Chemoinformatic Approaches in Finding a Representative Set for Chemical Data

A Dissertation submitted to the University of Hyderabad in partial fulfillment of the degree of

Master of Technology

in

Artificial Intelligence

By

Ajith Bhamidipati

09MCMI24

Department of Computer and Information Sciences

School of MCIS

University of Hyderabad

(P.O.) Central University, Gachibowli

Hyderabad-500 046

Andhra Pradesh,India.

April 30, 2011

CERTIFICATE

This is to certify that the dissertation entitled "**Chemoinformatic Approaches in Finding a Representative Set for Chemical Data**" being submitted to University of Hyderabad by **Ajith Bhamidipati**, bearing Reg. No. 09MCMI24, in partial fulfillment of the requirements for the award of Master of Technology in Computer Science, is a bonafide work carried out by him under my supervision and guidance. The dissertation has not been submitted previously in part or in full to this or any other University or Institution for the award of any degree or diploma.

Dr. S. Durga Bhavani
Project Supervisor,
Department of CIS,
University of Hyderabad.

Head of Department,
Department of CIS,
University of Hyderabad.

Dean,
School of MCIS,
University of Hyderabad.

To,

My dear amma, nanna and my sisters Ale and sravs

Acknowledgments

I take this opportunity to remember and acknowledge my guide **Dr.S.Durga Bhavani** whose advice, support and above all her patience helped me in completing the project.

I also wish to thank **Dr.T.Shobha Rani**, whose inquisite nature of approaching problems solved many an error.
I am glad to express my gratitude to **Dr.Atul Negi** who has been my mentor, and a person who helped me realize myself.

My thanks to Mr.P.Pavan kumar whose calm and systematic way to every task at hand inspired me in building my innate qualities.

Apart from all eminent personalities above, I thank my friends Pradeep, Yaswanth, Vaseem, Krishna, Meghana and Ananth Kiran, who stood beside me, in all the tenures I faced and guiding me at all bifurcations of life.

With Sincere Regards,
Ajith Bhamidipati.

Abstract

Chemoinformatics is an interface science aimed primarily at discovering novel chemical entities. In this field reducing chemical space is an important task for virtual screening and for drug discovery it is time consuming to search for lead molecules in a huge chemical space. The objective of the project is to construct a smaller representative set by retaining the feature diversity in the original data set using chemoinformatics and algorithmic approaches.

In the current work we proposed three methods to reduce the size of such a huge chemical space to a small representative set which retains the similarity and diversity of the original space.

First method uses the divide and conquer strategy that uses the structural cycle information present in the compounds to extract a representative set. Second method bins the entire space of data set (NPD) using the Tanimoto metric and the third method performs PCA on the data distribution to implement the cell based approach. Statistical parameters are used to analyse the results obtained and further, the work is validated by a comparative analysis of the resultant representative data sets with measures like KL-divergence.

List of Tables

Contents

Chapter 1

Introduction

Chemical space is a collection of theoretically available chemical compounds. It is unbounded and growing, even as you are reading this. But the space of druglike molecules and Naturally available molecules is bounded. However,even this reduced space is estimated to contain anything from 10^{12} - to 10^{180} molecules for Drug like and approximately 10^{6} for Natural Products.

Drug discovery generally follows a set of common stages. First, a biological target is identified that is screened against many thousands of molecules in parallel. The results from these screens are referred to as hits. A number of the hits will be followed up as leads with various profiling analysis to determine whether any of these molecules are suitable for the target of interest. The leads can then be converted to candidates by optimizing on the biological activity and other objectives of interest, such as the number of synthetic steps. Once a suitable candidate has been designed, the candidate enters preclinical development. Chemoinformatics is involved from initially designing screening libraries, to determining hits to take forward to lead optimization and the determination of candidates by application of a variety of tools[5].

Since the chemical space is very large, the task of searching for lead molecules is time intensive. Indeed, there are many significant challenges still open not only in chemical database searching, but also in the methods to analyze and access the related data, since the entire data is stored in large databases. In order to meet these challenges one has to minimize the chemical space to a small space without losing the original diversity. When-

ever a druggist or a chemist wants to find leads, s/he can search this small reduced chemical space, and so that search time can be saved. The main objective of the current work is:

To construct a smaller Representative set by retaining the feature Diversity in the original data set using Chemoinformatics and Algorithmic approaches.

In the current work we propose three methods discussed in chapter 3, which result in representative sets that are more similar and as diverse as the original data set. Designing of methods that generate representative sets are necessary since random sampling generated sets deviate severely if the size is too small and may also miss out some natural clusters/classes of the original distribution. As we will see,the methods designed yield a better representative set than a random set of same size extracted.

Most used format for storing data in chemical databases is SMILES[6]. SMILES is one of the many ways to write a chemical structure in a linear format. Linear formats have an advantage over 2D and 3D structures because they are easily processed through computer.

