BAŞKENT UNIVERSITY

INSTITUTE OF SCIENCE AND ENGINEERING

ASSOCIATING MICRORNA WITH ITS CHEMOTHERAPY RESISTANCE

MURATCAN İĞDELİ

MASTER OF SCIENCE THESIS

2016

BAŞKENT UNIVERSITY

INSTITUTE OF SCIENCE AND ENGINEERING

ASSOCIATING MICRORNA WITH ITS CHEMOTHERAPY RESISTANCE

MİKRORNALARIN KEMOTERAPİ DİRENCİ İLE İLİŞKİLENDİRİLMESİ

MURATCAN İĞDELİ

Thesis Submitted
in Partial Fulfillment of the Requirements
For the Degree of Master of Science
in Department of Computer Engineering
at Baskent University

2016

This thesis, titled: “ASSOCIATING MICRORNA WITH ITS CHEMOTHERAPY RESISTANCE”, has been approved in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER ENGINEERING**, by our jury, on 04/08/2016.

Chairman (Supervisor) : Prof. Dr. Hasan OĞUL

Member : Yrd. Doç. Dr. Mehmet DİKMEN

Member : Yrd. Doç. Dr. Mehmet Serdar GÜZEL

APPROVAL

../08/2016

Prof. Dr. Emin AKATA

Institute of Science and Engineering

BAŞKENT ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 31 / 08 / 2016

Öğrencinin Adı, Soyadı : MURATCAN İĞDELİ

Öğrencinin Numarası : 21420225

Anabilim Dalı : BİLGİSAYAR MÜHENDİSLİĞİ ANA BİLİM DALI

Programı : BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSAN PROGRAMI

Danışmanın Adı, Soyadı : PROF. DR. HASAN OĞUL

Tez Başlığı : ASSOCIATING MICRORNA WITH ITS CHEMOTHERAPY RESISTANCE

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 51 sayfalık kısmına ilişkin, 30/08/2016 tarihinde şahsım tarafından "Turnitin" adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı %9'dur.

Uygulanan filtrelemeler:

1. Kaynakça hariç

2. Alıntılar hariç

3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esasları"nı inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası

Onay

31 / 08 / 2016

Öğrenci Danışmanı Unvan, Ad, Soyad,

Prof. Dr. Hasan OĞUL

THANKS

To Prof. Dr. Hasan OĞUL (Thesis Advisor) for his helps to finish the thesis and handle with the problems, and his guides...

To Atıf YILMAZ, currently a computer science department undergraduate student, for his helps and our common research and studies...

To the Scientific and Technological Research Council of Turkey (TUBITAK) for the project grant 113E527.

ABSTRACT

ASSOCIATING MICRORNA WITH ITS CHEMOTHERAPY RESISTANCE

Muratcan İĞDELİ
Baskent University Institute of Science and Engineering
Department of Computer Engineering

Genes are regulated by several factors including tiny molecules, called microRNAs. This regulation affects several processes in the cell. Recent findings suggest that microRNAs play important role in resistance to certain chemotherapies. The knowledge of what microRNAs are potentially resistant to given chemotherapies is therefore a crucial knowledge on drug design and therapy scheduling activities. In this thesis, we attempt to predict the list of microRNAs which are resistant to given drug using solely their mature sequence information. With this objective, we employ three common approaches for sequence classification in bioinformatics, i.e. pairwise, generative and discriminative models. The experimental results on a knowledge-driven dataset promote the use of pairwise models as a complementary tool in association studies for microRNAs and drugs.

KEYWORDS: microRNA, chemotherapy resistance, prediction

Advisor: Prof. Dr. Hasan OĞUL, Baskent University, Department of Computer Engineering.

ÖZ

MİKRORNALARIN KEMOTERAPİ DİRENCİ İLE İLİŞKİLENDİRİLMESİ

Muratcan İğdeli

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Ana Bilim Dalı

Genler mikroRNA olarak adlandırılan küçük moleküller gibi birçok faktör tarafından düzenlenir. Bu düzenleme hücrelerde birçok süreci etkiler. Güncel buluşlar mikroRNA'ların kemoterapiye karşı dirençte önemli bir rol oynadığını göstermektedir. mikroRNA'ların verilen kemoterapiye karşı potansiyel dirençleri hakkındaki bilgi ilaç dizaynı ve terapi ayarlanması üzerinde çok önemli bir bilgidir. Bu tez çalışmasında, verilen ilaca karşı direnç gösteren mikroRNA'ların listesi, yalnızca mikroRNA'ların olgun dizilerinin bilgileri kullanılarak tahmin edilmeye çalışılmıştır. Bu açıdan, 3 farklı ve biyo-enformatik alanında yaygın olarak kullanılan dizi sınıflandırma metotları kullanılmıştır. Bunlar ikili karşılaştırma yöntemi, yayımlayıcı yöntem ve ayırt edici yöntemdir. Bilgi odaklı bir dataset üzerinden elde edilen deneysel sonuçlar ikili karşılaştırma modelini, mikroRNA'lar ve ilaçları ilişkilendirmede tamamlayıcı bir araç olarak göstermiştir.

ANAHTAR SÖZCÜKLER: mikroRNA, kemoterapi direnci, tahmin etme

DANIŞMAN: Prof. Dr. Hasan Oğul, Başkent Üniversitesi, Bilgisayar Mühendisliği Bölümü.

CONTENT

Page

SYMBOLS AND ABBREVIATIONS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
miRNA	MicroRNA
TUBITAK	The Scientific and Technological Research Council of Turkey
PST	Probabilistic Suffix Tree
SVM	Support Vector Machines
LWL	Locally Weighted Learning
ROC	Receiver Operating Characteristics

LIST OF FIGURES

Page

LIST OF TABLES

Page

1 INTRODUCTION

1.1. Motivation and Problem Definition

With rapidly growing biological techniques to identify and map the genes of human-beings and extensively using biotechnology and bioinformatics in the field lead to building new algorithms, new techniques and new studies that will provide the ability of exploring and managing clinical consequences of molecular systems. As an example to these biomedical informatics activities, findings about microRNAs and their roles in resistance to chemotherapies allow us to think about scheduling chemicals to their corresponding users.

