frac{c_i}{\sum_i c_i} \quad (2.2)$$

where c_i is the population of pharmacophore i in the whole reagent set. Shannon entropy is chosen as the fitness function for optimisation. A greater Shannon entropy value means that the pharmacophores for the selected reagent subset is more evenly distributed over the 30 bins. Hence, during the course of optimisation, a set of reagent compounds is sought to maximise the Shannon entropy.

A “greedy” search algorithm [48] was used as the optimisation engine to search for an optimal reagent subset. This was done by a first “greedy”-build up of the subset until the desired number of compounds is selected, followed by a second phase that re-evaluates each of the selected compounds in the subset to see if a better choice is available. The second stage continues until no improvement in the subset is possible.

2.1.3 Optimisation of the Pharmacophore Entropy and the Library Property Profile

Physico-chemical properties are an important aspect to consider in library design strategy. For example during lead optimisation stage, Chemists always try to not synthesize compounds with poor solubility or high hERG risk. Computational models for predicting solubility and hERG binding have been reported [49-51] and in many library design papers [52, 53] these physico-chemical properties were taken into account in the library design strategy as multiple constraint for optimisation. So a more realistic library design strategy would therefore be to extend the “ProSAR” concept further to include the library property profile. In order to calculate the properties of a library, a full enumeration has been done and the properties are calculated at the product level. At AstraZeneca, a set of stringent property criteria has been established for checking compound collection enhancement libraries [54]. Here in this study, the compound novelty (compared with in-house/external compounds to check if the compound is novel), aqueous solubility, predicted hERG liability and an in-house lead profile score [55, 56] are calculated as components of compounds property profile. All properties were calculated by in-house prediction tools.

An in-house library design tool GALOP has been developed at AstraZeneca. This uses a genetic algorithm (GA) optimisation method to optimise the reagent pharmacophore entropy and product properties simultaneously. The fitness function that the GA uses consists of two terms, one term represents the pharmacophore Shannon entropy for the reagents and the other term refers to the product properties. The fitness function formula is shown in Equation 2.3:

$$Score = w_p F + w_e \sum_j SE_j \quad (2.3)$$

Here, F means fraction of “good” compounds in the designed library and SE_j refers to the Shannon entropy for reagent set j . A compound is regarded as “good” only if it meets all the specified property criteria. w_p and w_e are weighting factors for property and entropy respectively.

2.1.4 Diversity Based Library Design Strategy

To provide a basis for evaluating the ProSAR method, a structural diversity based library design strategy was implemented in GALOP and tried out in this study. The fitness function for diversity optimisation is shown in Equation 2.4:

$$Score = 1 - \frac{2 \sum_{i=1}^n \sum_{j>i}^n S_{ij}}{n(n-1)} \quad (2.4)$$

Here, S_{ij} refer to the Tanimoto similarity index between reagent i and j . So the average pairwise Tanimoto similarity for reagents will be minimised during the optimisation.

2.2 Library Examples

For this validation work, five examples of chemical scaffold were used. They were taken from several publications from year 1992 to 2007 [57-61]. In these examples, one is for one dimension (1D) combinatorial library design and all the other 4 examples are for two dimensions (2D). 1D and 2D combinatorial library design correspond to one and two substitution group attached to the scaffold, respectively.

For 1D combinatorial library design, a scaffold from Adolor [57] was used as an example. This example is named as 1D_Lib1 in this work. It has the following structure in Figure 2.2 and the corresponding reaction scheme is shown in Figure 2.3. Aldehyde was used as a reagent for library enumeration.

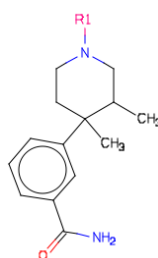


Figure 2.2: Combinatorial library example for 1D_Lib1

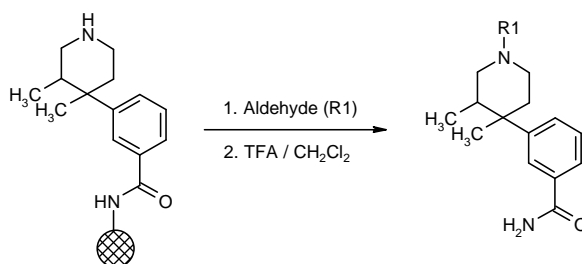


Figure 2.3: Reaction scheme for 1D_Lib1

As examples for 2D library design, several scaffolds were taken from the literature [58-61]. These examples are named 2D_Lib1, 2D_Lib2, 2D_Lib3 and 2D_Lib4 respectively. Figure 2.4 and Figure 2.5 show the scaffold and reaction scheme of 2D_Lib1. Aliphatic bromides were used as reagents for library enumeration.

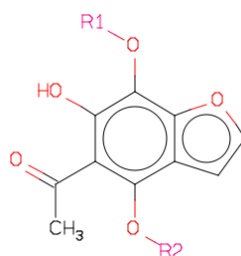


Figure 2.4: Combinatorial library example for 2D_Lib1

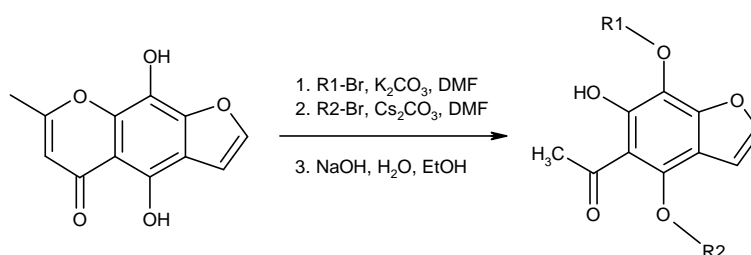


Figure 2.5: Reaction scheme for 2D_Lib1

The scaffold for 2D_Lib2 is shown in Figure 2.6 and its corresponding reaction scheme is shown in Figure 2.7. In this example, amino acids and aliphatic bromides were used as reagents.

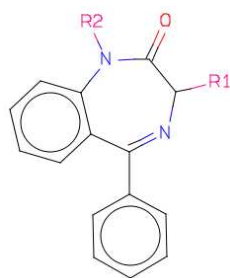


Figure 2.6: Combinatorial library example for 2D_Lib2

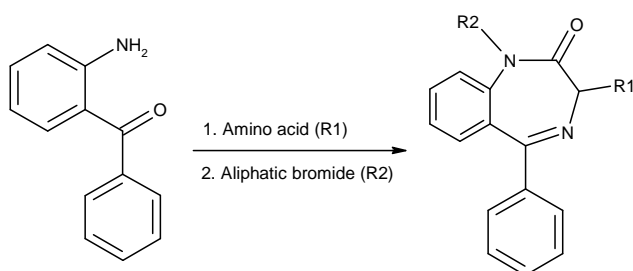


Figure 2.7: Reaction scheme for 2D_Lib2

Amine and alcohol were used as reagents in the example of 2D_Lib3. The 2D_Lib3 scaffold and reaction scheme are shown in Figure 2.8 and Figure 2.9 respectively.



Figure 2.8: Combinatorial library example for 2D_Lib3

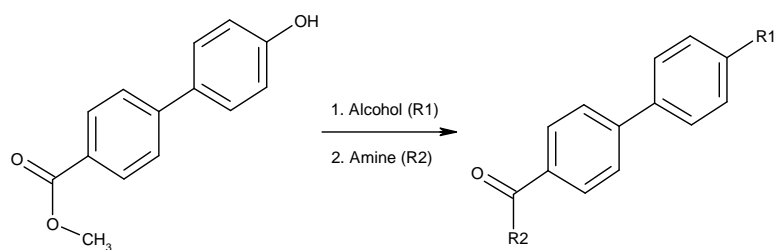


Figure 2.9: Reaction scheme for 2D_Lib3

In 2D_Lib4, one set of aldehydes and one set of acid chlorides were used in the reaction. The scaffold and reaction scheme for 2D_Lib4 are shown in Figure 2.10 and Figure 2.11.

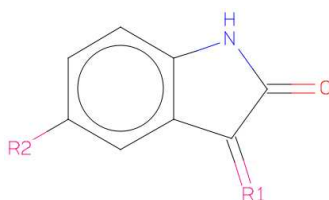


Figure 2.10: Combinatorial library example for 2D_Lib4

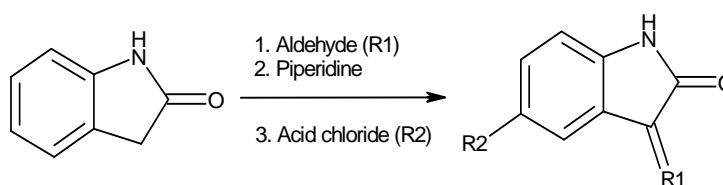


Figure 2.11: Reaction scheme for 2D_Lib4

A hypothetical library example is used to demonstrate the effectiveness of the design strategy that takes both reagent pharmacophore entropy and library property profile into account. This library has the reaction scheme shown in Figure 2.12.

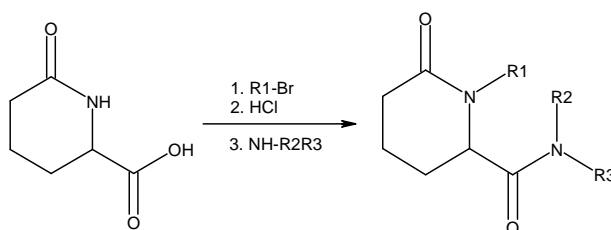


Figure 2.12: Combinatorial library example taking into accounts both pharmacophore entropy and property profile

2.3 Reagent Preparation

The reagents for all the above library examples were retrieved from the ACD database [62]. Before the library enumeration was carried out, a filtering procedure was applied to exclude several unwanted reagents. The filtering procedure was done by first

removing the salts and duplicates, followed by using an in-house program AZFParse [63] to exclude some compounds that have unwanted chemical features. Unwanted chemical features which are defined in the AstraZeneca in-house compiled AZFILTERS [64] list were used to filter out reagents. In this work, for all combinatorial library examples, the reagent sets were used are aldehyde, aliphatic bromide, amino acid, alcohol, primary amine and acid chloride. After the filtering procedure, the numbers of all reagents used for these combinatorial library designs are shown in Table 2.1.

Table 2.1: Reagents used in all library examples

Library Name	Reagent Used	
1D_Lib1	Aldehyde	2012
2D_Lib1	Aliphatic bromide	1105
	Aliphatic bromide	1105
2D_Lib2	Amino acid	741
	Aliphatic bromide	1105
2D_Lib3	Alcohol	588
	Primary amine	1062
2D_Lib4	Aldehyde	2012
	Acid Chloride	717

2.4 Computational Procedure

In 1D library design, 40 reagents were selected for each method, resulting 40 compound containing combinatorial libraries in each method. For 2D library design, 20 reagents were selected from each type of reagent pools giving a 20×20 combinatorial library in each of the methods.

2.4.1 ProSAR Library Design

The “ProSAR” library design strategy is to select a reagent set that can cover pharmacophore space as much as possible while keeping the pharmacophore elements distribution as even as possible. Reagent pharmacophore fingerprints were generated in a two-step procedure. First, two-point pharmacophores were created by an in-house tool TRUST [29]. A shell script was thereafter used to create the reagent pharmacophore fingerprints based on the TRUST output. The “greedy” search algorithm was implemented in Python [65] to calculate and optimise reagent pharmacophore entropy. Due to the deterministic nature of the “greedy” search algorithm, for each library design example that has been described, one PSAR library was generated.

2.4.2 Random Library Design

Random library design strategy is used for comparison. This strategy is to randomly pick up reagents from the reagent pool to enumerate library. For each library design example, 10 random libraries were generated to get statistical meaningful results.

2.4.3 Diversity Library Design

As another way of designing library, diversity based library design strategy is used. It selects reagents which has minimum average ensemble Tanimoto similarity (i.e. maximal Tanimoto dissimilarity, shown in Equation 2.4). The GA based library design tool GALOP is used for reagent selection. For each library example, GALOP was run 10 times to enumerate 10 different diversity based libraries for comparison.

2.4.4 Preparation of Validation Set

For each library example, a set of true active compounds are extracted from GVKBio database [66] (a comprehensive collection of active compounds published in a variety of journals and patents) by searching for active compounds that have the same scaffold. Those true active compounds are used as validation set. Side chain groups for those active compounds are extracted by using the AstraZeneca in-house program “Scaffrtab” and then TRUST is used to generate a pharmacophore fingerprint for the extracted side chain groups. These generated pharmacophore fingerprints represent the pharmacophore space for active compounds and are thereafter compared with the pharmacophore fingerprint of reagents selected from different library design strategies. Table 2.2 shows the number of retrieved active compounds from GVKBio database for each library example.

Table 2.2: The number of retrieved active compounds from GVKBio database

Library Example	Number of Retrieved Active Compounds
1D_Lib1	113
2D_Lib1	83
2D_Lib2	52
2D_Lib3	128
2D_Lib4	280

2.4.5 Similarity Search

As a part of performance comparison, compounds from enumerated libraries are used as query structure and similarity searches against GVKBio database are done to retrieve active compounds which have the same scaffold. By doing this, the number of retrieved active compounds can be regarded as a measurement of the effectiveness of generating active compounds for different libraries. The in-house structural fingerprint FOYFI is used for calculating Tanimoto similarity of two compounds. Tanimoto similarity cut-off is set at 0.85.

