Novel Algorithm for Elucidating Biologically Relevant Chemical Diversity Metrics

By

Bhargav Theertham

Submitted to the Department of Electrical Engineering and Computer Science and the faculty of the Graduate School of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science

Chairperson **Dr. John Gauch**

Committee members:

Dr. Arvin Agah

Dr. Luke Huan

Dr. Gerald Henry Lushington

Dr. Jianwen Fang

Date of Thesis Defense: _____

The Thesis Committee for Bhargav Theertham certifies that this is the approved version of the following thesis

Novel Algorithm for Elucidating Biologically Relevant Chemical Diversity Metrics

Chairperson **Dr. John Gauch**

Committee members:

Dr. Arvin Agah

Dr. Luke Huan

Dr. Gerald Henry Lushington

Dr. Jianwen Fang

Date Approved: _____

Acknowledgements

I would like to thank my thesis advisor Dr. John Gauch for the guidance and support that he provided throughout the thesis work. I would also like to thank Dr. Arvin Agah, Dr. Luke Huan, Dr .Gerald Henry Lushington and Dr. Jianwen Fang for agreeing to serve on the committee. I am deeply indebted to Dr. Lushington who provided me with an opportunity to work as a Graduate Research Assistant at the Molecular Graphics and Modeling lab when I came to the University of Kansas in 2004 and for the constant encouragement and support that he provided while working on the thesis. It was an amazing learning experience working for the Computational Chemistry group at the Molecular Graphics and Modeling Lab during the two years of my stay at the University of Kansas. I would like to furthermore thank Dr. Jianwen Fang for the ideas and suggestions that he provided while working on the thesis. I would like to thank my parents who have always been by my side and have been motivating me to accomplish my goals in life. I would also like to thank my sister Babitha and brother-in-law Bala for the help and support they provided me while pursuing my masters.

Finally, I would like to thank the almighty GOD for his blessings and for providing me with the strength to overcome difficulties in life.

Abstract

Despite great advances in the efficiency of analytical and synthetic chemistry, the number of unique compounds that can be practically synthesized and evaluated as prospective pharmaceuticals is still limited. Given a known bioactive species, it is valuable to be able to readily identify a small subset of compounds likely to have similar or better activity. Many popular chemical diversity metrics do not perform very well in this role. A new emphasis on identifying diversity metrics that also encode biological trend information is thus emerging as a desired tool for guiding the assembly of targeted screening libraries. This thesis aims at developing novel algorithm that seeks to permit simultaneous evaluation of compound collections according to chemical diversity and potential bioactivity. An extensive set of descriptors are thus evaluated herein according to ability to differentiate chemical and biological similarity trends within compound sets for which screening results exist, and low-dimensional subsets are identified that retain such differentiation capacities. Bioactivity differentiation capacity is quantified as the ability to co-localize known bioactives into bioactive-rich clusters derived from K-means clustering. The descriptors are sorted according to relative variance across a set of training compounds, and filtered by mining increasingly finer meshes for pockets of descriptors whose exclusion from the model induces drastic drops in relative bioactive colocalization. This scheme is found to yield reasonable bioactive enrichment (greater than 50% of all bioactive compounds collected into clusters with enriched positive/negative rates) for screening data sets of some biological targets.

Table of Contents

List of figures

Chapter 1

Introduction:

The primary basis for medicine is the discovery of chemical compounds that can correct the biochemical malfunctions in our body. The main challenge here is to identify such compounds from among the myriad available. It is often difficult to intuit which compounds might provide the best therapeutic effect. One of the most prevalent strategies is to purchase or synthesize compounds and biochemically test for their activities by a technique known as High Throughput Screening (HTS). The main disadvantage of the HTS process is that a huge number of compounds often need to be screened in order to find plausible active compounds. This process can be augmented computationally, however, by a variety of methods including molecular diversity analysis. For the latter, all the chemical compounds can be represented by a set of descriptors that uniquely identify them. These chemical compounds can be projected into an n-dimensional space where each dimension represents a descriptor, and where spatial distribution should reflect chemical similarity of compounds plotted in this space. This project deals with developing a strategy to identify the best set of chemical diversity descriptors that would produce clusters whereby compounds that are located in similar chemical space locations exhibit similar biological activity. The descriptors should ideally also produce clusters with substantially enriched active/inactive ratio by sequestering an optimal number of active compounds into them.

Several methods have been developed before to reduce the dimensionality of chemistry space like Principal Component Analysis (PCA), multidimensional scaling, etc., but they do not take bioactivity into consideration. So the reduced dimensions may provide a reasonable account for descriptor variance across the compound set, but might not be relevant to the particular assay and might result in wrongly identifying active compounds when examined as a function of Cartesian distance from known actives. Some effort has been devoted to validating metrics in terms of their relationship to bioactivity. The most basic question to be answered by any metric validation approach is to what extent they place compounds that show affinity towards a receptor together i.e. how close are the compounds clustered to each other. Pearlman and Smith presented a novel approach known as activity seeded structure based clustering for metric validation and they also considered activity data along with structural information of the compounds. They used auto choose algorithm implemented by DiverseSolutions software to select the best BCUT descriptors to represent the chemistry space. They clustered compounds using the activity seeded structure based algorithm and found that all the active compounds were found in just three clusters. However their approach validates diversity related metrics only. Another attempt to validate the choice of descriptors based on biological activity was made by Bayada et al. They used a set of 560 compounds which had activity values in 38 selected activity classes. They also chose an additional 480 compounds that did not have activity values. A total of 86 descriptors were calculated for all the

compounds. They performed Cluster Significance Analysis (CSA) for each descriptor and each of the 38 activity sets i.e., 86 descriptors x 38 sets=3286 analyses. They used a threshold of 0.05 to judge whether a CSA was significant or not and decided to keep only those descriptors that are significant in more than 10 out of the 38 tests. The drawback of this approach is that while validating metrics, they consider each descriptor in isolation which does not guarantee that they will still be meaningful in combination. Waller et al. proposed a novel algorithm called FRED (Fast Random Elimination of Descriptors). This algorithm iteratively generates offspring models, each model containing a set of fixed or variable number of descriptors. For each of the offspring models generated, a fitness function is calculated and once all the models are generated, they are sorted according to their fitness. The descriptors present in the low fit models and not present in the more fit models are considered detrimental and put in a taboo list. They are given a second chance in the next generation, but if they still produce sub optimal models, they are eliminated from the list of allowable descriptors. This approach however does not take bioactivity into consideration.

The aim of our thesis is to develop a strategy to identify the best set of chemical diversity descriptors that would produce clusters where compounds that are located in similar chemical space locations exhibit similar biological activity. The descriptors should also produce clusters with substantially enriched active/inactive ratio by sequestering an optimal number of active compounds into them. The approach

developed in this project helps in designing intelligent compound collection by enhancing the ability to map potentially interesting compounds (to be tested for their biochemical activity for a particular assay) into these active clusters. The mapping enables identification of compounds with enhanced potential for activity, which can then be tested in the wet lab for their activities. This greatly reduces the number of compounds to be screened and thus makes the screening process efficient.

We tested our approach on subsets of the NCI human tumor cell line growth inhibition assay data for the SK-OV-3 Ovarian cell and IGR-OV-1 Ovarian cell lines containing 7718 and 7986 compounds respectively. Our aim was to isolate the descriptors that would exhibit biological discrimination i.e., sequester optimal number of active compounds into clusters with substantially enriched active/inactive ratio. In the first phase, we clustered all the compounds using K-means clustering. The resulting clusters are evaluated for their purity by calculating the ratio of active to inactive compounds present in that cluster.

The clusters that have the maximum and minimum ratio are assumed to contain descriptors that have the best ability to segregate active and inactive compounds. Then using a ranking algorithm, these descriptors are ranked according to the relative variance in the two clusters. Datasets with different number of descriptors are formed taking equal number of top ranking descriptors from both the most positive and most negative clusters and the purity of the clusters again is calculated in each case using the K-means algorithm. The next phase is known as Drastic Drop Isolation where we try to filter descriptors by mining increasingly finer meshes for pockets of descriptors

whose exclusion from the model induces drastic drop in relative active to inactive ratio. We used leave out one fold (LOOF) cross – validation approach for training data. This method is a variation of k-fold cross validation, whereby in each fold, 'k-1' subsets are used for training and the remaining 1 subset as the test set. The dataset was randomly divided into five parts, each part having equal number of active and inactive compounds. In each fold of testing, the best model is identified using the above mentioned alogrithm and then is tested across all the other test sets. The model that has the best average over all the test sets is chosen as the final model. The optimal model for the first dataset consisted of four dimensions and was able to segregate 55.78% of actives in just one cluster. In case of second dataset, the optimal model was three dimensional and it was able to segregate 52.30 % of total active compounds in just three clusters. The results of clustering the compounds using the best models were visually verified using 2D plots. We also used a commonly used chemical diversity analysis program, Diverse Solutions for selecting the best set of descriptors and then use K- means clustering for segregating active compounds into distinct regions of chemical space and found out that our algorithm outperformed this program. We also demonstrated a practical example by mapping a set of compounds synthesized at KU onto the chemical space derived for the two assays to see which of those compounds are potential active compounds.

Chapter 2 gives the information about the drug discovery process, the high throughput screening process, bioactivity and bioassay, diversity analysis, combinatorial chemistry etc that would provide the required information

needed to understand the relevance of this problem in the pharmaceutical industry. Chapter 3 discusses in detail, the motivation for the work in this thesis, which is to identify the most biologically relevant descriptors that would result in effective segregation of the active compounds into clusters that have significantly enhanced active/inactive ratio. Chapter 4 discusses the previous work done relevant to the problem. Chapter 5 discusses the implementation part of the thesis. This section discusses in detail about the ranking algorithm that we used, the drastic drop isolation approach, and other relevant details. Chapter 6 is the results section that discusses the results of implementing the algorithm discussed in chapter 5 on two different datasets. Chapter 7 is the analysis section which analyzes the results by generating 2 dimensional plots for the best models from both the datasets that shows qualitatively how well the active compounds are clustered. Chapter 8 concludes the document with an overview of the work done in the thesis. Chapter 9 discusses the future work that can be done as an extension to the current thesis. The next section gives brief background information about some important concepts in pharmacology, drug discovery and some techniques used to identify compounds that are potential drugs.

Chapter 2

Background:

2.1 Concepts of pharmacology:

Pharmacology is the study of drugs – what they are, how they work and what they do. Drugs can be defined as chemical agents that interact with specific target biomolecules, thereby producing a biological effect. Pharmacology includes the study of the manner in which the function of living tissues and organisms is modified by chemical substances and the study of the effect of chemical agents on living processes. The important aspects of pharmacology are:

1. The use of drugs to study physiological mechanisms.
2. The use of drugs in medicine.
3. Understanding the biological effects of environmental chemicals.

Another important aspect of pharmacology is the study of the drugs once they enter the body using *ADME* analysis:

Absorption: How the medication is absorbed into an organism?

Distribution: How is it spread through different tissues in the organism?

Metabolism: Is the medication converted chemically once it is inside the body and into which substances? Are those substances active or toxic?

Excretion: How is the medication eliminated from the organism?

2.2 Drug Discovery:

Drug discovery and development is an important branch of pharmacological sciences. It is a process by which specific chemicals are designed for therapeutic activity, based largely on their capacity to inhibit, activate or modulate biochemical processes carried out by proteins, lipids or other biomolecules.

The modern approach to drug discovery involves an attempt to understand more about a disease and the elements that cause it, using this knowledge in order to intuit and screen suitable compounds. After initial identification of a candidate biomolecule, the process of drug discovery proceeds to the identification of possible drug candidates, followed by their synthesis, characterization, screening and assays for therapeutic efficacy. Once a compound has shown its value in these tests, it may be selected for clinical tests en route to possible use as a real pharmaceutical therapeutic.

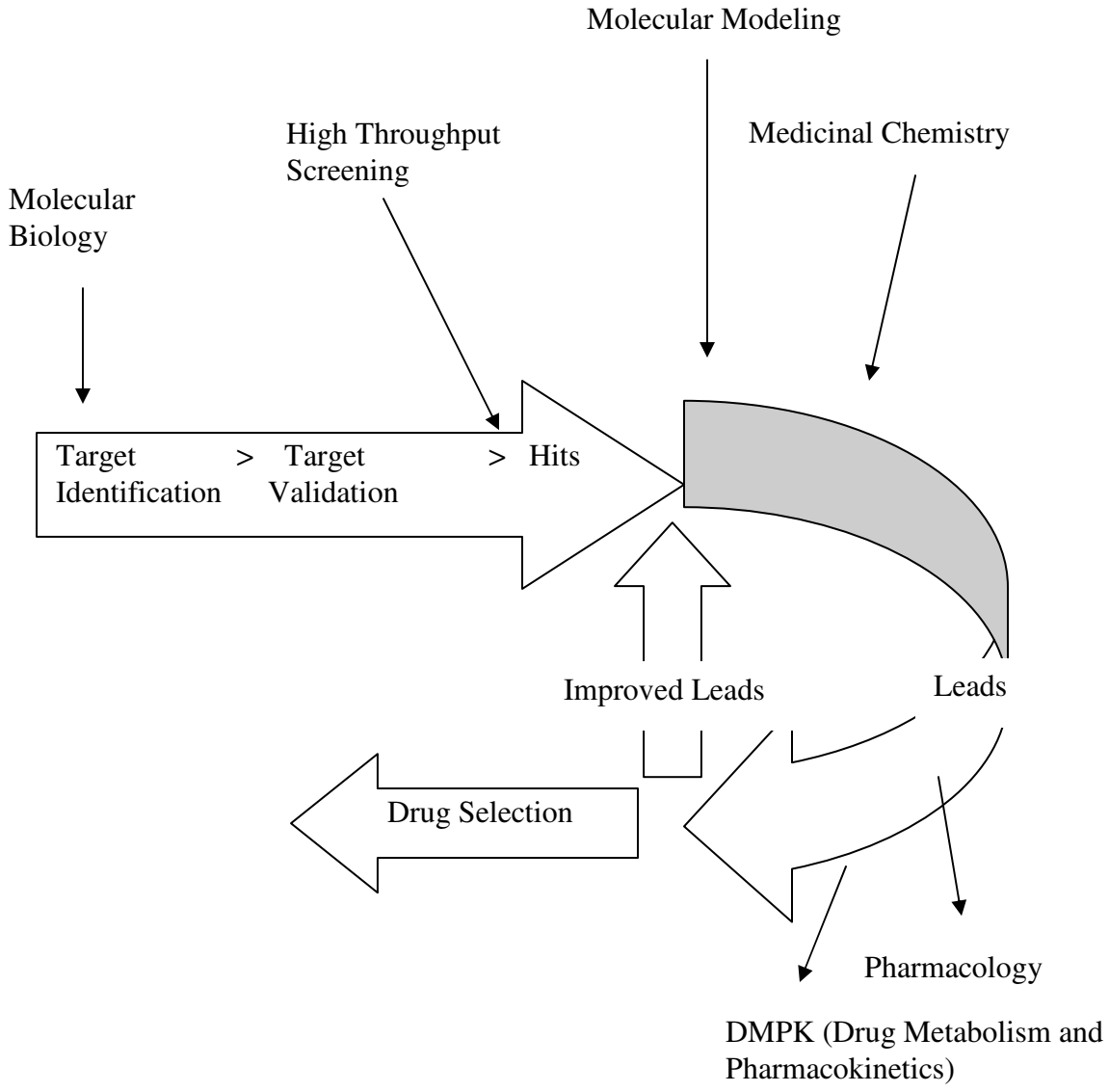


Figure 1.0 Life cycle of drug discovery [1]

The following sections will briefly describe the steps involved in drug discovery along with a brief description of different disciplines of science involved in the process.