MicroRNAs do not encode any proteins but regulate gene expressions post-transcriptionally. It is also known that there are links between cancer and the microRNAs. MicroRNAs not only act in a cell-specific manner but also influence drug resistance in a drug-specific way. MicroRNAs have been found to interfere with specific molecular targets blocked by medications [1]. It is also known that chemotherapy is a challenging treatment because of using strong chemicals. There is also chemotherapy resistance situation encountered during the treatment process. Chemotherapeutic drug treatment transforms predominant, fast-dividing cells into drug-resistant ones. The chemotherapy drug resistance arises from two main parts that are inherited (natural) resistance and acquired resistance [1]. Inherited resistance can be partially overcome by incorporating multiple agents into chemotherapy regimens, while acquired resistance to chemotherapeutic drug accounts for greater than 90% of unsuccessful treatments in advanced cancer patients. There is also another perspective about chemotherapy, that the chemotherapy processes on different patients have different results. The effects of chemotherapy differ from one person to another person, and the level of resistance is another variable among patient [2, 3]. The idea to schedule or manage the chemotherapy resistance level among patients can find an optimal way for a specific patient in his chemotherapy treatment process that decreases the negative impacts on chemicals on specific patient. For example, by using chemicals that are fitting to patient’s genomic structure can prevent losing weight fast in the treatment process. A specific patient treatment model that uses chemicals that are

fitting to the patient's genomic structure can be beneficial in the chemotherapy process. At that point, using the experimentally validated chemotherapy resistant microRNA sequences, we attempted to predict the list of new microRNA sequences that are related to resistance of chemotherapy chemicals. The relation between microRNA sequences, solely mature sequence information of the microRNA sequences, and the drugs used in the treatment may result in the form that the relation between some pairs whose microRNA's mature sequences include them with the 'microRNA-x' can cause the high resistance level to the 'drug-x' and trying another drug can accelerate the velocity of the chemotherapy treatment.

In this thesis, in order to predict the list of microRNA sequences that are resistant to given drug, we employ three main techniques which are frequently used in sequence classification problems in bioinformatics; i.e. pairwise, discriminative, and generative models. In the first approach, the aim basically is to find some similarities in a structural way in sequences to see the relation between the microRNA sequences and their resistance to given drugs. In the second approach, we closely look at sequential features of microRNAs, such as *k-mer* frequencies, and try to make discrimination between resistant and non-resistant microRNAs to a given drug by training support vector machines (SVM) and locally weighted learning (LWL) algorithms fed by those extracted features. In the third approach, i.e. the generative model, a variable-length Markov chains for sequences that are known to be resistant to a certain drug are built and new microRNA sequence is predicted by its probability of being generated from that Markov model. All the approaches are evaluated based on empirical studies on knowledge-driven dataset of chemicals and microRNAs.

1.2. Biological Background

1.2.1. RNA – Ribonucleic Acid

Ribonucleic acids are biologic macromolecules composed of four types of pyrimidine that are adenine, guanine, cytosine and uracil. The ribonucleic acid molecules are in the linear long chain molecule form. The nitrogenous base, ribose sugar and phosphate are its main

units of matter. The main difference between DNA standing for deoxyribonucleic acid and the RNA standing for ribonucleic acid is that RNA contains ribose sugar, while DNA contains deoxyribose sugar.

DNA is a nucleic acid which includes genetic knowledge or the instructions about functions and developments of all living organisms, and it is also called the ‘blueprint’ of the cells. Indeed, DNA contains the genetic information about cell growing, nutrition taking and so on. The other difference between DNA and RNA is that DNA includes base thymine but RNA includes uracil in place of thymine.

RNA is mainly responsible for the protein synthesis, the process is also called ‘central dogma’ that can be described as RNA creation from DNA and protein creation or synthesis from RNA. RNA’s main job in the process is transferring the genetic code from nucleus to the ribosome that is responsible for protein synthesis in a cell that is needed for the protein synthesis. With this process, DNA stands in the nucleus and its genetic information can be transferred to the ribosome to start the protein synthesis. It can also be said that with this aspect, DNA and its genetic information is protected in a safe way. Therefore, RNA has a crucial responsibility in protein synthesis and proteins cannot be produced without RNA.

RNA is produced or formed from DNA by a process named ‘transcription’ and there are 3 significant types of RNA that are named mRNA standing for Messenger ribonucleic acid, tRNA standing for transfer ribonucleic acid, and rRNA standing for ribosomal ribonucleic acid.

mRNA is responsible for carrying the DNA message to the cytoplasm that is the internal part of a cell. Protein is producing from amino acid sequences that is accordingly specified by the mRNA in the cytoplasm.

tRNA is responsible for linking the mRNA and the amino acid sequences. tRNA carries amino acids to the ribosome, tRNA is also the necessary part of translation process that is the protein creation process in the ribosomes. There are 20 types of tRNA as same quantity as the amino acid types.

rRNA is the RNA part that is in the ribosomes and responsible for the protein creation in the ribosome.

There is also miRNA stands for microRNA that are non-coding RNAs that means RNAs that are not responsible for protein synthesis but acting a crucial role in regulations of gene expressions.

1.2.2. MicroRNA

MicroRNAs are small non-coding RNA molecules which are not translated into proteins. MicroRNAs function in controlling of gene expressions at the RNA level, this controlling mechanism is also called ‘post transcriptional regulation’ or the post-transcriptional controlling. “MicroRNAs (miRNAs) are composed of 22-nucleotide, short, noncoding RNAs that are thought to regulate gene expression through sequence-specific base pairing with target mRNAs” [4].

Currently, more than 600 human microRNAs are identified [5]. “Accumulating evidence has linked the deregulated expression patterns of miRNAs to a variety of diseases, such as cancer, neurodegenerative diseases, cardiovascular diseases and viral infections” [5]. Substantially, miRNAs regulate or modulate other genes via binding to their complement sequences in the target gene.

Although more than six hundreds of microRNAs are identified in human, there are not currently many specific targets that are validated. There is a correlation between a number of clinically crucial diseases such as virus infection, Alzheimer’s diseases cancers, metabolic diseases and many others with the miRNA inadequacy or deficiencies or the overflow or excesses [6] [7] [8] [9] [10]. Regarding to many studies, it is also known that there is also a relationship between miRNAs and the primary human tumor [11].

There are some characteristics such as ability to grow in an increasing rate, losses in cellular identity and changes in cell death systems that all cancer types have. Many studies and researches demonstrate that miRNAs are capable of regulating or modulating

these kind of cellular processes, which also implies that there are relationships between cancer and miRNAs.

There are also relationships between heart diseases and the miRNAs. MiRNAs are increasingly recognized as crucial regulators of some of heart functions. In some heart disease situations, many miRNA expressions are changed and some different type of heart diseases are associated with distinguishable changes in miRNA expressions [5]. They all demonstrate that there are relationships between miRNAs and human diseases such as cancer, human tumor and heart diseases.

The mature microRNA sequence is highly responsible in determining the linking or binding capability between the microRNA and its specific, corresponding mRNAs. Therefore, identification of microRNA and its specific, corresponding mRNA targets is related to the mature miRNA sequences.