Chapter 3

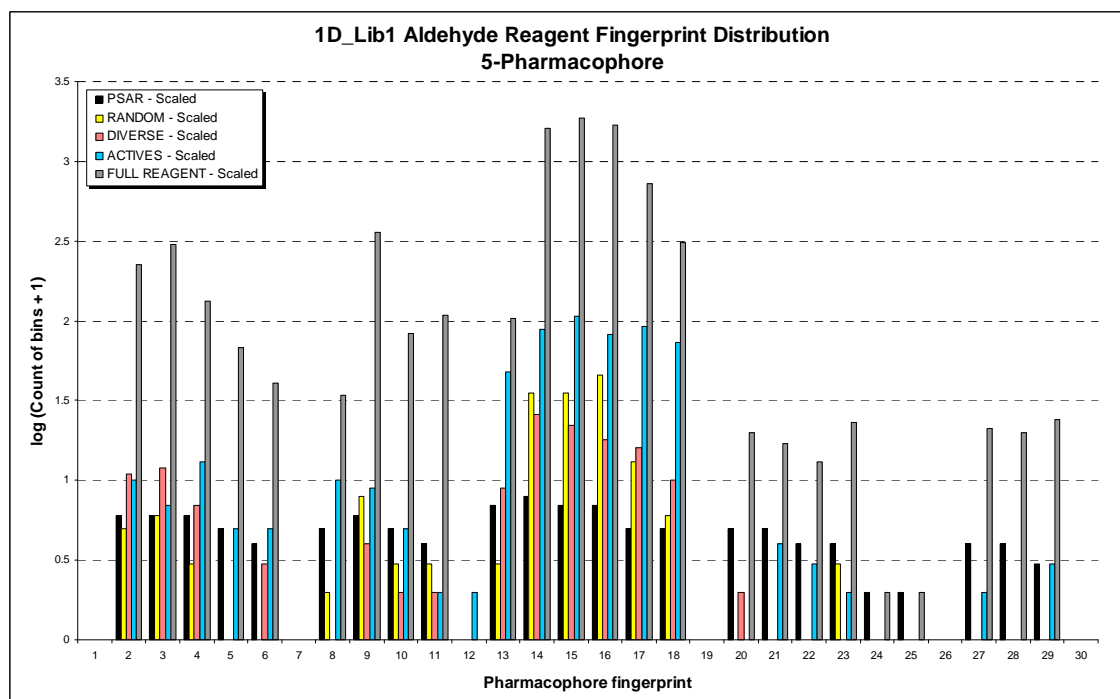
Results and Discussions

3.1 Comparison of Pharmacophore Coverage for Different Library Design Strategies

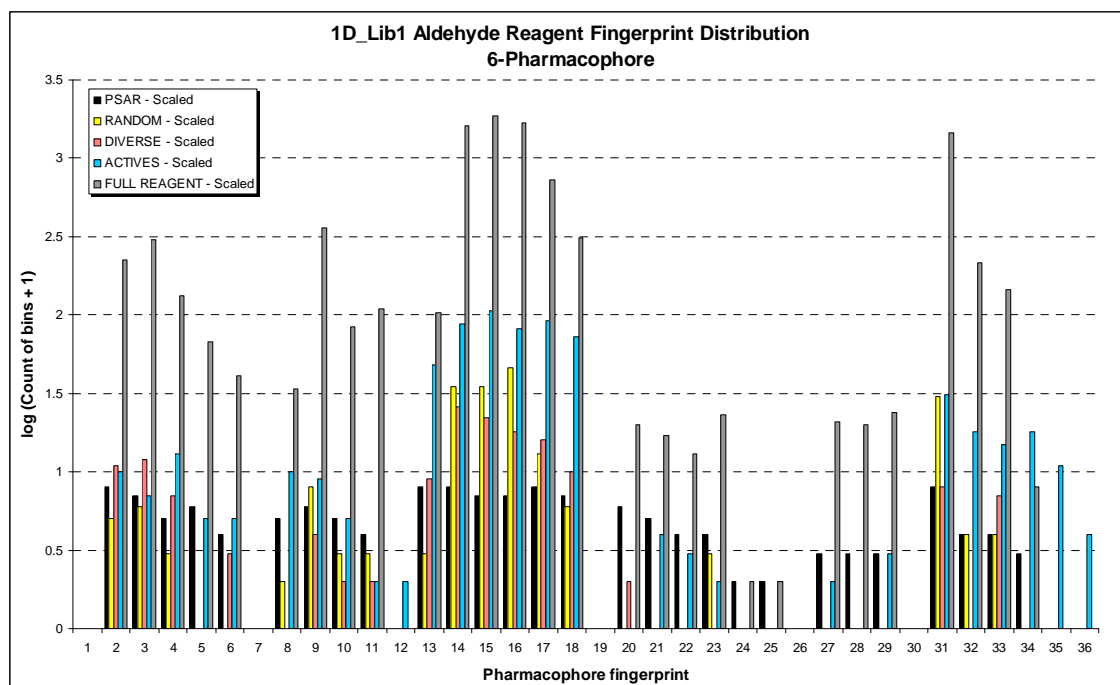
Altogether 5 library examples were selected from literature and used as test cases to show how the PSAR method works in practice. For the 1D library example, 40 reagents were selected to enumerate libraries and for all the other 2D library examples, several 20×20 libraries were enumerated. Besides using ProSAR strategy to select reagents, random library design and structural diversity library design scheme were used to compare their performance in pharmacophore coverage. For each example, 10 libraries were generated for random design and structural diversity library design and due to the deterministic nature of the greedy search algorithm only one PSAR library was enumerated.

The comparison of pharmacophore coverage of 5 pharmacophore type and 6 pharmacophore type ProSAR libraries with libraries from other design strategies are shown in Appendix A and Appendix B respectively. In the pharmacophore distribution plots of all five examples, ProSAR selected reagents clearly have most even distribution on the pharmacophore bins. In 1D_Lib1 example (as shown in Figure 3.1a and Figure 3.1b), the pharmacophore corresponding to bin number 5 is missing in both random and diversity selection but the PSAR selection contains the pharmacophore. Other missing pharmacophores in both random and diversity selection but not in PSAR are observed in bin number 21, 22, 24, 25, 27, 28 and 29 for the five pharmacophore and additional bin number 34 for the six pharmacophore type. The random and diversity libraries in the above example contain a lot of reagents with pharmacophoric features corresponding to bins 14 to 17. These bins correspond to lipophilic group in the reagents. The frequency of these bins has been dramatically reduced in the PSAR library. In return, the PSAR method is able to select bins number 20 to 29 (the positive and negative charge centres respectively) that are rarely selected by random and

diversity selection. The PSAR library doesn't contain bins such as 1, 7, 12, 19, 26, 30, 35 and 36, because the full reagent set doesn't have any reagents with these pharmacophores. This result shows that PSAR selection performs much better in covering pharmacophore space compared to the other two methods.



(a)



(b)

Figure 3.1: Comparison of pharmacophore coverage in 1D_Lib1 library example taken from Appendix A and Appendix B. (a) 5-pharmacophore types fingerprint (b) 6-pharmacophore types fingerprint.

This can also be seen from the Table 3.1, where the entropy values for ProSAR libraries, random libraries and diversity libraries are given. As seen from this table, PSAR reagent sets always have higher entropy values than reagent set which are selected by using other methods. For example in 1D_Lib1 case, the total entropy for five pharmacophore type PSAR selection is 4.46, random selection is 3.01 and diversity selection is 3.25. Based on the entropy definition (Equation 2.1, 2.2), the larger the entropy values is, the more even the distribution on pharmacophore bins is.

Table 3.1: Shannon entropy values (SE) of selected reagents from all library examples

Library Example		Pharm. Types ^a	ProSAR	Random ^b	Diversity ^b	Actives ^c	Full ^d
1D_Lib1	R1	5 Pharms	4.46	3.00	3.25	3.19	3.10
		6 Pharms	4.65	3.31	3.52	3.66	3.39
2D_Lib1	R1	5 Pharms	4.60	2.89	3.17	2.69	3.18
		6 Pharms	4.69	3.24	3.40	3.22	3.50
	R2	5 Pharms	4.60	3.03	3.21	2.38	3.18
		6 Pharms	4.69	3.37	3.42	2.86	3.50
2D_Lib2	R1	5 Pharms	4.62	2.84	3.62	3.40	3.18
		6 Pharms	4.71	3.11	3.62	3.87	3.43
	R2	5 Pharms	4.60	2.97	3.25	3.55	3.18
		6 Pharms	4.69	3.26	3.49	3.78	3.50
2D_Lib3	R1	5 Pharms	4.17	2.88	3.17	2.65	3.02
		6 Pharms	4.32	3.17	3.37	2.97	3.31
	R2	5 Pharms	4.32	3.09	3.24	3.05	3.26
		6 Pharms	4.52	3.32	3.39	3.42	3.50
2D_Lib4	R1	5 Pharms	4.46	2.95	3.15	3.26	3.10
		6 Pharms	4.65	3.29	3.41	3.61	3.39
	R2	5 Pharms	4.04	2.51	2.94	3.51	2.63
		6 Pharms	4.24	2.91	3.26	3.82	3.04

Note: a) Refers to the types of pharmacophore (5 pharmacophore and 6 pharmacophore).

b) Averaged values for 10 libraries.

c) The corresponding SE values of the R-reagent from the active compounds.

d) The corresponding SE values of the R-reagent from the full reagent set.

Pharmacophores for PSAR libraries, random libraries and diversity libraries are also compared with those of real active compounds. For each library example, active compounds for those libraries can be found in GVKBio databases by using the library scaffold as search queries. R1, R2 reagents (only R1 reagent in case of 1D library) could then be extracted from active compounds and encoded into pharmacophore fingerprint. Some of the reagents in the validation set have bond distance longer than 6. However, when the pharmacophore fingerprints were generated, only pharmacophores up to 6 bonds distance from the attachment point were considered. Pharmacophores further away than 6 bonds were ignored. This is due to the limitation used in pharmacophore encoding procedure to keep molecules complexity below a certain level and to avoid too long side chains. The pharmacophore distributions of active compounds are incorporated into the figures in Appendix A and Appendix B.

Another way of comparing pharmacophore coverage of libraries is to compare the pharmacophores of enumerated libraries with that of active compounds. The comparison figures are shown in Appendix C (5 pharmacophore type fingerprint) and Appendix D (6 pharmacophore type fingerprint). There are four colors present in those figures, which are green, yellow, red and grey. Those colored cells represent the pharmacophore comparisons between a particular enumerated library and the validation set (active compounds). Green cells represent pharmacophores that are present in both an enumerated library and the validation set. Yellow cells represent pharmacophores that are present in an enumerated library but not in the validation set, while red ones are not present in a designed library but are present in the validation set. In addition, pharmacophores that are both absent in a library design and the validation set are colored by grey. Therefore in those figures, more green cells and fewer red cells in a library design represents better performance of a library in covering pharmacophore space, compared to the validation set. However, it is also interesting to take note of the yellow cells. These yellow bins are the pharmacophore cells that do not exist in the validation set but appeared in library design, which correspond to some additional pharmacophores covered by designed library. So those yellow cells are also an indicator of potential pharmacophore coverage of a designed library. In contrast, grey bins can turn into red bins if more active compounds were available in validation set. Hence, green and yellow cells are positive indicators of pharmacophore coverage in a library design while red and grey bins are negative ones. As shown in figures in Appendices C and D, in all demonstrated examples, PSAR libraries show better coverage in terms of number of green cells and red cells. For example for the R1-reagent selection in 2D_Lib1 case (shown in figure 3.2), the PSAR selected R1 reagents have 0 red cells and 12 green cells in terms of 5 pharmacophore type fingerprint and randomly selected reagents have 6 red cells and 6 green cells, while diversity ones have 4 red cells and 8 green cells. For R2 reagents, PSAR selected R2 reagents have 0 red cells and 9 green cells, randomly selected reagents have 2 red cells and 7 green

cells, while diversity ones have 2 red cells and 6 green cells. This means that PSAR library miss fewer pharmacophores that appeared in active compounds and cover more potential pharmacophores compared with random library and diversity library.

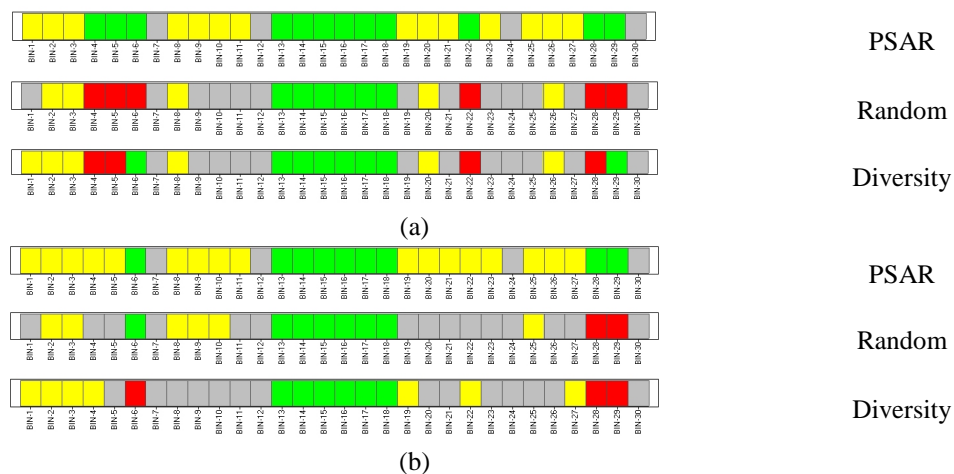


Figure 3.2: Comparison of pharmacophore (5 pharmacophore type) coverage of designed libraries in 2D_Lib1 example on (a) R1 reagents (b) R2 reagents.

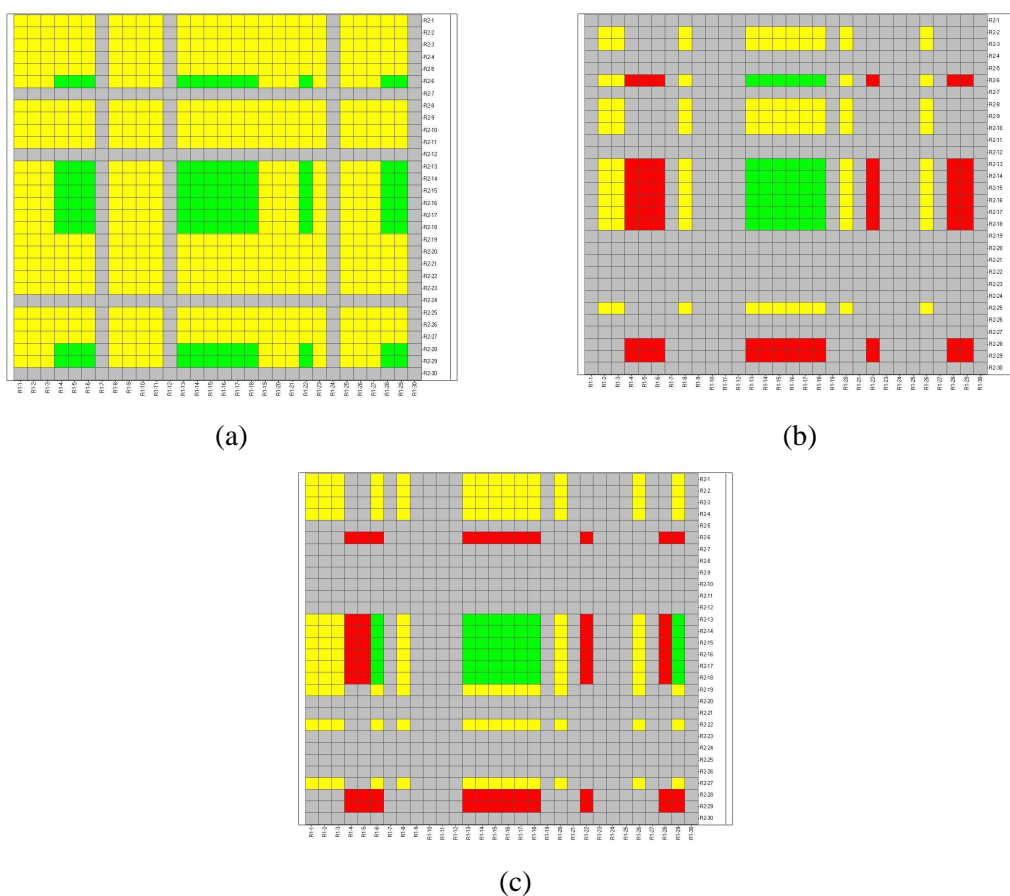


Figure 3.3: Comparison of pharmacophore (5 pharmacophore type) coverage in 2D heat maps for 2D_Lib1 examples. (a) PSAR library compared with validation set. (b) Random library compared with validation set. (c) Diversity library compared with validation set.

For two dimensional (2D) libraries, 2D heat maps can be used to represent the pharmacophore coverage for R1 and R2 reagents for different libraries. These are shown in Appendix E. Here the same color scheme is used. The same trend can be seen that PSAR libraries tend to cover more pharmacophores than other libraries in all the test cases. For example in 2D_Lib1 case, in the five pharmacophore type heat map (shown in figure 3.3), PSAR library has 0 red cell and 108 green cells, random library has 66 red cells and 42 green cells, while diversity library has 60 red cells and 48 green cells. This analysis shows that PSAR libraries cover most of the pharmacophores which are present among the active compounds.

3.2 Comparison of Retrieved Active Compounds for Different Library Design Strategies

Table 3.2: The number of retrieved active compounds from GVKBio database

Library Example	Number of Retrieved Active Compounds for Different Libraries			
	PSAR (5 pharm.)	PSAR (6 pharm.)	Random ^a	Diversity ^a
1D_Lib1	8	10	4.5	2.4
2D_Lib1	68	21	40.9	26
2D_Lib2	4	11	7.3	0.5
2D_Lib3	8	18	10.5	1.6
2D_Lib4	4	20	14.4	1.4

Note: a) Averaged value from 10 runs

So far the performance on pharmacophore coverage for difference library design strategies has been compared; it would be interesting to compare how many active compounds exist in libraries which are designed by using different design strategies. Although real biological activity data is not available for the compounds in the designed libraries, an estimation of the likelihood of obtaining actives from the libraries could be made by a similarity search against the GVKBio database with a high similarity cut-off (i.e. assessing the effectiveness of the libraries for generating leads). It is normally assumed that similar compounds have higher probability of having similar bioactivity than dissimilar ones (the similar property principle) [23] and a high retrieval rate from the GVKBio database could be taken as an indication that active molecules are present in the library. Compounds in all the designed libraries were therefore used as query structures for a similarity search against the entire GVKBio database to check how many active compounds which have the same scaffold

could be retrieved from the GVKBio database at a given similarity cut-off. In this study, a Tanimoto similarity of 0.85 was used as similarity cut-off and the similarity was calculated based on AstraZeneca in-house FOYFI fingerprint. Brown *et al.* [21, 67] reported that if a molecule has a Tanimoto similarity, based on Unity fingerprint, of ≥ 0.85 to an active compound, then the molecule has an 80% chance of itself being active in the same assay. Others [22, 42, 68] have reported that compounds having a Tanimoto similarity ≥ 0.85 usually share similar biological activities. Although the similarity cut-off is not transferable between datasets [29] and fingerprint types, the similarity cut-off value of 0.85 for FOYFI fingerprint works well to identify similar compounds in our experience [69].