The drug discovery is a cyclic process. The lead optimization stage is iterative as the compounds are modified, re-analyzed and tested [1].

2.2.1 Target Identification:

Drugs usually act on specific biochemicals (e.g., proteins, lipids, etc.) whose dysfunctional behavior in the body is believed to be associated with disease. Such biochemicals are generally referred to as drug targets. Scientists use a variety of techniques to identify, isolate and learn more about the target functions and how these influences diseases. Drug design strategies for a given target typically depend substantially on whether a target is new or established. Established targets are those for which there is a good scientific understanding, supported by lengthy publication history of how the target functions in normal physiology and how it is involved in human pathology. New targets are typically those that have been recently discovered and whose functions are still being determined through current basic scientific research.

2.2.2 Target Validation:

To select targets most likely to be useful in the development of new treatments for disease, researchers analyze and compare each prospective drug target to others based on their association with a specific disease and their ability to regulate biological processes and chemical compounds in the body. Tests are conducted to confirm that drug interactions with identified targets are associated with a desired change in the behavior of diseased cells.

2.2.3 Synthesizing compounds using Combinatorial Chemistry:

Combinatorial Chemistry is a process of efficiently synthesizing large number of chemical compounds, ideally with potentially drug-like properties that make them suitable for use in subsequent screening studies to detect specific chemical families that might be active against a given target. A more detailed description about this method is provided in a later section.

2.2.4 Hit Identification:

High Throughput Screening is a rapid bioactivity characterization technique employed to find an initial set of compounds of potential applicability to the target of interest. In this process, a large number (up to hundreds of thousands) of compounds from existing libraries or those synthesized using combinatorial chemistry process as mentioned above, are screened concurrently against the target, and the ones showing some effect on the target are identified as hits, and are subject to further scrutiny. The

effect is usually referred to as bioactivity and there are lots of methods that quantify it. A detailed description of bioactivity, assays and HTS process is given in the subsequent sections.

2.2.5 Lead Identification:

The empirical knowledge derived from screening studies may be augmented with molecular modeling studies to identify potential lead compounds (i.e., hits of substantial prospective value) and to focus further testing and trials on the best drug candidates. Molecular modeling refers to the theoretical and computational techniques to model or mimic the biological effect of molecules, including their propensity to interact with a given target, their potential side-effects, as well as their potential suitability for drug use (i.e., solubility in biological fluids, ability to be absorbed into tissue and distributed through the body, etc.).

2.2.6 Lead Optimization:

The next step after lead identification is lead optimization. It involves using medicinal chemistry techniques to make synthetic modifications to the hits identified in the previous step in order to improve the biological properties of the compounds, based on information derived from studying existing drugs, their biological properties and their structure activity relationships. The lead optimization process is iterative. The lead compounds are successively modified in some way, re-analyzed and tested.

2.2.7 Drug Development:

Drug development is generally considered to be the process during which a drug candidate is prepared for human clinical trials. The processes of pre-nominating a candidate drug and the final release of a drug are lengthy and governed by very strict rules. Though no drugs are ever rushed quickly to market, application of advanced technologies and careful analysis in the steps preceding clinical studies can greatly expedite the process up until this point and improve the likelihood of eventual successful approval.

The steps described above describe the various stages involved in the drug discovery process. The next section deals with some key concepts in the drug discovery process.

2.3 Bioactivity and Bioassay:

Bioactivity is an expression describing the beneficial or adverse effects of a drug on living matter. Activity is generally dosage dependent but the effects may range from beneficial to adverse when going from low dosage to high dosage. Pharmacological activity is usually taken to describe the beneficial effects of the drug candidates, whereas toxicity is the term that quantifies adverse effects.

A biological assay is a type of experiment conducted to measure the effects of a substance on a living organism. Screening is a term given to a process of analyzing the “effect” of a series of compounds on a given target. The “effect” can vary

depending on the stage of screening or the nature of the target. The nature of the physical measurement varies although fluorescence, radioactive isotope labeling and Liquid Chromatography / Mass spectrometry are the common methods used to quantify the "effect". IC_{50} is one quantity that is used as a measure of drug effectiveness. It represents the solvated concentration of an inhibitor that is required for 50% inhibition of the target bio-molecule. IC_{50} value depends on various conditions under which inhibition is measured. For example for enzymes, IC_{50} value increases as enzyme concentration increases. In this thesis, we have used systematically employed IC_{50} values derived from NIH-funded studies as the measure of activity for our analyses.

2.4 The High Throughput Screening Process:

High Throughput screening is a process where in hundreds and thousands of compounds are screened against a target and the ones that have measurable bioactivities are noted, and identified as 'hits'. These hits are passed on to the next stage for further evaluation as potential drug candidates. The High Throughput Screening process has led to a significant improvement in the lead identification process by broadening the spectrum of sample structures displaying some possible activity and quickly eliminating classes of compounds that have no effect. HTS is typically applied in the first phase after target identification and validation [1].

In HTS experiments, screens are run by loading compounds in multi well plates. Given a set of plates, automated screening systems have been developed to run the experimental measures. These systems employ robots for sampling wells on plates in the correct order and running appropriate analyses. There are two kinds of screening in HTS – primary screening and secondary screening.

In primary screening, HTS attempts to screen thousands of compounds at a time with a view of obtaining a single value for each sample at a single fixed known dose – whether it is "active" and has an effect on the target, or it is inactive. It is a simple "pass-fail" test that makes process of screening thousands of compounds at a time possible. The subsequent process of elimination constitutes the threshold between active and inactive. These kinds of screens are called primary screens. They are likely to be the first screens in a given study run against the target of interest. The measured actives from the primary screen are called "hits" and they move into the next stage of processing [1].

In the secondary screening, assays are run on the compounds obtained from primary screening with an aim to investigate the effect of compounds on the target in more detail. Each compound will have separate measurements taken on it, for example a dose-response curve. The process of secondary screening is also highly automated and aims to reduce the number of interesting compounds to some small fraction

(often about 5%) of the original set of hits, thus permitting the identification of a small number of leads that have potential for further development into drugs [1].

Since millions of compounds have been synthesized and are potentially available for HTS analysis, it is physically impossible to screen all of these compounds, even with the best available automation technology. An effective primary screen thus relies heavily on an analytical capability of isolating a small fraction of these compounds that are either representative of the set as a whole, or particularly likely to be of interest in application to a given target. Diversity analysis is a method that is used to select a smaller subset that best represents the whole set. As hits are discovered, focused virtual combinatorial libraries can be designed, through the synthesis of additional similar members. The next section describes the concept of diversity analysis in detail.

2.5 Diversity Analysis:

Diversity and similarity analysis typically involves defining N different molecular descriptors to serve as Cartesian-axes that allow you to plot the distribution of molecules in N-dimensional space, with molecules that are spatially close to each other considered to be similar, and those far from each other assumed to be dissimilar. This concept can be exploited for the selection of chemical library subsets that are either highly diverse (i.e., those composed of optimally dissimilar and minimally redundant species, as is ideal for primary screening [2]) or finely targeted (composed

of species all very similar to particular compounds of interest, as may be appropriate for secondary screening). Specifically, given a set of relevant axes, each defined by a given descriptor, there are various representativity and clustering methods that can be used to partition the library, and one may thus either make uniform selections from as many different clusters as possible to obtain a diverse subset, or make all selections from within a narrow range to obtain a targeted collection.

The descriptors used to define diversity space are simple molecular properties that encode physical or chemical information about the compounds. Diversity describes the ratio of dissimilarity to similarity within a set of chemical structures. The term diversity cannot be used in an absolute sense: two chemical structures might appear similar when seen from one frame of reference but might actually be very dissimilar when looked at from a different frame of reference. So the concept of diversity makes best sense when viewed within a frame of reference. This kind of reference-dependent diversity is known as 'functional diversity' as opposed to the more general concept of 'structural diversity' [2]. Structural diversity is 'internal and relative' because the structures are compared with one another, without any reference to behavior in an external system (i.e. an assay). In contrast to structural diversity, functional diversity can be defined as 'external and absolute', because the degree of diversity is measured according to an absolute benchmark [2]. In the case of drug design, the most important frame of reference is biological effect. In this frame of reference, particular axes might carry more weight than others and be important if

they correlate with observed bioactivity. If we do not use these biologically important axes when comparing the diversity of two compounds, we may wrongly conclude based on general structural similarity measures that two compounds have similar bioactivities, since the descriptors used may have very little relationship to trends observed in a given assay [2]. Functional diversity models constructed based on a bioactivity model are thus much more reliable tools for intuiting comparable therapeutic function among collections of molecules. Constructing them requires some algorithm for identifying optimal sets of axes based on relationships with activity trends.

2.6 Descriptors:

One of the important components in diversity analysis is molecular description. In order to generate meaningful chemical space representations, it is necessary to find mathematical representations that can encode molecular properties that discriminate compounds with different structures, shapes, compositions and physiochemical properties. A large number of descriptors are available that can encode molecular properties with various degrees of complexity and information content [4]. Descriptors are classified according to the type of information they represent, and their dimensionality (one-dimensional, two-dimensional and three-dimensional). Physicochemical properties such as molecular weight, molar refractivity, logP are often referred as one-dimensional (1D) descriptors as they are calculated based only on the molecular formula or constituent atom types, whereas two-dimensional

descriptors encode atomic connectivity information, and 3D descriptors account for properties relating to spatial effects. Both 1D and 2D are generally very computationally efficient, however 3D terms are often much more demanding to evaluate.

2.7 Combinatorial Chemistry:

Even for the most extensive compound collections available for HTS studies, diversity analysis reveals that there are substantial holes in the cartesian chemical diversity space (i.e., regions of property space for which there are very few constituent molecules in a given collection). While one may assemble collections with improved diversity by carefully sampling species purchased from a variety of different chemical vendors [4], the option also exists to plan massive chemical syntheses to fill the holes in diversity space, as is now being attempted through nationally coordinated efforts. Combinatorial chemistry is an important technology for efficiently accomplishing such directed syntheses.

Combinatorial Chemistry is technique whereby large numbers of compounds may be synthesized based on performing a defined set of reactions with a variety of reactants that differentially modify a given starting structure (a.k.a., scaffold) by substituting new chemical groups onto a variety of different chemically reactive sites within that scaffold [4]. Synthesis of compounds in a combinatorial fashion can quickly lead to a large number of molecules. Researchers attempting to optimize the activity profile of

a given compound may create a 'library' of many different but related compounds. Advances in robotics have lead to an industrial approach to combinatorial synthesis, enabling companies to produce large numbers of new and unique compounds every year. In order to assess the vast number of structural possibilities, researchers often create a 'virtual library': a computational enumeration of all possible structures of a given pharmacophore with all available reactants. Such a library may consist of thousands to millions of 'virtual' compounds. Diversity analysis is usually performed on both the target collection that one plans to contribute molecules to, as well as the proposed new virtual combinatorial libraries in advance of their synthesis, with the goal of identifying that subset of the proposed compounds that really do enhance the chemistry space of the target collection.

Once the HTS analysis has been done on the diverse set, the determined activities may pinpoint regions of chemical space that are particularly fertile. The next step would then be to screen many more compounds within this fertile region to see if there are other compounds in this area that are stronger inhibitors than the initial hits. Similarity analysis can be done to identify possible compounds to screen with respect to the initial hits.

The descriptor sets used for the current diversity methods have typically not been tailored for bioactivity recognition. We have tried herein to address this problem by developing an algorithmic approach that would identify receptor relevant subspace that would enhance the percentage of active compounds in the more active clusters.

The next section gives detailed information about descriptors that may be used in such diversity analyses.

