Spring 2011

# IMPROVING MOLECULAR FINGERPRINT SIMILARITY VIA ENHANCED FOLDING

Victor Chen

# IMPROVING MOLECULAR FINGERPRINT SIMILARITY VIA ENHANCED FOLDING

A Thesis

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

By

**Victor Chen**

April, 2011

SAN JOSÉ STATE UNIVERSITY

The Undersigned Project Committee Approves the Project Titled

# IMPROVING MOLECULAR FINGERPRINT SIMILARITY VIA ENHANCED FOLDING

by

Victor Chen

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

| | | |
|---|---|---|
| Dr. Teng Moh, | Department of Computer Science | Date |
| Dr. Robert Chun, | Department of Computer Science | Date |
| Mr. George Lane, | Gilead Sciences | Date |

# ABSTRACT

Drug discovery depends on scientists finding similarity in molecular fingerprints to the drug target. A new way to improve the accuracy of molecular fingerprint folding is presented. The goal is to alleviate a growing challenge due to excessively long fingerprints. This improved method generates a new shorter fingerprint that is more accurate than the basic folded fingerprint. Information gathered during preprocessing is used to determine an optimal attribute order. The most commonly used blocks of bits can then be organized and used to generate a new improved fingerprint for more optimal folding. We thenapply the widely usedTanimoto similarity search algorithm to benchmark our results. We show an improvement in the final results using this method to generate an improved fingerprint when compared against other traditional folding methods.

# ACKNOWLEDGEMENT

I would like to express my sincere appreciation to my project advisor, Dr. TengMoh, for his attention, guidance, and expertise in supporting me throughout my research and preparation for this thesis. Thanks for giving me the opportunity to be part of the research group. In addition, I would like to thank Dr. Moh for introducing me to the topic of clustering in CS 274 taken in Spring 2010. This course sparked my interest to conduct research towards applicable use of clustering within the life sciences and gave me the theoretical concepts to apply it towards this project.

I would also like to express gratitude to my committee member, Dr. Robert Chun, for serving on my thesis committee, and for his in depth knowledge in parallel processing in CS 159, taken in Spring 2010. His expertise in computer science provided great value towards my project.

Lastly, I would like to thank my colleague, Mr. George Lane, for his expertise in chemistry. Our conversations and work together helped me understand the fundamentals of molecular fingerprinting and were the basis for my project work.

## Table of Contents

List of Figures and Tables ..... vii

1. Introduction ..... 1

2. Molecular Fingerprint Background ..... 3

   2.1 Structural Based Fingerprints ..... 4

   2.2 Chemical Based Fingerprints ..... 5

3. Similarity Matching ..... 6

   3.1 Ochiai/Cosine ..... 7

   3.2 Jaccard/Tanimoto ..... 7

   3.3 Russell/Rao ..... 7

   3.4 Baroni-Urbani/Buser ..... 7

   3.5 Simple Matching ..... 8

   3.6 Kulczynski ..... 8

   3.7 Fossum ..... 8

   3.8 Forbes ..... 8

   3.9 Simpson ..... 8

   3.10 Pearson ..... 9

   3.11 Yule ..... 9

   3.12 Stiles ..... 9

   3.13 Dennis ..... 9

4. Molecular Fingerprints and Similarity Matching ..... 10

5. Growing Problem with Molecular Fingerprints ..... 11

6. Prior Work ..... 12

   6.1 Create an Integer Based Fingerprint ..... 12

   6.2 Pruning the Database ..... 12

   6.3 Fingerprints with Entropy Codes ..... 12

   6.4 Similarity Search Analysis ..... 13

7. Project Goal ..... 13
8. New Approach ..... 14
8.1 Splitting Into Smaller Blocks ..... 15
8.2 Determining a Better Bit Order ..... 16
8.3 Ordering and Folding the Bits ..... 17
9. Implementation ..... 17
9.1 Parsing the SDF File ..... 18
9.2 Folding One Time ..... 19
9.3 Folding Two Times ..... 19
9.4 Calculating Tanimoto Similarity ..... 20
9.5 Calculating Statistics ..... 20
9.6 Using Statistics to Create an Improved Fingerprint ..... 20
9.7 Clustering the Resulting Data ..... 20
9.8 Application Correctness ..... 21
10. Results ..... 21
10.1 Number of Unique 32 Bit Blocks ..... 21
10.2 Original Fingerprint vs. Folded 1x ..... 22
10.3 Original Fingerprint vs. Folded 2x ..... 23
10.4 Improved Fingerprint vs. Original Fingerprint ..... 24
10.5 Improved Fingerprint vs. Folded 1x Fingerprint ..... 25
10.6 Timing ..... 26
11. Future Work ..... 27
11.1 Further Enhanced Pre-processing Step ..... 27
11.2 Similarity Algorithm ..... 28
12. Conclusion ..... 28
13. References ..... 29

## List of Figures and Tables

### Figures

Figure 1. Image of a molecule represented as a 3D image. Trying to compare different molecules is extremely challenging. ..... 3
Figure 2. This is an example of one way a molecule (Ethylbenzene) can be decomposed to create structural bits. The very top molecule is the target molecule. Each of the six decomposed molecules are parts of the original. Since they are available, they have a bit of 1 in the fingerprint. ..... 5
Figure 3. Example of Folding 1x. In this example, the original fingerprint is split into two pieces. The bits from the first half of the fingerprint and second half of the fingerprint are folded once to produce a fingerprint half of the original size. . 11
Figure 4. An example on the detrimental effects that folding a fingerprint multiple times has on a fingerprints actual result. Although some information may be maintained, there is also a higher chance that zeroes may become ones, thus misinterpreting the original fingerprint. ..... 14
Figure 5. Information is retained with 1. Information is lost with 0. The best outcome is when the values being compared are either 1 or 0. ..... 15
Figure 6. How to determine the number of occurrences at each index. The example above shows the result of four fingerprints of length 8. For each bit, we take a count on the number of occurrences. The results above show that the first two and last two bits have the highest occurrence. ..... 16
Figure 7 With the ordered bits in the best case scenario, there is no information loss. ..... 17
Figure 8. An example of an SDF file. Although there is not a standard on expected data within a file, the portions of interest in these SDF files are found in the ><fingerprint_1> sections. ..... 19
Figure 9: Number of Unique 32 Bit Blocks. As the number of fingerprints increases, there are more fingerprints available. ..... 22

## Tables

Table 1: Comparing results from the Original Fingerprint against the Folded 1x Fingerprint 23
Table 2: Comparing results from the Original Fingerprint against the Folded 2x Fingerprint 24
Table 3: Comparing results from the Original Fingerprint against the Improved Fingerprint 25
Table 4. Comparing results from the Improved Fingerprint against the Folded 1x Fingerprint 26

## 1. Introduction

In 2009, prescription pharmaceutical sales in the world's major markets totaled $555 billion, with the highest sales from the United States and the European Union [1]. The U.S. pharmaceutical market generated half of the total sales in 2009, and continues to grow at an expected rate of 3-5.0% [2]. The single best-selling drug's annual sales were $12.9 billion [3]. While pharmaceuticals can enjoy the high profits from a new drug, companies also face imminent patent expirations, in which the drug formulations are subjected to generic forms. In this industry, it is critical for a pharmaceutical company to constantly innovate to develop new drugs. On average, only 25 truly novel compounds are annually approved by the US Food and Drug Administration (FDA) [4]. The process for scientists to research and develop such compounds is lengthy, costly, and does not necessarily guarantee approval from the FDA or any return on investment.

Scientists usually begin their research through previous experiments or exposure to familiar compounds. They must examine and perform tests on thousands of molecules in hope of discovering a therapeutic efficacy. An enormous number of molecules are routinely assayed through a series of tests in the laboratory with the goal of creating a marketable drug. Scientists must start with a drug target, an existing cellular or molecular structure that is directly involved with the drug-in-development, and perform high-throughput screening (HTS). The HTS process compares the drug target with vast libraries of chemicals that are tested for their ability to modify the target. Because this screening process is so expensive and time consuming, it is in the interest of pharmaceutical companies to limit the number of molecules for assaying. One part of the process involves scientifically eliminating molecules that do not fit attributes of the target. This task is also a time consuming process. Once assays are run against a set of selected molecules, they can be analyzed and the entire process may repeat based on newfound data.