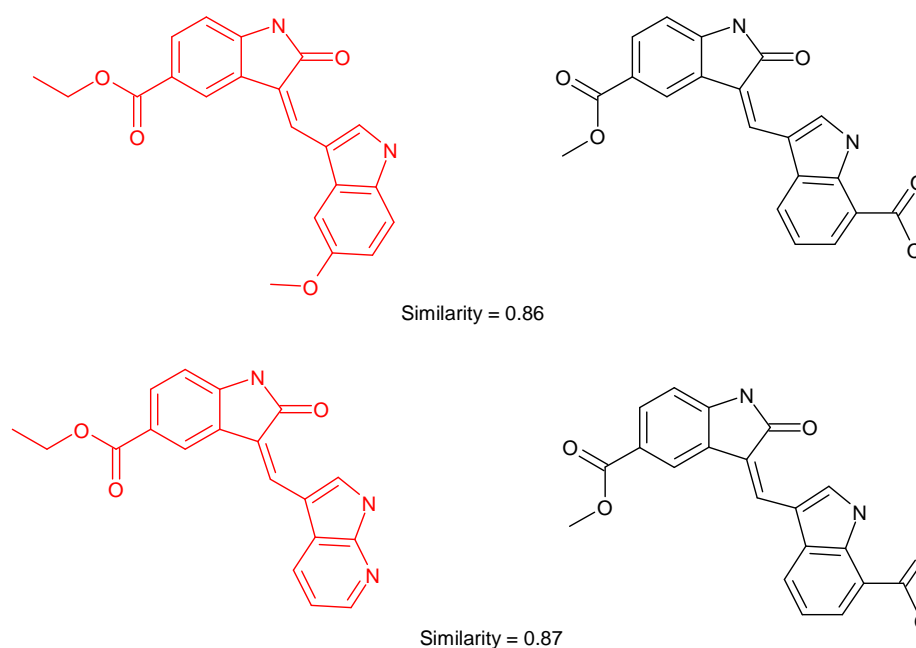


Figure 3.4: Similarity search examples. The molecules on left hand side are retrieved active compounds and the molecules on right hand side are query molecules

The numbers of retrieved active compounds for all 5 test cases are shown in Table 3.2. The results show that, in most of cases, PSAR designed libraries retrieve most active compounds and this indicates that PSAR design strategy has greater potential to find active compounds than other design strategies. Generally 6 pharmacophore type PSAR libraries perform better than the 5 pharmacophore type PSAR libraries. This is probably due to the fact that the lipophilic atoms are separated into aliphatic and aromatic parts in 6 pharmacophore type fingerprint and therefore can make a more balanced distribution among aliphatic groups and aromatic rings than with 5 pharmacophore type fingerprint. For example, in the 2D_Lib4, the active compounds in Figure 3.4 aren't retrieved by 5 pharmacophore fingerprint based PSAR library but are retrieved by 6 pharmacophore fingerprint based PSAR library. The query compound is on the right hand side in Figure 3.4 and it has an aromatic indole ring

which contributes to the high similarity indexes for the retrieved active compounds. The indole ring containing reagents are not selected in the 5 pharmacophore type based PSAR library.

3.3 Extension of ProSAR to Include Property Profile Optimisation

In the examples that are shown so far, only reagent pharmacophore space diversity is considered in the library design and compound physico-chemical properties aren't taken into consideration. This is somehow unrealistic in an industrial setting. There are some properties that need to be taken into account when designing a library. First, the aqueous solubility is a crucial property for a compound and should be considered during library design since it is desirable to synthesis highly soluble compounds. Secondly, risk assessment needs to be done to make sure only non-toxic compounds are synthesized. For example, hERG [70] risk assessment is very important. Therefore, the PSAR strategy needs to be further extended to include the compound property profile into the optimization process. To solve these problems, an in-house genetic algorithm optimizer GALOP was used specifically to design compound libraries with multiple constraints. GALOP program is capable to optimise the reagent matrix with both reagent based constraints (the reagent pharmacophore coverage) and product based constraint (product property profile).

Table 3.3: Results for the GA optimized libraries using 5 pharmacophore types

Libraries	Percent of good compound (%)	Shannon entropy		Diversity		Covered bins	
		R1	R2	R1	R2	R1	R2
PSAR + Property	99	3.79	3.64	0.69	0.67	17	16
Diversity + Property	99	2.74	2.68	0.78	0.79	9	11
Property	100	2.14	2.56	0.62	0.68	8	12
Random	38	2.49	2.75	0.74	0.70	9	13
Full library	40	2.83	2.95	0.74	0.73	21	20

As shown in Equation 2.3, both pharmacophore entropy and compound property profile were included in the GA fitness function. Based on experience, when a weight ratio (w_e/w_p) equal to 2.0 is applied, a reasonable and balanced library can be obtained. In the algorithm implementation, several properties were considered: (1) novelty check (2) *in silico* predicted solubility (3) *in silico* predicted hERG liability (4) in-house lead-like criteria. A “good compound” has to pass all these four criteria. This extended library design strategy is applied in a hypothetical example as seen in Figure 2.12. This library synthesis consists of

two reaction steps, first aliphatic bromide (R1) reagents were added to a scaffold and the product then reacted with amine (R2) reagents to form a two dimensional combinatorial library. For comparison, other libraries are also generated. They are a random library, a diversity combined with property optimization library and an property optimised only library. For the diversity driven library design, the diversity is characterized by the Tanimoto dissimilarity of the reagents based on the in-house FOYFI fingerprint. The results are shown in Table 3.3. This optimization was based on 5 pharmacophore fingerprints.

Table 3.4: Results for the GA optimized libraries using 6 pharmacophore types

Libraries	Percent of good compound (%)	Shannon entropy		Diversity		Covered bins	
		R1	R2	R1	R2	R1	R2
PSAR + Property	99	3.87	3.72	0.72	0.69	18	17
Diversity + Property	99	3.14	2.97	0.78	0.79	12	13
Property	100	2.80	2.58	0.58	0.67	9	13
Random	38	2.87	2.58	0.65	0.63	11	13
Full library	40	3.20	3.11	0.74	0.73	25	23

PSAR combined with property calculation optimized library has the best reagent entropy among all four libraries and has a high percentage of good compounds (99%). Diversity combining property optimized library has the best reagent diversity and its percentage of good compounds is also high (99%). For the library which is optimized by the property only, the entropy and diversity values are worse than the others. As a baseline, the fully enumerated library only has 40% good compounds in total and medium entropy and diversity values. Random library has medium entropy and diversity values and contains only 38% good compounds. As expected, PSAR library covers most number of pharmacophore bins on both R1 and R2 substituent. Another GA optimization is also done by using six pharmacophore fingerprints. The result is shown in Table 3.4 and the same pattern is observed. PSAR combined with property optimization performs best in entropy value and has a high percentage of good compounds (99%) compared to others. Regarding the compound properties for these design libraries, since the property control is included in GA optimization, all three optimized libraries have high percentage of good compounds. The percentage of good compounds in PSAR combined with property optimized library and diversity combined with optimized library are 99% and 100% which is comparable with that of the property optimized only library. These results show that by using the extended PSAR strategy, libraries which have both good pharmacophore coverage and good property profile can be obtained.

Chapter 4

Conclusions

Case studies were carried out to validate the “ProSAR” (PSAR) methodology, an AstraZeneca in-house developed library design strategy. “ProSAR” method is capable to design combinatorial library which has optimal coverage and even distribution of the pharmacophore elements among the reagents by encoding the pharmacophore element on reagents into fingerprint and further optimising the Shannon entropy of pharmacophore fingerprint distribution. This methodology was applied on five different library design examples and the results show that most of the pharmacophore features that appear in active compounds could be covered by the PSAR derived library. Furthermore in those examples, PSAR libraries include more compounds which are structurally similar to true active compounds than random libraries and diversity libraries which are optimised by average ensemble Tanimoto similarity. The PSAR strategy can be further expanded to include compound properties to design a library which has not only good pharmacophore coverage of side chains but also desirable physico-chemical properties by using a GA optimisation method. This extended PSAR strategy was tried out on an illustrative test case, in which the aim was to design a 400 compound two-dimensional combinatorial library. The PSAR libraries were contrasted with libraries constructed by other design strategies such as diversity (characterised by the average ensemble Tanimoto similarity in this study) driven and property driven library design. The results demonstrate that “ProSAR” designed libraries are clearly superior in covering pharmacophore space and create more even distribution of the side chain pharmacophore elements than other methods, while at the same time a good compound property profiles are obtained.

Bibliography

1. Gallop, M.A., Barrett, R.W., Dower, W.J., Fodor, S.P., Gordon, E.M. 1994. Applications of combinatorial technologies to drug discovery. 1. Background and peptide combinatorial libraries. *J Med Chem*, 37: 1233-51.
2. Gordon, E.M., Barrett, R.W., Dower, W.J., Fodor, S.P., Gallop, M.A. 1994. Applications of combinatorial technologies to drug discovery. 2. Combinatorial organic synthesis, library screening strategies, and future directions. *J Med Chem*, 37: 1385-401.
3. Terrett, N.K., Gardner, M., Gordon, D.W., Kobylecki, R.J., Steele, J. 1995. Combinatorial Synthesis - the Design of Compound Libraries and Their Application to Drug Discovery. *Tetrahedron*, 51: 8135-8173.
4. Merrifield, R.B. 1964. Solid-Phase Peptide Synthesis. 3. An Improved Synthesis of Bradykinin. *Biochemistry*, 3: 1385-90.
5. Merrifield, R.B. 1963. Solid Phase Peptide Synthesis .1. Synthesis of a Tetrapeptide. *Journal of the American Chemical Society*, 85: 2149-&.
6. Merrifield, R.B. 1964. Solid Phase Peptide Synthesis .2. Synthesis of Bradykinin. *Journal of the American Chemical Society*, 86: 304-&.
7. Merrifield, R.B. 1964. Solid Phase Peptide Synthesis .4. Synthesis of Methionyl-Lysyl-Bradykinin. *Journal of Organic Chemistry*, 29: 3100-&.
8. Geysen, H.M., Meloen, R.H., Barteling, S.J. 1984. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc Natl Acad Sci U S A*, 81: 3998-4002.
9. Selway, C.N., Terrett, N.K. 1996. Parallel-compound synthesis: Methodology for accelerating drug discovery. *Bioorganic & Medicinal Chemistry*, 4: 645-654.
10. Lajiness, M., Watson, I. 2008. Dissimilarity-based approaches to compound acquisition. *Curr Opin Chem Biol*, 12: 366-71.
11. Olah, M.M., Bologa, C.G., Oprea, T.I. 2004. Strategies for compound selection. *Curr Drug Discov Technol*, 1: 211-20.
12. Brown, R.D., Hassan, M., Waldman, M. 2000. Combinatorial library design for diversity, cost efficiency, and drug-like character. *J Mol Graph Model*, 18: 427-37, 537.
13. Schnecke, V., Bostrom, J. 2006. Computational chemistry-driven decision making in lead generation. *Drug Discovery Today*, 11: 43-50.
14. Walters, W.P., Stahl, M.T., Murcko, M.A. 1998. Virtual screening - an overview. *Drug Discovery Today*, 3: 160-178.
15. Drewry, D.H., Young, S.S. 1999. Approaches to the design of combinatorial libraries. *Chemometrics and Intelligent Laboratory Systems*, 48: 1-20.
16. Gillet, V.J., Willett, P., Bradshaw, J. 1997. The effectiveness of reactant pools for generating structurally-diverse combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 37: 731-740.
17. Gillet, V.J., Willett, P., Fleming, P.J., Green, D.V.S. 2002. Designing focused libraries using MoSELECT. *Journal of Molecular Graphics & Modelling*, 20: 491-498.
18. Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J. 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev*, 46: 3-26.
19. Jamois, E.A., Hassan, M., Waldman, M. 2000. Evaluation of reagent-based and product-based strategies in the design of combinatorial library subsets. *J Chem Inf Comput Sci*, 40: 63-70.
20. Gillet, V.J. 2002. Reactant- and product-based approaches to the design of combinatorial libraries. *Journal of Computer-Aided Molecular Design*, 16: 371-380.

21. Brown, R.D., Martin, Y.C. 1998. An evaluation of structural descriptors and clustering methods for use in diversity selection. *Sar and Qsar in Environmental Research*, 8: 23-39.
22. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Weinberger, L.E. 1996. Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors. *Journal of Medicinal Chemistry*, 39: 3049-3059.
23. Maggiora, G.M., Johnson, M.A. 1990. Introduction to Similarity in Chemistry. *Concepts and Applications of Molecular Similarity*: 1-13.
24. Maldonado, A.G., Doucet, J.P., Petitjean, M., Fan, B.T. 2006. Molecular similarity and diversity in chemoinformatics: from theory to applications. *Mol Divers*, 10: 39-79.
25. Maccs II, Molecular Design. In. 14600 Catalina St., San Leandro, CA 94577, (510)895-1313.
26. Unity Chemical Information Software, Tripos Associates. In. 1699 S. Hanley Rd., Suite 303, St. Louis, MO 63144, 1-800-323-2960.
27. Daylight Chemical Information Software, Daylight Chemical Information. In. 18500 Von Karman, #450, Irvine, CA, (714)476-0451.
28. Daylight Theory Manual, Daylight Chemical Information Systems, Inc. In. 120 Vantis - Suite 550 - Aliso Viejo, CA 92656.
29. Kogej, T., Engkvist, O., Blomberg, N., Muresan, S. 2006. Multifingerprint based similarity searches for targeted class compound selection. *Journal of Chemical Information and Modeling*, 46: 1201-1213.
30. McGregor, M.J., Muskal, S.M. 1999. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *Journal of Chemical Information and Computer Sciences*, 39: 569-574.
31. McGregor, M.J., Muskal, S.M. 2000. Pharmacophore fingerprinting. 2. Application to primary library design. *Journal of Chemical Information and Computer Sciences*, 40: 117-125.
32. Bath, P.A., Morris, C.A., Willett, P. 1993. Effect of Standardization on Fragment-Based Measures of Structural Similarity. *Journal of Chemometrics*, 7: 543-550.
33. Fisanick, W., Cross, K.P., Rusinko, A. 1992. Similarity Searching on Cas Registry Substances .1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *Journal of Chemical Information and Computer Sciences*, 32: 664-674.
34. Hamming, R.W. 1950. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29: 147-160.
35. Willett, P., Barnard, J.M., Downs, G.M. 1998. Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38: 983-996.
36. Kubinyi, H. 1998. Similarity and dissimilarity: A medicinal chemist's view. *Perspectives in Drug Discovery and Design*, 9-11: 225-252.
37. Spellmeyer, D.C., Grootenhuys, P.D.J. 1999. Recent developments in molecular diversity: Computational approaches to combinatorial chemistry. *Annual Reports in Medicinal Chemistry*, Vol 34, 34: 287-296.
38. Beno, B.R., Mason, J.S. 2001. The design of combinatorial libraries using properties and 3D pharmacophore fingerprints. *Drug Discovery Today*, 6: 251-258.
39. Willett, P. 2000. Chemoinformatics - similarity and diversity in chemical libraries. *Current Opinion in Biotechnology*, 11: 85-88.
40. Turner, D.B., Tyrrell, S.M., Willett, P. 1997. Rapid quantification of molecular diversity for selective database acquisition. *Journal of Chemical Information and Computer Sciences*, 37: 18-22.
41. Jamois, E.A. 2003. Reagent-based and product-based computational approaches in library design. *Current Opinion in Chemical Biology*, 7: 326-330.
42. Potter, T., Matter, H. 1998. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *Journal of Medicinal Chemistry*, 41: 478-488.