Fingerprint representation of a molecule is another most widely used approach for chemical database mining in similarity searching. Fingerprints are bit string encoding of structural features or calculated molecular properties or both. Fingerprints quantify the similarity measure based on some molecular features called descriptors since different molecular features capture different aspect of molecular structures. Each bit position in a fingerprint corresponds to one molecular feature and it monitors the presence or absence of a particular molecular property

A variety of similarity metrics are available for quantitatively comparing fingerprint overlap between reference and database molecules. Tanimoto Coefficient continues to be the most popular one because of its ability to quantify the similarity between two molecules. The range of Tanimoto Coefficient is from 0 to 1. For highly similar molecules Tanimoto Coefficient is very near to 1 and for any pair of dissimilar molecules it is close to 0.

Kullback-Leibler divergence is a measure used to calculate the distance between two distributions, which we used to compute and compare the distances between the random

generated representative set and original data set to that of between the designed representative set and the original data set.

Thus the designed representative sets are validated using Kullback-Leibler Divergence and by comparing the statistical measures like the mean, Variance and standard deviation of the distributions.

Chapter 2

Review of Literature

Chemoinformatics [5] is a large scientific discipline that deals with the storage, organization, management retrieval, analysis, dissemination, visualization, and use of chemical information available in a chemical space. It has emerged from several older disciplines such as computational chemistry, computer chemistry, chemometrics, QSAR, chemical information, etc. Cheminformatics involves the use of computer technologies to process chemical data. In Chemoinformatics, the term Chemical space referred to the collection of theoretically available all chemical molecules. Since the chemical space itself theoretically unbounded, it is a time consuming task to search for lead molecules in the process of drug discovery. It is necessary to reduce such huge chemical space to a small manageable subspace so that a chemist or a druggist can accelerate his research effectively. So the problem of reducing the size of a chemical space of millions of molecules is an important task for virtual screening in drug discovery. The field of Chemoinformatics provides many approaches to beat this problem. An extensive research has been done in this direction.

2.1 Virtual Screening

Virtual and high-throughput screening are time saving techniques that have been successfully applied to identify novel chemo-types in biologically active molecules. Both methods require the ability to aptly handle large numbers of chemicals prior to an experiment or acquisition. The main goal of virtual screening is to select activity-enriched sets of molecules exhibiting desired activity from the space of all synthetically accessible structures. Currently the most advanced High Throughput Screening techniques allow for testing of hun-

dreds of compounds per day, and a typical corporate screening library contains several hundred thousand samples.

2.2 Similarity Searching

Similarity searching is a virtual screening method, that targets at " Which molecules in a database are similar to the query molecule(s)". The limitations posed by sub structure searching such as the user who is posing a database query must already have formed a fairly clear view of the types of structure that will be retrieved, have led to the to the development of the alternative, and complementary,access mechanism known as similarity searching. Chemical similarity searching is to find the most similar chemical compounds in a large database to a molecule that is known to exhibit a certain activity.This assertion is based on the structure-similarity principle [1] which states that two structurally similar molecules are likely to exhibit a similar activity.Molecular similarity searching is an integral part of the early stages of the drug discovery process. The idea is to take an active molecule and to find other compounds with similar structure in a database. Usually, a certain fraction of the most similar molecules are considered for further testing.

2.2.1 Importance of Similarity Searching

The entirety of Chemical databases (often comprising millions of entities) can not be screened at a reasonable cost even by current technology, and computer-based methods can be used to suggest subsets of compounds that are most representative to the database, thus making targeted screening more efficient than a test everything approach.

2.2.2 Encoding: Finger Prints

Molecular fingerprints are a way of encoding the structure of a molecule. Fingerprint [11] representation of molecules is widely used in similarity searching. The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the molecule. Comparing fingerprints allows you to determine the similarity between two molecules, to find matches to a query substructure

2.3 Similarity Coefficients

Calculation of the similarity or dissimilarity between two molecules is a standard practice in the chemoinformatics field. One of the factors that affects these calculations is the similarity coefficient used to compare the molecules. The coefficients [13] fall into three categories: association coefficients, commonly used with binary representations and often normalized to lie within the range zero (no common features) and unity (identical features); correlation coefficients, which measure the degree of correlation between the characterizations; and distance coefficients, which quantify the degree of dissimilarity between the characterizations and, when normalized, have the range zero (identity) and unity (no common features).The most widely used similarity coefficient is the Tanimoto coefficient.

2.3.1 Tanimoto Coefficient

Tanimoto Coefficient [13] continues to be most popular similarity coefficient because of its ability to quantify the similarity between two molecules. The range of Tanimoto Coefficient is from 0 to 1. For highly similar molecules, Tanimoto Coefficient value is very near to 1 and for any pair of dissimilar molecules its value is near to 0.

The formula for Tanimoto Coefficient is as follows.

For any two given molecular fingerprints $A = (A_i)$ and $B = (B_i)$, we have

$$TC(A, B) = \frac{\sum_{i=0}^{n} A_i B_i}{\sum_{i=0}^{n} A_i^2 + \sum_{i=0}^{n} B_i^2 - \sum_{i=0}^{n} A_i B_i}$$

if $TC(A, B) \equiv 1$ then A, B are similar and
if $TC(A, B) \equiv 0$ then A, B are dissimilar.