The first step to this drug discovery is for scientists to focus on understanding how a disease interacts in the body. Once this interaction is understood, biologists and chemists try to de-active parts of the interaction with a specially design and chosen molecule. Ideally, scientists hope to use selective features to find molecules that interact most with a chosen target. This process is iterative to

narrow down an extremely large database of molecules to a smaller more manageable list of molecules that shares the most common chemical features. Scientists see if the matching similarities against the chosen targets will also negatively interact with other closely related targets. This is to determine if these other related targets will create a toxic mixture. The screening process aids scientists to find compounds that share similar chemical properties, which ultimately brings them a step closer towards a lead compound.

It is the lead compounds that are then moved into drug development. These molecules need to be tested to insure that at a specific dosage, the drug is still safe and effective, both in the lab and with test animals such as rats or monkeys. Once this data has been approved by the FDA, the process makes its way into human test trials. Human test trials are even more expensive to run than in the lab and in animal and can directly impact a company's reputation. It is estimated that the cost to research and develop a new molecular entity is around $1.8 billion [5].

Another method that scientists have used to determine molecule similarity include various visualization tools including the radial clustergram [6], where molecules are grouped together and viewed via a dendrogram or other 3$^{rd}$ party software including major companies like Spotfire to aid in the decision-making process.

Our research primarily focuses on the process of narrowing the extremely large database of molecules that scientists must use to screen their targets. Given a particular drug target, each molecule has different measureable attributes that are used to compose a molecular fingerprint. With the help of molecular fingerprints and existing methods to find similar compounds, we have developed a more efficient and accurate screening method that better narrows the list of lead compounds.

A high level overview of our work is to shorten the fingerprint while maintaining accuracy. The first step is tocreate, or preprocess, a list of molecular fingerprints of interest. Based on analysis of the position of its fingerprint properties from preprocessing, we determine an optimum sort order to best regenerate all fingerprints.

In the sections to follow, we will provide the basic background to molecular fingerprints and how they are generated. Next, we review various similarity matching algorithms. With molecular fingerprints, we explain how similarity matching algorithms are used to determine molecular similarities and identify the bottleneck in the process. We discuss prior work in this area from other authors who have introduced improvement methods. Finally, we discuss our new approach, its implementation, and results.

## 2. Molecular Fingerprint Background

A molecular fingerprint is a method in chemical informatics where a molecule's attributes are represented as a string of binary bits of variable length [7]. Binary encoding of these fingerprints detect a 1 for attribute presence or 0 for attribute absence [8]. A fingerprint is the simplest way for a computer to compare for similarity because comparing strings of bits is far simpler than comparing 3D molecules.

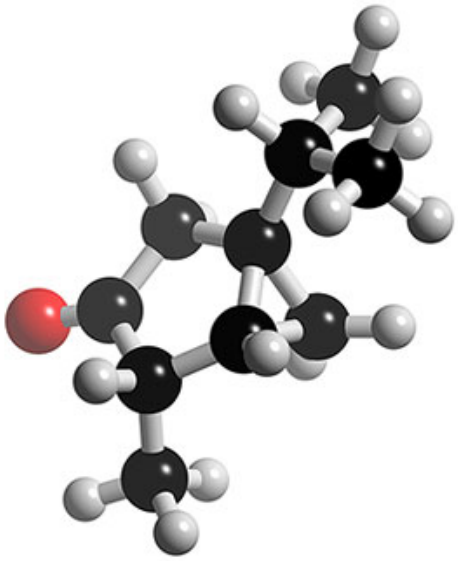

**Figure 1.**Image of a molecule represented as a 3D image. Trying to compare different molecules is extremely challenging.

There is no maximum length of a fingerprint and no rules on which attributes are required to be included in the fingerprint. There can therefore be a limitless number of ways to represent the same molecule, depending on which the attribute bits are defined. There are several large companies who create and maintain commercially available databases of fingerprints for the pharmaceutical

industry. The downside of commercially available databases is that they are expensive and available through restricted interfaces. Two popular databases include Molecular Design Limited's (MDL) Available Chemical Directory (ACD) and the American Chemical Societies (ACS) CAS Registry, which contains all molecules reported in chemical literature. Other available databases include ChemSpider, which contains compound information from Web sources, and PubChem, which stores compound information from governmental and academic sources related to data from bioactivity [9, 10]. An approximate of twelve thousand molecules is added each day to public databases, but private ones could contain more [11].

In addition to commercially available companies, other educational sources such as the University of California, Irvine (UCI) maintains a publically available source of molecules for research purposes only called ChemDB [12].

Fingerprints can be created to support various types of attributes, which include structural, chemical, and various combinations.

## 2.1 Structural Based Fingerprints

The goal of structural fingerprints is to find molecules that have the most matching physical properties as a target molecule. Based on the Similarity Property Principle, which says that structurally similar compounds are likely to exhibit similar biological activity, scientists can use known structural attributes to guide pathway manipulations [13]. Therefore, the intent of a structural fingerprint is to identify molecules with a similar molecular shape. Each bit of fingerprint represents whether the structural fragment of the molecule exists.The set of fragments are small substructures of the molecule to describe a phenyl, nitro, and carboxylic acid group [14].

An example of a structural fingerprint takes a molecule and decomposes it into its smallest atoms. Through a series of iterations, every possible fragment of the original molecule is added as an attribute to the fingerprint. The fingerprint does not take into account the number of each fragment, but whether or not the fragment exists. This resulted in a linear speedup during similarity searching [15].

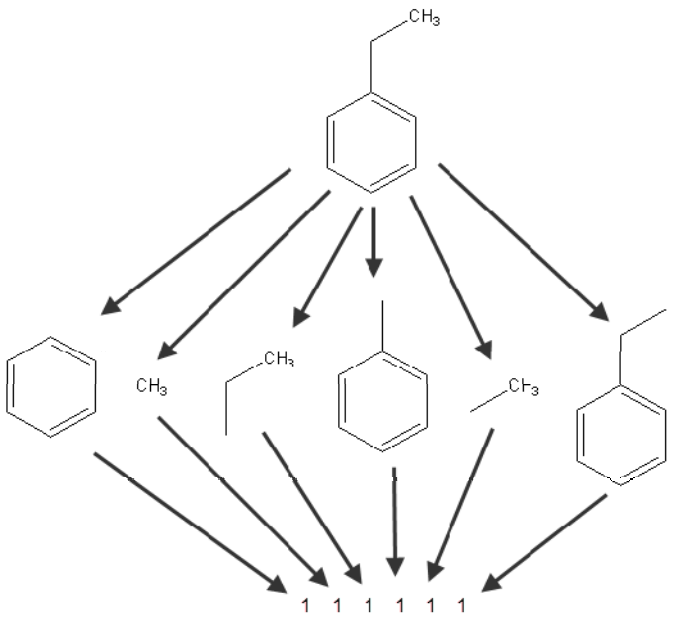


**Figure 2.This is an example of one way a molecule (Ethylbenzene) can be decomposed to create structural bits. The very top molecule is the target molecule. Each of the six decomposed molecules are parts of the original. Since they are available, they have a bit of 1 in the fingerprint.**

Other methods of structural fingerprinting enforce a maximum path length on the fragment size of the molecule. This limits the length of the fingerprint for larger molecules, but results in poorly represented chemical structures [16].

The major challenge with structural fingerprints is that there are too many ways to represent a single molecule. Based on the example in the figure above, we can see there are at least six common chemical attributes, but it can potentially have many more, especially if the molecule is larger and more complicated [17].

## 2.2 Chemical Based Fingerprints

The chemical fingerprint consists of the chemical properties of a particular molecule. Some properties can be obtained through lab experiments. Other properties can be obtained through other software that attempts to predict chemical properties. The intent of a chemical fingerprint is to match a molecule with the same chemical properties, but perhaps a completely different structural shape.