43. Zheng, W.F., Cho, S.J., Tropsha, A. 1998. Rational combinatorial library design. 1. Focus-2D: A new approach to the design of targeted combinatorial chemical libraries. *Journal of Chemical Information and Computer Sciences*, 38: 251-258.
44. Leach, A.R., Green, D.V.S., Hann, M.M., Judd, D.B., Good, A.C. 2000. Where are the GaPs? A rational approach to monomer acquisition and selection. *Journal of Chemical Information and Computer Sciences*, 40: 1262-1269.
45. Chen, H., Börjesson, U., Engkvist, O., Kogej, T., Svensson, M., Blomberg, N., Weigelt, D., Burrwos, N.J., Lange, T. *Journal of Chemical Information and Modeling* in press.
46. Shannon, C.E., Weaver, W. 1963. The mathematical theory of communication. University of Illinois Press, Urbana.
47. Godden, J.W., Stahura, F.L., Bajorath, J. 2000. Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. *Journal of Chemical Information and Computer Sciences*, 40: 796-800.
48. Miller, J.L., Bradley, E.K., Teig, S.L. 2003. Luddite: An information-theoretic library design tool. *Journal of Chemical Information and Computer Sciences*, 43: 47-54.
49. Cavalli, A., Poluzzi, E., De Ponti, F., Recanatini, M. 2002. Toward a pharmacophore for drugs inducing the long QT syndrome: Insights from a CoMFA study of HERG K⁺ channel blockers. *Journal of Medicinal Chemistry*, 45: 3844-3853.
50. Pearlstein, R.A., Vaz, R.J., Kang, J.S., Chen, X.L., Preobrazhenskaya, M., Shchekotikhin, A.E., Korolev, A.M., Lysenkova, L.N., Miroshnikova, O.V., Hendrix, J. *et al.* 2003. Characterization of HERG potassium channel inhibition using CoMSiA 3D QSAR and homology modeling approaches. *Bioorganic & Medicinal Chemistry Letters*, 13: 1829-1835.
51. Jouyban, A., Soltanpour, S., Soltani, S., Chan, H.K., Acree, W.E. 2007. Solubility prediction of drugs in water-cosolvent mixtures using Abraham solvation parameters. *Journal of Pharmacy and Pharmaceutical Sciences*, 10: 263-277.
52. Egan, W.J., Merz, K.M., Baldwin, J.J. 2000. Prediction of drug absorption using multivariate statistics. *Journal of Medicinal Chemistry*, 43: 3867-3877.
53. Darvas, F., Dorman, G., Papp, A. 2000. Diversity measures for enhancing ADME admissibility of combinatorial libraries. *Journal of Chemical Information and Computer Sciences*, 40: 314-322.
54. Muresan, S., Kocis, P., Chen, H., Steele, J., Li, J. Manuscript in preparation. In.
55. Oprea, T.I. 2002. Current trends in lead discovery: Are we looking for the appropriate properties? *Journal of Computer-Aided Molecular Design*, 16: 325-334.
56. Oprea, T.I., Davis, A.M., Teague, S.J., Leeson, P.D. 2001. Is there a difference between leads and drugs? A historical perspective. *Journal of Chemical Information and Computer Sciences*, 41: 1308-1315.
57. Le Bourdonnec, B., Belanger, S., Cassel, J.A., Stabley, G.J., DeHaven, R.N., Dolle, R.E. 2003. trans-3,4-Dimethyl-4-(3-carboxamidophenyl)piperidines: A novel class of mu-selective opioid antagonists. *Bioorganic & Medicinal Chemistry Letters*, 13: 4459-4462.
58. Harvey, A.J., Baell, J.B., Toovey, N., Homerick, D., Wulff, H. 2006. A new class of blockers of the voltage-gated potassium channel Kv1.3 via modification of the 4- or 7-position of khellinone. *Journal of Medicinal Chemistry*, 49: 1433-1441.
59. Bunin, B.A., Plunkett, M.J., Ellman, J.A. 1996. Synthesis and evaluation of 1,4-benzodiazepine libraries. *Combinatorial Chemistry*, 267: 448-465.
60. Faghih, R., Dwight, W., Pan, J.B., Fox, G.B., Krueger, K.M., Esbenshade, T.A., McVey, J.M., Marsh, K., Bennani, Y.L., Hancock, A.A. 2003. Synthesis and SAR of aminoalkoxy-biaryl-4-carboxamides: Novel and selective histamine H-3 receptor antagonists. *Bioorganic & Medicinal Chemistry Letters*, 13: 1325-1328.
61. Adams, C., Aldous, D.J., Amendola, S., Bamborough, P., Bright, C., Crowe, S., Eastwood, P., Fenton, G., Foster, M., Harrison, T.K.P. *et al.* 2003. Mapping the kinase domain of Janus Kinase 3. *Bioorganic & Medicinal Chemistry Letters*, 13: 3105-3110.

62. MDL Available Chemicals Directory database 2007, S.T., Inc., Santa Clara, CA 95051.
63. A AstraZeneca in-house program which classify compounds based on user defined property filters and chemical structure filters.
64. A set of AstraZeneca in-house compiled unwanted chemical substructures.
65. Rossum, G.v. 1991. Python. In. Python Software Foundation,
<http://www.python.org/>.
66. GVKBio Database, GVK Bioscience Private Ltd, Hyderabad 500016, India.
67. Brown, R.D., Bures, M.G., Martin, Y.C. 1995. Similarity and Cluster-Analysis Applied to Molecular Diversity. *Abstracts of Papers of the American Chemical Society*, 209: 3-Comp.
68. Matter, H. 1997. Selecting optimally diverse compounds from structure databases: A validation study of two-dimensional and three-dimensional molecular descriptors. *Journal of Medicinal Chemistry*, 40: 1219-1229.
69. Steffan, A., Kogej, T., Hoppe, C., Engkvist, O. Comparison of Molecular Fingerprint Methods on the Basis of Biological Profile Data. *Journal of Chemical Information and Modeling*, in press.
70. Keating, M.T., Sanguinetti, M.C. 1996. Molecular genetic insights into cardiovascular disease. *Science*, 272: 681-5.

Appendix A

Comparisons of Pharmacophore Distribution Based on 5 Pharmacophore Type Fingerprint.

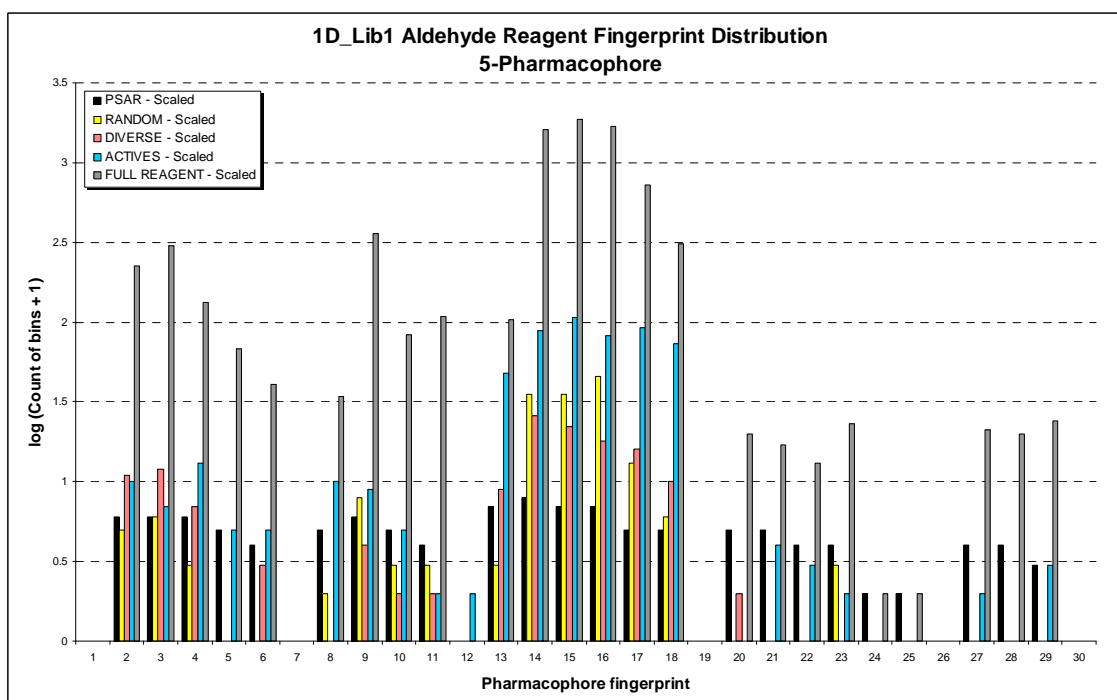


Figure A-1: Pharmacophore fingerprint distribution of R1-reagents in 1D_Lib1 library example.

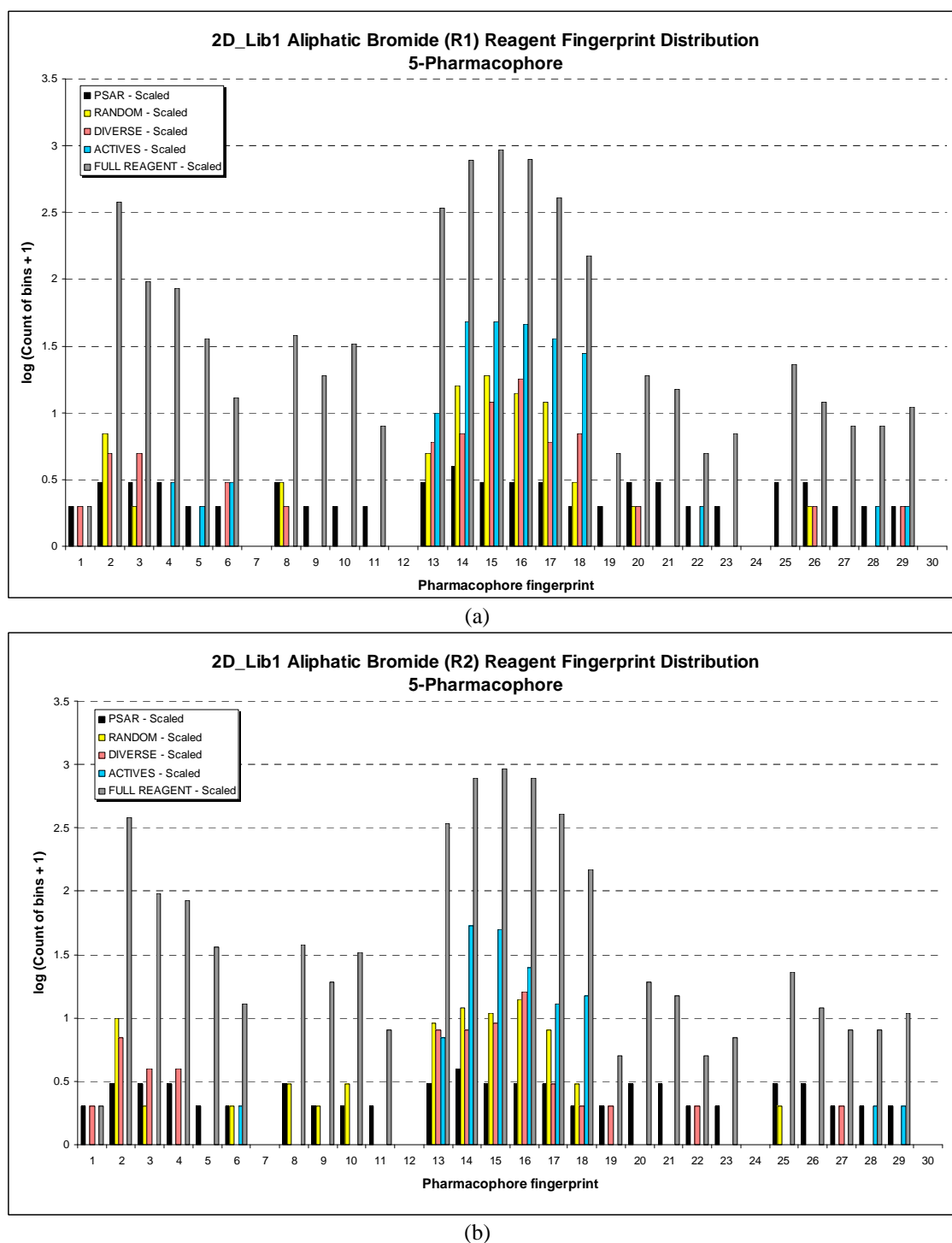
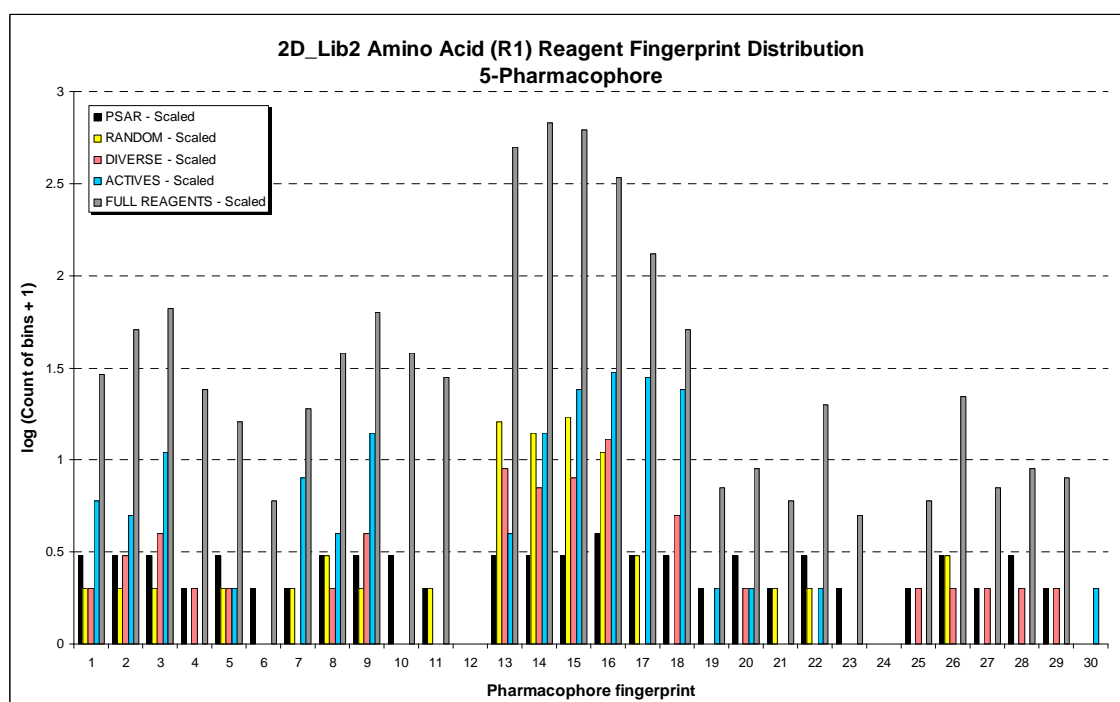
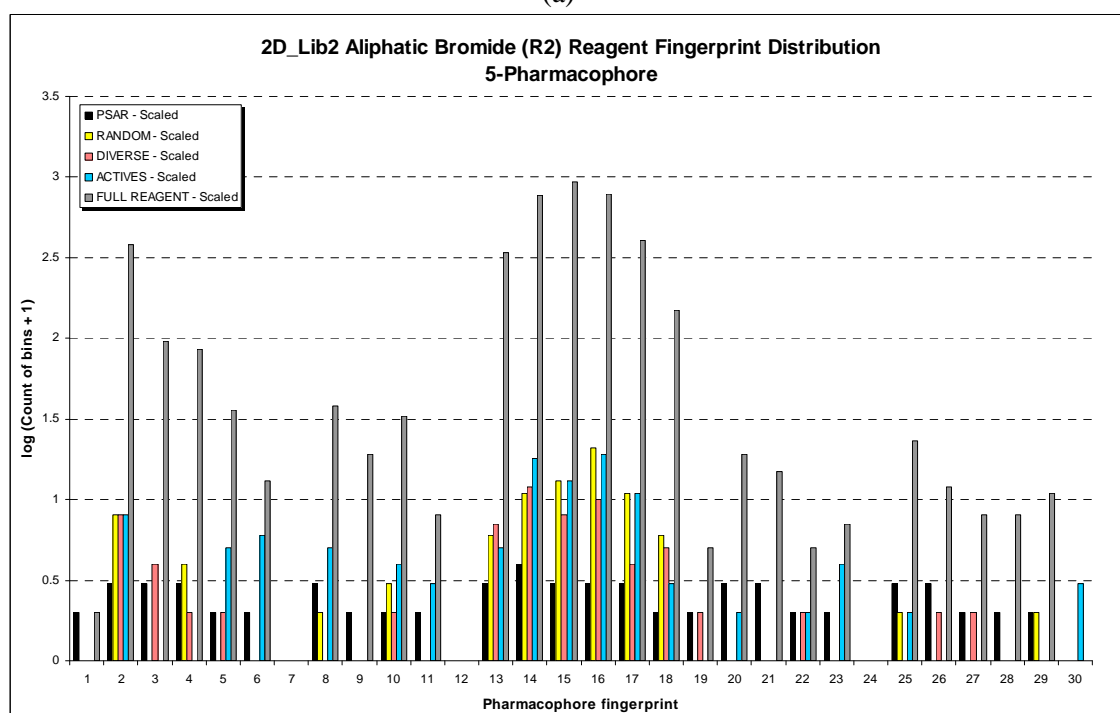


Figure A-2: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib1 library example.