Here A_i and B_i are binary variables representing i^{th} bits in fingerprints A and B respectively and $A_i B_i$ represents their product.

2.3.2 Cosine Coefficient

It is another useful similarity coefficient [15]. For any two molecular fingerprints $A = (A_i)$ and $B = (B_i)$, we have ,

$$\text{Cosine Coefficient}(A, B) = \frac{\sum_{i=0}^{n} A_i B_i}{\sqrt{(\sum_{i=0}^{n} A_i^2)(\sum_{i=0}^{n} B_i^2)}}$$

2.3.3 Dice Coefficient

For given any two molecular fingerprints $A = (A_i)$ and $B = (B_i)$, we have

$$\text{Dice Coefficient}(A, B) = \frac{2\sum_{i=0}^{n} A_i B_i}{\sum_{i=0}^{n} A_i^2 + \sum_{i=0}^{n} B_i^2}$$

2.4 Kullback-Leibler Divergence

Relative entropy between element sets E_i and E_j subset of E of a discrete random variable is Kullback-Leibler divergence between their distributions in the feature domain.The KullbackLeibler divergence is always non-negative. It is the average of the logarithmic difference between the probabilities P and Q, where the average is taken using the probabilities P. The K-L divergence is only defined if P and Q both sum to 1 and if $Q(i) > 0$ for any i such that $P(i) > 0$. If the quantity 0log0 appears in the formula, it is interpreted as zero.

2.4.1 Discrete Distributions

For probability distributions P and Q of a discrete random variable their KL divergence is defined to be

$$D_{KL}[E_i||E_j] = \sum_{f \in F} P(f|E_i) \log(\frac{P(f|E_i)}{P(f|E_j)})$$

The KullbackLeibler divergence is additive for independent distributions and is a non-symmetric measure of the difference between two probability distributions.It also does not satisfy the triangular inequality. KL divergence is a special case of a broader class of divergences called f-divergences.

Most instances a random sample of a relatively smaller size is used as a representative set which may contain many pit falls as:

- As the size of the random sample decreases the resulting set deviates severely from the original set.
- It may miss some natural clusters/ classes.

The following chapter describes the methods/ approaches we use to design the representative set, which implicitly address the above pit falls.

Chapter 3

Proposed Methods for Generating a Representative Set

3.1 About the Data Used

The source of data is ZINC [15] which is an on line public data base.The following methods are experimented and validated on Natural Products data base. The data set contains compounds which have a zinc id, along with the SMILES notation of the compound. ZINC data base brings virtual screening libraries to a wide community of druggists and biologists so that drug discovery task becomes fast and efficient. Each compound is described using a set of 58 feature descriptors[15]. The NPD set contains 89217 compounds where as the Drug like comprises of 8783230. It is worth mentioning that with chemical data bases the execution for an algorithm of even O(n) complexity, is bound to consume more time due to the large size of these data sets. The features[15] of these data sets are listed in the below table.

Physical(9)	Atom Count(10)	Structural(9)
1. Mol. Weight	10. Br Count	20. Cyclic
2. logP	11. C Count	21. Acyclic
3. De_apolar	12. Cl Count	22. Mono Cyclic
4. De_polar	13. F Count	23. Bi Cyclic
5. HBD 14.	I Count	24. Tri Cyclic
6. HBA	15. N Count	25. Tetra Cyclic
7. tPSA	16. Na Count	26. Hi cyclic(>5 cycles)
8. Charge	17. O Count	27. Hetero cyclic
9. NRB	18. P Count	28. Chiral Centers
	19. S Count	
Functional Groups(30)		
29. -Cl	39. -COOH	49. -CHO
30. -Br	40. -COOR	50. Ketone
31. -F	41. -COOCl	51. Thiketone
32. -O	42. Cyano	52. Peptide
33. -S	43. Isocynate	53. Nitroso
34. -N	44. –C=N-R	54. Nitro
35. Alkylamino	45. Acetyne	55. Furon
36. Dialkylamino	46. Ethylene	56. Pynol
37. Amide	47. Azo N#N	57. Aromatic S
38. Amide2	48. Phenol	58. Phenyl

With the survey of literature and introduction to the domain in previous chapters , we now introduce three methods that construct representative sets. This resulting set size can be specified by the chemist and the results are compared with a random sampling of the same size. These methods rather than immediately jumping, to randomly select the compounds, we postpone this process a few steps further so that the data now sampled is more close to the original. Each of the approaches are compared and validated using statistical measures in the below sections and as a further step of validation we use Kullback-Leibler divergence measure in the next chapter.