A downside of chemical properties is that certain assays can only measure up to a certain range, depending on the sensitivity of a certain assay. With different situations, different assays can be created to measure. Therefore, it is possible

that a chemical result can have a result containing a greater than or equal to value, indicating it was higher than the assay is able to measure. This can create a noise which is also a problem [18, 19].

## 3. Similarity Matching

With a list of molecular fingerprints from a single source, a target's fingerprint can be compared against all other fingerprints in the database. The results determine how similar the attributes of the fingerprints are compared to a target molecule. This informatics portion of drug discovery helps scientists to quickly narrow down their search results.

Aside from the molecular fingerprint comparison, similarity matching can also be done through weighting schemes that show the frequency that a certain fragment occurs. This provides greater detail than the 1 (present) and 0 (absent) encoding with molecular fingerprints. By prioritizing how some fragments are weighted more heavily than others, scientists can get better results on similarity between compounds.

Other areas of similarity matching included pre-processing the fingerprints to determine which molecules need to be scanned based on the number of 1 bits relative to the number of 1 bits in the target molecule. Pre-processing in this manner was concluded to linearly improve the similarity checking process [20].

Within similarity matching, algorithms are used to check by similarity coefficients. There are various algorithms for checking fingerprint similarity. The algorithms result in coefficients that fall into a range of possible values, such as the range from 0 to 1. These coefficients are categorized as association coefficients, correlation coefficients, and distance coefficients. Association coefficients are described as values that range from zero, shows no common features, to unity, shows identical features. For our project work, we will focus on association coefficients. The most commonly used indexes, range, and formula are described below. In the algorithms below,“a”equals the number of bits that are in common to the target molecule and the molecule being compared.“b”equals the number of bits that are ones in the target molecule. “c”equals the number ofone bits in the molecule being compared.“d” equals the number of bits found in neither molecule. “N” is the fingerprint length. [21]

### 3.1 Ochiai/Cosine

The Cosine index is the ratio of the bits in common to the geometric mean of the number of on bits in the two items. The range of the cosine method is 0 to 1.

$$\frac{a}{\sqrt{(a+c)*(a+c)}}$$

### 3.2 Jaccard/Tanimoto

The Tanimoto and Jaccard indexes are the same, and have a range from 0 to 1. The Tanimoto index selects smaller compounds, and has an association coefficient type. The Tanimoto method is the most widely used algorithm because of its proven superiority over other algorithms.

$$\frac{a}{a+b+c}$$

### 3.3 Russell/Rao

The Russell/Rao index has a range from 0 to 1. It has the worst performing coefficient when tested on a 20k data with 2000 cluster levels. Through various test trials, it is found that Russell/Rao is biased toward larger compounds. Performance improves when the active compounds retrieved are relatively large in size. The Russell/Rao has an association coefficient type.

$$\frac{a}{n}$$

### 3.4 Baroni-Urbani/Buser

The Baroni-Urbani/Buser was tested in a clustering technique and performed as well as the Pearson, Stiles, and Yule index. It is categorized within the 10% best performers with rankings at levels of 2000 clusters, 1000 clusters, and 500

clusters. In mapping the standard deviation in size of compounds selected, Baroni-Urbani/Buser is most suitable for compound selection.

$$\frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}$$

### 3.5 Simple Matching

The Simple Matching algorithm has a poor-performing coefficient when tested against a 2000 cluster level. It has a tendency to select larger (in terms of bit density) compounds and retrieve smaller ones. Simple Matching has an association coefficient type.

$$\frac{a + d}{n}$$

### 3.6 Kulczynski

The Kulczynski index is the mean of the individual substructure similarities and provides average rankings by itself or in combination. It has a range from 0 to 1.

$$\frac{a}{2(2a + b + c)}{(a + c)(a + b)}$$

### 3.7 Fossum

$$\frac{n\left(a\left[-\frac{1}{2}\right)\right]^2}{(a + b)(a + c)}$$

### 3.8 Forbes

The Forbes index has no upper limit. It has a poor performing coefficient when tested against the 2000 cluster level. It has an improved performance with retrieving smaller compounds.

$$\frac{n * a}{(a + b)(a + c)}$$

### 3.9 Simpson

The Simpson index is the best of the individual substructure similarities, however, when tested among the other indexes, it performed as poorly as the Simple Match, Forbes, and Russell/Rao.

$$\frac{a}{\min(a+b, a+c)}$$

### 3.10 Pearson

The Simpson index is the best of the individual substructure similarities, however, when tested among the other indexes, it performed as poorly as the Simple Match, Forbes, and Russell/Rao.

The Pearson index is a superior clustering technique when tested with various levels of clustering.

$$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$$

### 3.11 Yule

The Yule index is a superior clustering technique when tested with various levels of clustering.

$$\frac{ad-bc}{ad+bc}$$

### 3.12 Stiles

$$log_{10}\frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(a+c)(b+c)(c+d)}$$

### 3.13 Dennis

$$\frac{ad-bc}{\sqrt{n(a+bc)(a+c)}}$$

In providing an overall analysis of the various indexes, the Simple Match, Russell/Rao and Forbes algorithms weighted differently in terms of size distributions, which lowers the chances of finding a similarity match based only on structural chemical properties. The Baroni-Urbani/Buser coefficient selects compounds that show a similar size distribution as the compounds from the database. The Tanimoto coefficient performs the best, but it can be improved by combining selections of complementary coefficients through various combining

data ranking techniques. Pearson, Yule, Stiles, and Dennis have correlation coefficient types [22].

While each algorithm may prove more effective depending on the data and sample size, it is impractical to use a different algorithm each time a molecular fingerprint needs comparison. This leads to further examination and analysis of compound search methods. However, since the Tanimotoapproach is the most widely used similarity algorithm, we will use this as our benchmarking algorithm throughout the rest of this paper.

Also, the final similarity score that is retrieved for a molecule is often times confused as a general score for the entire molecule. The fact that the similarity score is highly dependent on the attributes presented in the fingerprint is often overlooked. If attributes are completely irrelevant to the target profile, the resulting similarity score will have less meaning.

## 4. Molecular Fingerprints and Similarity Matching

With a molecular fingerprint, scientists who understand the pathways of a particular disease start with a general concept of how to manipulate the disease. This pathway information can be continued research in novel areas of an existing drug, where data is published and replicated. Another situation is research data from an academic institution who have published research papers. A third way is a scientist may have some understanding on a specific target and begin running experimenting with molecules whose shape and characteristics fit the target's profile.

Once a scientist has laboratory data that positively support their hypothesis, similarity searching becomes a reality. They can take other well-known attributes from the molecules of interest and assigned them to clearly identify the attributes they want to find from other molecules that may have been previously overlooked.

The goal of similarity matching is to assign a similarity score for molecules so scientists can focus more attention on molecules that are more similar to molecules of interest. This way, a scientist is not required to dedicate time to individually examine molecules that do not match. Additionally, money otherwise

spent on poor molecules can be spent on molecules with a higher chance of success.

## 5. Growing Problem with Molecular Fingerprints

A problem with using molecular fingerprints is that scientific technology is becoming more advanced. Two advances are making molecular fingerprint comparisons more challenging. With the introduction of high throughput screening, the list of known molecules and fingerprints are growing larger and at a faster rate. This presents a challenge to fingerprint similarity because it increases the number of possible comparisons. In addition to more comparisons, an improved science allows each molecule to have more known attributes available for comparison. This results in longer fingerprints. Therefore, the problem is two sided. Fingerprints are longer and more numerous.

To alleviate extremely long fingerprints, a technique known as “Folding” is available. This method manages to keep the important attribute information while minimizing the size of the fingerprint. In short, it increases information density, with the assumption that most bits of a fingerprint are expected to be zeros by assuming more information in a single bit [23]. Therefore, each bit in the original fingerprint is related to a single attribute. In a single folded fingerprint, each bit contains data for two attributes.

There is no limit to the number of times a fingerprint can be folded. In the folding example below, the binary string below is folded a single time.