1.3. Contribution of the Thesis

In this thesis, a comprehensive evaluation of three different sequence classification approaches is represented for the microRNA classification task. These approaches are pairwise model, generative model and discriminative model that are correspondingly comparing one input microRNA sequence with the other microRNAs in the resistant set of the drug under consideration, taking the set of microRNAs that are resistant to the corresponding to chemotherapy drug, and calculating a probabilistic model of generating those microRNA sequences in the same resistance characteristics, and using two sets of microRNAs that are resistant to its corresponding chemotherapy chemical and non-resistant to its corresponding chemotherapy chemical.

The results obtained may shed light on understanding the biological mechanism behind the occurrences of drug resistance by microRNA molecules. The statistics from empirical results obtained from three common approaches for sequence classification in bioinformatics may be used to comment on association between microRNAs and chemotherapy resistance.

To our knowledge, this research is the first study that evaluates the predictability of specific chemotherapy resistance of microRNA by using only the sequence information. The results obtained from the study may be the source for the new studies in many specific areas in the bioinformatics field.

1.4. Organization of the Thesis

The thesis is organized in a manner that it starts with the general overview of the methodologies and the scientific view about three approaches. Then, the approaches are described in the order that pairwise approach is described first, then the generative approach is described, the discriminative approach is the last approach that is going to be described.

In the pairwise approach part of the thesis, the information about the sequence alignment is going to be given and the pairwise chemotherapy resistance prediction and its methodology is described.

In the generative approach part of the thesis, probabilistic suffix trees, its building mechanism and the generative chemotherapy resistance prediction and its methodology are described.

In the discriminative approach part of the thesis, feature extraction, support vector machines (SVM), locally weighted learning (LWL) are going to be described and the discriminative chemotherapy resistance prediction is derived .

Last two sections are the results and the conclusion parts. Results section contains the information about the datasets, experimental setup and the evaluation and the empirical results of three approaches and the comparisons between results. The conclusion includes the comments on the results and some predictions about the future studies about the topic.

2. METHODS

2.1. General Overview

We consider the problem of predicting chemotherapy resistance of a microRNA to a certain drug as a binary classification task for each distinct chemotherapy chemical. The task is to classify an unknown microRNA sequence as resistant or not-resistant to given drug. Therefore, we first desire a model learnt from a set of known microRNAs which has been already shown experimentally to be resistant to the drug in question. Then, the unknown sequence is assigned to a resistant or non-resistant class through that learnt model. We study the problem in three different ways. In pairwise approach, we consider each microRNA in a resistant set separately and evaluate its similarity with input sequence in the hope of getting similarity in their resistance behaviors. In generative approach, we train a probabilistic model through only positive set of resistant microRNAs. In discriminative approach, on the other hand, we consider both positive and negative (i.e. no-resistant) sets when learning a separating decision function.

2.2. Pairwise Method

2.2.1. Sequence alignment

Sequence alignment for DNA, RNA, or protein sequence is arranging to determine the structural, similar or evolutionary relationships between the sequences [12]. Sequence alignment is a powerful technique in bioinformatics, but sequence analysis is also used in linguistics to determine the edit distance cost between words or the strings. Sequence alignment can actually be performed by hand in small sequences or similar sequences. However, sequences that have such a big number of characters, high numbers of variables cannot be aligned by hand. At that moment, the computational power is used to align the sequences. Computational approaches to the sequence alignment process are grouped into two main classes that are ‘global alignments’ and ‘local alignment’. ‘Global alignment’ is kind of a ‘global optimization’ method that performs on the whole sequence. ‘Local alignment’ is about identifying the parts or the regions of the similarities between

sequences. Sequence alignment methods and algorithms are generally based on the dynamic programming methodology.

Sequence alignment is mainly divided into 3 groups that are pairwise alignment, multiple sequence alignment, and structural alignment. Pairwise sequence alignment is mainly trying to find the best local alignment score between two sequences.

The pairwise sequence alignment using is related to finding the best matching local alignment score between two sequences. ‘Dot-matrix’ method, ‘dynamic programming’ and ‘word methods’ can be applied to perform pairwise sequence alignment.

Multiple sequence alignment is kind of an extended version of pairwise sequence alignment that gives opportunity to align more than two sequences uncooperatively at a specific time. Indeed, multiple sequence alignment is responsible for aligning all of the sequences in a given training set. Multiple sequence alignment is considered as one of the computationally important and popular problem in the computational biology field [13]. Multiple sequence alignments are considered that they are computationally hard to solve and formulate them, also most of them are considered to be in the NP-complete problems [14] [15]. ‘Dynamic programming’, progressive methods, iterative methods, and motif finding techniques are used in multiple sequence alignment process.

Structural alignment is an alignment method that uses the secondary or the tertiary structure of the RNA molecules or the protein molecules to align the sequences structurally. Structural alignment is generally used for local alignment and applied on two or more sequences, but the restriction for the structural alignment is that the structural information about the sequence or sequences are needed to be known. The structural alignment is powerful than the other alignment models, because it also has power to align sequences structurally. The ‘DALI’ method, distance-matrix alignment, ‘SSAP’, sequential structure alignment program, methods are used to perform structural alignment.

2.2.1.1. Pairwise Alignment

Pairwise alignment in the early stages was used for finding the differences among amino acid sequences between different species. Because the number of amino acid types are known, that is 20, and the number of pyrimidine are also know, that is 4, the pairwise alignment was used to find differences among them.

However, in recent years pairwise alignment is used to find the best-matching local alignment score between two sequences. On the two sequences of symbols given on an alphabet, the pairwise alignment of subsequences can be constructed by adding spaces to the subsequences to make them in the order of same length. After making the two subsequences having the same length, these modified subsequences can be considered as two rows with the same length. Then, columns that have no space is considered as matches or substitutions.

The sequences are called ‘homologous’, when the sequences share common ancestor, when any two of the sequences have similarities that are measurable by the aligning them, homology, in this sense, implies that the similarity coming from the alignment is kind of a specific consequence of sharing common ancestor [16]. The sequence alignment also provides information or evidence that the sequences can be diverged from a common ancestor by some evolutionary processes [17].

As mentioned above, alignments are based on two main classes that are global alignment and the local alignment. The first global alignment method or the algorithm was proposed by Needleman and Wunsch in 1970 [18]. After proposing the global alignment algorithm, Smith and Waterman proposed a new alignment algorithm that was the first local algorithm that was used dynamic programming as a variation of the algorithm of Needleman and Wunsch [19]. The main difference between the global alignment algorithm of Needleman and Wunsch and the local alignment algorithm of Smith and Waterman is mainly that the alignment process can end everywhere in the matrix of alignment, and the matrix refers to a two dimensional array to define the similarities and the differences between the sequences [20].