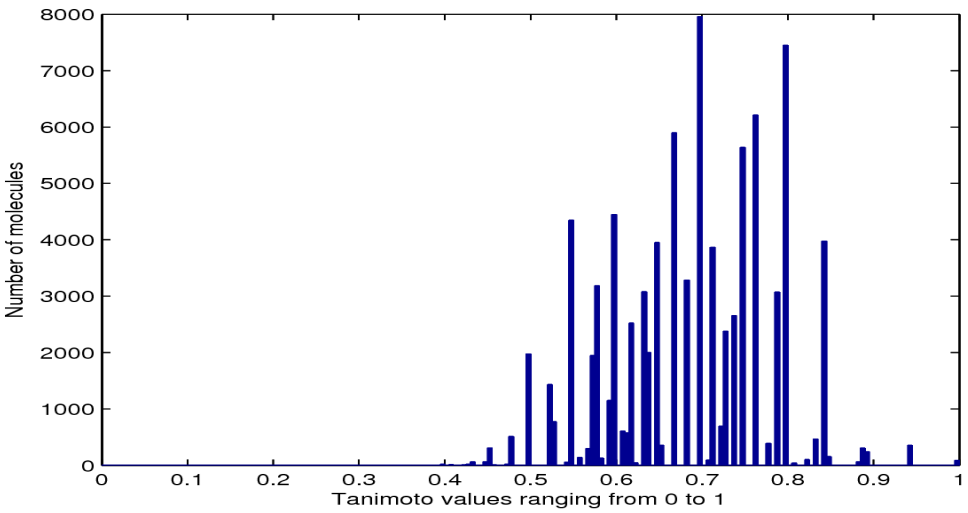
Statistical measures form the basic step in knowing the correctness of a any result. Similarly for the methods in this chapter, we provide statistical measures that prove the advantage of using these methods instead of a random selection in generating a represen-

tative set. We have to review that a representative set is also a subset of the original set(NPD). So if the distribution of representative set is similar to the original distribution then the Means of these two distributions will be close,i.e for our data set since each of the features is independent, we come up with a 58 length vector comprising of Means of each feature. Similarly for the representative set (CS/RS), we generate a Mean vector of the same length. To compare the random and designed representative sets we compute the Euclidean distances from each of these representative sets to the original distribution. Similarly the distances of variance and standard deviation vectors from NPD are also compared.

3.2 Method: Binning based on Tanimoto metric

3.2.1 Approach

The Tanimoto metric is a similarity measure used to compute the similarity between two compounds. This value lies between 0 and 1 indicating identical compounds if the value is 1. As an initial step of this method we select a reference compound which is one of the compounds of the data set. Further distances are computed between this reference molecule and every molecule of the data set which results in tanimoto similarity ranging all over the stretch between 0 and 1. This entire distribution of compounds is binned based on the tanimoto values.



Depending on the representative set size, compounds are drawn from each of the bins in proportion to the bin's size. The resulting set is our representative set which contains at least one compound from each bin. This generated representative set forms a distribution for which we compute the Mean. The distance of this Mean from the Mean of the original distribution(NPD) is determined and compared with that of the Random representative set's Mean and Mean of NPD. In the next chapter we validate these statistical measures using Kullback-Leibler formulation.

The bins are described by the starting and ending tanimoto values. The Table 3.1 describes one such experiment with reference molecule corresponding to ZINC04090485. Here each bin comprises of compounds which are at most 0.025 similarity away. It also indicates the 'size', which is the number of compounds of NP data set that go in to the corresponding bin along with the number of data points from each bin that contribute to the respective representative set of various sizes.

Table 3.1: Binning based on Tanimoto metric: Number of compounds from each cluster for various sizes of representative set

bin	size	5000	4000	3000	2000	1000
0.335 to 0.360	2	1	1	0	0	0
0.360 to 0.385	4	1	1	1	0	0
0.385 to 0.410	3	1	1	1	1	1
0.410 to 0.435	90	5	4	3	2	1
0.435 to 0.460	381	21	17	13	8	4
0.460 to 0.485	527	30	24	18	12	6
0.485 to 0.510	1974	111	88	66	44	22
0.510 to 0.535	2206	123	99	74	49	25
0.535 to 0.560	4540	254	203	152	102	50
0.560 to 0.585	5558	311	246	186	124	62
0.585 to 0.610	6206	348	278	208	139	69
0.610 to 0.635	6215	347	278	208	139	70

0.635 to 0.660	6309	353	282	211	141	70
0.660 to 0.685	9180	514	411	308	205	103
0.685 to 0.710	8053	450	360	270	180	90
0.710 to 0.735	6934	388	310	233	155	77
0.735 to 0.760	8298	464	371	278	186	93
0.760 to 0.785	6603	369	295	222	147	73
0.785 to 0.810	10563	589	473	354	236	118
0.810 to 0.835	565	32	25	19	13	6
0.835 to 0.860	4128	231	185	138	92	46
0.860 to 0.885	61	3	3	2	1	1
0.885 to 0.910	545	31	24	18	12	6
0.935 to 0.960	354	20	16	12	8	4
0.985 to 1.0	89	5	4	3	2	1

3.2.2 Results

To determine whether our designed set is a better representative, we calculate the Euclidean distance between NPD and our CS. The Table **??** shows the distances between the Mean vectors of NPD and the CS of various sizes.