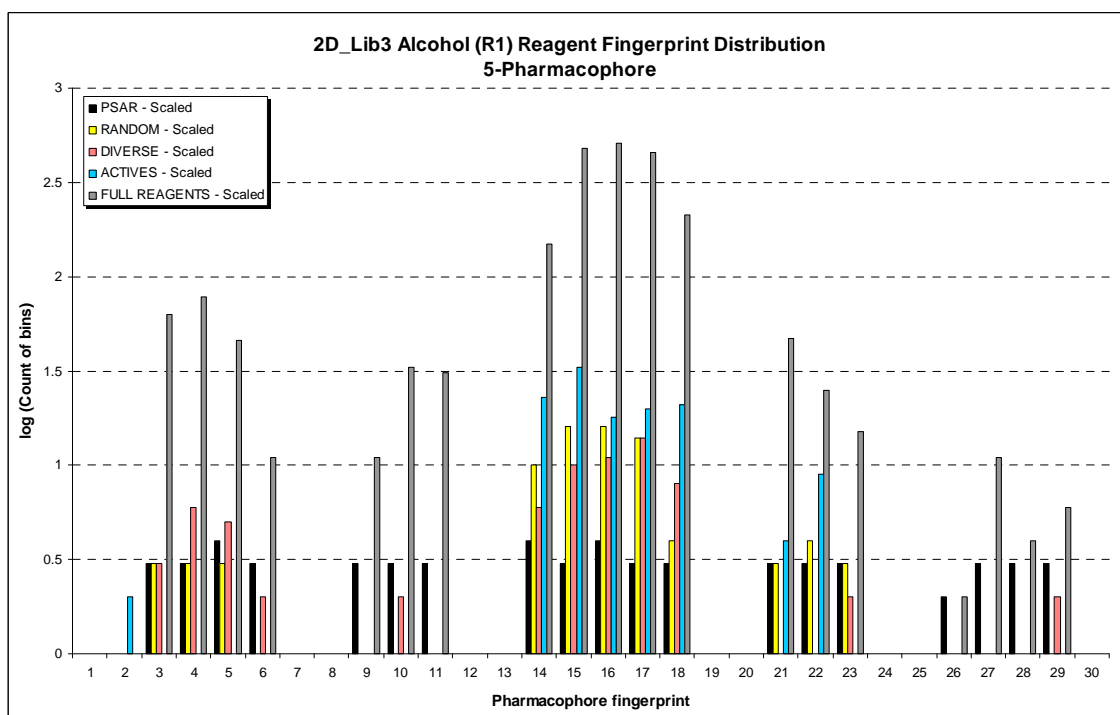


(a)

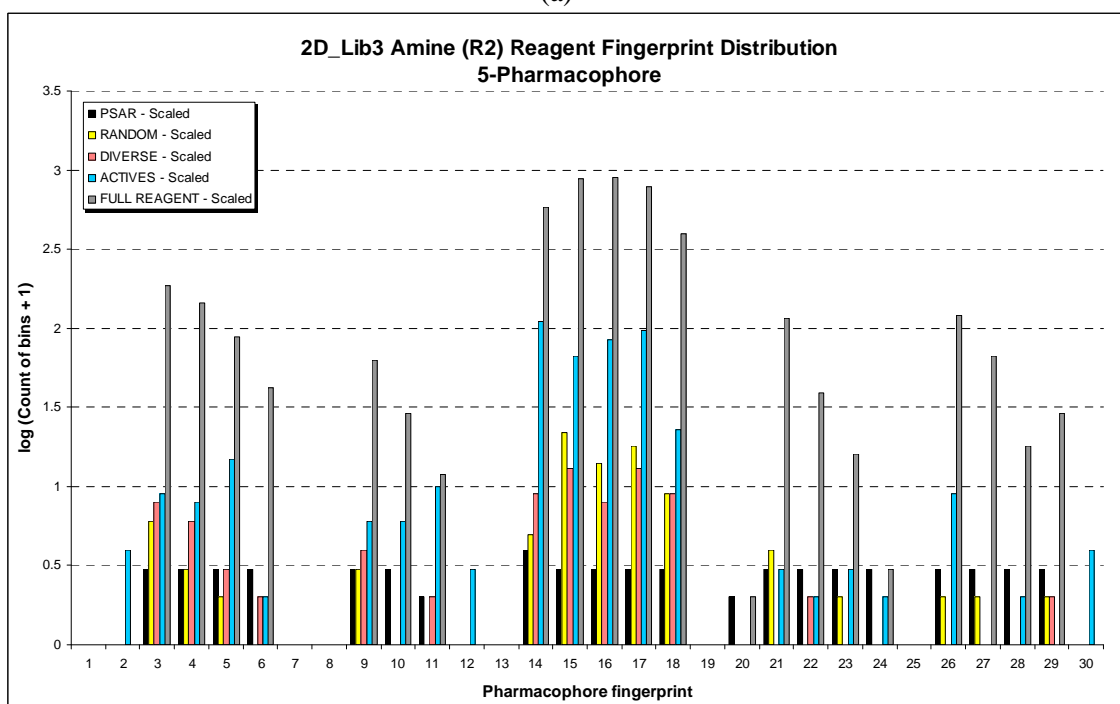


(b)

Figure A-3: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib2 library example.

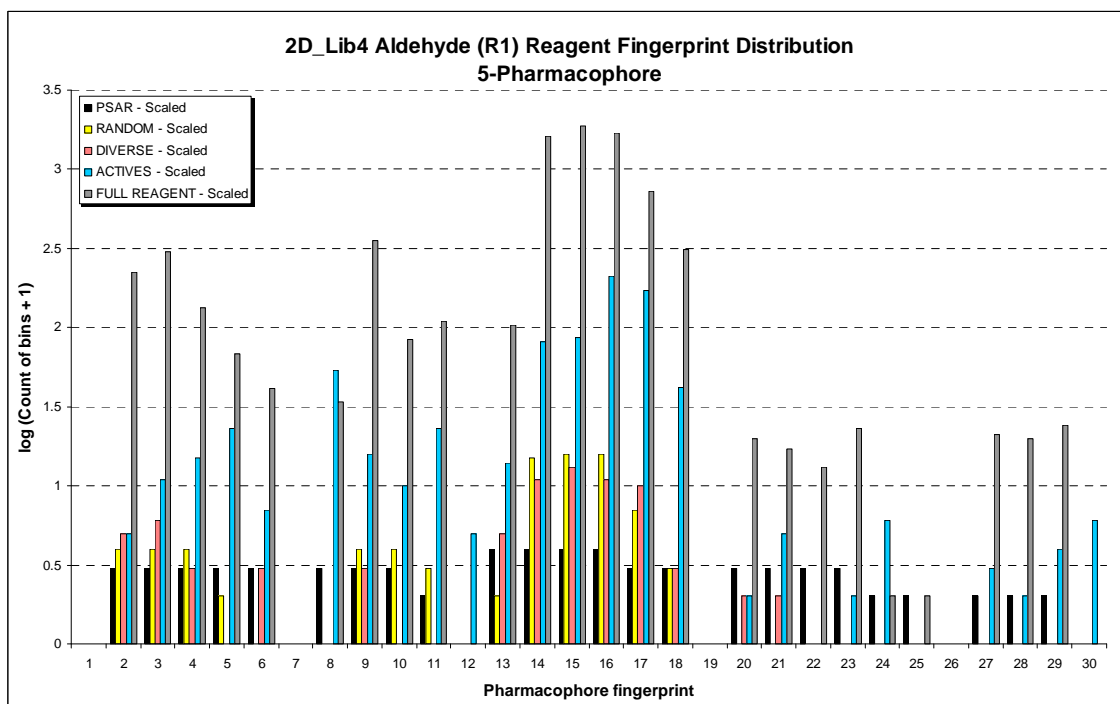


(a)

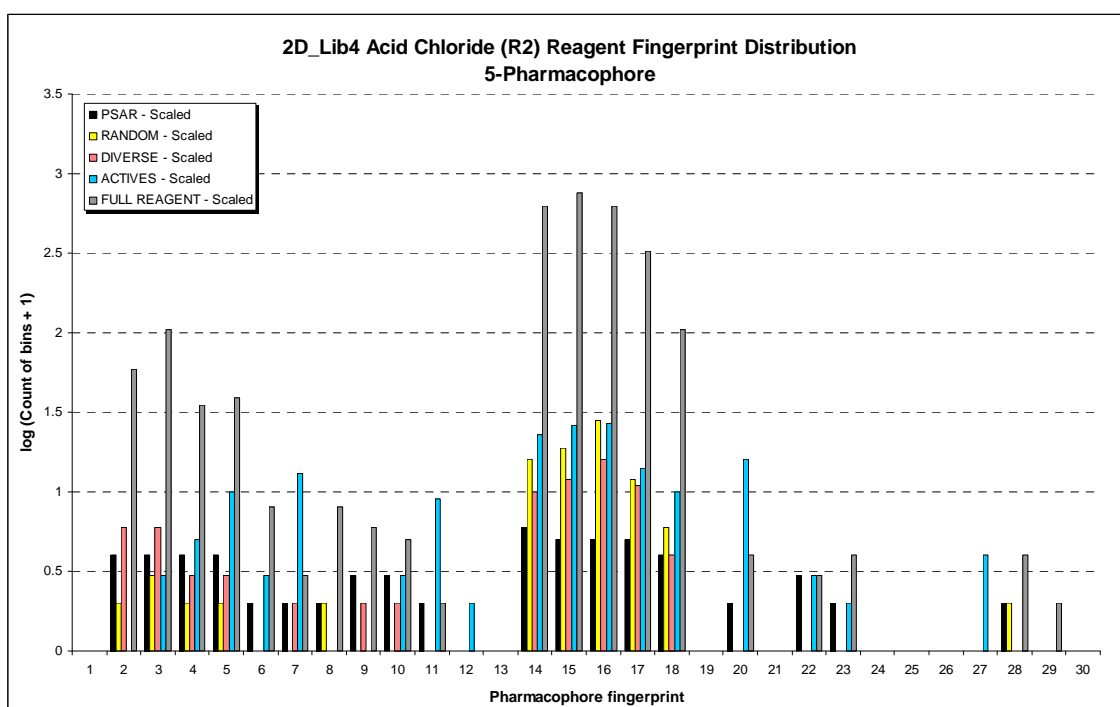


(b)

Figure A-4: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib3 library example.



(a)



(b)

Figure A-5: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib4 library example.

Appendix B

Comparisons of Pharmacophore Distribution Based on 6 Pharmacophore Type Fingerprint.

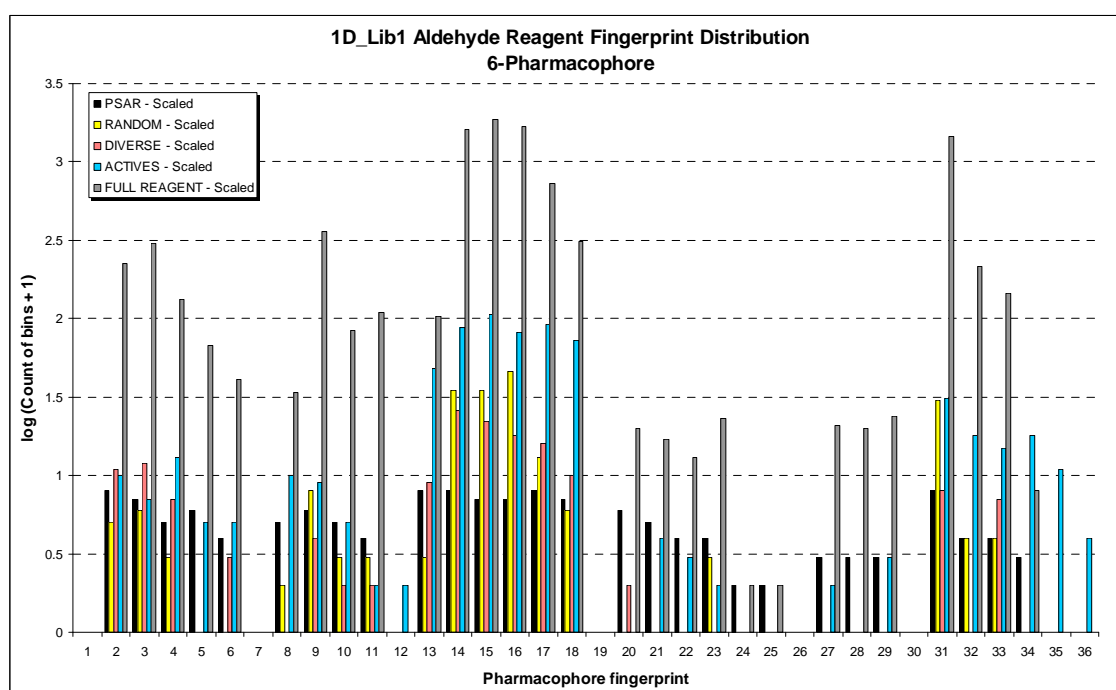
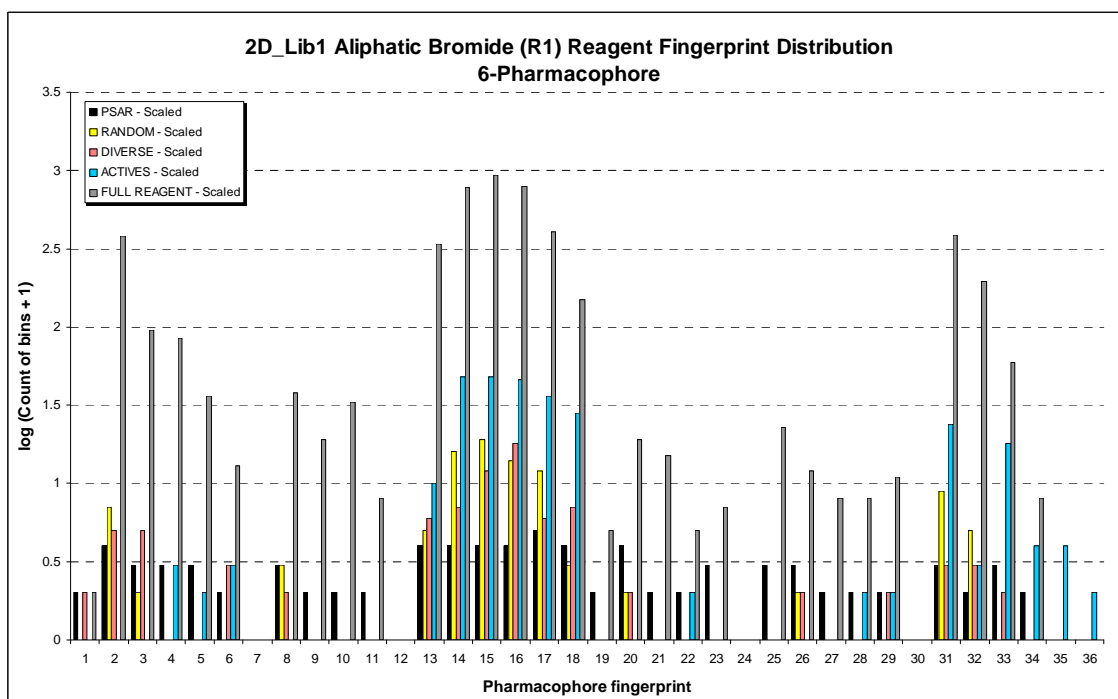
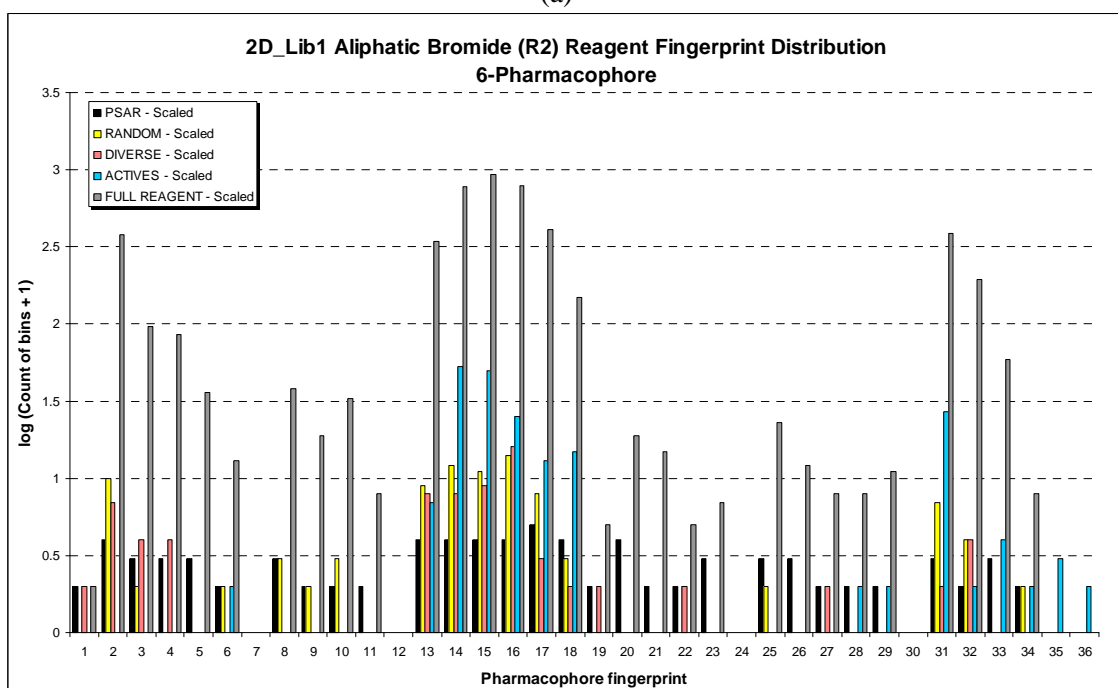


Figure B-1: Pharmacophore fingerprint distribution of R1-reagents used in 1D_Lib1 library example.