Given two sequences S and T over the alphabet ∑, their alignment score is calculated as follows:

$$H(a\,,b) = max\begin{cases} 0, \\ max_{l\geq 1}\{H\,(a, b-l) + W_l\}\,, Insertion \\ max_{k\geq 1}\{H\,(a-k, b) + W_k\}\,, Deletion \\ H(a-1, b-1) + s(S_a\;\;, T_b), \quad Match\,, Mismatch \end{cases}$$

$$\text{such that } 1 \leq \text{a} \leq \text{m}\;\;, 1 \leq \text{b} \leq \text{n}$$

$$\text{where}$$

S and T are strings over the alphabet ∑

M = length (S) and n = length (T)

S(S, T) is the similarity function on the alphabet

H (a, b) is the maximum similarity score between a suffix of S [1....i] and suffix of T [1...j]

W_k is the gap scoring scheme

Needleman and Wunsch pairwise alignment calculation method is shown at Table-1.

Table 1 Needleman and Wunsch pairwise alignment

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5

A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Smith and Waterman pairwise alignment calculation method is shown at Table-2.

Table 2 Smith and Waterman pairwise alignment

		G	C	A	C	G	C	T	G
	0	0	0	0	0	0	0	0	0
G	0	2	1	0	0	2	1	0	2
A	0	1	0	2	1	0	0	0	0
C	0	0	2	1	3	2	4	3	2

G	0	2	1	0	2	4	3	2	4
C	0	1	3	2	4	3	5	4	3
G	0	3	2	1	3	5	4	3	5
C	0	2	4	3	5	4	6	5	4
G	0	4	3	2	1	6	5	4	6

2.2.1.2. Multiple Sequence Alignment

“Multiple sequence alignment involves global alignment of more than two sequences. Given that pairwise alignment tries to find the best path in a matrix, multiple sequence alignment can be conceived as a multidimensional problem” [20]. A solution to that problem has some computation time and space complexities [21]. The ‘progressive alignment’ algorithm is used for the multiple sequence alignment due to its performance factors, referred by Feng and Doolittle [22].

There are several multiple alignment software such as ClustalW [23], ClustalX [24], TCoffee [25].

2.2.2. Pairwise chemotherapy resistance prediction

The pairwise approach takes an input microRNA and compares it with the other microRNAs in the resistant set of the drug under consideration. The comparison is made through the alignment of their mature sequences. Then, the model uses the annotations of the microRNA that has the highest alignment score to predict the resistance level of the input microRNA (Figure 1). Since it is already known that the regulatory behaviors of

microRNAs are guided by their mature sequences [16], we may reasonably expect a similarity between matures sequences of two microRNAs that give resistance to same chemotherapy drug.

Pairwise alignment in biological sequence analysis is basically used for finding sequence similarities in sequences in such a way that high sequence similarity implies crucial structural and functional similarity between sequences. Sequence analysis is simply a procedure of comparing two or more sequences by looking for a series of individual characters or patterns combined from characters that have the same order in sequences. Pairwise alignment of sequences takes two sequences and compares individual characters or patterns between them to find some similarities.

The aim in this study to use sequence analysis is to predict the relations between chemotherapy resistance and microRNA sequences. To solve that problem, it is needed to find the ‘best’ alignment between two sequences that are the mature sequences of the microRNA with known resistance and the microRNA that will be tested. To do that, we need a method for scoring alignments and an algorithm to find the alignment that has the ‘best’ score. The alignment score is found by dynamic programming algorithm using matrix and gap penalties [26]. Dynamic programming is also an efficient way of finding an optimal alignment score for relatively short sequences. Since microRNA mature sequences are of 22-25 nucleotides length, in this context, using a dynamic programming technique is a reasonable choice. Alignment by using dynamic programming ensures that the resulting alignment is optimal while it provides a cost effective solution [26].

To predict chemotherapy resistance of an input microRNA, we first perform a dynamic programming based local alignment between the input and other microRNAs in the training set. Then, most similar microRNA is selected to annotate the input such that their chemotherapy resistance is same. In other words, if the most similar microRNA is resistant, the query is also predicted as ‘resistant’.

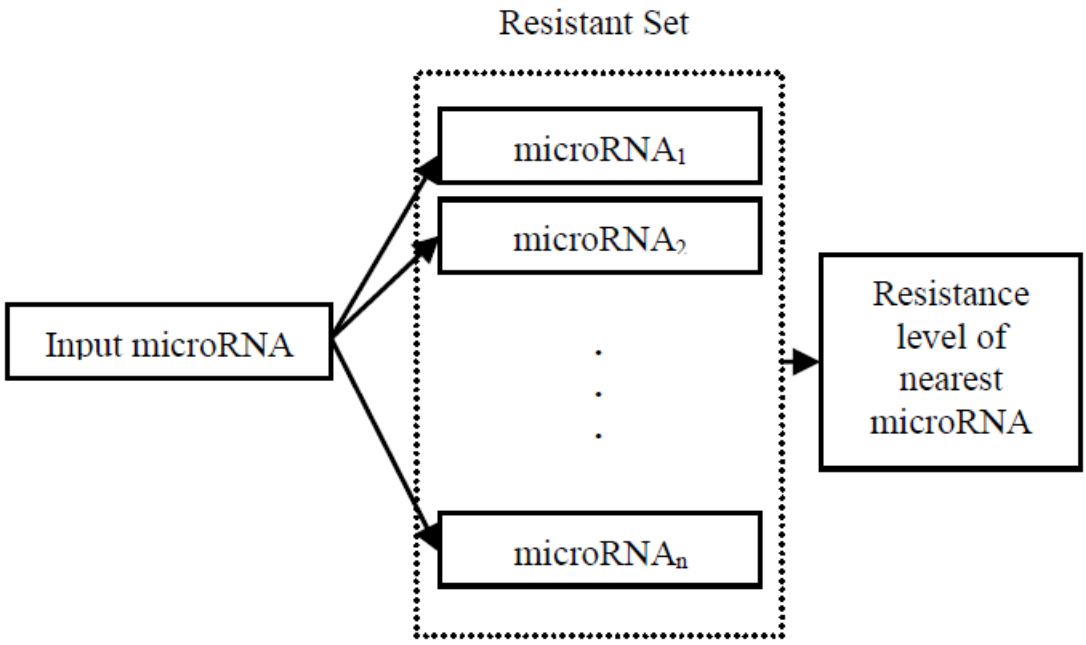

Figure 1 Pairwise model for predicting microRNA resistance to a certain chemotherapy drug

2.3. Generative Method

2.3.1. VLMC - Variable Length Markov Chains

Sequence classification and clustering is complex, important and has significance in bioinformatics field. Attaining a training set of sequences, and fitting them with a generative probabilistic model that captivates statistical correlativity from the sequences in the set [27]. Variable length Markov chains gives us a capability to model a selected set of sequences or contexts that can have different lengths [27]. Variable length Markov chains (VLMM) are also powerfully capable of optimizing memory length locally within a model, and that special quality of VLMC's ability capture the long term dependencies of some parts of the sequences and short term dependencies of the sequences [28].