Table 3.2: Euclidean distances between Mean vectors of NPD and CS(RS) of various sizes

size	E(NPD-CS)	E(NPD-RS)
5000	0.63	2.22
4000	0.89	0.974
3000	1.17	0.978
2000	1.84	2.46
1000	2.44	1.82

3.3 Method: Cell-Based approach

Literature tells us that distance-based algorithms consider only inter compound distances, but not the absolute positions of compounds in chemistry-space. As a result, distance-based algorithms are inherently limited and are ill-suited for many of the other diversity-related tasks. In contrast, by dividing the multi-dimensional space into a grid, cell-based diversity algorithms partition chemistry-space into a lattice of multi-dimensional cubes and there by, considering not only inter-compound distance, but also absolute position of compounds in chemistry-space. These cell based approaches perform well for low dimensional chemical spaces and hence perfectly suits our space with 58 dimensions/ feature descriptors.

3.3.1 Approach

In this approach we construct a representative set using the following steps.

- Principal Component Analysis is used for dimensionality reduction of the data set from 58 dimensions to the number of required dimensions.
- The entire data is distributed in the space of these principal components and a grid dividing all the dimensions is placed over this distribution
- The grid is adjusted over each dimension such that when at least one data point from each cell of the grid is taken, it results in the representative set.(CS)
- The divergence of this set (CS) is compared with that of a random sample(RS) taken before placing the grid.

The number of dimensions to which we reduce the dimensions using the PCA also effects the representativeness of the set generated. Through several experiments it is observed that there is an increase in the closeness of the representative set generated, to the original set as the number of Principal components increase that is as the dimensionality of the space in which the data is distributed increases. The previous statement is validated using the KL-Divergence which is presented in the next chapter. Further as the number of principal components increase from 2 to 30 the representativeness increases and as we move above 30, it is observed that there is no significant variation.

3.3.2 Results

In this section we try to observe the Euclidean distances between the Means of NPD and the designed representative sets of various sizes, and compare them with those of the random generated representative sets. The standard deviations are also listed.

Table 3.3: Euclidean distances between Mean vectors of NPD and CS(RS) of various sizes

size	E(NPD-CS)	E(NPD-RS)
5000	0.6251	2.22
4000	0.53	2.947
3000	1.613	0.887
2000	1.209	1.231
1000	1.730	2.367

From the Table **??** we can observe that the Mean of designed set is closer than the random representative set.

3.4 Method: Structure based division

3.4.1 Classification

The data set comprises of compounds which in their structural representation contain a specific number of cycles. Based on the number of cycles present the data points are divided into nine classes where all the acyclic go into one subset, mono-cyclic into another and so on.

3.4.2 Approach

1. The data set is split into subsets based on the number of structural cycles present in the compound[15].
2. Random samples are taken from each of the above generated subsets in proportion to their sizes.
3. These samples are combined to form a Representative set(CS).

4. KL-Divergence between this combined representative set(CS) and original data set is compared to that of a randomly generated representative set(RS).

Table 3.4: Division of NPD into subsets based on the presence of number of structural cycles

category	Frequency	Percentage	Tetra cyclic	25961	29.03
Acyclic	639	0.71	5 cyclic	17036	19.05
Mono cyclic	4718	5.28	6 cyclic	5423	6.06
Bi cyclic	13147	14.7	7 cyclic	505	0.56
Tri cyclic	21966	24.57	8 cyclic	22	0.02

3.4.3 Results

This method uses the domain knowledge in contrast with the previous methods which make no use of domain knowledge in their algorithms to design a representative set. The Table **??** shows the Euclidean distances between the Mean of CS and NPD and also between RS and NPD, observing which we can infer that the Mean of designed set(distribution) is more close to the Mean of NPD distribution and thus indicating this method as a better approach to design a representative set.

Table 3.5: Structure based division: Number of molecules from each subset required to construct representative sets of various sizes

Category	Frequency	5000	4000	3000	2000	1000
Acyclic	639	36	29	21	14	7
Mono cyclic	4718	264	211	158	106	53
Bicyclic	13147	735	588	441	294	147
Tricyclic	21966	1229	983	737	491	246
Tetracyclic	25961	1451	1161	871	580	290
5-cyclic	17036	952	762	572	381	191
6-cyclic	5423	303	243	182	121	61
7-cyclic	505	28	23	17	11	6
8-cyclic	22	2	1	1	1	1

Table 3.6: Euclidean distances between Mean vectors of NPD and CS(RS) of various sizes

size	E(NPD-CS)	E(NPD-RS)
5000	0.932444	2.22077
4000	0.95911	0.97443
3000	1.21552	0.97823
2000	0.88926	1.62919
1000	1.30635	3.18863

Chapter 4

Results Validation

The previous chapter concludes by listing the results which we observed to be encouraging for the designed methods. The results were verified by computing the statistical measures of the representative distributions which indicated that all the methods performed better in generative a representative to the original distribution in comparison with the traditional random sampling. Now as a step further we validate the results using Kullback-Leible divergence and also see the coverage property of the various distributions.