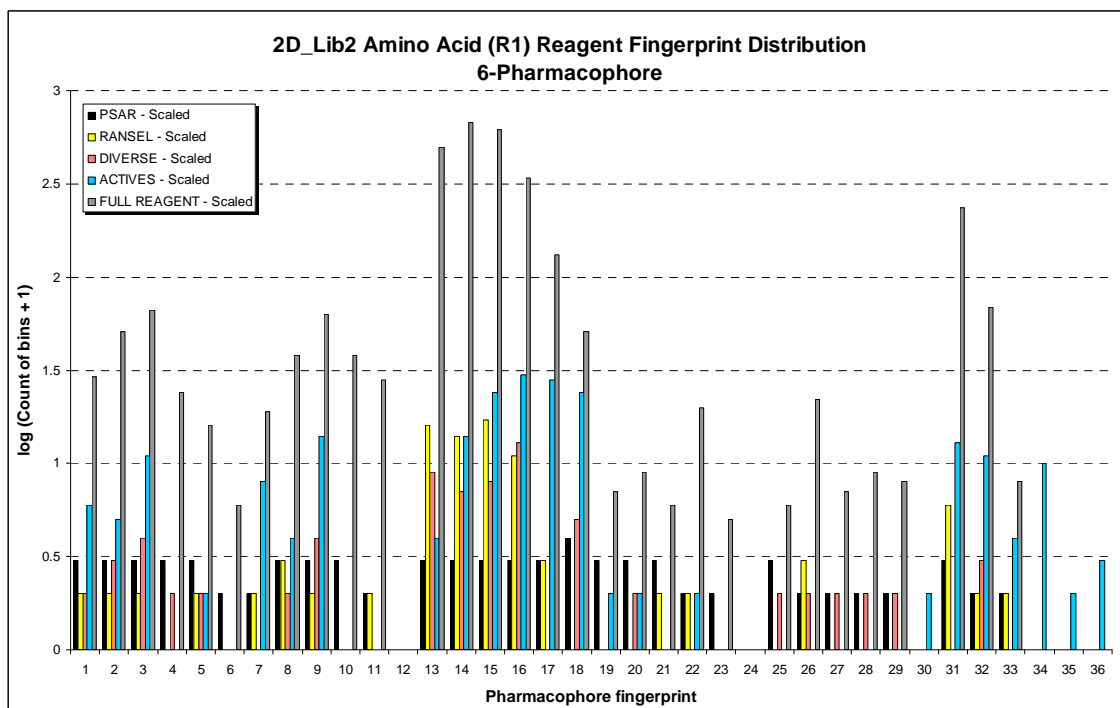


(a)

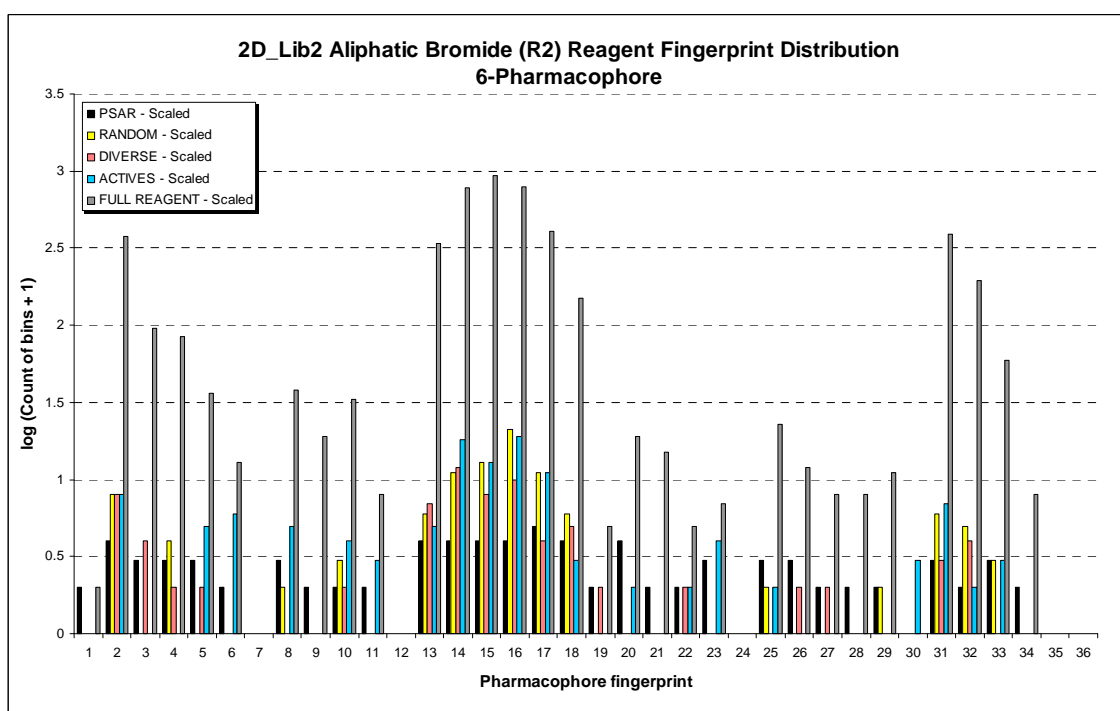


(b)

Figure B-2: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib1 library example.

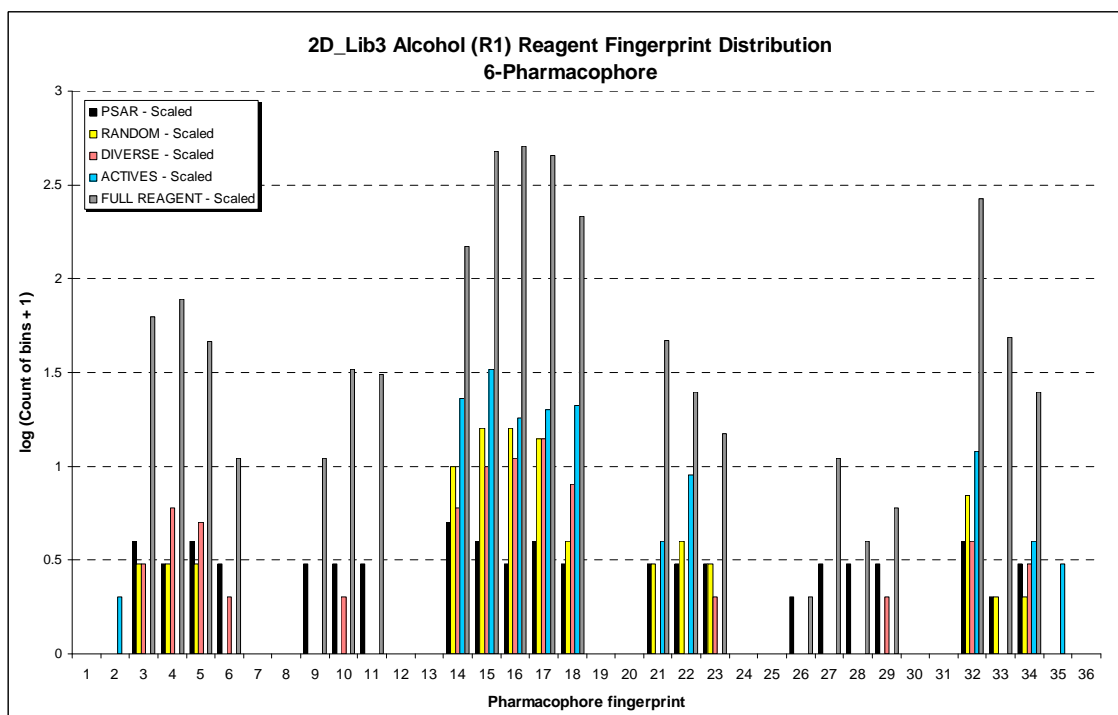


(a)

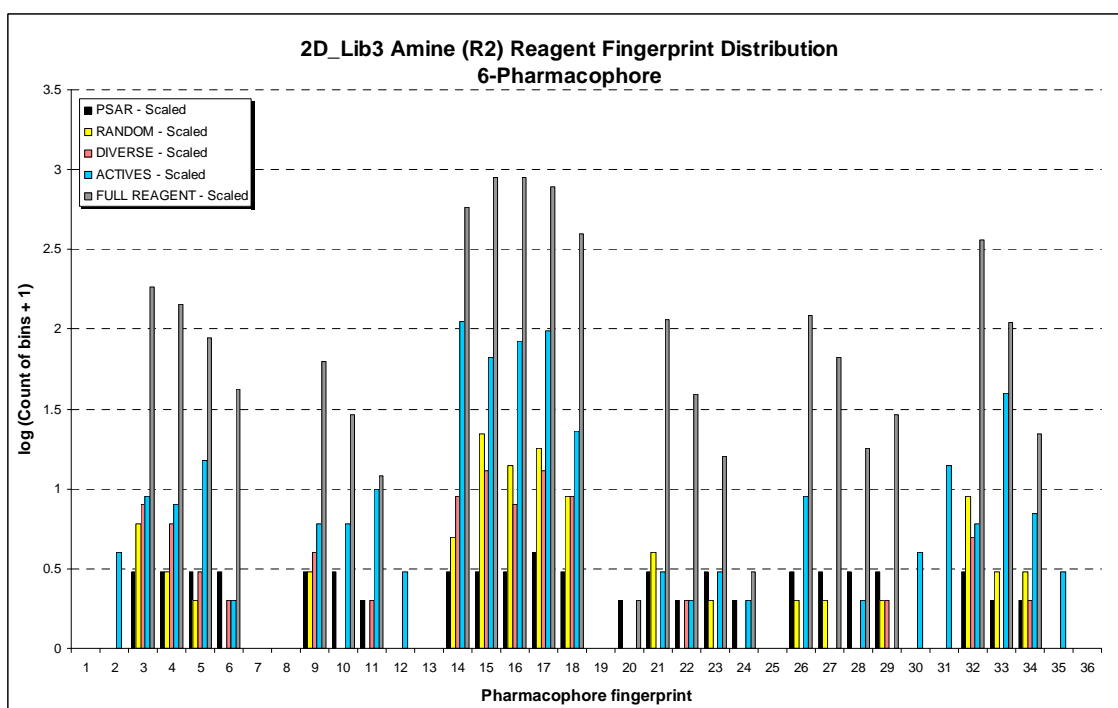


(b)

Figure B-3: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib2 library example.



(a)



(b)

Figure B-4: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib3 library example.

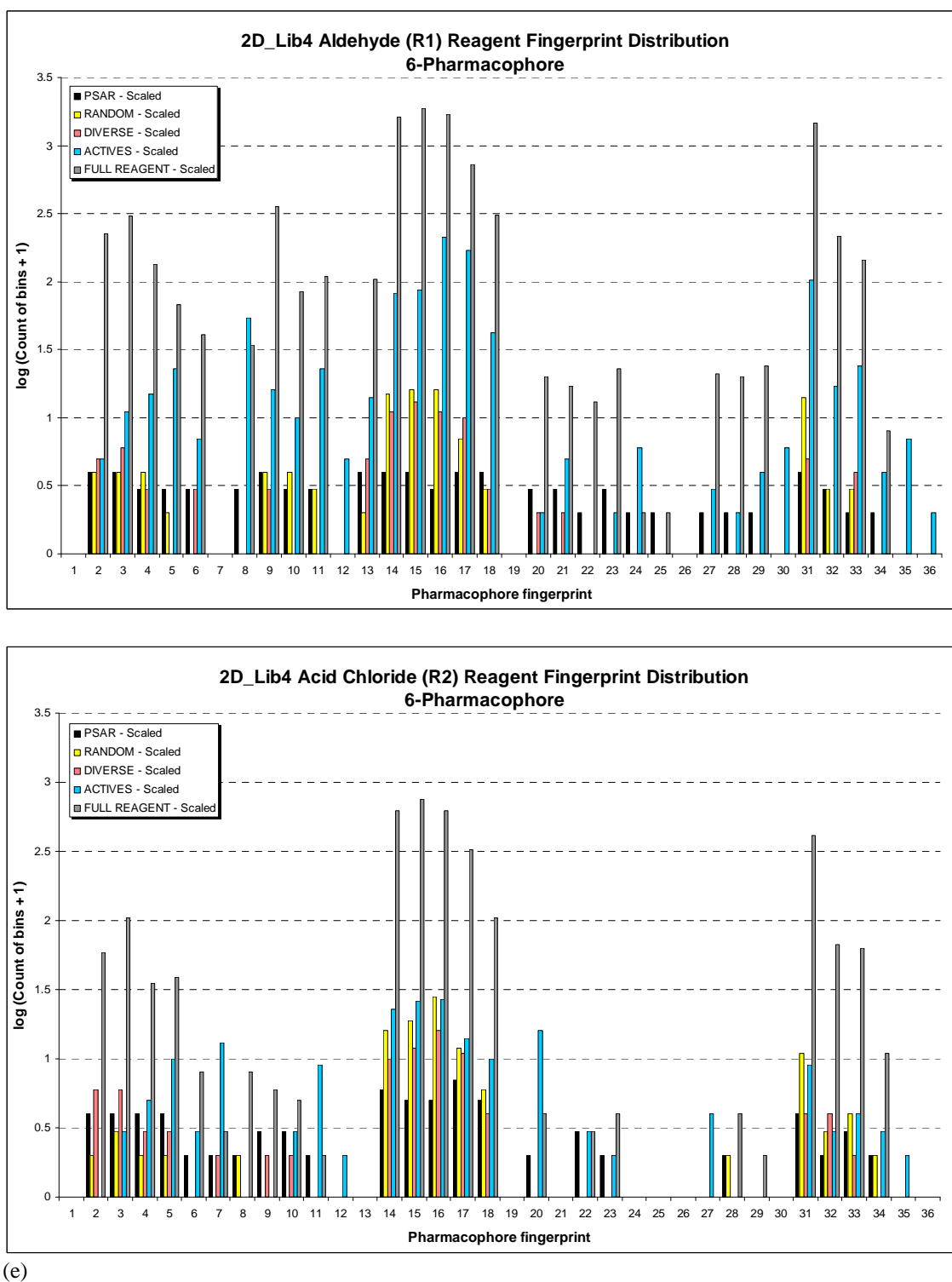


Figure B-5: Pharmacophore fingerprint distribution of R1-reagents (a) and R2-reagent (b) in 2D_Lib4 library example.