Although n-th order Markov models are used to model the memory of fixed lengths, VLMMs are used for modeling the processes that has the varying memory length. That feature helps us about optimizing the length of memory that is needed locally.

Variable length Markov chains are sort of Markov chains that also have a structure that memories of chains depend on a variable number of lagged or delayed values [29].

For an input mature sequence *S*, indexed between 1 and *N* its likelihood for a specific VLMC is given by;

$$P(S_1^N) = \prod_{j=1}^{N} P(S_j = s_j \mid S_{j-L_j}^{j-1} = s_{j-L_j}^{j-1}) \quad (1)$$

Where L_j is the optimal length of preceding subsequence and $s_{j-L_j}^{j-1}$ is that subsequence [30].

2.3.2. PST – Probabilistic Suffix Trees

A probabilistic suffix tree is an efficient data structure to implement VLMC. The probabilistic suffix tree was proposed firstly by 'Ron', 'Singer' and 'Thishby' in 1996 as 'probabilistic suffix automata' [31]. The aim to propose this method was creating a new learning method. This method is used by many fields such as pattern recognition, machine learning. Its first usage in bioinformatics is classifying the protein families [32]. Probabilistic suffix trees had several variations to be used for protein sequencing. Biological probabilistic suffix tree approach was proposed by 'Bejenaro' and 'Yona' [33] which stands for a type of several probabilistic suffix models.

Probabilistic suffix tree is an index-based suffix tree storing the probabilistic values regarding to the subsequences and using probabilistic models [33]. It can be said that this method is based on the characteristic named 'short memory' that is common in biological sequences. Before probabilistic suffix tree is proposed, Markov chains and Hidden Markov models were used for modeling the sequences. However, both Markov chain and Hidden Markov model have some restrictions in their practical usage. Markov chain grows exponentially regarding of the length, because of that the less length Markov chains can work efficiently. In Hidden Markov model, the learning difficulty problem occurs.

Therefore, there was a need for a new probabilistic learning model to get rid of the restrictions. Although probabilistic suffix tree has some restrictions like learning difficulty, it can be used efficiently on more length sequences. Probabilistic suffix tree was used firstly in 2000 by ‘Bejenaro’ and ‘Yona’ for classification of protein sequences.

“A PST over an alphabet is a non-empty tree, whose nodes vary in degree between zero (for leaves) and the size of the alphabet” [34]. The tree is composed from nodes and edges, and each is marked by a single symbol or the character from the alphabet that the probabilistic suffix tree is applied on and any symbol or character cannot be used more than one in the edges, which ensures that the degree of nodes are same and the bounded by the alphabet that the probabilistic suffix tree applied on. Since edges are marked or labeled by a character or a symbol, nodes are marked by strings that can be generated from that node to the root. Nodes also contain probability distribution vectors regarding to the alphabet. These probability distribution vectors act a significant role in predicting important subsequences or the patterns in strings.

The difference between the suffix trees and the probabilistic suffix trees is the structuring direction.

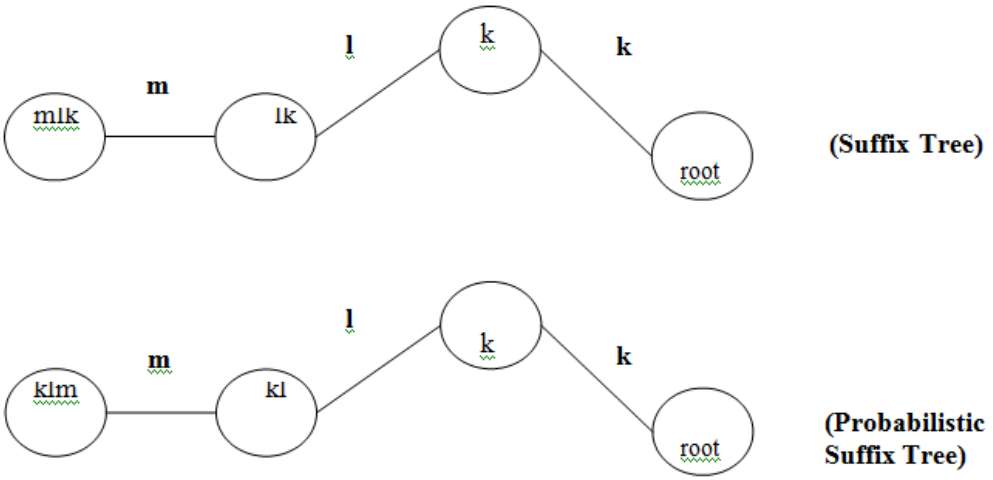

Figure 2 Examples of a suffix tree and a probabilistic suffix tree

Take the above simulation as an example, in the suffix tree model when the symbol ‘k’ is added to the root, the first node is demonstrated as ‘k’, then after adding the symbol ‘l’, the second node is demonstrated as ‘lk’. However in the probabilistic suffix tree model, when the symbol or the character ‘k’ is added to the root, the first node is demonstrated as ‘k’, then after adding the symbol or the character ‘l’, the second node is demonstrated as ‘kl’. The difference between suffix trees and the probabilistic suffix trees is about structuring direction.