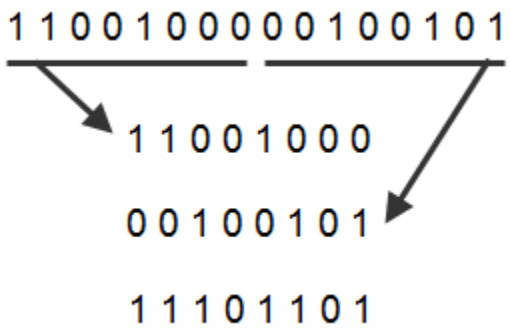


**Figure 3.Example of Folding 1x. In this example, the original fingerprint is split into two pieces. The bits from the first half of the fingerprint and second half of the fingerprint are folded once to produce a fingerprint half of the original size.**

The result is a binary string half of its original size, but retains the known information about a molecule. If a molecule is extremely long, a fingerprint can always be folded multiple times to make the molecular fingerprint shorter. Since fingerprints have more 0 bits than 1 bits, this does not drastically affect the result. However, with each additional fold, the rate of false positives increases since more bits that were previously zero are converted to one bits.

## 6. Prior Work

The challenges surrounding increasingly long and numerous fingerprints has been acknowledged and reviewed by many other papers. Each paper presents a unique method to improve the situation.

### 6.1 Create an Integer Based Fingerprint

With this method, the original author attempts to shorten a fingerprint by completely changing its signature type [24]. Instead of using a simple one or zero bit for the fingerprint, the author creates an integer based fingerprint. This fingerprint print is created by splitting the original fingerprint a variable number of groups. Based on modulo hashing, the members of each hashed group are summed and the resulting value is the integer value for the signature of the new fingerprint. With this approach, the author is able to shorten the fingerprint [25].

### 6.2 Pruning the Database

Research has been conducted to prune the database. This process involves a preprocessing step and the basic fingerprint that contains only ones and zeroes. The research sums the number of one's bits in the target molecule's fingerprint. Instead of checking every single molecule in the database, researchers only check those fingerprints that have a similar number of ones in their fingerprints [26]. Therefore, the research needs to preprocess all molecules and determine the number of ones. Using this method, the research is able to gain a speed up during similarity comparison. This is due to the fact that there are less molecules requiring similarity searching.

### 6.3 Fingerprints with Entropy Codes

In another related method,research was conducted to improve ways to store fingerprints. The researching methods are to improve storage and retrieval times

of fingerprints [27]. Instead of simply storing the fingerprints as a string of binary bits, the researchers also calculate a new fingerprint representation based on Golomb and Golomb-Rice Codes. This method uses statistics to reduce the length of a fingerprint by reassigning the most popular sections of bits to a shorter value. The goal of this paper was specifically for the storage of fingerprints in a lossless fashion.

### 6.4 Similarity Search Analysis

The similarity search analysis study looked into how the different similarity searches performed with different fingerprint types. Researchers analyzed four fingerprints with different attributes by listing the pros and cons of each type. Additionally, they obtained fingerprints for each and looked at the result of how different types of fingerprints affected the similarity results. What the researchers found were the descriptor fingerprints are more defining that the behavioral based fingerprints [28].

## 7. Project Goal

Since a common method of molecular folding is popular, we want to compare this method of fingerprint shortening and investigate if there is an alternative method that both shortens the fingerprint and maintains the molecular attribute information. The goal is to create a method that is faster than and as accurate as a linear search.

We begin accuracy in the folding methods. We will gather data to compare how folding one time and folding two times is faster and retain some accuracy than doing a linear search on the original database [29], which we will call the “original.”. Because theTanimoto Similarity score is the most widely used method for comparison, we decided to use it as our benchmark similarity method to measure how accurate the search results are to the original.  After we have gathered this data, we will compare how our new, modified algorithm performs against the folding benchmarks.

At the end, we want our fingerprint results to be more accurate than the folding, but also achieve similar or better timing performance than the similarity search. It is critical that at the very minimum, the molecules that are present in the original fingerprint similarity results also appear in our statistic based fingerprint result.

## 8. New Approach

The new approach is designed to enhance the folding process that results in more accuracy than the standard fold. Currently, the folding once method takes a molecular fingerprint, and folds exactly in half. This folding method rounds up the number of similarities between compounds, and thus provides a larger list of common compounds than the original. Similarly, folding twice creates an even larger list of common compounds because more zero bits are converted to one bits.

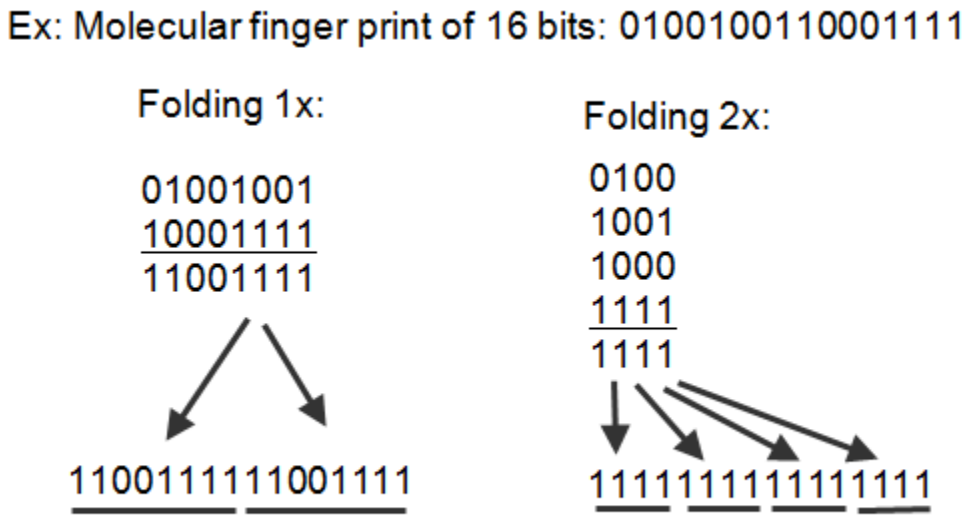


**Figure 4.An example on the detrimental effects that folding a fingerprint multiple times has on a fingerprints actual result. Although some information may be maintained, there is also a higher chance that zeroes may become ones, thus misinterpreting the original fingerprint.**

Since we expect a fingerprint to contain more zeros than ones, we want to give a fold the best chance of resulting in a one bit only when both bits in the fold are one, and result in a zero bit when both bits in the fold are zeros. In the figure below, the first and fourth case yield the best results. The partial goal is to minimize the amount of information that is lost during the fold. As seen in the figure, case 2 and 3 contain the most lost information.

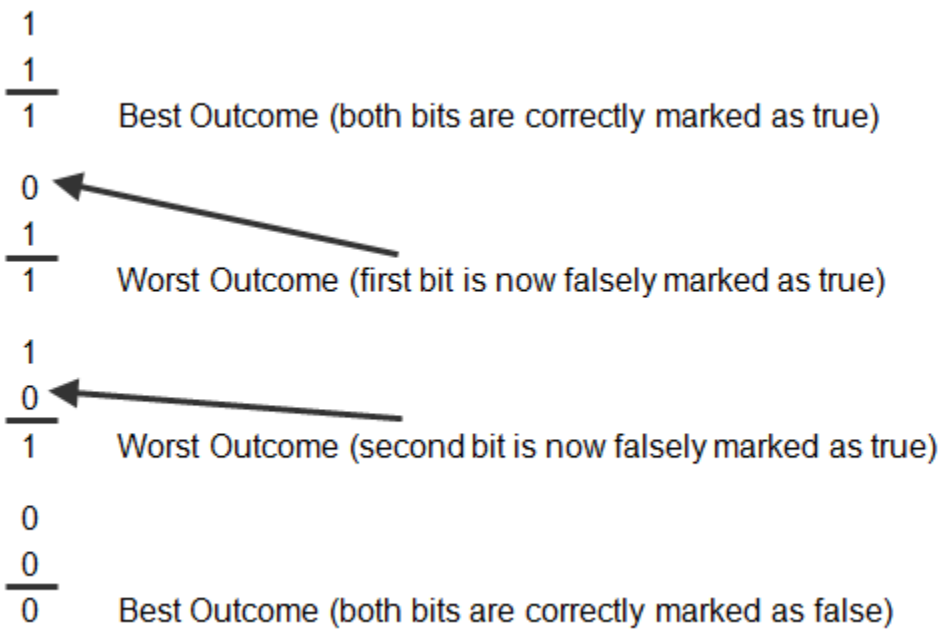


**Figure 5. Information is retained with 1. Information is lost with 0. The best outcome is when the values being compared are either 1 or 0.**

We accomplish this by performing a pre-processing step which loops through every fingerprint and analyzing the index of a shorter block bits that have a value of one to determine a new bit order. Note that reordering the bits of the fingerprint is not harmful as long as the reordering occurs uniformly throughout all fingerprints. Reordering with an entirely new string of bits that have no relation to the original string of bits is also unacceptable, as the attributes would be incorrectly grouped together. This is the case because the benchmarking similarity algorithm, Tanimoto, takes each individual bit into consideration when calculating the similarity coefficient.