Appendix C

Comparison of Pharmacophore Coverage for Different Library Design Strategies by using 5 Pharmacophore Type Fingerprints

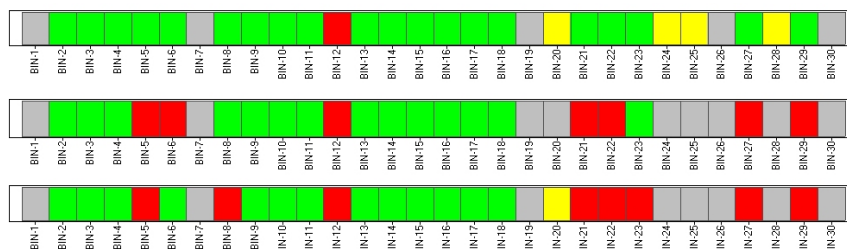


Figure C-1: Pharmacophore coverage comparison for R1-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 1D_Lib1 library example. Green colour represents bins that are both covered in the library design and active compounds; yellow colour represents bins that are present in the library design but not in active compounds; red colour represents bins that are absent in library design but present in active compounds while grey colour are for bins that are both absent in the library design and active compounds.

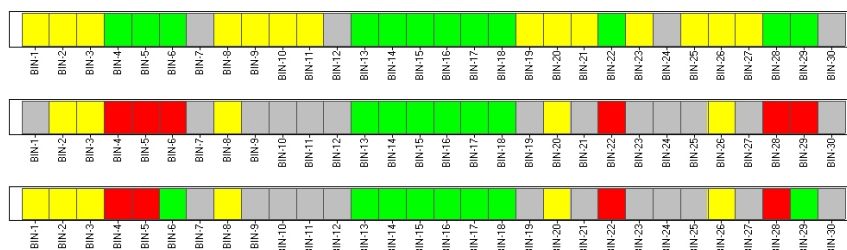


Figure C-2: Pharmacophore coverage comparison for R1-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib1 library example. Colour scheme is the same with Figure C-1.

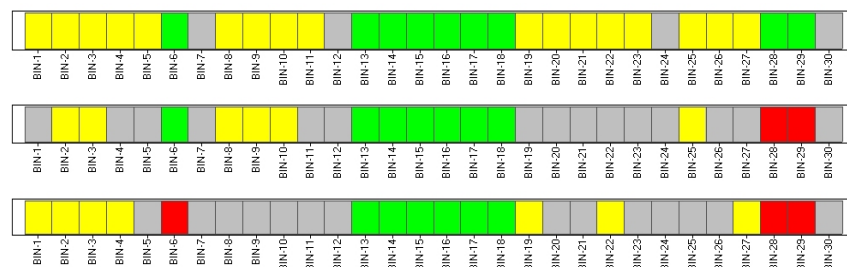


Figure C-3: Pharmacophore coverage comparison for R2-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib1 library example. Colour scheme is the same with Figure C-1.

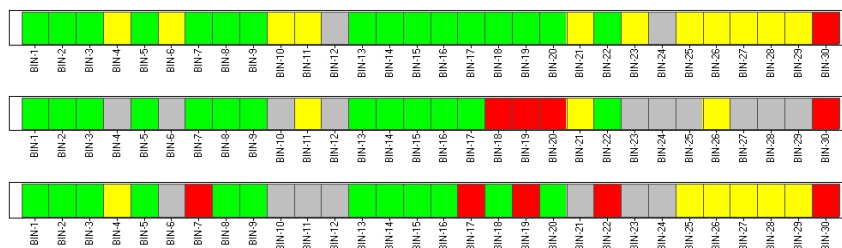


Figure C-4: Pharmacophore coverage comparison for R1-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib2 library example. Colour scheme is the same with Figure C-1.

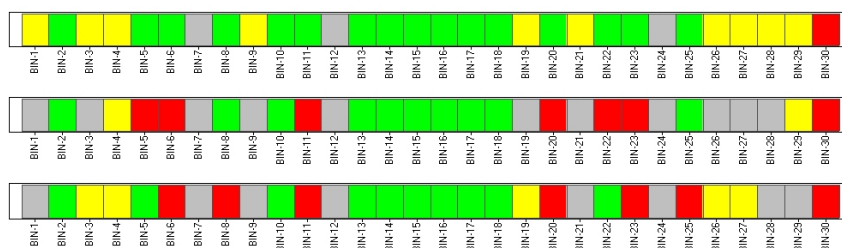


Figure C-5: Pharmacophore coverage comparison for R2-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib2 library example. Colour scheme is the same with Figure C-1.

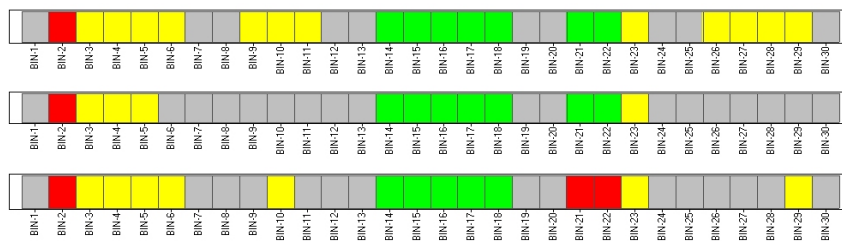


Figure C-6: Pharmacophore coverage comparison for R1-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib3 library example. Colour scheme is the same with Figure C-1.



Figure C-7: Pharmacophore coverage comparison for R2-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib3 library example. Colour scheme is the same with Figure C-1.



Figure C-8: Pharmacophore coverage comparison for R1-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib4 library example. Colour scheme is the same with Figure C-1.

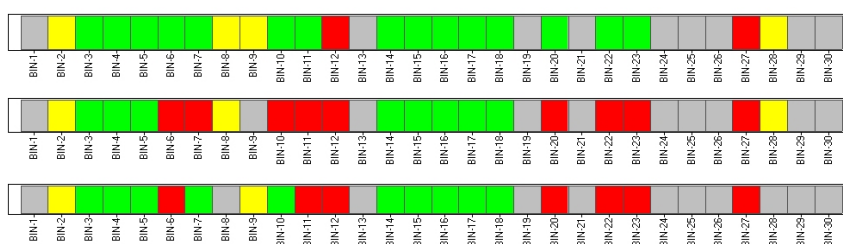


Figure C-9: Pharmacophore coverage comparison for R2-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib4 library example. Colour scheme is the same with Figure C-1.

Appendix D

Comparison of Pharmacophore Coverage for Different Library Design Strategies by using 6 Pharmacophore Type Fingerprints

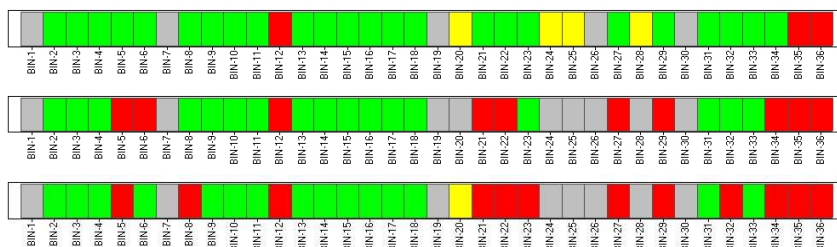


Figure D-1: Pharmacophore coverage comparison for R1-reagent among ProSAR (top), random (middle) and diversity (bottom) library in 1D_Lib1 library example. Colour scheme is the same with Figure C-1.

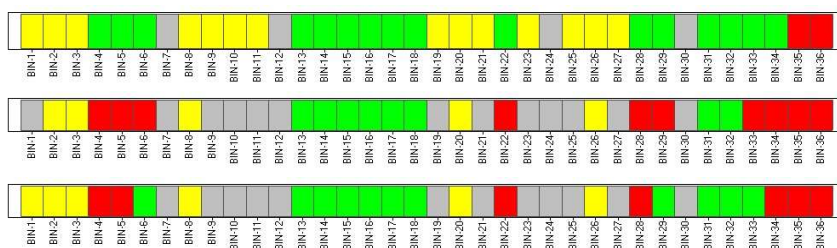


Figure D-2: Pharmacophore coverage comparison for R1-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib1 library example. Colour scheme is the same with Figure C-1.

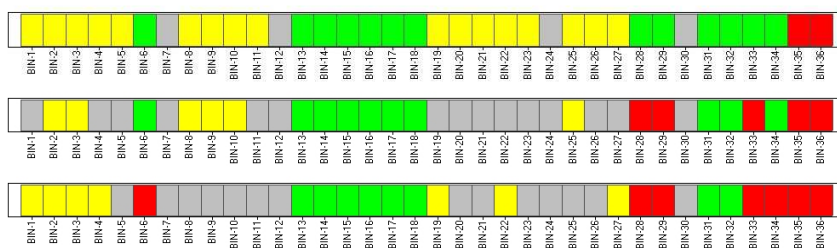


Figure D-3: Pharmacophore coverage comparison for R2-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib1 library example. Colour scheme is the same with Figure C-1.

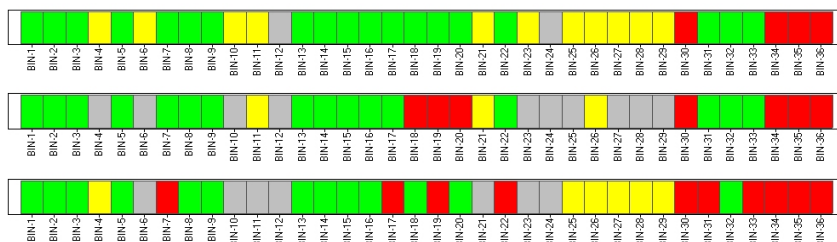


Figure D-4: Pharmacophore coverage comparison for R1-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib2 library example. Colour scheme is the same with Figure C-1.

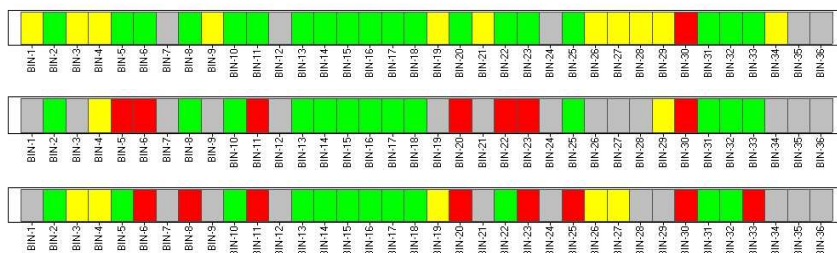


Figure D-5: Pharmacophore coverage comparison for R2-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib2 library example. Colour scheme is the same with Figure C-1.

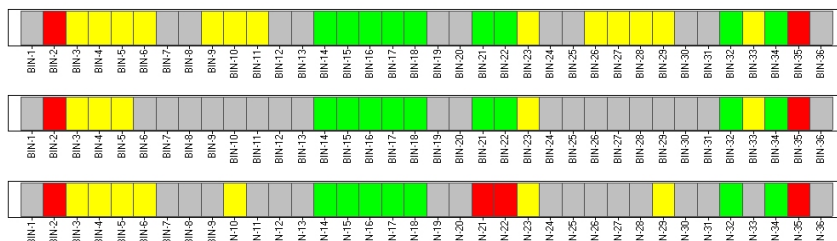


Figure D-6: Pharmacophore coverage comparison for R1-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib3 library example. Colour scheme is the same with Figure C-1.

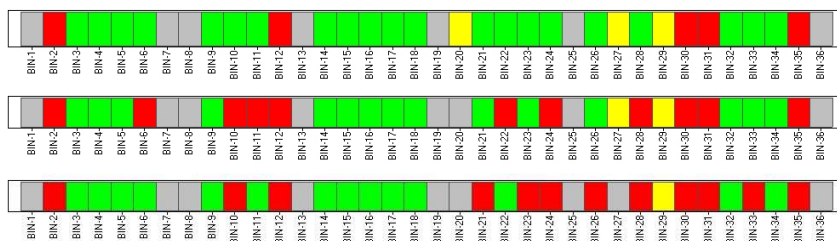


Figure D-7: Pharmacophore coverage comparison for R2-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib3 library example. Colour scheme is the same with Figure C-1.



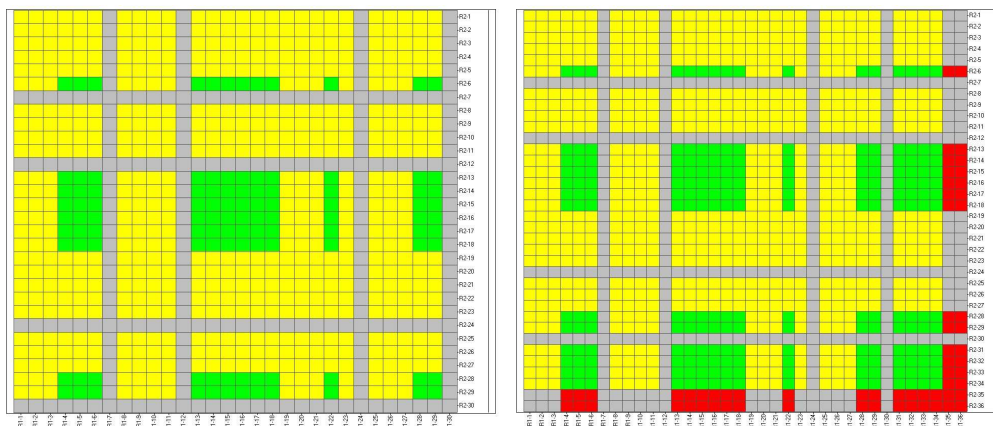
Figure D-8: Pharmacophore coverage comparison for R1-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib4 library example. Colour scheme is the same with Figure C-1.



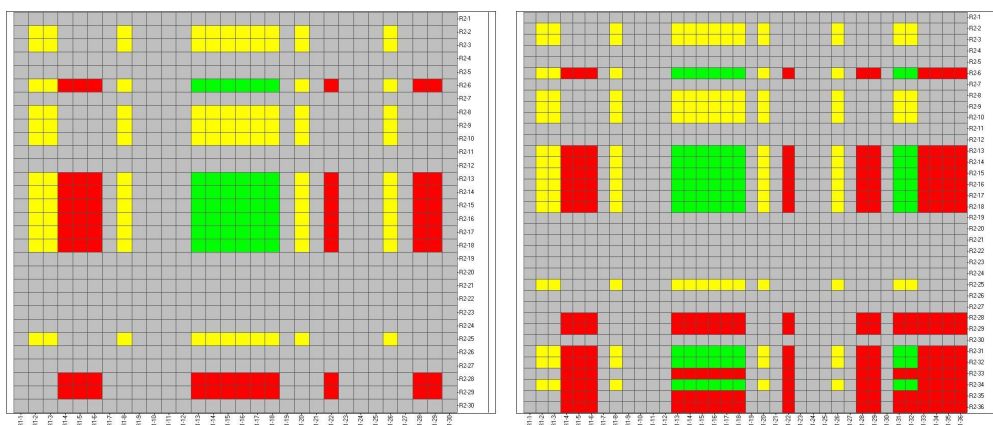
Figure D-9: Pharmacophore coverage comparison for R2-reagent between ProSAR (top), random (middle) and diversity (bottom) library in 2D_Lib4 library example. Colour scheme is the same with Figure C-1.