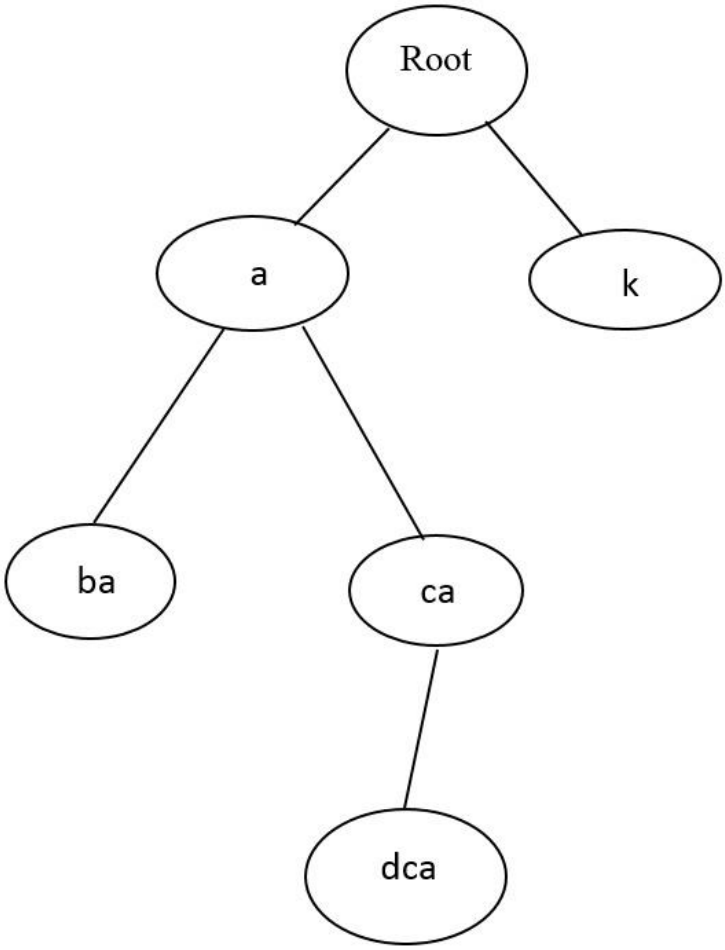

Figure 3 An example of Probabilistic Suffix Tree with length = 3

An example of a probabilistic suffix tree is given in the Figure 3. This example is constructed on the alphabet = {a, b, c, d, k}. The probability distribution vectors of nodes are given that (.2 , .2 , .2 , .2 , .2) is the probability distribution vector of the root node, (.05 , .5 , .15 , .2 , .1) is the probability distribution vector of the node that has ‘a’ in it, (.6 , .1 , .1 , .1 , .1) is the probability distribution vector of the node that has ‘k’ in it, (.05 , .4 , .05 , .4 , .1) is the probability distribution vector of the node that has ‘ba’ in it, (.05 , .25 , .4 , .25 , .05) is the probability distribution vector of the node that has ‘ca’ in it, and (.1 , .1 , .35 , .35 , .1) is the probability distribution vector of the node that has ‘dca’ in it. The probabilistic distribution vector of the node demonstrates the probability distribution over the next symbol. For instance, the probability distribution of the node that contains ‘ca’ as a substring is 0.05 for the symbol ‘a’, 0.25 for the symbol ‘b’, 0.4 for the symbol ‘c’, 0.25 for the symbol ‘d’ and 0.05 for the symbol ‘k’ correspondingly.

$\sum$ is the symbol that demonstrates the alphabet to construct a probabilistic suffix tree. The alphabet can be amino acids that have 20 different types of amino acids, or the nucleotides of DNA or RNA that are adenine, guanine, cytosine, thymine or adenine, guanine, cytosine, uracil correspondingly.

$$r_1, r_2, r_3 \dots\dots\dots r_n \; are \; the \; sample \; sets \; of \; n \; strings \; over \; the \; alphabet$$

$$i \; is \; the \; in \; the \; range \; of \; [1 \, , m] \; and$$

$$r^i = r_1^i r_2^i r_3^i \dots\dots\dots r_m^i$$

Then, the empirical probability is defined over a subsequence ‘s’. The empirical probability is defined in a way that the number that the subsequence is observed in the sample set is divided by the maximum number of occurrences of a pattern with the same size. Let ‘l’ defines the length of the subsequence ‘s’;

$$s = s_1 s_2 s_3 \dots\dots\dots s_l$$

$$X_s^{i,j} = \begin{cases} 1\ , if\ s_1 s_2 s_3 \dots\dots\dots s_l = \ r_j^i r_{j+1}^i r_{j+2}^i \dots\dots\dots r_{j+l-1}^i \\ 0\ ,\ otherwise \end{cases}$$

The actual empirical probability is related to the number of occurrences of “s” in the sample set of sequences. After defining the empirical probability, conditional empirical probability of a symbol that is the next symbol or right after the sequence. It is also defined in a way that the number of occurrences that the desired symbol is next to the given sequence or right after the given sequence divided by the total number of occurrences in the sequence.

2.3.2.1. PST Building

The length of the PST, memory length of the PST, is defined as the starting point in the PST construction process. The length can be the maximum length of a string that will be shown in the tree. Given an empirical probability threshold, sequences of length 1 through the length of the sequence is constructed in a probabilistic suffix tree up to reaching the empirical probability threshold or up to reaching to the maximal length boundary to the sequence. Empirical probability threshold is demonstrated as P_{min} and it is responsible for avoiding the exponential growing. As explained above in the probabilistic suffix tree part, Markov chains and Hidden Markov models have disadvantages about exponentially growing and the probabilistic suffix tree solves the problem by using the empirical probability threshold demonstrated as P_{min}.

Constructing a probabilistic suffix tree starts with defining a root node. Then for every subsequences of the sequence, the empirical probability of observing the symbol that is next to the sequence or right after to the sequence is checked, and if the checked empirical probability is not negligible and if it is also importantly have different value from the empirical probability value of observing that symbol next to the sequence or right after to the sequence that is derived from removing the leftmost character of the subsequence, the subsequence is added to the PST [35]. The steps processed on the subsequences to be added to the PST is called pruning. In addition to the adding process, if the prediction function of a leaf node is identical or similar to its parent node’s prediction

function, then the leaf node is considered as useless. This also prevents the having almost identical or similar nodes in a PST.

The PST constructing process on a sample set includes five parameters. They are memory length as 'L', empirical probability threshold or the minimal probability as 'P_{min}', the difference measured between the parent node and the current node as 'r', the smoothing factor as γ_{min}, and the α parameter that is used together with the smoothing factor to define the threshold probability. Smoothing process assures that symbols will not have zero probabilities. γ_{min} defines the minimum probability value of a symbol or the character, and also the empirical probability values need to satisfy the smoothing factor.

2.3.2.2. Prediction with PST

For a given string called 's', the prediction process of it using a probabilistic suffix tree is done character by character or symbol by symbol. It is done by "calculating the probability of each character or symbol by scanning the tree in search of the longest suffix that appears in the tree and ends just before that letter" [35]. Then, the conditional probability of that character or the symbol is given by the probability distribution associated with the corresponding node in the PST [35].

2.3.2.3. The Complexity of PST

"Denote the length of the training set by n, the depth bound on the resulting PST by L, and the length of a generic query sequence by m. In these terms, the learning phase of the algorithm can be bounded by O (Ln^2) time and O (Ln) space, as there may be O (Ln) different subsequences of lengths 1 through L, each of which can be searched for, in principle, in time proportional to the training set length, while each contributes but a single node beyond its father node to the resulting structure" [35]. Given the length of the training set is 'n', and the depth of the PST is 'L', and the sequence length is 'm', the complexity of training and the prediction periods are given in the Table-3.