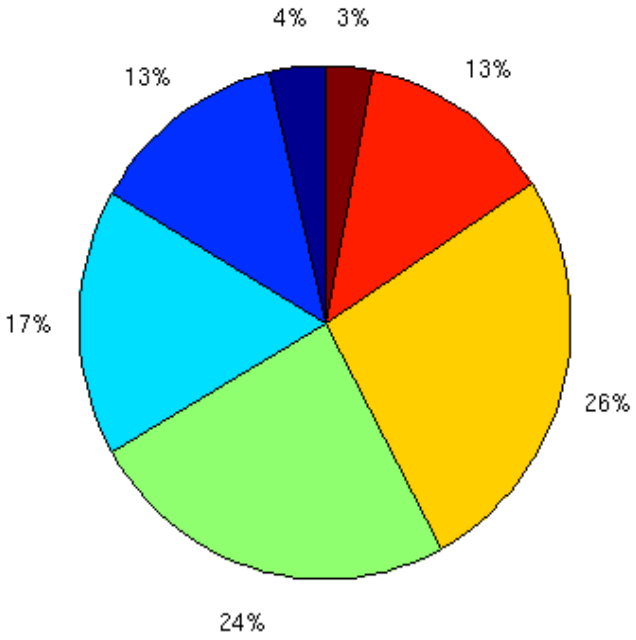
4.1 Coverage of the Data Set

As mentioned earlier the NPD set contains many types of classes one of which is the presence of number of structural cycles in the molecules. Based on this, the NPD is divided into nine subsets which indicates that, a set forms a good representative of the original distribution if it covers all these classes. We say that a class is covered when at least one element from the class is present in the final representative set. In the case of random generation of representative sets if the size of the subset/class is small then it is most likely that the class will not be covered in the representative set. But it is not the same for the designed representative sets generated from each of the methods above, where in each class is covered most of the times. Figure 4.2a validates the previous statement.

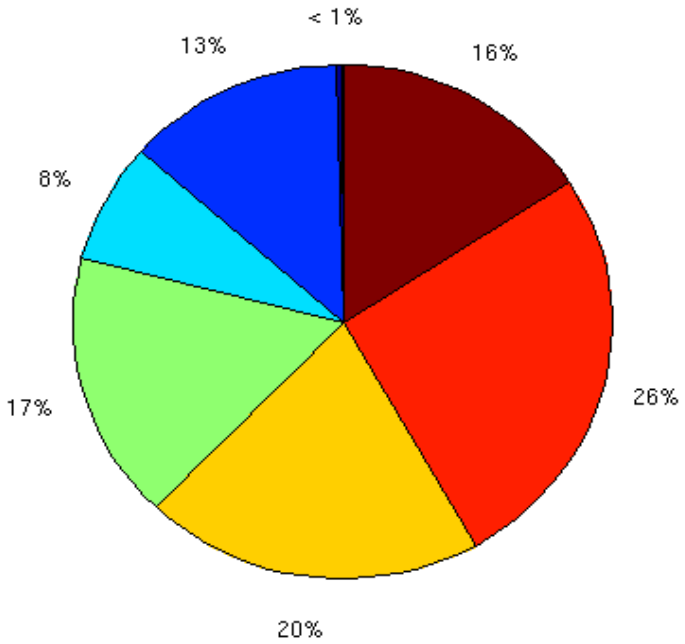
Figures 4.1a and 4.1b show RS of size 5000 and 3000 respectively, generated as part of Binning and Cell-Based methods which do not contain all the classes i.e miss out 9-cyclic, 8&9 cyclic compounds respectively, and thus result in less coverage, where as the designed representative sets as shown in Figures 4.2a and 4.2b result in higher coverage.

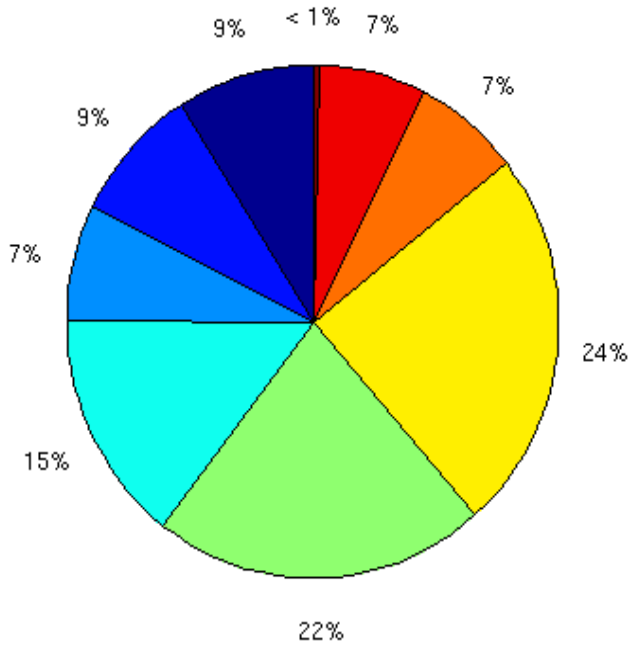
Figure 4.1: Coverage of classes by Random and designed representative sets