We accomplish this by performing a pre-processing step which loops through every fingerprint and analyzes the index of a shorter block bits that have a value of one to determine a new bit order.

### 8.1 Splitting Into Smaller Blocks

In a pre-processes phase, we split all the fingerprints in the database up into equally sized smaller blocks of bits. We then create a list of these smaller blocks of bits and count the number of unique occurrences. This list is sorted descending by number of occurrences. The purpose of pre-processing is to determine the blocks of bits with the most number of occurrences. Though all

blocks of bits are important, we do not want the blocks of bits that occurred less frequently to negatively weight our results. Given that we are analyzing a molecular fingerprint, we expect there to be an extensive list of repeated blocks of bits, since the occurrences of zeros are much higher than ones.

### 8.2 Determining a Better Bit Order

Using the pre-processed list with only the highest occurrence of blocks of bits, we want to create a rule to determine a new ordering of the bits within those blocks of bits that would best improve the folding process.For this to occur, we want the two indexes with the first and second highest number of one bit occurrences to be folded together. With this reasoning, we want to put the index with the highest number of one occurrence in the zero indexes, and the index with the second highest number of ones occurrence in exactly the middle of the block of bits. In this manner, when the block of bits are folded, those two bits will fold together, and no information about the attributes will be lost.

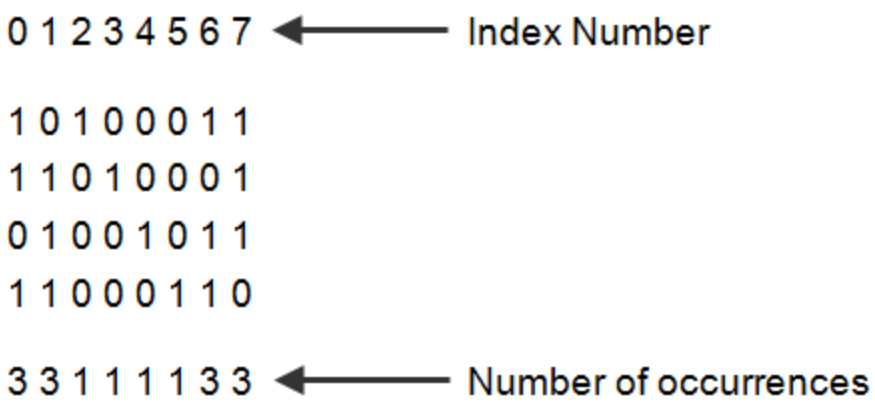


**Figure 6.How to determine the number of occurrences at each index. The example above shows the result of four fingerprints of length 8. For each bit, we take a count on the number of occurrences. The results above show that the first two and last two bits have the highest occurrence.**

In the figure above, we assume that there are four fingerprints. The goal is to determine if any of the property indexes are populated more frequently than the other property indexes. In this example, we discover that indexes 0, 1, 6, and 7 are the most frequently used columns. Therefore, those are the columns we can determine should be matched together.

### 8.3 Ordering and Folding the Bits

After defining a better bit order rule, we are now ready to create our new fingerprint. We take the bit re-ordering rule previously generated and apply it to the shorter blocks of bits. Next, with the new order, we fold the block of bits one time. After all small blocks of bits from one fingerprint have been reordered and folded, they are concatenated back into a new fingerprint whose length is half of the original fingerprint's length. This process is repeated to cover all blocks of bits for every fingerprint.

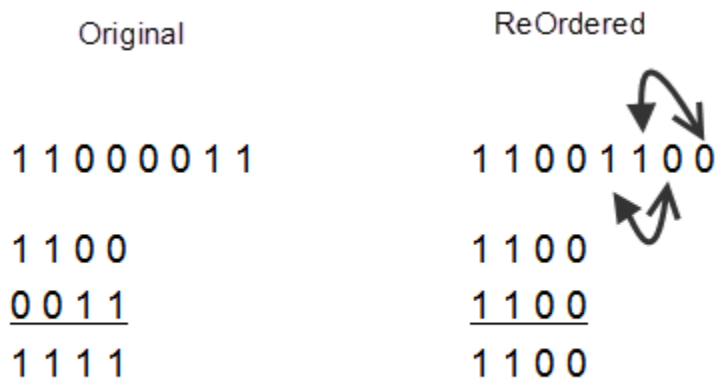


**Figure 7With the ordered bits in the best case scenario, there is no information loss.**

The figure above is an example of the best case situation for reordering. Again, we assume that indexes 0, 1, 6, and 7 have the most frequent occurrences. Therefore, we want the most commonly indexed to be matched together. In the original case (without reordering), the resulting folded fingerprint will lose half of its information because the bits are not be in the right position. In the re-ordered bit, we swap the fingerprint attributes in the second half the fingerprint so that the folding process will not lose information.

## 9. Implementation

Creating a system to experiment with ways to better improve the elimination process of molecules was implemented from scratch. The first step was to obtain an open source list of fingerprints. ChemDB is a readily available database of files in a structure-data file (SDF) format, which included a molecular fingerprint. ChemDB's molecular fingerprint includes values calculated from other third party

applications like CORINA, OEChem, and vendor annotations [29]. The SDF file is downloaded once outside of the actual system.

The main system itself was programmed as a console application in Microsoft Visual Studio 2010 in C#. The application was run locally on a Core 2 Duo laptop with 4GB of RAM.

We decided it was best to use the Tanimoto Similarity due to its widely accepted popularity and to benchmark the new fingerprint method against Folding 1x, Folding 2x, and the original fingerprint.

Additionally, the final version of the code does not try to optimize for speed. There are several sections in the code which perform correctness checking to help insure the system is bug free to avoid a bias in the results that may otherwise go unnoticed.

### 9.1 Parsing the SDF File

Parsing the SDF file was added to improve runtime. Since the SDF files are batched in groups of 100,000 molecules, we obtained around twenty files for 2 million compounds. There was approximately 11 GB worth of data. That was far too much to parse each time we ran the application. Therefore, we parsed the 11 GB of data to keep only the fingerprint in a simple CSV file format. This portion of the application had read only access to the location of SDF files. We extracted only the fingerprint and wrote the data out to an intermediate file to speed up future processing of the files. When writing out to the intermediate file, we also wanted to insure that all fingerprints were 1024 bits for this experiment. Any duplicates were removed to keep analysis of the fingerprint simple. An abridged example of a block of SDF file below shows the types of information contained for each section.