Appendix E

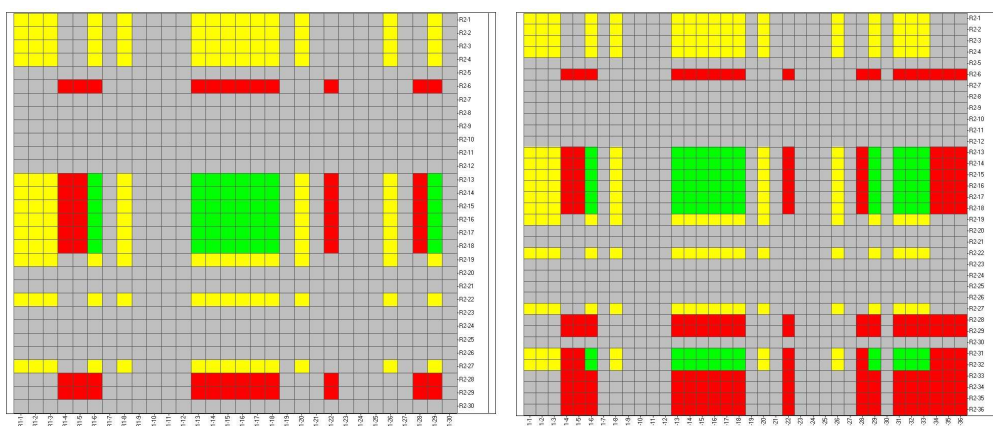
Comparison of Pharmacophore Coverage for Different Library Design Strategies in 2D Heat Map



(a) PSAR vs. Actives (2D_Lib1)



(b) RANDOM vs. Active (2D_Lib1)



(c) DIVERSITY vs. Active (2D_Lib1)

Figure E-1: Pharmacophore coverage comparison among ProSAR (a) random (b) and diversity (c) library in 2D_Lib1 library example. The maps on the left side are using the 5-pharmacophore type fingerprint and the ones on the right side are using the 6-pharmacophore type fingerprint. Color scheme is the same with Figure C-1.

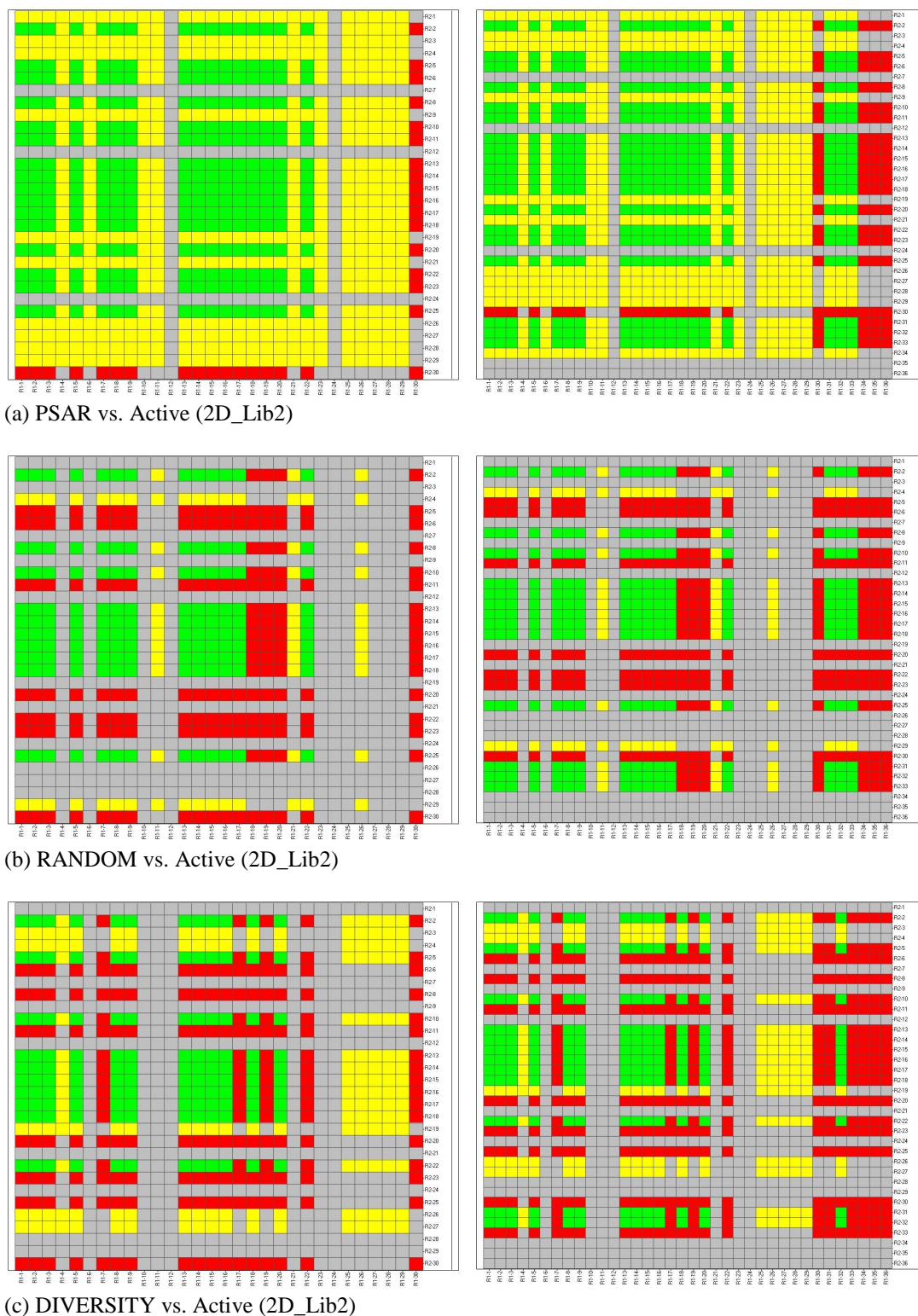


Figure E-2: Pharmacophore coverage comparison among ProSAR (a) random (b) and diversity (c) library in 2D_Lib2 library example. The maps on the left side are using the 5-pharmacophore type fingerprint and the ones on the right side are using the 6-pharmacophore type fingerprint. Color scheme is the same with Figure C-1.

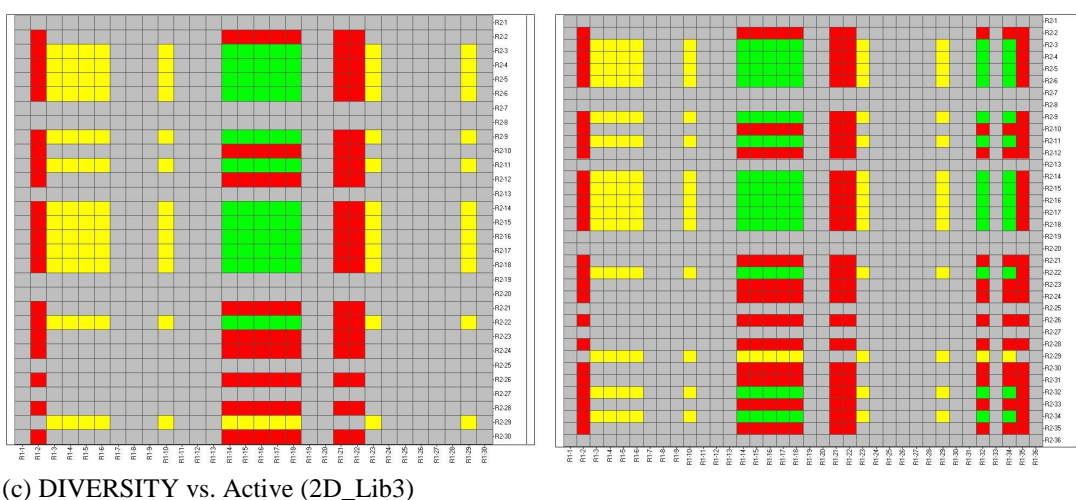
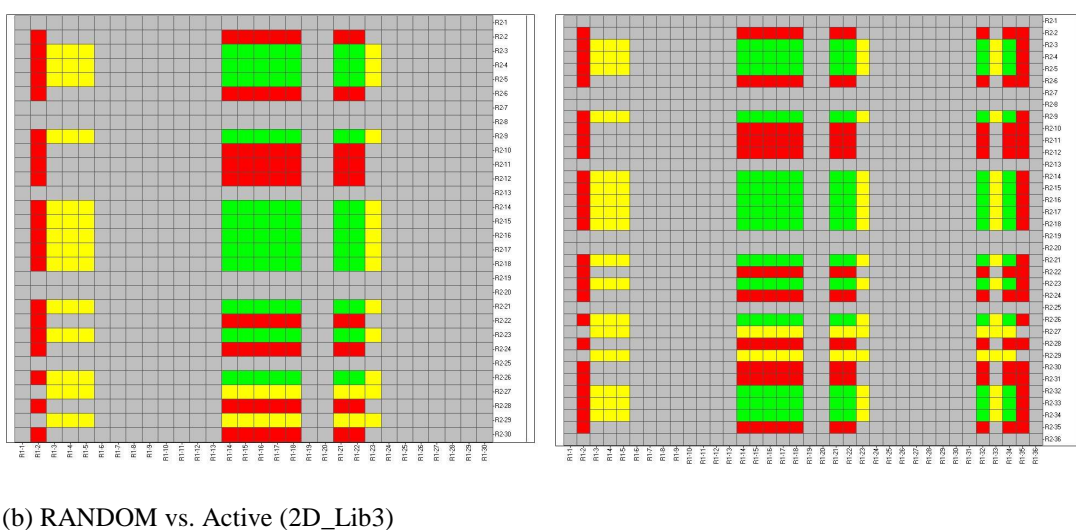
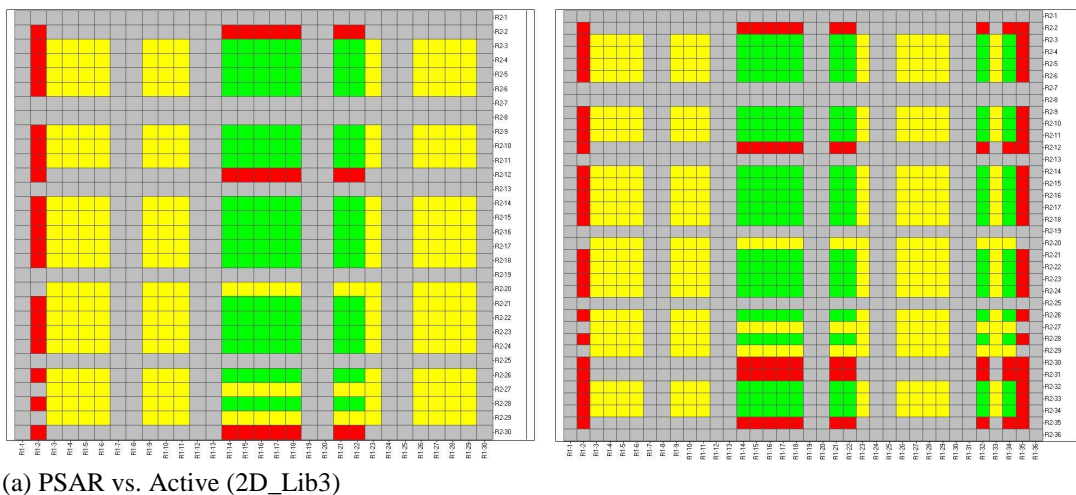
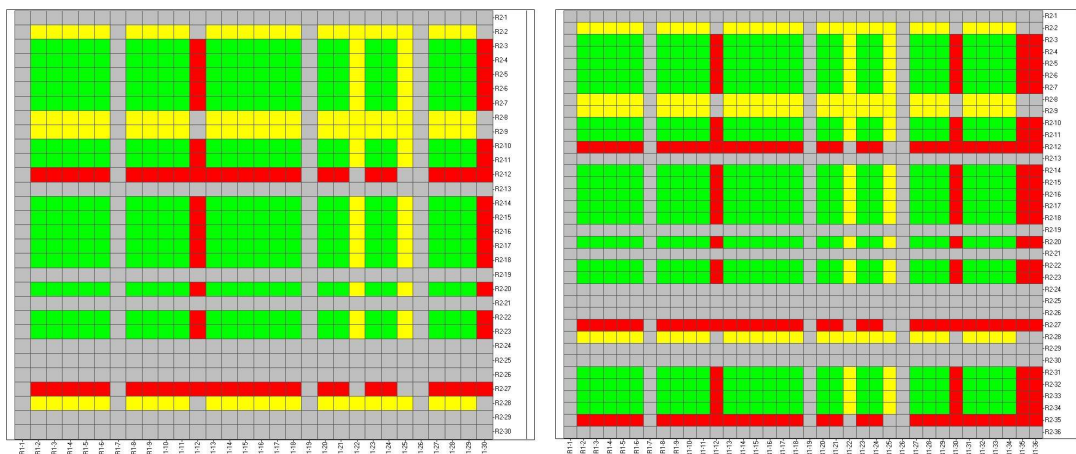
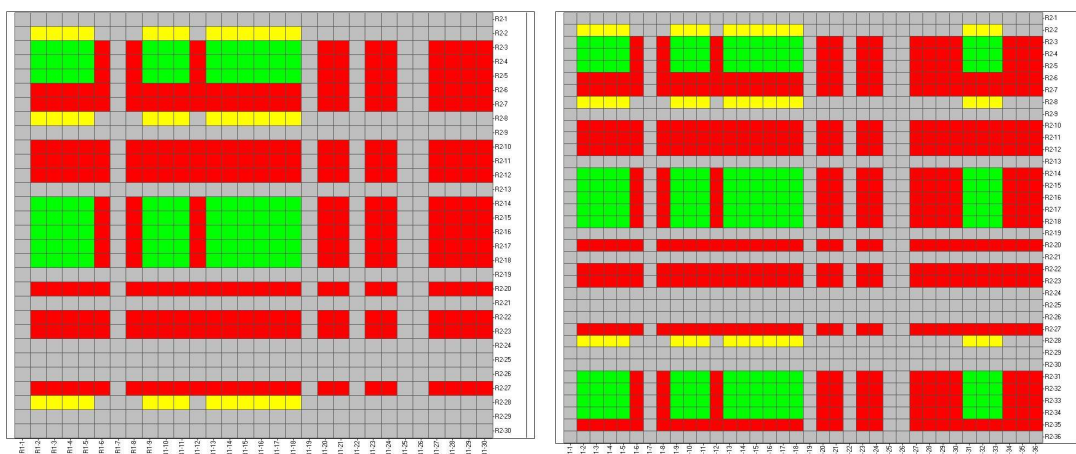


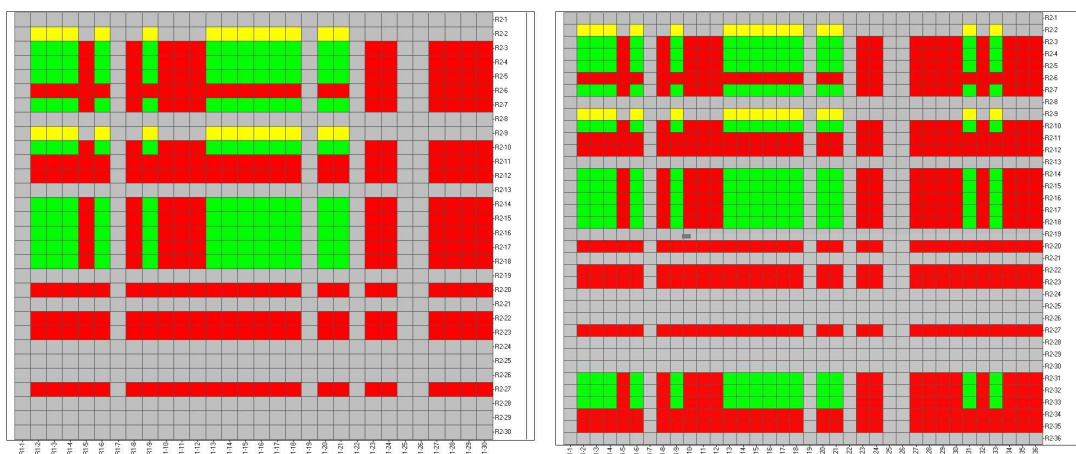
Figure E-3: Pharmacophore coverage comparison among ProSAR (a) random (b) and diversity (c) library in 2D_Lib3 library example. The maps on the left side are using the 5-pharmacophore type fingerprint and the ones on the right side are using the 6-pharmacophore type fingerprint. Color scheme is the same with Figure C-1.



(a) PSAR vs. Active (2D_Lib4)



(b) RANDOM vs. Active (2D_Lib4)



(c) DIVERSITY vs. Active (2D_Lib4)

Figure E-4: Pharmacophore coverage comparison among ProSAR (a) random (b) and diversity (c) library in 2D_Lib4 library example. The maps on the left side are using the 5-pharmacophore type fingerprint and the ones on the right side are using the 6-pharmacophore type fingerprint. Color scheme is the same with Figure C-1.