Chapter 3

Motivation:

3.1 Metric Selection:

Computational expediency is a major consideration in choosing descriptors for treating large numbers of molecules, thus 1D and 2D descriptors are typically favored for diversity analysis (wherein one often processes large libraries) while 3D properties are frequently omitted [5]. However the three-dimensional spatial information that the latter encode often plays a very important role in determining interactions with biological systems. As a result, existing diversity and similarity analyses frequently do not correlate well with observed bioactivity and the descriptors that are chosen in these models are not always the best for resolving activity trends. It is our opinion that a tangible increase in computational expense is significantly less of a hindrance to successful drug design than is an improperly selected screening set. Thus, the overall goal of our effort was to find the most biologically relevant descriptors, regardless of dimensionality, that would result in more effective segregation of active compounds into distinct clusters for a significant number of assays. Specifically, this thesis attempts to elucidate mechanism for specifying a Cartesian framework consisting of axes whose corresponding properties would concentrate the greatest number of active compounds into one or more clusters with a substantially enriched active/inactive ratio. Since the properties that best correlate to bioactivity vary as a function of the specific biological target, our analysis was

performed for two different assays. The simple Cartesian framework produced by this analysis allows one to detect and highlight active regions of chemical space. The practical benefit of this is that virtual libraries can then be mapped onto these active regions to identify which compounds from the virtual library demonstrate the best chance of yielding biological activity for a particular assay.

A huge number of possible descriptors are available for any diversity analysis. When designing a chemical diversity space model for a given assay, there is great value in being able to consider as many descriptors as are computationally feasible, in that such broad consideration affords the best chance of finding a subset that optimally correlates to the results of that assay. The drawback to the consideration of a large number of descriptors, however, is that those with no tangible correlation to observed activity interfere with the ability to use diversity concepts such as chemical space distances to perceive similar or dissimilar compounds. It is thus crucial to eliminate superfluous dimensions from the chemical space and thus distill the chemical space model down to those dimensions from which one can meaningfully, by use of Cartesian distances, identify neighbors of lead compounds, permit informed comparison of different libraries and other diversity related tasks.

The main computational challenge in developing a functional diversity model is the process of dimension reduction by which one can identify receptor relevant metrics that optimally cluster active compounds together and discard those with lesser or no

value. Given a broad enough starting set of descriptors, it is reasonable to hope that for any given assay there will be a relatively manageable number of relevant descriptors or metrics that will accurately represent activity considerations and tightly cluster active compounds together, however the process of perceiving biological relevance is challenging. Given a data set covering a substantial number of compounds of varying affinity towards a receptor, we have thus developed and pursued an algorithmic process to perform such descriptor discrimination.

Chapter 4

Related work

4.1 Dimensionality reduction of descriptor spaces:

There are several approaches proposed to reduce the dimensionality of the chemistry space. This section will provide an overview of some of them. Descriptors that are correlated increase the dimensionality but provide no additional information, so these descriptors can be eliminated without any loss of information. The first step to prune descriptors is thus to remove those that are significantly intercorrelated [4]. Whitley et al.used an unsupervised forward selection method to automate this process [6]. In the second step, statistical analysis is performed on the remaining descriptors. The distribution of the molecules as represented by different descriptors is analyzed according mean and median comparisons etc. The choice between descriptors of equivalent statistical measures is based on their chemical interpretability and computational complexity [7]. Other analysis like Shannon entropy [8], differential Shannon entropy [9], and a combination of both have also been proposed to identify informative and sensitive descriptors.

Principal Component Analysis (PCA) is a common dimension reduction technique. PCA uses variance of the dataset to form new orthogonal variables called the principal components. The first principal component accounts for as much of the

variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. While the number of theoretical principal components that can be formed is equivalent to the original number of descriptors, the rank ordering of principal components permits a convenient mechanism for discarding lower rank components.

Non-linear methods such as multidimensional scaling (MDS) and Non-linear Scaling [10] have also been used for dimension reduction. MDS attempts to map compounds into lower dimensional space such that their distances in the original high-dimensional space is preserved. The drawback of these methods is that they are computationally intensive. The non-linear mapping algorithm scales inversely with the increase in size of dataset. Kohonen self-organizing maps are a type of unsupervised learning networks method where data from high n-dimensional spaces are projected onto two-dimensional regular lattice of neurons. In the learning process, compounds are randomly presented to the neurons. The winning neuron is determined based on its similarity to the presented compound and its weights are updated along with the neighbor neurons. This performs an ordering of the neurons so that the topology is preserved, and similar compounds in the n-dimensional descriptors space are projected onto nearby units of the lattice [4].

All the approaches described above try to reduce the dimensions but do not take the bioactivity into consideration. So the reduced dimensions may provide a reasonable account for descriptor variance across the compound set, but might not be relevant to the particular assay and might result in wrongly identifying active compounds when examined as a function of Cartesian distance from known actives. The next section describes some of the previous work done in validating/identifying relevant metrics that also take bioactivity into consideration.

4.2 Metric validation:

Some effort has been devoted to the validation of descriptors in terms of their relationship to biological activity. The most fundamental question to be answered by any metric validation approach is to what extent they place compounds that show affinity towards a receptor together. This amounts to the extent to which the above compounds are clustered near each other in chemistry space [11]. Pearlman and Smith presented a novel approach for metric validation known as activity-seeded structure based clustering [11]. This approach requires consideration of the activity data along with the structural information of the compounds and is only intended for diversity related metric validation. The algorithm was based on the valid assumption that given a set of compounds which all bind to a receptor in the same way, it is reasonable to expect that these compounds will be placed near to each other within a valid chemical space. Such spatial proximity can usually be verified visually if the number of metrics is three or less. The activity seeded, structure based algorithm

provides a method to prove this in the case where the dimensionality is more than three. Pearlman and Smith used as metrics BCUT descriptors that incorporate both connectivity information and atomic properties relevant to intermolecular interactions as metrics. Since there are many choices of specific connectivity information, atomic information and scaling factors that can be computed, there can be many BCUT values generated for each chemical structure. Thus there is a need for some algorithm to decide which of those descriptors to use as chemistry space. Pearlman and Smith used the χ -squared based "auto choose algorithm" implemented by DiverseSolutions for this purpose. The auto choose algorithm provides a method for tailoring a chemistry space to best represent diversity and select the best BCUT descriptors from amongst the numerous possible combinations. The activity seeded, structure based algorithm for validating the BCUT descriptors selected above consists of the following procedure:

1. Choose a unit-cluster radius and center a sphere of that radius on the most active compound.
2. Assign an active compound located within that sphere to that cluster.
3. Center another sphere on the next most active compound that has not been already assigned to any cluster.
4. Repeat steps 2 and 3 until all the compounds have been assigned to some cluster.

5. Coalesce adjoining clusters and record the number of unit cluster spheres per coalesced cluster and the minimum inter-unit-cluster distance between each pair of coalesced cluster.

In choosing the BCUT descriptors they computed the positions of 191 prospective ACE inhibitors in chemistry space. The set had 74 active compounds. If the BCUT descriptors chosen were irrelevant, then the active compounds would be randomly distributed throughout the chemistry space. But using the activity seeded, structure based clustering algorithm, they found that the active compounds were contained in just three coalesced clusters whose total volumes occupy less than 0.02% of the entire chemistry space and less than 0.19% of the occupied chemistry space. The three clusters were found to be closer to each other and the largest inter cluster distance being 3.2 R, where R is the unit-cluster radius.

Bayada et al. also attempted to validate the choice of descriptors based on biological activity [12]. They used a total of 560 known compound drugs comprising 38 different biological activities. They also selected a set of 480 compounds for which none of the 38 selected activity classes were reported. For the whole set of 1040 compounds, a set of 38 descriptors were calculated. After that a Cluster Significance Analysis (CSA) was done for each molecular descriptor and each of the 38 sets, i.e., 86 descriptors x 38 sets=3286 analyses. For each of the analyses, a CSA was performed. The CSA for a given set of size S involves comparing the tightness of

cluster of descriptor values of the compounds in S against the tightness of the cluster of values for random sets of size S. They used a threshold of 0.05 to judge whether a CSA was significant. They decided to retain only those descriptors that were significant in more than 10 out of 38 tests. The main drawback in their approach is that they considered descriptors individually, which does not guarantee that these descriptors will still be meaningful in combination [12]. They used the Available Chemicals Directory database as being the most diverse database of chemicals available to them. After eliminating some compounds from these, 45 selected descriptors were calculated for the remaining compounds. They applied PCA and arbitrarily chose to select the first 10 principal components. The first 10 PCs explained 87% of the variance, with the first PC explaining 45% and the tenth PC 1.7%. Bayada et al. evaluated the efficiency of diversity selection methods like max-min method and several representativity techniques like Ward's clustering, Kohenen neural networks, partitioning method and found that Ward's clustering method using the 10 principal components gave the best result [12].

Waller et.al proposed a novel algorithm called FRED (fast random elimination of descriptors) which is a simple random strategy that implements iterative generations of models to rapidly identify from a pool of allowable variables those which are more closely associated with a given response variable [13]. Descriptors were iteratively eliminated to generate more fit offspring models. Only those descriptors determined

to contribute to the genetic make up of less fit offspring models were eliminated. After every generation, a new random search of the remaining descriptors results in the next generation of randomly constructed models. The user has to supply a kill factor prior to the generation of offspring models. Once all the offspring models have been generated, the models are sorted according to their fitness. The fitness function is the q^2_{LOO} (leave-one-out cross validated correlation) value for the offspring model. The kill factor is used to divide the offspring model population into less fit models and more fit models. The descriptors from less fit models are compared with those from more fit models and those not found in the more fit models are considered detrimental and put in a taboo list. The descriptors are given a second chance by including them in the next generation of offspring models, but if any descriptor produces a suboptimal model in the subsequent offspring, it is removed from the list of allowable descriptors. The algorithm is set to terminate if the standard deviation of the fitness function (q^2_{LOO}) for the entire population is less than a set threshold value [13].

Chapter 5

Implementation

5.1 Description of the datasets used:

The compound structures and their activities were downloaded from the PubChem website [16]. The section below gives a brief description of the compounds and their activities in two different assays that were used in our analysis.

The compounds were downloaded in the structure data file (SDF) format from the PubChem website. The descriptors for these compounds were generated using software called Dragon [17] which is an application for the calculation of molecular descriptors. For each compound, Dragon calculates a set of 1664 descriptors [18]. Since this descriptor dimension was huge, we proceeded to eliminate the following non-informative variables.

1. **Constant or indeterminate variables:** Descriptors with standard deviation lower than 0.0001 across the set of compounds, or those descriptors that have any missing values.
2. **Near Constant variables:** Descriptors no more than two distinct values across the set of compounds.

3. **Highly correlated pair variables:** in cases where two or more descriptors had a correlation coefficient equal or greater than 0.95, only one of the correlating descriptors was retained.

Dataset 1: SK-OV3 Ovarian Cell Line

The first dataset initially consisted of 10,000 compounds selected at random from the full PubChem dataset. After eliminating the compounds that were not processed by the Dragon software and those that did not have an activity value for the assays, the dataset was reduced to 7718 compounds. From the 1664 descriptors generated by the Dragon software, after eliminating non-informative descriptors mentioned in the previous section, the total number of descriptors was reduced to 732. The activities for this dataset are from the NCI human tumor cell line growth inhibition assay data for the SK-OV-3 Ovarian cell.

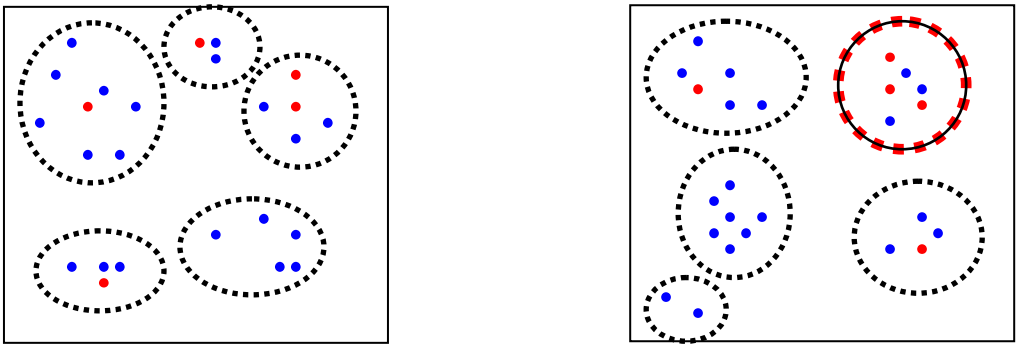
Dataset 2: IGR-OV1 Ovarian Cell Line

The second dataset consists of a set of 10,000 compounds selected at random from the full PubChem dataset. After eliminating the compounds based on the criterion mentioned for dataset 1, the dataset was reduced to 7986 compounds. The elimination of the non – informative descriptors reduced the number of descriptors to 582. We can see that there were a lot more redundant descriptors in the case of second dataset

when compared to the first one. The activities for this dataset are from the NCI human tumor cell line growth inhibition assay data for the IGR-OV-1 Ovarian cell line.

5.2 Methodology:

Our main objective was to isolate simple descriptors that would exhibit biological discrimination, i.e., sequester an optimal number of active compounds into clusters with a substantially enhanced active/inactive ratio.



Scattered **Sequestered**

Figure 2.0 Clusters with scattered and sequestered active compounds

An enriched positive cluster was defined to be a cluster where the ratio of active to inactive compounds is 1.2 times that of the active to inactive compounds in the whole dataset, a value that we determined empirically, over the course of our studies, to provide a reasonable and statistically significant enrichment target. In order to

identify the descriptors, it is first important to identify the number of clusters the dataset should be divided into. This should ideally be equal to the number of natural chemical families present in the data set and is very difficult to guess in advance. After experimenting different number of clusters we came to the conclusion that there were around 10 natural families of compounds in the dataset because specifying a greater number of clusters generally produced surplus clusters with few or no constituents. So we used 10 clusters for all the subsequent experiments. We used k-means clustering to do the initial clustering. The next section explains the k – means clustering algorithm used.