The very top of the SDF file contains a mol block, which is a textual representation of the molecule that can be converted to a 2d representation via various software applications. Next, any additional properties and property values can be added as necessary. There are no guidelines stating what properties are required. The portion that we are most interested in is the fingerprint_1 section. This value is what the rest of the application will consider as the “original” fingerprint.

```
 14 35  1  0  0  0  0
 15 20  2  0  0  0  0
 15 16  1  0  0  0  0
 16 17  2  0  0  0  0
 17 18  1  0  0  0  0
 17 36  1  0  0  0  0
 18 19  2  0  0  0  0
 18 37  1  0  0  0  0
 19 20  1  0  0  0  0
 19 38  1  0  0  0  0
 20 39  1  0  0  0  0
M  END
> <isomer>
2

> <h_acceptors>
2

> <num_atoms>
39

> <num_sg_bonds>
16

> <num_db_bonds>
6

> <fingerprint_1>
1024.-211-y-Nfx3j9003---7-f-v7-7zx-91b-fa-z-11LDo-vf-1-BAwH-74z14di10oj---nf---300gn17-vwffw-j2-wg137--fL1-fE2h-3x00---
3-v9vjbg-403vf313-751L3f3i7-d9bg-5xf-1v1771-1p3DdN3v-F---

> <zap_total>
9.02345

> <can_smiles>
Cc1cc(c(o1)C)C(=O)NC2CCCc3c2cccc3

> <oepassed_filter>
t

> <zap_area>
460.377

> <zap_solvation>
-2.48598

> <xlogp>
3.09

> <num_h>
19

$$$$
```

**Figure 8. An example of an SDF file. Although there is not a standard on expected data within a file, the portions of interest in these SDF files are found in the ><fingerprint_1> sections.**

## 9.2 Folding One Time

To create a shorter fingerprint, the most popular method is to fold the fingerprint one time. Doing so involved splitting the fingerprint up exactly in half. A logical OR is then performed bit by bit with each half to generate a now more densely populated variable. The resulting variable was stored along with the original fingerprint.

## 9.3 Folding Two Times

As a second analysis to see how a more compressed fingerprint would affect results, we folded the molecule a second time. This involved splitting the molecule up into quarters and then performing the logical OR on all portions together. This folded two times fingerprint was also stored in memory. Based on the results gained from Folding 2x, we conclude that it is unnecessary to Fold a third time.

## 9.4 Calculating Tanimoto Similarity

A generic Tanimoto Similarity function was created which accepts a bit array of any length and returns the final similarity score between 0 and 1. Additional overhead was added to insure the incoming parameters and return values were correct. This means checked that both fingerprints are of the same length, and make sure the resulting value was is truly between 0 inclusive and 1 inclusive.

## 9.5 Calculating Statistics

We want to pre-process the original fingerprint to generate statistics on the fingerprint. For each of the approximately one million molecular fingerprints, we split the fingerprint up into 32 bit blocks. Since the fingerprints are each 1024 bits, each fingerprint will contain exactly 32 equally sized blocks. With a list of 32 bit block, we count the occurrence of each unique block. We expect there to be an overlap of 32 bit blocks over the course of one million molecules. The final output of this step takes all fingerprints available and returns a dictionary of 32 bit block and the number of occurrences of each of the 32 bit blocks. The dictionary is kept sorted by the highest number of occurrences. These statistics can now be utilized in a later step to generate a statistics based molecular fingerprint.

## 9.6 Using Statistics to Create an Improved Fingerprint

Here, we take the resulting bits and create the reordering rules to create a new fingerprint definition that is of the same length. The first step of this process is to take any 32 bit block that has at least 10000 occurrences and count the number of occurrences at each bit index. Given the result, we can create a rule for reordering the 32 bit blocks. Once the rules have been determined, we loop through all of the unique 32 bit blocks of bits and create a new 32 bit block and fold it. Therefore, the new dictionary of 32 bit blocks contains the original 32 bit blocks along with the new 16 bit block. Finally, the third step is to go through all the original fingerprints again and replace 32 bit blocks with 16 blocks.

## 9.7 Clustering the Resulting Data

With similarity measures from each of the various types of fingerprints, we categorized the results into various “bins” to determine how well the folding one time, folding two times, and the improved fingerprint based fingerprint compared against the original fingerprint. Given that the similarity score is expected to be between 0 and 1, we place the scores into their respective bins, which are twenty

equally sized bins with a range of 0.05.For example, a similarity score of 0.78 falls into the bin labeled 0.75 to 0.80. The resulting numbers from the binning operation is then a benchmark to determine how well each new fingerprint method improved against the original fingerprint.

In addition to binning, the fingerprints that were returned from each operation were written out to a file for manual inspection. This is to manually verify the results from the code that runs the binning process.

We experimented with various numbers of equally sized bins to determine what sized bins results in the most similar information each fingerprint resembles the original fingerprint. We decided that twenty equally sized bins would be sufficient to show improvement for folded one time, folded two times, and the improved fingerprints.

### 9.8 Application Correctness

To insure basic correctness of the application, we also implemented various checks throughout the application. These checks were designed to verify at critical portions of the application to immediately detect unexpected input that may bias the results. Ultimately, the application would stop the binning process when an error was detected.

## 10. Results

The results we have obtained from over 1 million unique fingerprints are depicted in the chart below. At various points of the application, state information is gathered and aggregated. While reviewing the tables of results, it is critical to remember when the scientists review similarity matches, they will only be interested in molecules with the highest scores. The greater majority of the molecules will be disregarded.

### 10.1 Number of Unique 32 Bit Blocks

Before creating new fingerprints, we analyzed the number of unique 32 bit blocks that existed to determine how much overlap in bits existed. What we have observed is that as more original fingerprints were analyzed, the fewer unique 32 bit blocks were generated with each additional fingerprint and more occurrences of the 32 bit blocks fell into preexisting blocks. This shows us that as the number

of fingerprints to be preprocessed increases, we expect to see fewer unique 32 bit blocks, but increased overlapping 32 bit blocks.

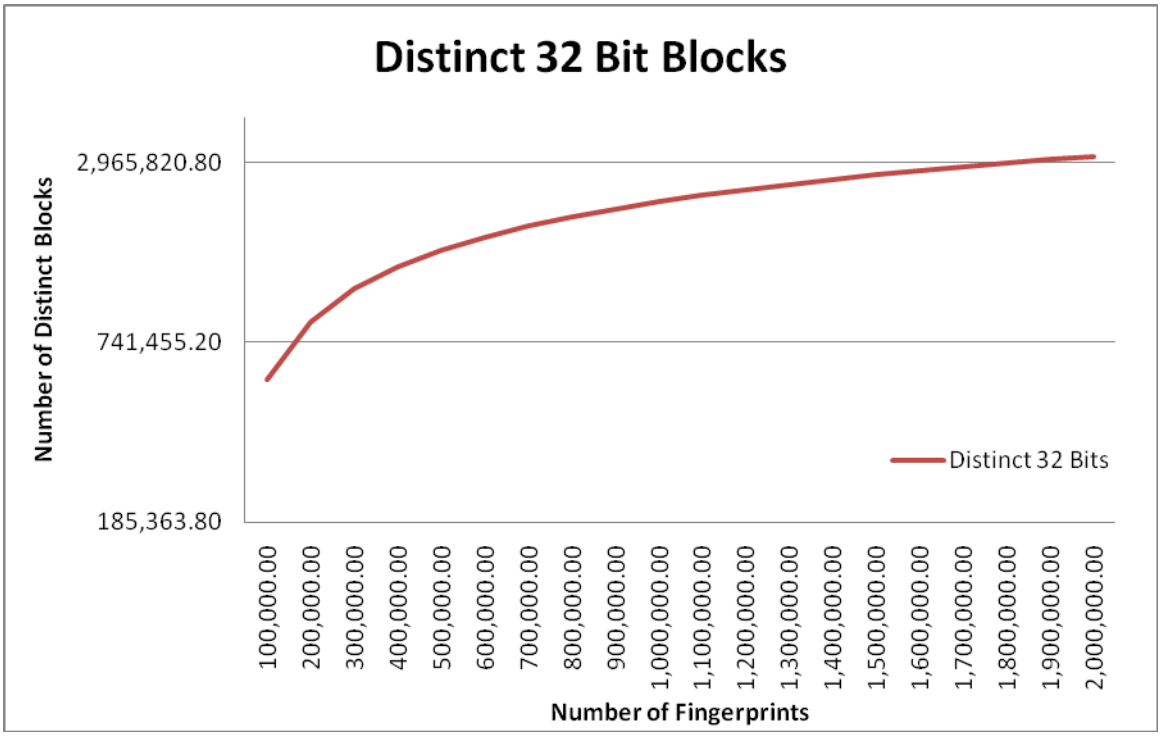


**Figure 9: Number of Unique 32 Bit Blocks. As the number of fingerprints increases, there are more fingerprints available.**