Table 3 Time and Space Complexity of the PST processes

	Time Complexity	Space Complexity
Training Period	$O(Ln^2)$	$O(Ln)$
Prediction Period	$O(Lm)$	$O(Ln^2)$
Total (Training and Prediction)	$O(Ln^2 + Lm)$	$O(Ln^2 + Ln)$

2.3.2.4. Generative Chemotherapy Resistance Prediction

The generative approach that we consider here to predict chemotherapy resistance takes the set that has the microRNAs that are resistant to the corresponding to chemotherapy drug, and calculates a probabilistic model of generating those microRNA sequences in the same resistance characteristics. The model predicts the resistance level for an unknown input microRNA based on its probability of being generated from that model (Figure 4). The intuition behind this scheme is to infer the main sequential characteristics that may lead to resistance of microRNAs to given chemotherapy drug.

Owing to the sequential nature of our data, we adopt here a Markov model to build a probabilistic generative model for each chemotherapy drug. Because of the computational and storage limitations of Markov Chain of order L and the need for an excessive number of learning parameters for Hidden Markov Models, we use Variable Length Markov Chain in this context. Because its storage cost is less and performance is better [36], PSTs are used in this approach to implement VLMC for predicting the relations between chemotherapy resistance and microRNA sequences. The method is based on finding and identifying a set of significant patterns in microRNA sequences corresponding to chemotherapy chemicals. VLMC enables us to calculate the likelihood of whole

sequence for corresponding drug by simply multiplying local probabilities [30]. We simply compare the likelihood of an input microRNA being resistant or not to identify its class.

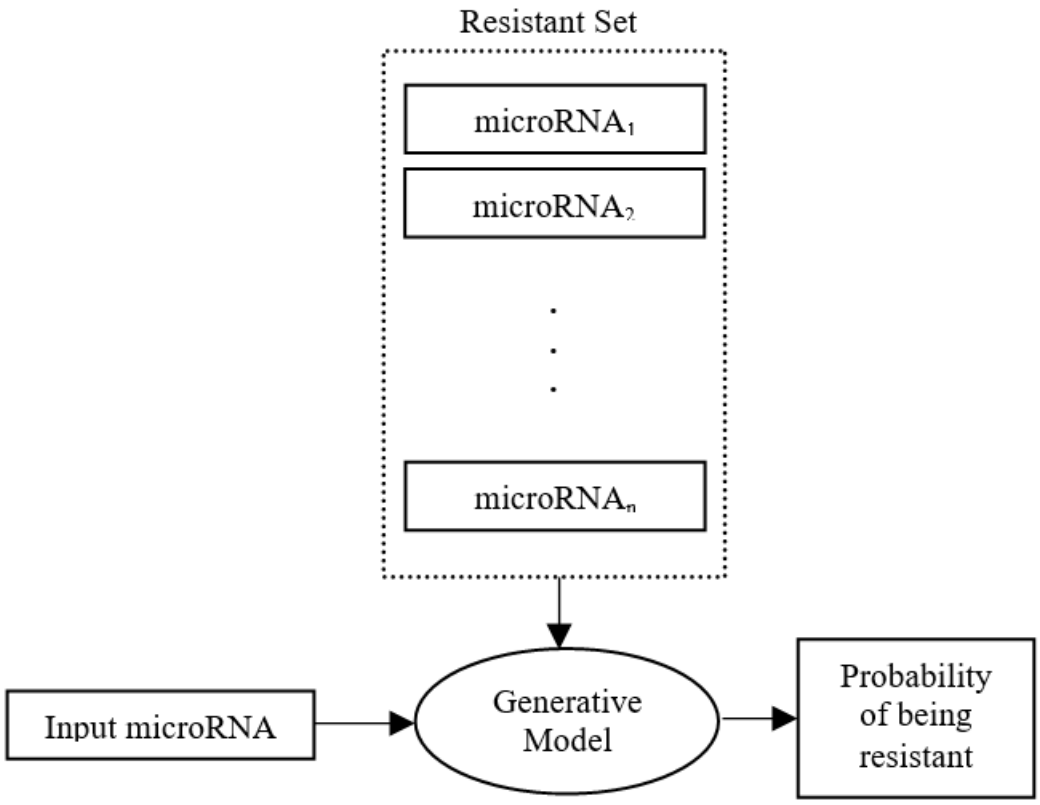

Figure 4 Generative model for predicting microRNA resistance to a certain chemotherapy drug.

2.4. Discriminative Method

2.4.1. Feature extraction

Feature extraction is the technique that is used for diminishing the quantity of resources that are required to be set of data by analyzing the data and clustering and grouping them according to some characteristic specialties. Feature extraction is populously used in the machine learning, image analysis, pattern recognition, data mining and business intelligence areas.

Feature extraction basics are described by Guyon and Elisseeff as predictive modelling, feature construction and feature selection [37]. The machine learning process is about encountering relationships between the patterns among the training examples, the predictive modelling is about the categorization or the determining the classification technique [37]. Feature construction is also described as preprocessing that includes standardization, normalization, signal enhancement, extraction of local features, linear and non-linear space embedding methods, non-linear expansions and feature discretization methods according to Guyon and Elisseeff [37]. Feature construction is also considered as one of the keys steps in the data analyzing process [37].

Feature selection, also variable selection, attribute selection, is basically selecting subsets from the pertinent big set of data.

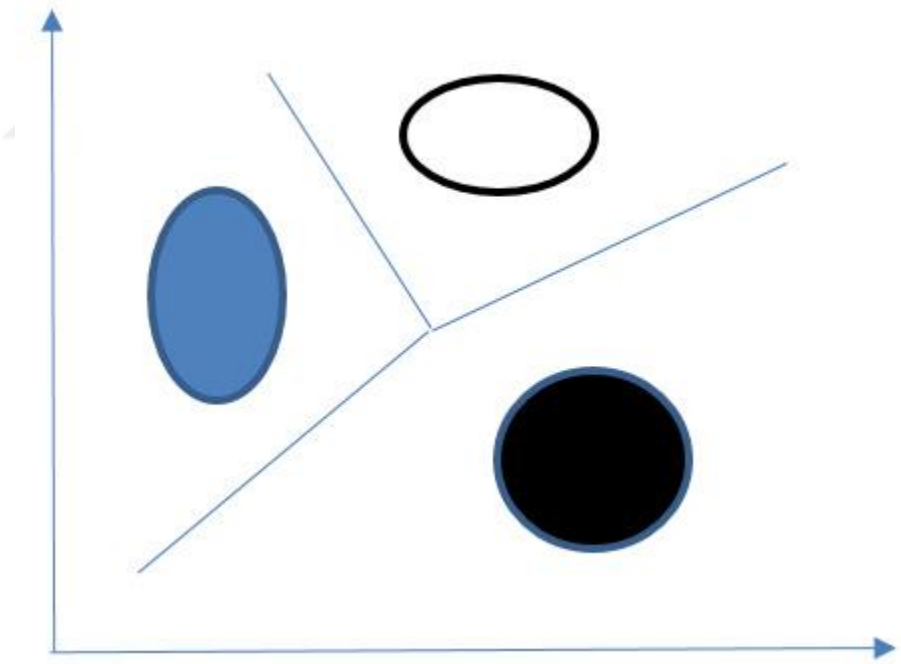

Figure 5 Clustered objects

Figure 5 includes three clusters that contains groups of points that are clustered or classified with some specific characteristics. Feature extraction is the key and main process in the classification process.