5.3 K – Means Clustering:

The procedure follows a simple way to classify a set of data through a certain number of clusters fixed *a priori*. The steps involved in k – means clustering are:

1. Choose K compounds randomly from the set of compounds that are being clustered. These K compounds will be the initial cluster centroids.
2. Assign each compound to the group that has the closest centroid.
3. When all the objects have been assigned, recalculate the positions of the K centroids.
4. Repeat steps 2 and 3 until the centroids no longer move.

After the compounds have been clustered, the clusters are evaluated for their purity. For each cluster, the number of active and inactive compounds and the P/N ratio (ratio of active and inactive compounds) is noted. The clusters that have the highest P/N ratio and the least P/N ratio are identified. These two clusters are assumed to contain descriptors with the best ability to properly segregate active and inactive compounds, and are thus taken as the basis to rank descriptors for preferential selection in smaller-dimension models.

5.4 Ranking algorithm:

After identifying the clusters, for each of the clusters that has the highest and lowest P/N ratio respectively, the descriptors are sorted and ranked in the ascending order of their mean standard deviation with respect to the distribution of values within the cluster. The higher the mean standard deviation is, the more distracter the descriptor is. Distracter descriptors are those that are less likely to help in segregating active compounds together. The smaller the mean standard deviation, the less distracter the descriptor is and the more important it is in separating actives from inactives. But such smaller standard deviations provide descriptors that are information poor i.e. they have little variation in their value and may do little to discriminate actives from inactives. So the best descriptors are neither those with the largest standard deviations nor the ones with the least standard deviations. In the experiments that we have performed, we found out that the best descriptor models consistently seemed to lay in the100^{th}-200^{th} smallest standard deviation which supports the above argument. Identification of these ranges was accomplished by considering the descriptors in

incrementally smaller sets of 700, 650, 600, 550, 500 … 50, reducing those 50 descriptors with the largest standard deviations each time. For a descriptor set of size 'n' above, the first 'n/2' descriptors are top ranking descriptors (i.e., smallest standard deviation) from the original cluster that has the highest P/N and the remaining 'n/2' descriptors are from the cluster that has the least P/N ratio.

5.5 Drastic Drop Isolation:

For each incrementally smaller set of descriptors obtained above, the k-means clustering is performed again to evaluate how effectively these descriptors place the active compounds into enriched positive clusters. After clustering, the purity of the cluster is evaluated by calculating the percentage of the total active compounds in the predominately positive cluster. This value is noted as a function of different descriptor set sizes and the results are compared in order starting from the largest size of descriptors (i.e. 700) all the way down to 50 descriptors. Any drop in the percentage of actives while going from descriptor set of size 'n' to descriptor set of size 'n-50' is noted, and the case where the maximum drop in the percentage of the total actives is identified. Since eliminating these 50 descriptors in the latter case leads to a drastic drop in the percentage of total active compounds in the positively enriched clusters, these descriptors are judged to be very important in placing the actives in enriched positive clusters. From these 50 important descriptors it is possible to further condense and refine the model in a similar fashion. For

example, if there were a major drop in the percentage of total positives while going from 150 descriptors to 100 descriptors, we would consider descriptor sets of sizes 110, 120, 130, 140, repeat the clustering, calculate the percentage of total actives as we did above and determine more accurately where the drop occurred. If the drop occurred while going down from 140 descriptors to 130 descriptors, we would assume these 10 descriptors to be of primary importance in yielding the enriched positive clusters, and identify them as kernel from which to find a subset that would even further improve the segregation into enriched clusters.

5.6 Refining the kernel descriptors:

The kernel descriptors obtained above are further refined to obtain the minimal set of descriptors that would concentrate most of the active compounds into the enriched positive clusters. The elimination process proceeds as follows. If eliminating a descriptor leads to a drop of more than 10% in the percentage of active compounds in the predominately positive cluster, then the descriptor is considered to be important and is added back. Otherwise it is discarded permanently. Starting with 10 descriptors, the highest standard deviation descriptor is eliminated first to see if the resulting 9d model leads to an improvement in the percentage of actives or doesn't drop the percentage by more than 10%. If so, then the 9d model is taken for further consideration. Otherwise other 9d models obtained by eliminating other descriptors (again ranked by standard deviation) are evaluated to see which one best satisfies the above criterion for elimination. The process is repeated for the resulting 8d model and so forth until we ascertain the smallest set of descriptors beyond which further

elimination doesn't improve the percentage of actives. This process is summarized in Figure 1, below.

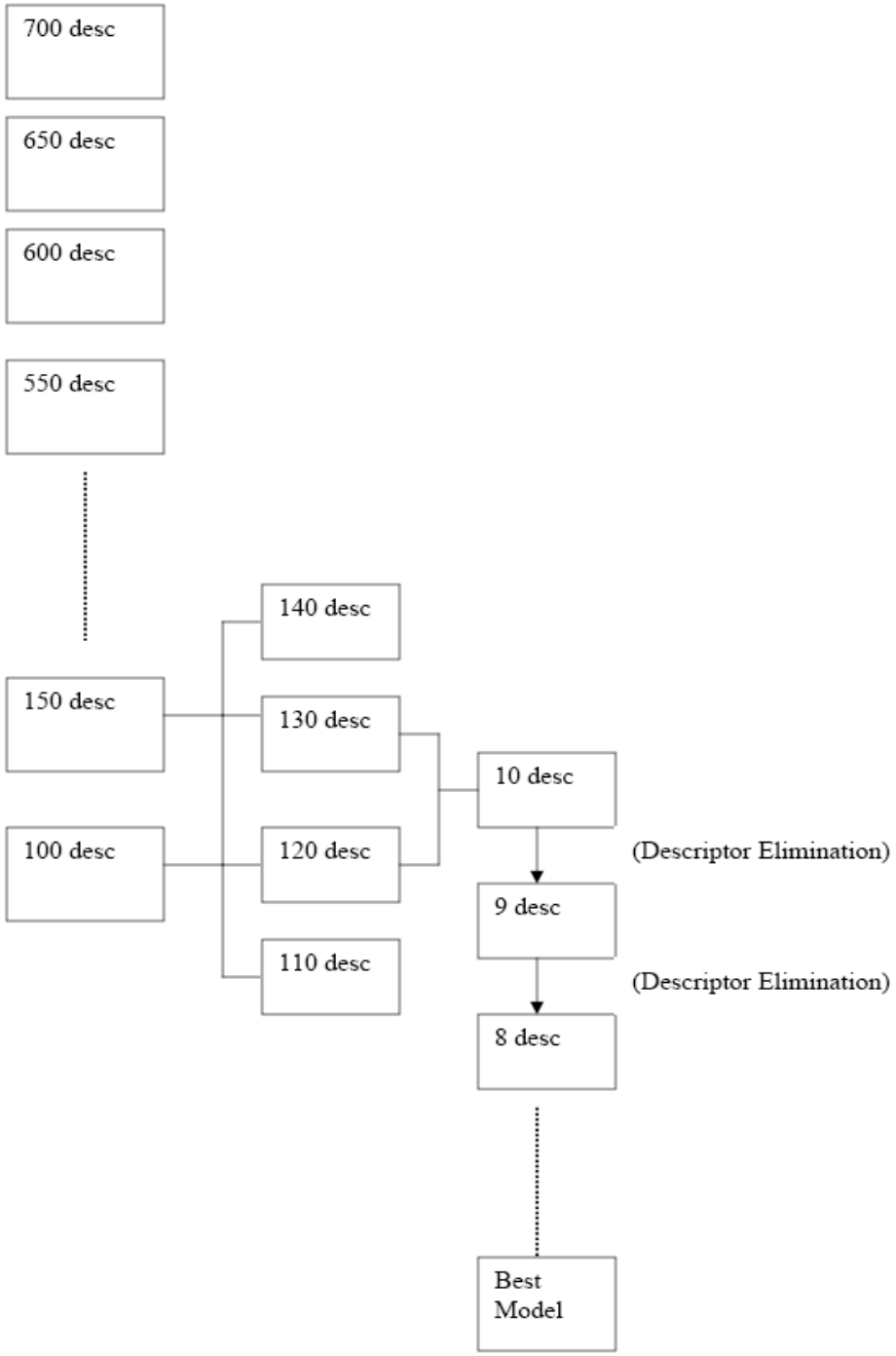


Figure 3.0 Block diagram depicting the idea of identifying the best model from a larger set of descriptors.

Leave-out-one fold cross validation:

In order to ensure that our descriptor selection was both representative of the data set as a whole and extrapolatively predictive for other compound sets, we used the leave-out-onefold (LOOF) cross validation approach to find the best descriptors. The dataset was randomly divided into five subsets each having an equal number of active and inactive compounds. The method is a variation of k-fold cross validation, whereby in each fold 'k-i' subsets are used as the training set and the remaining 'i' subsets as the test set. In the case of LOOF cross validation, i=1 which means that in each fold, exactly one subset is used as testing set and the others as training set.

In each fold, the set of best descriptors are identified using the procedure described in sections above, and is then tested across all the other test sets. The set of descriptors that has the best average over all the test sets is chosen as the final model. Set construction for this scheme is outlined below in figure 2.

Fold1

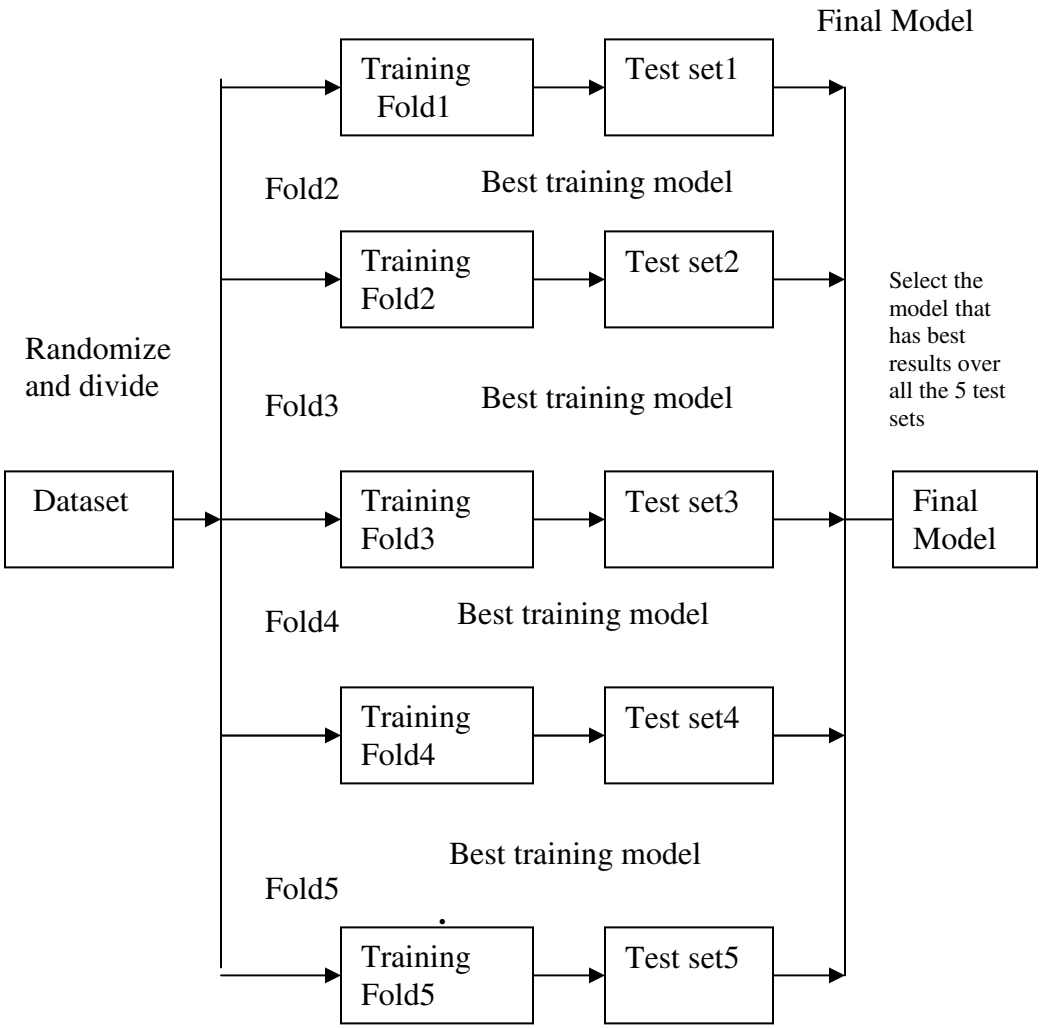


Figure 4.0 Overall view of the model building process

Chapter 6

Results:

6.1 Dataset1:

Total number of compounds: 7718

Active Compounds: 1384

Inactive Compounds: 6334

Total number of initial descriptors: 732

Out of the 7718 compounds, all the compounds that had log (IC_{50}) value <= -4.001 were considered actives (1384 compounds) and those > -4.001 were considered inactive. The compounds were randomly divided into 5 sets each having equal number of active and inactive compounds. In each fold, four of the sets were used for training and the remaining one for testing. The model is trained using the training set and the best model derived is then applied on the test set for that fold. The final model obtained in each fold is then applied on each of the 5 test sets and the one that has the best average is chosen as the final model.

Fold - 1

The table shows the descriptors and the percentage of actives in the positively enriched clusters.

Descriptors	% of active compounds in the positively enriched clusters
732	43.77
700	43.14
650	44.13
600	48.82
550	50.73
500	51.62
450	47.47
400	54.42
350	54.69
300	50.90
250	53.51
200	55.86
150	**66.87**
100	**20.39**

Table 1.1

Descriptors	% of active compounds in the positively enriched clusters
140	68.59
130	65.52
120	**63.35**
110	**19.13**

Table 1.2

From the results in Table 1.1, we can see that there was a drastic drop in the percentage of sequestered active compounds while going from 150d to 100d, which indicates that this set contains descriptors that are vital to placing biologically relevant compounds together. We can further delve into the descriptor set to identify where exactly the drop occurred. From Table 1.2, we can see that the actual drop actually occurred while going down from 120d to 110d. So we pick these 10 descriptors as kernel and refine it to obtain a minimal set of descriptors that would place at least 50% of the compounds in enriched active clusters. In cases where the

descriptors are not able to achieve the target of 50%, the model that gives the highest percentage is taken as the final model.