## 10.2 Original Fingerprint vs. Folded 1x

Our results from comparing the original fingerprint to that of a folded one time fingerprint is acceptable in the most important 0.95 to 1.00 bin range. However, as shown in the very next bin from 0.90 to 0.95 and 0.85 to 0.90, the number of false positive increases. Additionally, the resulting fingerprints from Original fingerprint clusters matched those found in the Folded 1x fingerprint clusters. This further backs up our conclusion that the folding algorithm was working as we expected.

| | Original Fingerprint | Folded 1x Fingerprint |
|---|---|---|
| **0.95 - 1.00** | 3 | 3 |
| **0.90 - 0.95** | 2 | 4 |
| **0.85 - 0.90** | 8 | 32 |
| **0.80 - 0.85** | 118 | 673 |
| **0.75 - 0.80** | 1,375 | 91,771 |
| **0.70 - 0.75** | 68,543 | 506,918 |
| **0.65 -0.70** | 302,509 | 67,931 |
| **0.60 - 0.65** | 399,924 | 16 |
| **0.55 - 0.60** | 191,158 | 5 |
| **0.50 - 0.55** | 19,981 | 0 |
| **0.45 - 0.50** | 5 | 0 |
| **0.40 - 0.45** | 0 | 0 |
| **0.35 - 0.40** | 0 | 0 |
| **0.30 - 0.35** | 0 | 0 |
| **0.25 - 0.30** | 0 | 0 |
| **0.20 - 0.25** | 0 | 0 |
| **0.15 - 0.20** | 0 | 0 |
| **0.10 - 0.15** | 0 | 0 |
| **0.05 - 0.10** | 0 | 0 |
| **0.00 - 0.05** | 0 | 0 |

**Table 1: Comparing results from the Original Fingerprint against the Folded 1x Fingerprint**

### 10.3 Original Fingerprint vs. Folded 2x

When compared to the Folded 2x with the original fingerprint, it is shown from the very start that the statistics are too poor to be of any use. In the very first bin, the three true similarity matches are grouped into a bin of 450 molecules. These results are very inconclusive and insufficient as there are too many compounds for a scientist to compare with its drug target. This leads to the time and cost

constraints that we aim to improve. Additionally, with the results gained from the Folded 2x, we found it would be unnecessary to Fold a third time as it would be impossible for the results to improve. In addition, in the case similar to that found in the Original Fingerprint vs Folded 1x fingerprint, the data here also shows the fingerprints are properly clustered.

| | Original Fingerprint | Folded 2x Fingerprint |
|---|---|---|
| **0.95 - 1.00** | 3 | 3 |
| **0.90 - 0.95** | 2 | 40,114 |
| **0.85 - 0.90** | 8 | 868,777 |
| **0.80 - 0.85** | 118 | 64,626 |
| **0.75 - 0.80** | 1,375 | 7,884 |
| **0.70 - 0.75** | 68,543 | 1,965 |
| **0.65 -0.70** | 302,509 | 256 |
| **0.60 - 0.65** | 399,924 | 1 |
| **0.55 - 0.60** | 191,158 | 2 |
| **0.50 - 0.55** | 19,981 | 0 |
| **0.45 - 0.50** | 5 | 0 |
| **0.40 - 0.45** | 0 | 0 |
| **0.35 - 0.40** | 0 | 0 |
| **0.30 - 0.35** | 0 | 0 |
| **0.25 - 0.30** | 0 | 0 |
| **0.20 - 0.25** | 0 | 0 |
| **0.15 - 0.20** | 0 | 0 |
| **0.10 - 0.15** | 0 | 0 |
| **0.05 - 0.10** | 0 | 0 |
| **0.00 - 0.05** | 0 | 0 |

**Table 2: Comparing results from the Original Fingerprint against the Folded 2x Fingerprint**

## 10.4 Improved Fingerprint vs. Original Fingerprint

Comparing the improved fingerprint against the original fingerprint, we find that the first two clustered bins match very closely with the original. In fact, it also incrementally improves on the matches in the 0.85 to 0.90 bin.

| | Original Fingerprint | Improved Fingerprint |
|---|---|---|
| **0.95 - 1.00** | 3 | 3 |
| **0.90 - 0.95** | 2 | 3 |
| **0.85 - 0.90** | 8 | 25 |
| **0.80 - 0.85** | 118 | 1,071 |
| **0.75 - 0.80** | 1,375 | 163,145 |
| **0.70 - 0.75** | 68,543 | 529,638 |
| **0.65 -0.70** | 302,509 | 281,122 |
| **0.60 - 0.65** | 399,924 | 8,617 |
| **0.55 - 0.60** | 191,158 | 2 |
| **0.50 - 0.55** | 19,981 | 0 |
| **0.45 - 0.50** | 5 | 0 |
| **0.40 - 0.45** | 0 | 0 |
| **0.35 - 0.40** | 0 | 0 |
| **0.30 - 0.35** | 0 | 0 |
| **0.25 - 0.30** | 0 | 0 |
| **0.20 - 0.25** | 0 | 0 |
| **0.15 - 0.20** | 0 | 0 |
| **0.10 - 0.15** | 0 | 0 |
| **0.05 - 0.10** | 0 | 0 |
| **0.00 - 0.05** | 0 | 0 |

**Table 3: Comparing results from the Original Fingerprint against the Improved Fingerprint**

### 10.5 Improved Fingerprint vs. Folded 1x Fingerprint

The comparison between the improved fingerprint and the Folding 1xfingerprint is the most interesting. What can be shown is that the data between the improved fingerprint and Folded 1x Fingerprint are similar. But when the improved fingerprint is specifically compared with the Folded 1x fingerprint in the 0.85 to 0.90 bin, the improved fingerprint is 20% more accurate than the Folded 1x

fingerprint. While we see less accuracy in a few of the lower bins, this is acceptable because these values are not weighted very heavily since the similarity score is too low. Scientists will not consider the compounds within bins lower than the 0.85 – 0.90 range as there are too many results to further analyze with the drug target. Therefore, bins of interest as stated above, we have found a 20.5% improvement in the number of returned molecules.

| | Folded 1x Fingerprint | Improved Fingerprint |
|---|---|---|
| **0.95 - 1.00** | 3 | 3 |
| **0.90 - 0.95** | 4 | 3 |
| **0.85 - 0.90** | 32 | 25 |
| **0.80 - 0.85** | 673 | 1,071 |
| **0.75 - 0.80** | 91,771 | 163,145 |
| **0.70 - 0.75** | 506,918 | 529,638 |
| **0.65 -0.70** | 67,931 | 281,122 |
| **0.60 - 0.65** | 16 | 8,617 |
| **0.55 - 0.60** | 5 | 2 |
| **0.50 - 0.55** | 0 | 0 |
| **0.45 - 0.50** | 0 | 0 |
| **0.40 - 0.45** | 0 | 0 |
| **0.35 - 0.40** | 0 | 0 |
| **0.30 - 0.35** | 0 | 0 |
| **0.25 - 0.30** | 0 | 0 |
| **0.20 - 0.25** | 0 | 0 |
| **0.15 - 0.20** | 0 | 0 |
| **0.10 - 0.15** | 0 | 0 |
| **0.05 - 0.10** | 0 | 0 |
| **0.00 - 0.05** | 0 | 0 |

**Table 4. Comparing results from the Improved Fingerprint against the Folded 1x Fingerprint**

## 10.6 Timing

The amount of time it took to calculate the similarity step very closely matched that of the single folded fingerprints. These results are accurate since the length

of these fingerprints is the same. There was time spent pre-processing, however once the rules were set, the time spent creating the fingerprints was not noticeablylonger. The Tanimoto Similarity measurement (and all other similarity measurements) run on a bit by bit manner, and therefore are expected to run in linear time. Therefore, any differences in the similarity calculation times are directly linked to the number of bits in the fingerprint.