(a) RS of size 5000

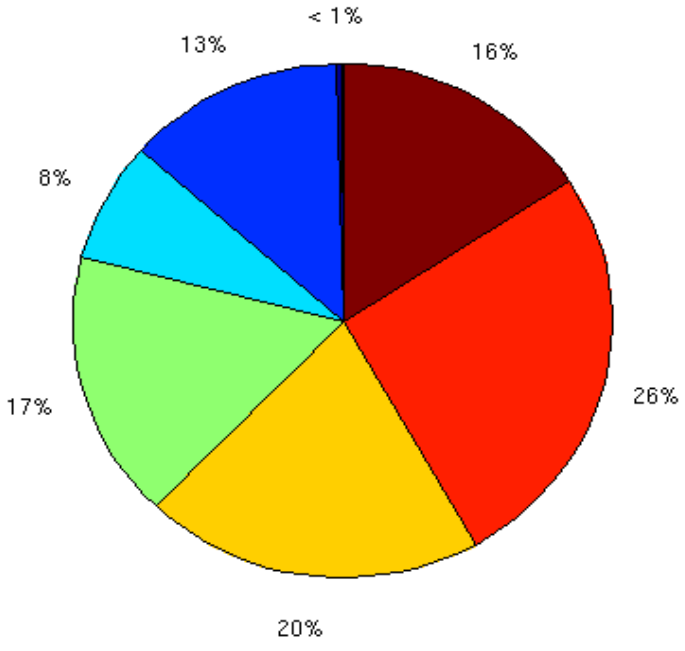


(b) RS of size 3000





(a) CS of size 5000: Binning method



(b) CS of size 3000: Binning method

Table 4.1: Coverage percentage of random versus designed samples of sizes 5000 and 3000

size	RS	CS
5000	77.7%	100%
3000	77.7%	88.8%

The Table 4.1 shows the coverage percentages, where we observe that there is 30% increase in the case of 5000 sized representative set and 10% for a representative set of size 3000, generated using the methods mentioned.

4.2 Validation using Kullback Leibler Divergence

KL-divergence is used to compute the relative entropy between two distributions and since our NPD and the representative sets generated from(RS and CS) represent respective distributions, we use this measure to validate the results of statistical measures and thus comply that the design approaches mentioned in the previous chapter will generate a better representative set than the usual random sampling. Here we present all the divergences in the direction of representative set from NPD, which implies that the divergence in the opposite direction may be different.

4.2.1 Binning method

In the following table we realise that the divergence of **C**ombined **S**ample(CS) from **NPD** is less than that of **R**andom **S**ample(RS) for a given size of the representative set. This indicates that the sets generated using binning method are more close to NPD and thus form better representatives.

size	D(NPD-RS)	D(NPD-CS)
1000	1.00057	0.89921
2000	0.88764	0.52846
3000	0.56821	1.23378
4000	0.15555	0.02741
5000	0.06454	0.05689

4.2.2 Cell-Based approach

In this method we observe that the divergence of the representative set is not influenced only by the size but also the number of Principal components used to transform the original distribution into the orthogonal space.

Table 4.2: Distribution over Principle components: Divergences of various representative sets from NPD

size	2-D		10-D		20-D	
	D(NPD-RS)	D(NPD-CS)	D(NPD-RS)	D(NPD-CS)	D(NPD-RS)	D(NPD-CS)
1000	1.27516	0.98462	1.00910	0.92114	0.47471	0.46335
2000	1.24789	0.85476	0.57652	0.63117	0.32811	0.28342
3000	0.88741	0.40005	0.23444	0.20881	0.17774	0.17200
4000	0.45681	0.28455	0.20978	0.17689	0.13639	0.12227
5000	0.11854	0.21249	0.11751	0.17459	0.10771	0.13632
size	30-D		40-D		58-D	
	D(NPD-RS)	D(NPD-CS)	D(NPD-RS)	D(NPD-CS)	D(NPD-RS)	D(NPD-CS
1000	0.36741	0.22432	0.32663	0.31476	0.31449	0.30627
2000	0.19631	0.20663	0.18111	0.17249	0.18012	0.18314
3000	0.09432	0.08367	0.08777	0.08350	0.07999	0.08731
4000	0.06697	0.06934	0.06419	0.06327	0.06422	0.06428
5000	0.06384	0.06001	0.06279	0.05921	0.06240	0.05910

From the divergences listed in Table 4.2 it is clear that mostly if not all times, the designed sample forms a better representative than any random sample. We can also deduce that as we increase the number of principal components from 30 to above, there is no much variation in the divergence values which indicates that the representativeness of the set generated does not significantly change. Geometrically there is no much variation in the slope of the lines joining the divergence values.