Descriptors	% of actives
10d	65.97
9d	67.32
8d	70.75
7d	75.45
6d	78.15

The 6d model obtained above was the best model that could give more than 50% of actives in the predominately active cluster. This model was then used on the test set to see what percentage of actives it places in the predominately active clusters.

Descriptors	% of actives in positively enriched clusters for test set
6d	41.30

Fold 2

The table shows the descriptors and the percentage of actives in enriched positive clusters.

Descriptors	% of active compounds in the positively enriched clusters
732	48.23
700	48.87
650	51.67
600	43.99
550	54.01
500	50.22
450	42.18
400	63.41
350	42.27
300	53.74
250	65.76
200	66.75
150	**62.14**
100	**32.06**

Table 2.1

Descriptors	% of active compounds in the positively enriched clusters
140	68.29
130	65.76
120	**63.41**
110	**19.15**

Table 2.2

Here we can observe that the drastic drop occurred while going from 150d to 100. We can drill down even further to find out exactly where the drop occurred.

We can see that the drop actually occurred while going down from 120d to 110d. So we take these 10 descriptors as kernel and further reduce the set to derive the best model.

Descriptors	% of actives
10d	22.49
9d	23.21
8d	27.64
7d	43.54
6d	42.18
5d	52.75
4d	50.22
3d	55.28
2d	52.21

The 2d model obtained above was the best model that could give more than 50% of actives in the predominately active cluster. This model was then used on the test set to see what percentage of actives it places in the predominately active clusters.

Descriptors	% of actives in positively enriched clusters for test set
2d	59.2

Fold3

The table shows the descriptors and the percentage of actives in enriched positive clusters for fold 3.

Descriptors	% of active compounds in the positively enriched clusters
700	50.49
650	46.07
600	47.51
550	45.52
500	33.51
450	55.91
400	51.03
350	62.24
300	42.90
250	56.63
200	66.66
150	**71.63**
100	**22.85**

Table 3.1

Descriptors	% of active compounds in the positively enriched clusters
140	**57.54**
130	**27.10**
120	13.55
110	17.07

Table 3.2

We can see that the major drop in the percentage of actives was found while going down from 150d to 100d. We can further delve into the set to find out more precisely where the drop actually occurred.

From the above table we can see that the major drop off occurred while going from 140d down to 130d. So we take these 10 descriptors as the kernel and try to refine them to obtain the minimal set of descriptors.

Descriptors	% of actives
10d	45.25
9d	43.54
8d	40.83
7d	44.26

6d	43.36
5d	43.54
4d	49.86
3d	49.86

The best model obtained in this fold was 3d and gave almost 50% of actives in the predominately active clusters. We now use this model to see how it performs on the test set.

Descriptors	**% of actives in positively enriched clusters for test set**
3d	52.70

Fold4:

The table shows the descriptors and the percentage of actives in enriched positive clusters for fold 4.

Descriptors	% of active compounds in the positively enriched clusters
700	35.23
650	34.50
600	46.34
550	32.88
500	43.90
450	57.18
400	69.64
350	59.89
300	61.42
250	70.00
200	70.82
150	**64.67**
100	**21.86**

Table 4.1

Descriptors	% of active compounds in the positively enriched clusters
150	**64.67**
140	**31.25**
130	24.84
120	26.01
110	25.20

Table 4.2

From **Table 4.1,** we can see that the major drop off occurred while going down from 150d to 100d. On further analysis, we can observe that the actual drop actually occurred while going from 150d to 140d. So we take these 10 descriptors as kernel and try to refine it.

Descriptors	% of actives
10d	29.9
9d	37.48
8d	39.02
7d	63.86
6d	59.71
5d	69.10
4d	63.23

In this case the 4d model was the best one that gave 63.23% of actives in the predominately active clusters.

Applying this model to the test set in Fold 4, yields:

Descriptors	% of actives in positively enriched clusters for test set
4d	51.98

Fold5:

The table shows the descriptors and the percentage of actives in enriched positive clusters for fold 5.

Descriptors	% of active compounds in the positively enriched clusters
700	45.07
650	62.60
600	62.60
550	56.27
500	52.66
450	50.40
400	43.81
350	42.45
300	68.02
250	56.91
200	64.58
150	**66.21**
100	**18.33**

Table 5.1

Descriptors	% of active compounds in the positively enriched clusters
140	71.00
130	**65.04**
120	**24.93**
110	24.75

Table 5.2

We can see from table 5.1 that the drastic drop off occurs while going from 150d to 100d. So we further delve into the descriptor set in order to find where the drop actually occurred. We can see from table 5.2 that the drop actually occurred while going from 130d to 120d. So we take these descriptors as kernel descriptors and further refine them.

Descriptors	**% of actives**
10d	32.52
9d	35.77
8d	42.72
7d	72.80
6d	64.31
5d	62.24
4d	58.71

The best model in this case was 4d, which gave 58.71 % of actives in the positively enriched clusters. We apply this model on the test set to see how it performs.

Descriptors	**% of actives in positively enriched clusters for test set**
4d	65.34

We now take all the best models obtained in each fold (i.e. the 6d model in fold1, 2d model in fold 2, the 3d model in fold 3, the 6d model in fold 4 and the 4d model in fold 5), and apply each of them to all the test sets to see which model gives better performance over all the test sets .

Best model from Fold #	Test set1	Test set2	Test set3	Test set4	Test set5	Average Over all test sets
1 (6d)	41.30	37.54	35.74	51.98	27.43	38.79
2 (2d)	46.37	59.20	50.54	63.17	35.74	51.00
3 (3d)	53.26	48.73	52.70	54.15	24.90	46.74
4 (4d)	41.30	26.35	20.57	51.98	28.51	33.74
5 (4d)	60.14	58.12	40.07	49.81	65.34	54.69

We can see that the average performance of the model obtained in fold 5 is 54.69 %, which is better than any other model. So we take this model as the final model.

The descriptors present in the final model were:

1. ATS8m: Broto-Moreau 8^{th}-nearest neighbor mass correlation -- $\sum_i \sqrt{(M_i M_{i=>8})}$
2. R_{6u}^+: Max. unweighted distance between 6^{th} nearest neighbors
3. R_{4p}^+: Max. 4^{th} nearest neighbors dist. weighted by polarizability – $\text{Max } |r_i P_i - r_{i=>4} P_{i=>4}|$

4. nN_q: Number of quaternary N

The final 4d model is now applied on the entire dataset of 7718 compounds. It was found that this model was able to cluster 63.36% of the total active compounds into the predominately positive cluster with one cluster alone having 55.36%.

6.2 Dataset2:

Total number of compounds: 7986

Active Compounds: 1256

Inactive Compounds: 6730

Total initial number of descriptors: 582

The same approach that was used for dataset 1 was used for this dataset. The compounds in this dataset were randomly divided into 5 sets each having equal number of active and inactive compounds as done for dataset1. In each fold, four of the sets were used for training and the remaining one for testing. The model is trained using the training set and the best model derived is then applied on the test set for that fold. The final model obtained in each fold is then applied on each of the 5 test sets and the one that has the better average is chosen as the final model.

Fold1

The following table shows the descriptors and the percentage of actives in the predominately positive cluster.

Descriptors	% of active compounds in the positively enriched clusters
582	30.44
550	32.33
500	31.34
450	29.55
400	29.55
350	31.34
300	38.50
250	**36.31**
200	**24.07**
150	33.13
100	22.88

Table 6.1

Descriptors	% of active compounds in the positively enriched clusters
250	**36.31**
240	**15.02**
230	33.53
220	20.39
210	25.07

Table 6.2

From the Table 6.1, we can see that the major drop off occurred while going down from 250d to 200d. We can delve further to identify where the drop actually occurred. From Table 6.2, we can see that the drop actually occurred while going down from 250d to 240d. So we take these 10descriptors as the kernel and try to further refine it.

Descriptors	% of actives
10d	55.92

In this case, the best model obtained was 10d that gave 55.92% of actives in the predominately positive cluster. Applying this model on the test set.

Descriptors	% of actives in positively enriched clusters for test set
10d	58.56

Fold2:

Descriptors	% of active compounds in the positively enriched clusters
550	32.13
500	32.03
450	33.23
400	36.61
350	33.03
300	**34.92**
250	**16.41**
200	17.11
150	16.61
100	12.53

Table 7.1

Descriptors	% of active compounds in the positively enriched clusters
290	36.01
280	**32.83**
270	**16.61**
260	16.51

Table 7.2

We can see from Table 7.1 that the major drop off occurred while going down from 300d to 250d. On further drilling, we notice from Table 7.2 that the drop actually occurred between 280d and 270d. So, we take these 10 descriptors as kernel and try to further refine them.

Descriptors	% of actives
10d	25.67
9d	27.76
8d	34.42
7d	29.75
6d	26.76
5d	26.36
4d	32.32
3d	36.91

The best model in this case is a 3d model, but the % of actives is 36.91 %. We now apply this 3d model on the test set.

Descriptors	% of actives in positively enriched clusters for test set
3d	21.11

Fold3:

The following table shows the number of descriptors and the percentage of active compounds in the positively enriched clusters.

Descriptors	% of active compounds in the positively enriched clusters
550	32.33
500	31.74
450	31.94
400	32.63
350	30.84
300	31.44
250	**35.32**
200	**17.91**
150	25.77
100	21.19

Table 8.1

Descriptors	% of active compounds in the positively enriched clusters
240	34.02
230	20.00
220	**35.02**
210	**18.40**

Table 8.2

From Table 8.1, we can see that the drastic drop off occurred while going down from 250d to 200d. We can drill further into it to identify exactly where the drop occurred. From Table 8.2, we can see that the major drop actually occurred while going down from 220d to 210d. We take these 10 descriptors as kernel descriptors and try to refine them further.

Descriptors	% of actives
10d	20.09
9d	21.09
8d	20.79
7d	58.80
6d	61.00
5d	55.12
4d	50.07
3d	58.10

In this case, the best model obtained was 3d which was able to give segregation where the percentage of active compounds was 58.1%. We now apply this model on the set to see how well it performs.

Descriptors	% of actives in positively enriched clusters for test set
3d	52.98

Fold 4:

The following table shows the number of descriptors and the percentage of active compounds in the positively enriched clusters.

Descriptors	% of active compounds in the positively enriched clusters
550	29.68
500	31.20
450	31.77
400	**43.52**
350	**15.73**
300	15.93
250	28.08
200	36.85
150	36.95
100	24.90

Table 9.1

Descriptors	% of active compounds in the positively enriched clusters
390	34.02
380	20.00
370	**35.02**
360	**18.40**

Table 9.2

From the Table 9.1, we can see that there was a drastic drop in the percentage of active compounds between 400d - 350d. From Table 9.2, we see that in the first case, the drop actually occurred between 360d-370d. We now take these 10 descriptors as kernel and try to further refine them.

Descriptors	% of active compounds in the positively enriched clusters
10d	33.76
9d	35.05
8d	33.46
7d	36.45
6d	42.23
5d	49.1
4d	46.71
3d	50.69

The best model obtained in this case is 3d which gave a percentage of 50.69%. We now apply this model on the test set to see how it performs.

Descriptors	% of actives in positively enriched clusters for test set
3d	26.1

Fold 5

The following table shows the number of descriptors and the percentage of active compounds in the positively enriched clusters.

Descriptors	% of active compounds in the positively enriched clusters
550	29.65
500	31.24
450	28.35
400	35.92
350	30.14
300	**29.65**
250	**15.42**
200	15.42
150	16.11
100	25.17

Descriptors	% of active compounds in the positively enriched clusters
300	**29.65**
290	**15.62**
280	17.21
270	15.62
260	15.52

Table 10.2

Table 10.1

From Table 10.1, we can see that the drastic drop in percentage occurred while going from 300d to 250d. On further drill down, we found that the drop actually occurred between 300d-290d. So, we take the set of these 10 descriptors as kernel and try to further refine them.

Descriptors	% of actives
10d	26.96
9d	25.27
8d	33.73
7d	38.30
6d	41.59
5d	44.37

The best model obtained in this fold was 5d. We now apply this model on the test set.

Descriptors	% of actives in positively enriched clusters for test set
5d	38.24

We now take all the best models obtained in each fold (i.e. the 10d model in fold1, 3d model in fold 2, the 3d model in fold 3, the 3d model in fold 4 and the 5d model in

fold 5), and apply each of them to all the test sets to see which model gives better performance over all the test sets.

Best model from Fold #	Test set1	Test set2	Test set3	Test set4	Test set5	Average Over all test sets
1 (10d)	58.56	33.86	45.41	49.80	52.19	47.96
2 (3d)	23.90	21.11	27.88	27.88	21.11	24.37
3 (3d)	49.00	43.82	52.98	43.02	49.40	47.64
4 (3d)	47.8	37.84	49.8	26.1	39.84	40.27
5 (5d)	45.01	17.92	41.03	39.84	38.24	36.40

We can see from the table that the models obtained in the fold 3 and fold1 are almost identical in performance but the model in fold3 has less number of descriptors. For the sake of obtaining a simple and intuitive model, we thus take the 3d model as the final one.

The descriptors present in the final model are:

1. $HATS_0u$: measure of molecular anisotropy .