Also, the implementation of the individual algorithms, system efficiency, and memory management was not a key aspect of this project.

| | Time (ms) | Fingerprint Length (number of bits) |
|---|---|---|
| **Original Fingerprint** | 31597 | 1024 |
| **Folded 1x Fingerprint** | 15533 | 512 |
| **Folded 2x Fingerprint** | 7164 | 256 |
| **Improved Fingerprint** | 15088 | 512 |

**Table X: Table of the similarity run times plus the resulting fingerprint length**

## 11. Future Work

There were some variables in the current project that were briefly tested with various lengths that may make for an interesting enhancement. By fine tuning the combinations of variables, it would be interesting to see if additional gains in result improvements can be achieved.

### 11.1 Further Enhanced Pre-processing Step

Another step that may further improve the algorithm is to improve the pre-processing section by adding into the algorithm of reordering a function to remove the very rarely used attributes of the fingerprint. By removing the least commonly used fingerprints, the number of similarity searches performed after the pre-processing step can be reduced and the fingerprint of each can be shorter. Additionally, by removing rarely used fingerprints, we would expect to see an improvement in the folding rate. A specific rule on determining when an attribute can be removed based on the overall pre-processed results would be critical. But more importantly, it would be interesting to find a rule that would minimize the differences in final results, but speed up the process.

### 11.2 Similarity Algorithm

Given that there are many other types of similarity algorithms, it is possible to try analyzing various combinationsof algorithms to see if other algorithms yield different results. If some algorithm is found to perform better given certain situations, given the pre-processing phase, there may be an opportunity to allow the pre-processing algorithm to determine at run time the most appropriate algorithm to apply.

## 12. Conclusion

We find that partitioning into a smaller equally sized block of bits, and then reordering bits for an optimal setting does aid in creating a better result. Although the results are still not optimal, it is an incremental on the folded 1x fingerprint. By dividing the fingerprint into 32 bits, we were able to isolate the blocks of bits that were most common and use those to build up our new fingerprint. The improved preprocessing methodthat reorganizes properties to best take advantage of a fold meets our goal of improving the similarity. Given that time spent on preprocessing, the amount of time spent calculating the similarity was the same. Additionally, the pre-processed fingerprintnot only kept the relevant fingerprints from the original fingerprint results, but also our result set contained fewer false positives than both the folded 1x and folded 2x fingerprints, which is critical to any folding process. In all, there was a 20% increase in the most relevant bins.

## 13. References

[1] Pharmaceutical Key Trends 2011 - Prescription Pharmaceutical Sales Overview http://www.datamonitor.com/store/product/pharmaceutical_key_trends_2011_prescription_pharmaceutical_sales_overview?productid=HC00062-005, retrieved 4/10/2011.

[2] Gatyas, Gary. IMS Health Forecasts Global Pharmaceutical Market Growth of 5-7 Percent in 2011, Reaching $880 Billion. http://www.imshealth.com/portal/site/imshealth/menuitem.a46c6d4df3db4b3d88f611019418c22a/?vgnextoid=119717f27128b210VgnVCM100000ed152ca2RCRD&vgnextchannel=41a67900b55a5110VgnVCM10000071812ca2RCRD&vgnextfmt=default, retrieved on 4/11/2011.

[3] Wikipedia. http://en.wikipedia.org/wiki/Pharmaceutical_industry, retrieved on 4/11/2011.

[4] Wikipedia. http://en.wikipedia.org/wiki/Drug_discovery, retrieved on 4/11/2011.

[5] Steven M. Paul, Daniel S. Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R. Lindborg& Aaron L. Schacht (2010). "How to improve R&D productivity: the pharmaceutical industry's grand challenge". Nature Reviews Drug Discovery9 (3): 203–214.

[6] Dimitris K. Agrafiotis,, Deepak Bandyopadhyay, and, Michael Farnum. Radial Clustergrams: Visualizing the Aggregate Properties of Hierarchical Clusters. Journal of Chemical Information and Modeling 2007 47 (1), 69-75.

[7] Molecular fingerprints, background. http://www.dalkescientific.com/writings/diary/archive/2008/06/26/fingerprint_background.html, retrieved 4/16/2011.

[8] Fara, Dan C., Oprea, Tudor I. Cheminformatics - Basics: Molecular Descriptors and Fingerprints.

http://biocomp.health.unm.edu/biomed505/Course/Cheminformatics/basic/descs_fingers/molec_descs_fingerprints.htm, retrieved 4/16/2011.

[9] PubChem. http://pubchem.ncbi.nlm.nih.gov/, retrieved 4/16/2011.

[10] ChemSpider. http://www.chemspider.com/, retrieved 4/16/2011.

[11] Willett, Peter. Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. 241-51.

[12] ChemDB: a public database of small molecules and related chemoinformatics resources. http://cdb.ics.uci.edu/, retrieved 4/16/2011.

[13] MaciejHaranczyk and, John Holliday. Comparison of Similarity Coefficients for Clustering and Compound Selection. Journal of Chemical Information and Modeling 2008 48 (3), 498-508.

[14] Willett, Peter. Similarity-based data mining in files of two-dimensional chemical structures using fingerprint measures of molecular resemblance. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2011. 241-51.

[15] Swamiidass, SJ., Baldi, P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. Journal of Chemical Information and Modeling. 2007 Mar-Apr; 47(2): 302-17.

[16] Darren R. Flower. On the Properties of Bit String-Based Measures of Chemical Similarity. Journal of Chemical Information and Computer Sciences 1998 38 (3), 379-386.

[17] Andreas Bender, Jeremy L. Jenkins, Josef Scheiber, SaiChetan K. Sukuru, Meir Glick, John W. Davies. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. Journal of Chemical Information and Modeling 2009 49 (1), 108-119.

[18] Andreas Steffen, Thierry Kogej, Christian Tyrchan, Ola Engkvist. Comparison of Molecular Fingerprint Methods on the Basis of Biological Profile Data. Journal of Chemical Information and Modeling 2009 49 (2), 338-347.

[19] Balid, P. and Hirschberg, D. An Intersection Inequality Sharper than the Tanimoto Triangle Inequality for Efficiently Searching Large Database. J. Chem. Inf. Model. 2009, 49, 1866-1870.

[20] Stackebrandt, Erko. Molecular identification, systematics, and population structure of prokaryotes. Birkhauser, 2006.

[21] Holliday J, Hu C, Willett P. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Combinatorial Chemistry & High Throughput Screening* [serial online]. March 2002;5(2):155. Available from: Academic Search Premier, Ipswich, MA.

[22] Haranczyk, M., Holliday, J. Comparison of Similarity Coefficients for Clustering and Compound Selection. American Chemical Society. 2008, 48, 498-508.

[23] Daylight. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html, retrieved 4/10/2011.

[24] Balid, P. and Hirschberg, D. An Intersection Inequality Sharper than the Tanimoto Triangle Inequality for Efficiently Searching Large Database. J. Chem. Inf. Model. 2009, 49, 1866-1870.

[25] Kharbutli, M., Solihin, Y., Lee, J. Eliminating Conflict Misses Using Prime Number-Based Cache Indexing. http://www.ece.ncsu.edu/arpers/Papers/pmod_ieee_tc05.pdf, retrievied 4/16/2011.

[26]Lo Ghemtio, Marie-Dominique Devignes, MalikaSmal-Tabbone, Michel Souchet, Vincent Leroux, Bernard Maigret. Comparison of Three Preprocessing Filters Efficiency in Virtual Screening: Identification of New Putative LXRβ Regulators As a Test Case. Journal of Chemical Information and Modeling 2010 50 (5), 701-715.

[27]Baldi, P., Benz, RW., Hirschberg, D.S., Swamidass, SJ. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. Journal of Chemistry Information and Modeling. 2007, 47(6); 2098-109.

[28]Pierre Baldi, Ramzi Nasr. When is Chemical Similarity Significant? The Statistical Distribution of Chemical Similarity Scores and Its Extreme Value. Journal of Chemical Information and Modeling 2010 50 (7), 1205-1222.

[29] ChemDB: a public database of small molecules and related chemoinformatics resources. http://cdb.ics.uci.edu/, retrieved 4/16/2011.