4.2.3 Splitting based on cycles

There is less computation in this method when compared to the previous methods and this procedure of generating a representative set for NPD works efficiently for smaller sets. We observe the same from the results in the Table 4.3 where the divergence of CS for sizes larger than 1000 is more compared to that of the RS of corresponding size. But as we decrease the representative set size this divergence almost equals the RS's divergence. But the advantage of this method even at the cost of a small increase in divergence is that its coverage is 100%.

Table 4.3: Structure based Division: Divergences of CS and RS from NPD

size	D(NPD-RS)	D(NPD-CS)
1000	1.44887	1.34551
2000	0.74657	1.55684
3000	0.58721	1.71298
4000	0.07542	1.08534
5000	0.05488	0.91587

Chapter 5

Future Work

Bibliography

[1] Jun Xu and Arnold Hagler:Chemoinformatics and Drug Discovery, Molecules, 7, (2002), 566–600.

[2] S. Joshua Swamidass and Pierre Baldi:Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time, J. Chem. Inf. Model, 47, (2007), 302–317.

[3] Robert S. Pearlman and K.M. Smith : Novel Software Tools for Chemical Diversity, Perspectives in Drug Discovery and Design, 9, (1998), 339–353.

[4] Martin Vogt and Jurgen Bajorath:Predicting the Performance of Fingerprint Similarity Searching, Chemoinformatics and Computational Chemical Biology, Methods in Molecular Biology, 49, (2011), 1369–1376.

[5] Brown Nathan:Chemoinformatics an introduction for computer scientists, ACM Comput. Surv., 41:2, (2009), 1–38.

[6] Weininger David : SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci., 28:1, (1988), 31–36.

[7] Andreas Bender , Jeremy L. Jenkins , Josef Scheiber , Sai Chetan K. Sukuru , Meir Glick , John W. Davies: How Similar are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space, J. Chem. Inf. Model, 49, (2009), 1–38.

[8] Peter Willett, John M. Barnard , Geoffrey M. Downs : Chemical Similarity Searching, J. Chem. Inf. Compu. Sci, 38, (1998), 983–996.

[9] Dimitris K. Agrafiotis, Deepak Bandyopadhyay Jorg K. Wegner , Herman van Vlijmen : Recent Advances in Chemoinformatics, J. Chem. Inf. Model, 47, (2007), 1279–1293.

[10] Pierre Baldi, Ryan W. Benz : BLASTing small moleculesstatistics and extreme statistics of chemical similarity scores, J. Chem. Inf. Model, 24, (2008), i357–i365.

[11] Peter Willett,Jurgen Bajorath (ed.) : Similarity Searching Using 2D Structural Fingerprints, Chemoinformatics and Computational Chemical Biology, Methods in Molecular Biology, 672,(2011).

[12] Pierre Baldi, and Ramzi Nasr : When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Values, World Academy of Science, Engineering and Technology.

[13] Naomie Salim, John Holliday and Peter Willett : Combination of Fingerprint-Based Similarity Coefficients Using Data Fusion, J. Chem. Inf. Comput. Sci, 43, (2003), 435–442.

[14] John J. Irvin and Brain K. Shoichet : ZINC- A Free Database of Commercially Available Compounds for Virtual screening, J. Chem. Inf. Model, 45, (2005), 177–182.

[15] Sankara Rao Ambati: Chemoinformatics Approaches in Classification and Screening of Databases,(2010).

[16] Valerie J. Gillet : Diversity selection algorithms, 2011.

Appendix

These pages in the book are intended to introduce the environment in which the project was experimented along with few facts and finally we conclude by providing few results for further investigation.

Environment

- All the programs are written in PERL.
- They are executed on i686 processor with0 Linux 2.6.34 as kernel.
- The system has a RAM of 2GB and a cache of 8MB.

Facts

The domain of Chemoinformatics is very new to us and initially it was difficult to understand the work done by our senior. But as we were determined to continue, we familiarized the subject and after a couple of months we were able to start of from the base provided by our senior. All the data is present in csv files and the size of raw data excluding the refinements made to it is 1.7GB in which the data is arranged the form of zincid, SMILES notation and the feature descriptors for every compound.

- Most of the programs are written with complexity $O(n^2)$
- It is important to organize the data along with the related files and programs and figures for every method experimented

- The calculation of KL-divergence is the most time consuming program with an average of 4 hours for every computation and the complexty of the program is $O(n^2)$

- It is worth redirecting results generated at every step in any experiment into files with an eye to, further usage of these results.

- It is very important to understand all the properties of the data such as the number of data points, the number of feature descriptors e.t.c as the names of the files may initially mislead.

- We used Matlab to perform numerical calculations and modifications on the results obtained and also for the generation of graphs and figures. The entire method of Cell-Based approach is implemented in Matlab.