2. HTm: Relative interatomic accessibility weighted by mass - $\sum_i \sum_j h_{ij} m_i m_j$

3. $R_{2e+:}$ Max. 2nd nearest neighbors dist. weighted by e-negativities –

   $Max|r_i e_i - r_{i=>2} e_{i=>2}|$

We applied this final model on the entire dataset to see how well it segregated active compounds. This model was able to cluster 52.3% of the total active compounds into enriched positive clusters.

6.3 Comparing with clustering from DVS descriptors:

We used the DiverseSolutions software [11] to generate BCUT [11] descriptors for both the above datasets. When instructed to do so, DiverseSolution auto chooses a set number of descriptors that it considers to provide an optimal representation of chemical diversity within the set, with all redundant descriptors eliminated from consideration. In our case, we requested five optimal BCUT, which the program reduced to four in light of linear dependency considerations. These BCUT descriptors are used for diversity related tasks and are frequently used for trying to identify compounds of potentially similar bioactivity to compounds of known function, by presuming that active compounds will be clustered together in distinct regions of chemistry space. We used the k-means clustering method described above to validate these four metrics by calculating the percentage of total active compounds segregated into the positively enriched clusters. We found that the percentage of actives segregated by the chemistry space derived using our own method significantly outperformed that obtained by using a model derived from auto-chooses BCUT descriptors from DiverseSolutions.

The following are the four BCUT descriptors generated by the auto-choose algorithm of Diverse Solutions software [11].

1. bcut_gastchrg_burden_000.100_K_H
2. bcut_haccept_burden_001.000_K_H
3. bcut_tabpolar_burden_000.400_K_H
4. bcut_tabpolar_burden_000.500_K_L

We used the above descriptors and applied the k-means clustering to see how well these descriptor are able to segregate actives. We found that for the dataset 1, the percentage of total active compounds was 49.68 % and for the dataset 2 it was 29.04 %, both of which constitute poorer performance as compared to the percentages obtained using our approach.

Chapter 7

Analysis

7.1 Visualization of clusters:

This section shows the plots generated using the best models obtained in the two datasets.

In the case of first dataset, the best model obtained was 4d. Since it is difficult to visualize a 4 dimensional plot, we considered all possible 2 dimensional plots using the 4 descriptors to see which one gives a more detailed view of the active clusters. Figure 11.1 is the plot generated for dataset 1 with ATS_{8m} on the x-axis and R_{6u+} on the y-axis.

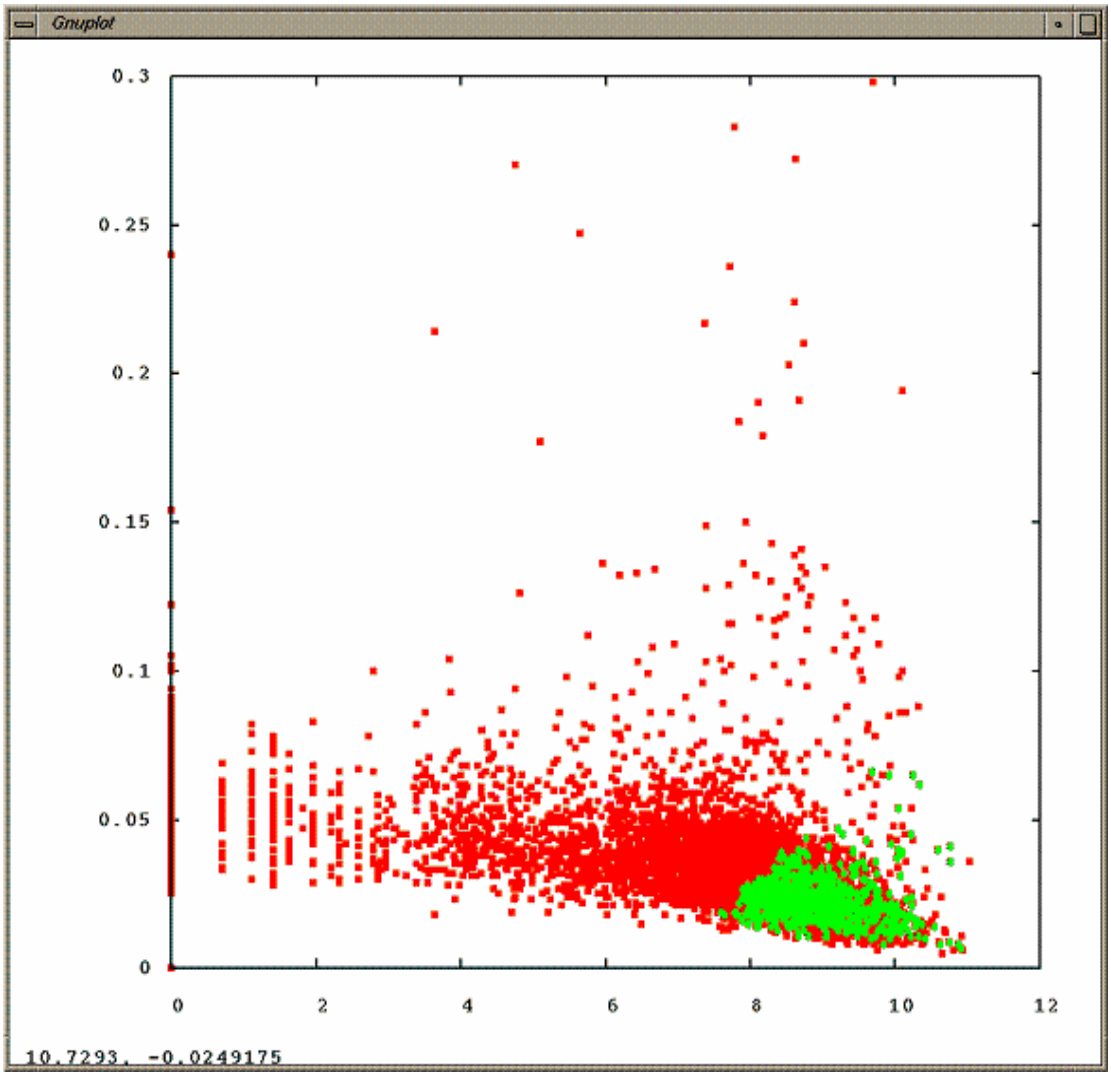


Figure 11.1

x-axis : $ATS_{8m,}$ y-axis : R_{6u+} GREEN : active compounds RED : inactive compounds

The compounds marked in green are the active compounds. We can see from the plot that the active compounds are clustered together tightly and the cluster contains 55.7 % of the total active compounds. Figure 11.2 shows the plot generated for dataset 2 with $HATS_0u$ on the x-axis and R_{2e+} on the y-axis.

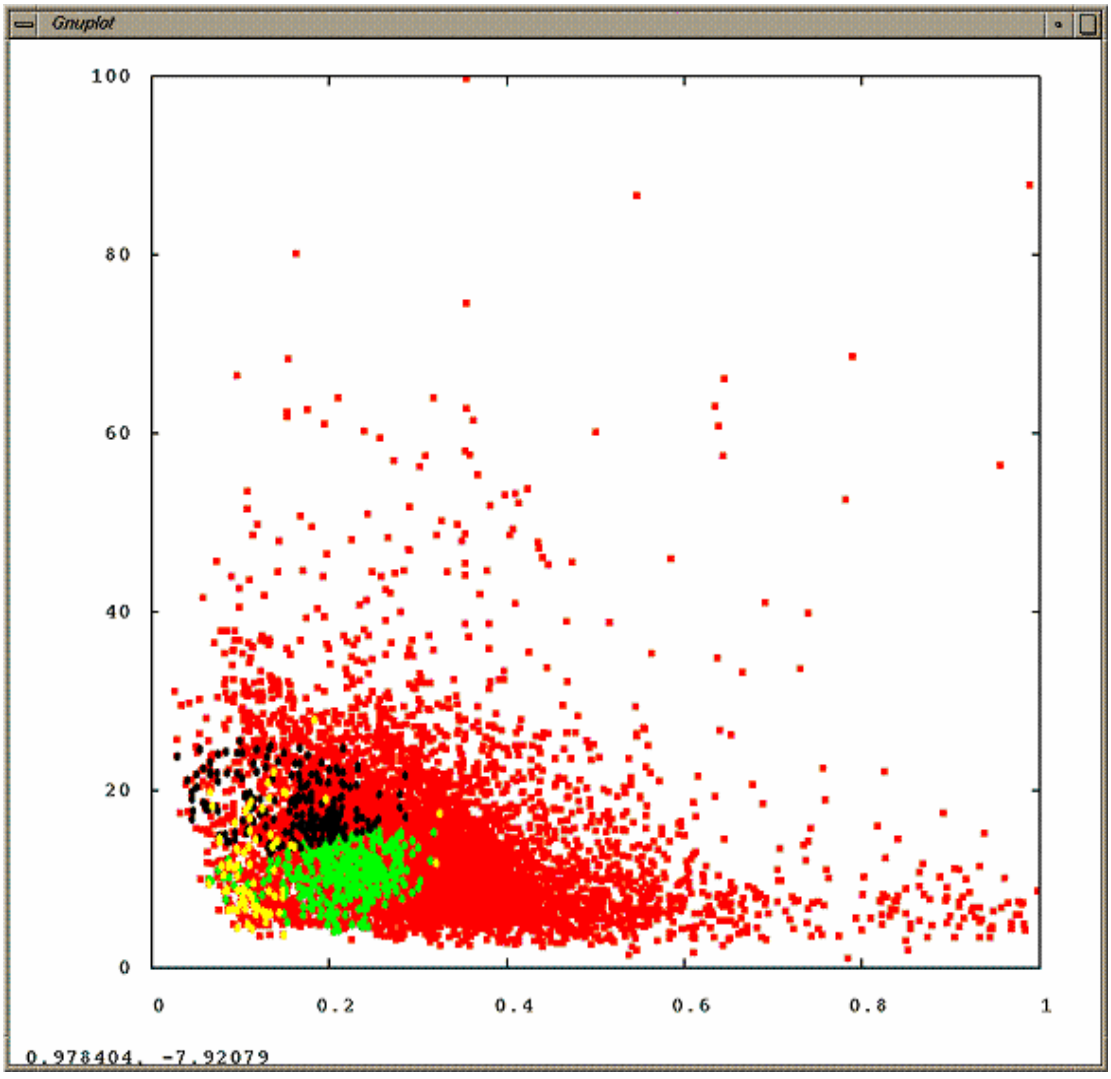


Figure 11.2

x-axis: HATS$_0$u , y-axis: R$_{2e+}$; GREEN, BLACK, YELLOW - active compounds, RED - inactive compounds.

The above plot shows three clusters that have tightly clustered active compounds.

These three clusters together contain 53.02 % of the active compounds.

7.2 Application: Identifying potential active compounds from a new dataset:

The practical purpose of identifying the active clusters is that we can map new compounds whose activity for a particular assay is not known into the chemistry space defined by the descriptors obtained from the best models and see if they map onto the active cluster. This would give us an idea of whether the compound is a potential active compound or not for that particular assay. We modeled a set of 316 compounds recently synthesized as part of a KU effort funded by the National Institutes of Health to develop pilot scale combinatorial libraries for HTS testing and mapped them onto the chemical spaces defined for both the datasets. Figure 11.3 shows the three dimensional plot of the compounds from the dataset 1 with ATS_{8m} on the x-axis, R_{6u+} on the y-axis and R_{4p+} on the z-axis. The KU-NIH compounds are marked in blue.

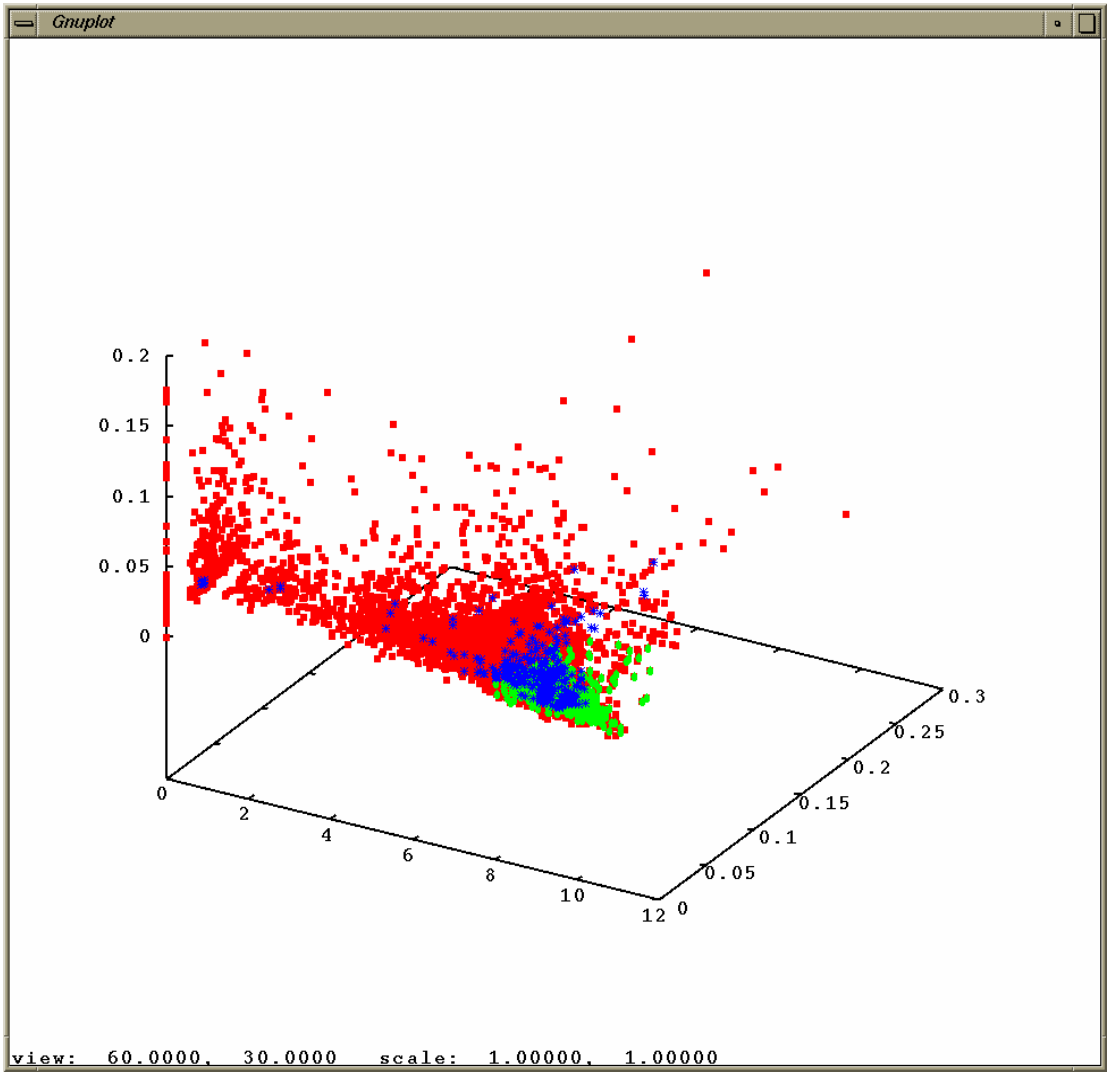


Figure 11.3

x-axis: ATS_{8m}, y-axis: R_{6u+}, z-axis: R_{4p+} ; BLUE, GREEN – active compounds, RED – inactive compounds

From the above plot, we can see that some of the KU-NIH compounds overlap with the compounds from the active cluster. The 2 dimensional plots in Figure 11.4 (ATS_{8m} on the x-axis vs R_{6u}+ on the y-axis) and Figure 11.5 (ATS_{8m} on the x-axis vs R_{4p+} on the y-axis) generated from these three descriptors confirm the same.

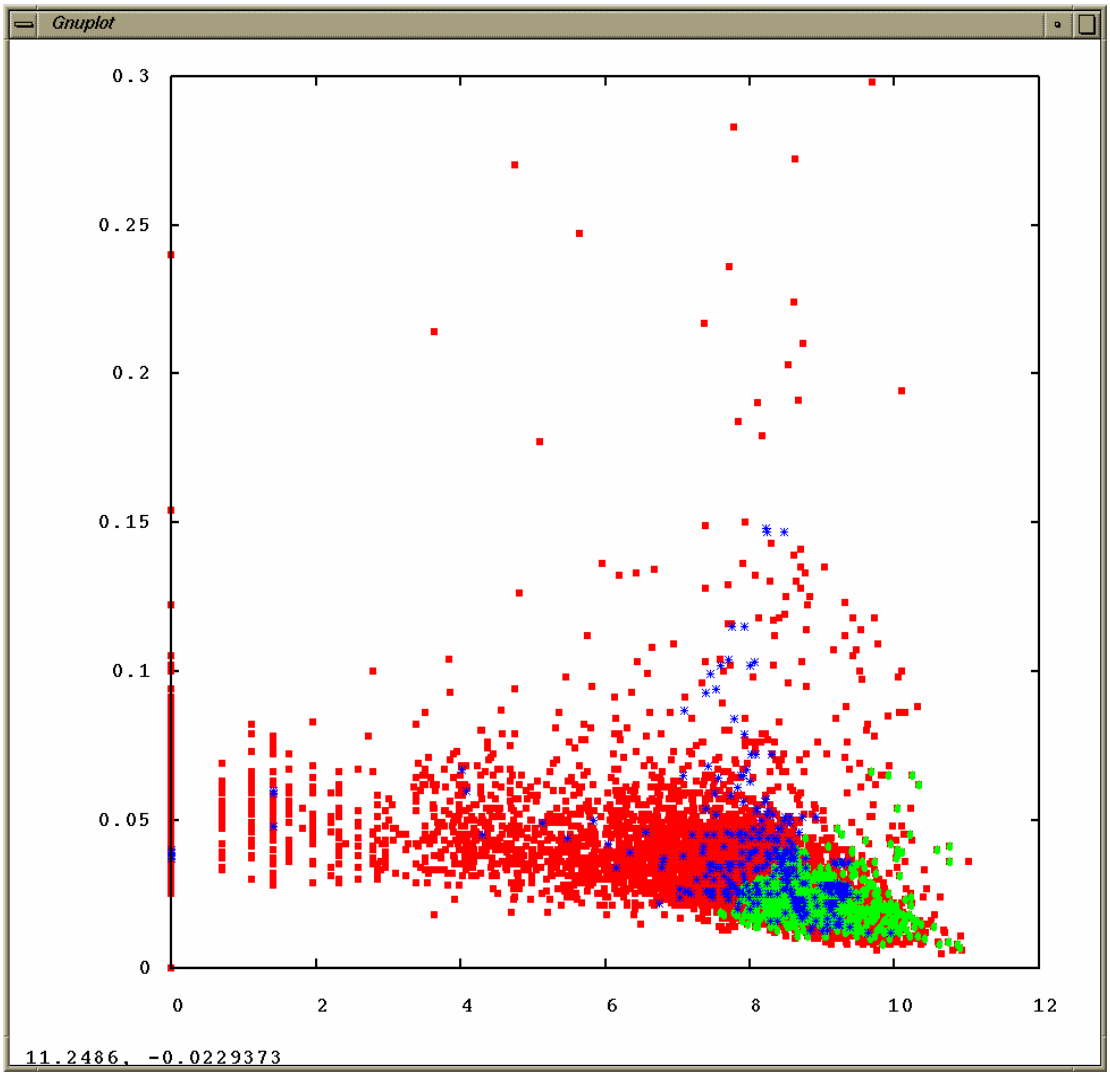


Figure 11.4

x-axis: ATS_{8m}, y-axis: R_{6u}+; BLUE, GREEN – active compounds, RED – inactive compounds

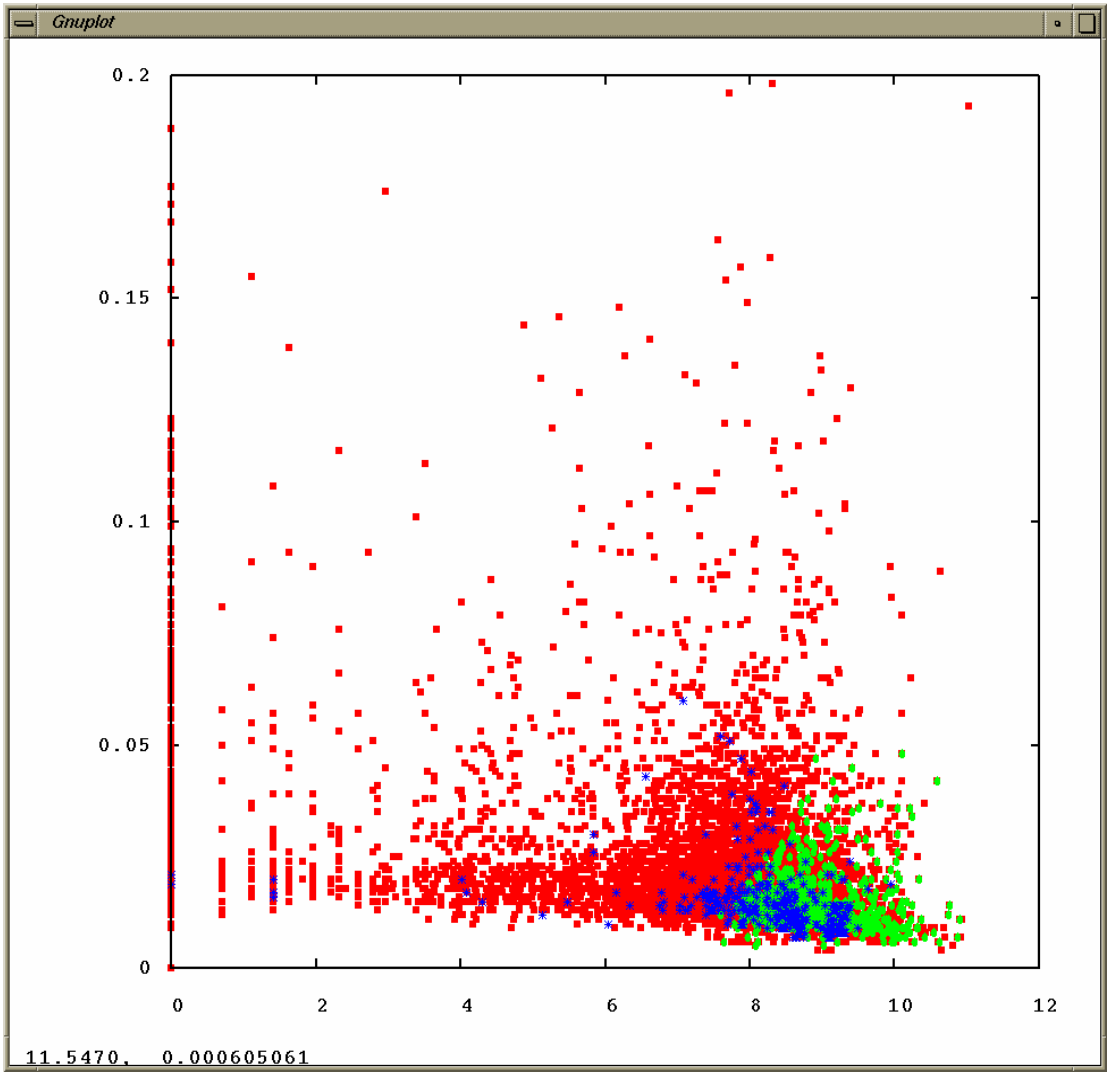


Figure 11.5

x-axis: ATS_{8m}, y-axis: R_{4p+} ; BLUE, GREEN – active compounds, RED – inactive compounds

We performed a similar analysis for the same set of KU-NIH compounds mapping it onto the compounds from dataset 2. Figure 11.6 shows the plot generated for the dataset2 with $HATS_0u$ on the x-axis, HTm on the y-axis and R_{2e+} on the z-axis. From

the plot, we can see that some of the KU-NIH compounds overlap with the compounds from the active clusters. The 2 dimensional plots in Figure 11.7 ($HATS_0u$ on the x-axis and HTm on the y-axis) and Figure 11.8 ($HATS_0u$ on the x-axis and $R2e_+$ on the y-axis) generated from the four descriptors also confirm the same.

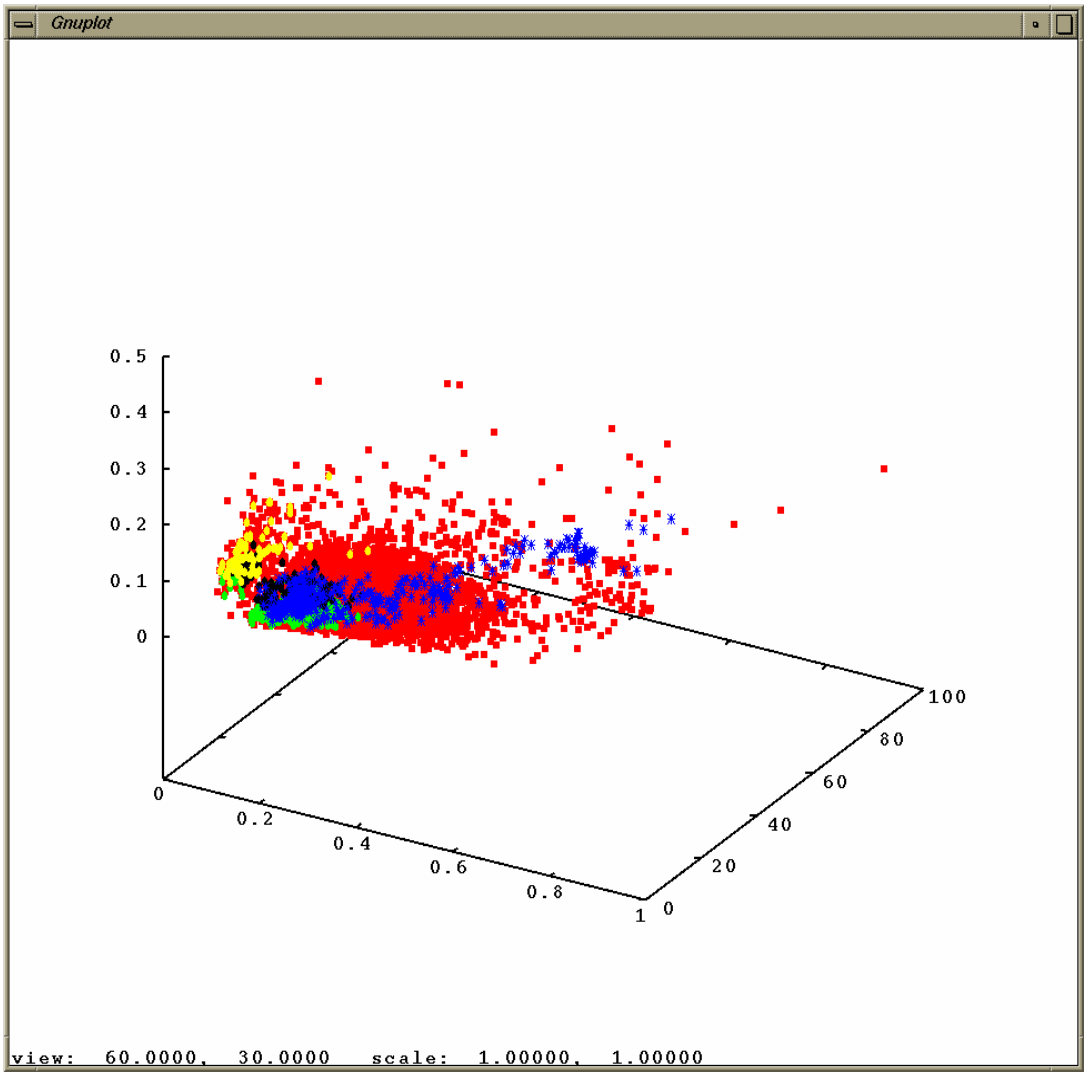


Figure 11.6

x-axis: $HATS_0u$, y-axis: HTm, z-axis: R_{2e+}; GREEN, BLACK, YELLOW, BLUE - active compounds, RED - inactive compounds.

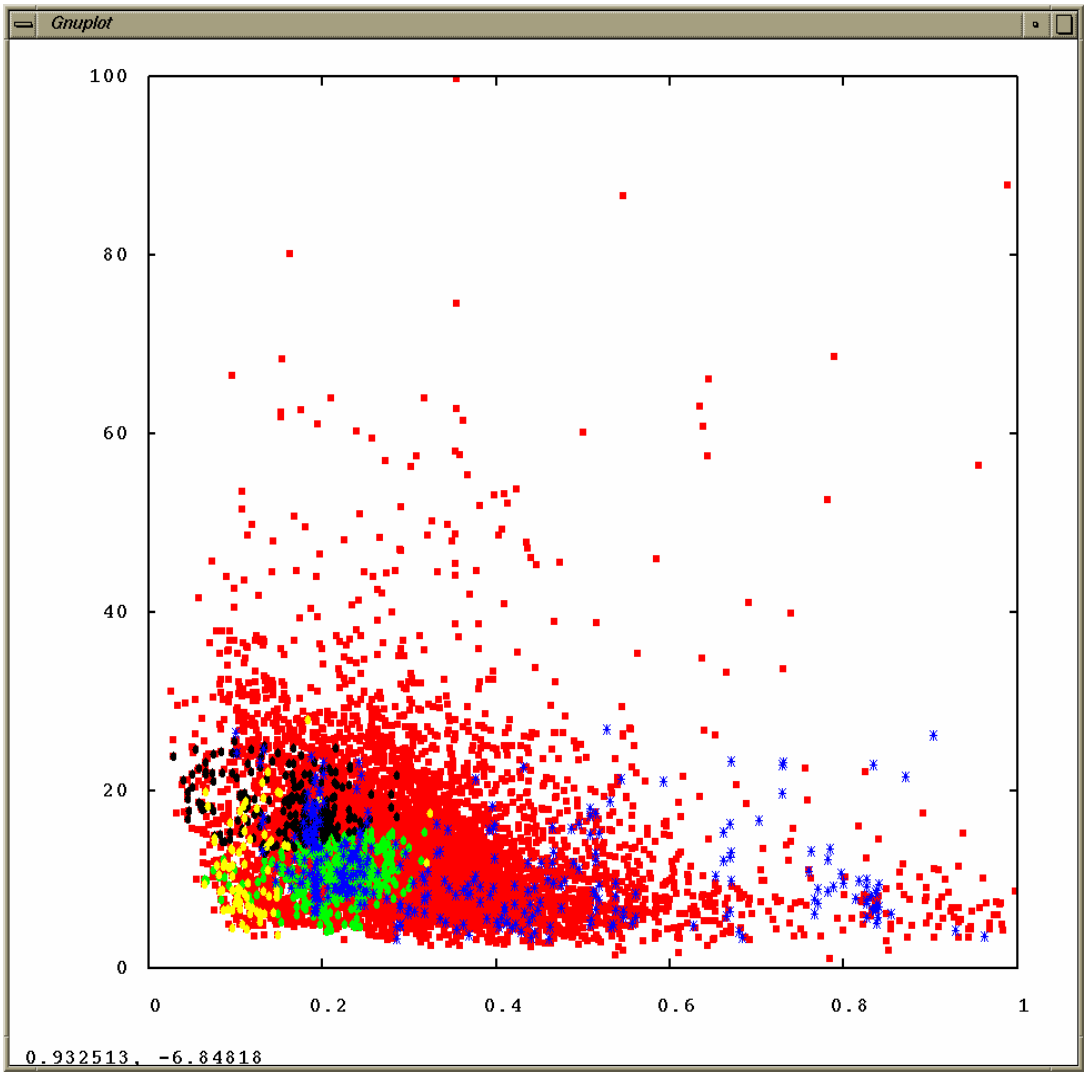


Figure 11.7

x-axis: $HATS_0u$, y-axis: HTm ; GREEN, BLACK, YELLOW, BLUE - active compounds, RED - inactive compounds.

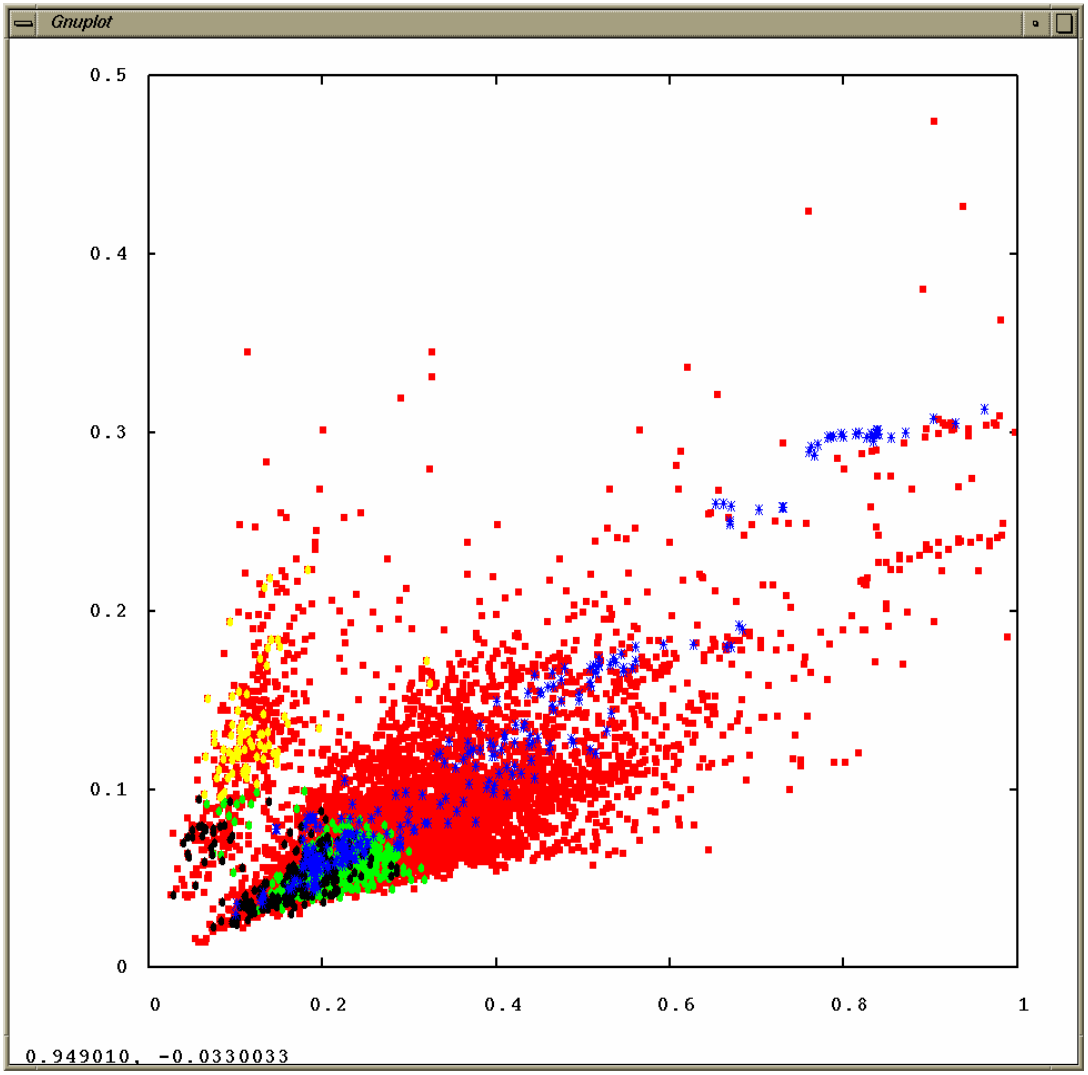


Figure 11.8

x-axis: HATS$_0$u, y-axis: R2e$_+$; GREEN, BLACK, YELLOW, BLUE - active compounds, RED - inactive compounds.

Chapter 8

Conclusion

The main goal of this thesis was to develop a computational method for choosing chemical diversity space metrics wherein compounds located in similar chemical space locations are likely to exhibit similar biological activity for a given assay. By isolating small subsets of descriptors within which the active compounds are optimally clustered, this permits us to identify the receptor relevant subspace within which one has the greatest chance of finding actives. We developed an algorithmic approach aimed at identifying small collections of descriptors that would support such active compound clustering. We tested our algorithmic approach on two different datasets and used "leave-out one-fifth" cross validation to identify the optimal model.

The optimal model in the case of first dataset was four dimensional and was able to segregate 55.78% of actives in just one cluster. The optimal model in the second case was three dimensional and it was able to segregate 52.30 % of the total active compounds in just 3 clusters. We also visually verified the clustering of active compounds using 2D plots, and quantitatively verified that our method substantially outperforms the most commonly used chemical diversity analysis program, DiverseSolutions, in terms of segregating active compounds into distinct regions of chemical space. We believe that our method will prove to be a useful tool in the design and planning of new chemical libraries, or in the purchase of existing libraries

for the purpose of targeted screens. Specifically, one may readily calculate the chemical space descriptors for these proposed compounds and thus overlay the prospective library over a plot of compounds of known activities, and thus identify those proposed compounds that lie within or close to biologically fertile regions. To demonstrate this application, we mapped a set of recently synthesized KU chemical compounds onto the chemical spaces derived for the two ovarian cancer assays that we had modeled. This overlay permits facile identification of potentially active compounds.

Chapter 9

Future Work

In this thesis we developed a strategy that would identify the receptor relevant chemical subspace for two assays for which extensive data was provided by the PubChem database. Similar methodology can be applied to the remaining (approximately 40) data sets on the PubChem site, as well as to other screening studies published in the literature or in other databases. Such a collection of receptor relevant chemical subspaces would provide us with an array of tools that can be used as inexpensive, rapid and insightful virtual prescreens that will help researchers to plan combinatorial library development, library purchases, and HTS experiment tailoring. The method can be intuitively extended by seeking to identify one single subspace set that would optimally cluster active compounds for a wide range of possible assays, i.e., to derive a Cartesian framework that would take into account broad trends in biological activity in such a way as to cluster an optimal number of generally biologically active compounds into enriched positive clusters, thus affording a tool that will enhance our understanding of what chemical characteristics are most likely to make compounds with some sort of biological activity, versus those characteristics that tend to yield biochemical neutrality.

References

1. Mike Rippin, Tesella Support services [Online] Available from (http://www.tessella.com/Literature/Supplements/PDF/HighThroughputScreening.pdf) [Accessed 22 October 2006].
2. Hans-Jorg Roth : There is no such thing as diversity, Current Opinion in Chemical Biology 2005, 9:293-295.
3. Dominique Gorse: Functional Diversity of compound libraries, Current Opinion in Chemical Biology 2000, 4:287-294.
4. Dominique Gorse: Diversity in medicinal space, Current topics in medicinal chemistry, 2001, 1873-4294.
5. Lemmen C, Lengauer T.: Computational methods for structural alignments of molecules, Journal of computer-aided molecular design 2000; 14(3):215-232.
6. Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. J Chem Inf Comput Sci 2000; 40(5):1160-1168.
7. Gorse D, Rees A, Kaczorek M, Lahana R. Molecular diversity and its analysis, Drug Discovery Today 1999; 4(6):257-264
8. Godden JW, Stahura FL, Bajorath J. :Variability of molecular descriptors in compound databases revealed by Shannon entropy calculations. J Chem Inf Comput Sci 2000; 40(3):796-800.

9. Stahura FL, Godden JW, Bajorath J.: Differential Shannon entropy analysis identifies molecular property descriptors that predict aqueous solubility of synthetic compounds with high accuracy in binary QSAR calculations. J Chem Inf Comput Sci 2002; 42(3):550-558.

10. Agrafiotis DK, Lobanov VS.: Nonlinear mapping networks. J Chem Inf Comput Sci 2000; 40(6):1356-1362.

11. R. S. Pearlman and K. M. Smith : J. Chem. Inf. Comput. Sci 1999, 39, 28-35.

12. Dennis M.Bayada : Molecular Diversity and Representativity in Chemical Databases, J. Chem. Inf. Comput. Sci 1999, 39, 1- 10.

13. Chris. L . Waller, Mary P. Bradley : Development and Validation of a Novel Variable Selection Technique with Application to multidimensional Quantitative Structure - Activity Relationship Studies, J. Chem. Inf. Sci. 1999, 39, 345-355.

14. Yvone C. Martin: Do structurally similar molecules have similar biological activity? J. Med. Chem 2002, 45, 4350-4358

15. Thorsten Potter: Random or Rational Design? Evaluation of Diverse Compound Subsets from Chemical Structure Databases, J. Med. Chem. 1998, 41, 478-488.

16. Databases of molecular data on the NCBI FTP site

17. [online] Available from : (ftp://ftp.ncbi.nlm.nih.gov/pubchem/) [Accessed 22 October 2006].
18. 17. Dragon software: [online] Available from : (http://www.talete.mi.it/main_net.htm) [Accessed 24th October 2006]
19. 18. Descriptors calculated using dragon: [online] Available (http://www.talete.mi.it/help/dragon_help/index.html) [Accessed 24th October].
20. Dominique Gorse, Anthony Rees, Michael Kaczorek and Roger Lahana: Molecular diversity and its analysis: Drug Discovery Today, DDT Vol. 4, No 6 June 1999.