2.4.2. Support Vector Machines (SVM)

Machine learning is widely used to identify structural patterns, groups from the data. We need to have training sets and some prediction models for computer to predict the class of an object. The system needs to generalize the objects for a given and previously seen and recognized set and needs to find an appropriate cluster or group for the new object. The computer can learn the recognition by a function that performs well in the given set that is also named the training set. Empirically, there is also chance to do it in the wrong way and it generates a 'training error'. There is also an overall risk that can be caused from many functions used for the recognition, called 'the test error'. In order to get a sufficient result, we also need to minimize the test error. Vapnik and Chervonenkis reported that there is an upper bound for the testing error is less than or equal to the summation of training error and the complexity of the models [38].

$$\text{Test Error} \leq (\text{Training Error}) + (\text{Complexity of the Models}) \quad (2)$$

Then, we need to have a function that is going to minimize the summation of the 'training error' and the 'complexity of the models'. Then, the support vector machines becomes one ıf the solution for this minimization problem.

Support vector machines (SVM) is a supervised learning algorithms that are widely used in regression methods, clustering and classification. Classification of images and objects, text recognition, biological areas are the major areas that the support vector machine algorithms widely used. The first and the original support vector machine, known as linear classification, algorithm was reported by Vapnik and Chervonenkis. SVM is grouped into two main categories that are linear SVM and non-linear SVM.

The LibSVM library of the Weka, Waikato Environment for Knowledge Analysis, software is used in this thesis and results for each chemical is collected as ROC values.

2.4.3. Locally Weighted Learning (LWL)

Locally weighted learning algorithms (LWL) are a form of memory based learning algorithms and lazy learning algorithms. Lazy learning algorithms can be described as delaying the processing of the training data until a process query needs to be processed or answered [35]. Locally weighted learning is a set of function or method approximation techniques that the prediction is processing by using approximated local models that are around the interested points [39]. The main goal of the function or the method approximation is to recognize or find relationships between the inputs and outputs. In supervised learning processes, inputs are associated with the outputs, specifically one input for one output, in order to constitute a model that predicts some values that are used to estimate the closeness to the true modeling function [39]. Locally weighted algorithms are prediction methods that are done by local functions on some subsets of data, instead of using global functions on sets of global data. In other words, the LWL algorithms is about building local models for the whole function sets instead of building global models [39]. As mentioned above, LWL is a form of a lazy learning because it is characteristic that is delaying the training data until a query needs to be answered. This characteristic makes LWL algorithms more accurate approximation function

2.4.4. Discriminative chemotherapy resistance prediction

We adapt in this thesis the discriminative approach to predict chemotherapy resistance such that it uses two sets of microRNAs that are resistant to its corresponding chemotherapy chemical and non-resistant to its corresponding chemotherapy chemical. Discriminative model learns a separating decision function for resistant and non-resistant sequence features. The trained model predicts the resistance level of the input microRNA based on that function (Figure 6). Here, we consider two alternative methods for training a discriminative model; Support Vector Machine (SVM) and Locally Weighted Learning (LWL).

SVM is a powerful machine learning tool that we consider to classify microRNA sequences. An SVM classifier is generated by a two-step procedure: first, the high

dimensional input space of the SVM is non-linearly mapped into a higher dimensional feature space. In the second step, a linear hyper-plane is constructed in this feature space with the largest possible margin separating the classes of the data. The points classified by SVM are of two types: support vectors and non-support vectors. Non-support vectors are perfectly classified by the hyper-plane and are located outside of the separating margin. The parameters of the SVM do not depend on them, even if their positions are changed, provided that these points will stay outside the margin. The major advantage of the SVM classification is its better generalization ability owing to the fact that it finds the separating hyper plane with the largest margin using support vectors, as opposed to Neural Networks, at which all possible hyper-planes are evaluated. Thus, SVM is considered to be less prone to over fitting than other classifiers.

Our second option for machine learning classifier is the LWL algorithm that incorporates the idea of localization [35]. The main idea of localization is to assign a weight to each training observation that regulates its influence on the training process. This weight depends upon the location of the training point in the input variable space relative to that of the point to be predicted. Weights are bigger for the data points that are closer to the data you are trying to predict, thus 'local' in the name. While instance-based classifiers are often successful in linearly separable case of input data, when the data set is not linear, these methods tend to under fit the training data. The local approach here alleviates this problem by assigning weights to training data. Weighting the data can also be viewed as replicating relevant instances and discarding irrelevant instances.

A machine learning classifier, either SVM or LWL requires that the input must be a fixed length vector. Since the sequences that we study here are naturally of different lengths, we need to have an encoding scheme to represent them in a fixed length vector. Here, we use a common representation based on *k-mer* frequencies. A *k-mer* frequency refers to the frequency of appearance of a word with a length of k inside the sequence. In our case, the RNA alphabet is composed of 4 types of nucleotides: A, G, C, U. Therefore, *1-mer* representation includes the frequencies of 4 nucleotide elements in mature microRNA sequence;

$$S1 = \{p_1 , p_2 , p_3 , p_4 \ \} \quad (3)$$

Where *pi* corresponds to frequency of Adenine, Guanine, Cytosine, or Uracil.

2-mer representation of them in the microRNAs' mature part takes 16 (4*4) vectors:

$$S2 = \{p_1 p_1 , p_2 \ p_2 , p_3 \ p_3, p_4 p_4, \dots \} \quad (4)$$

3-mer representation of them in the microRNAs' mature part takes 64 (4*4*4) vectors:

$$S3 = \{p_1 p_1 p_1, p_2 \ p_2 \ p_2 \ , p_3 \ p_3 p_3 \ , p_4 p_4 \ p_4, \dots \} \quad (5)$$

We integrate three k-mer representations for k=1,2, and 3 to build a single feature vector having a length of 84. We use then these vectors to feed SVM or LWM classifiers.

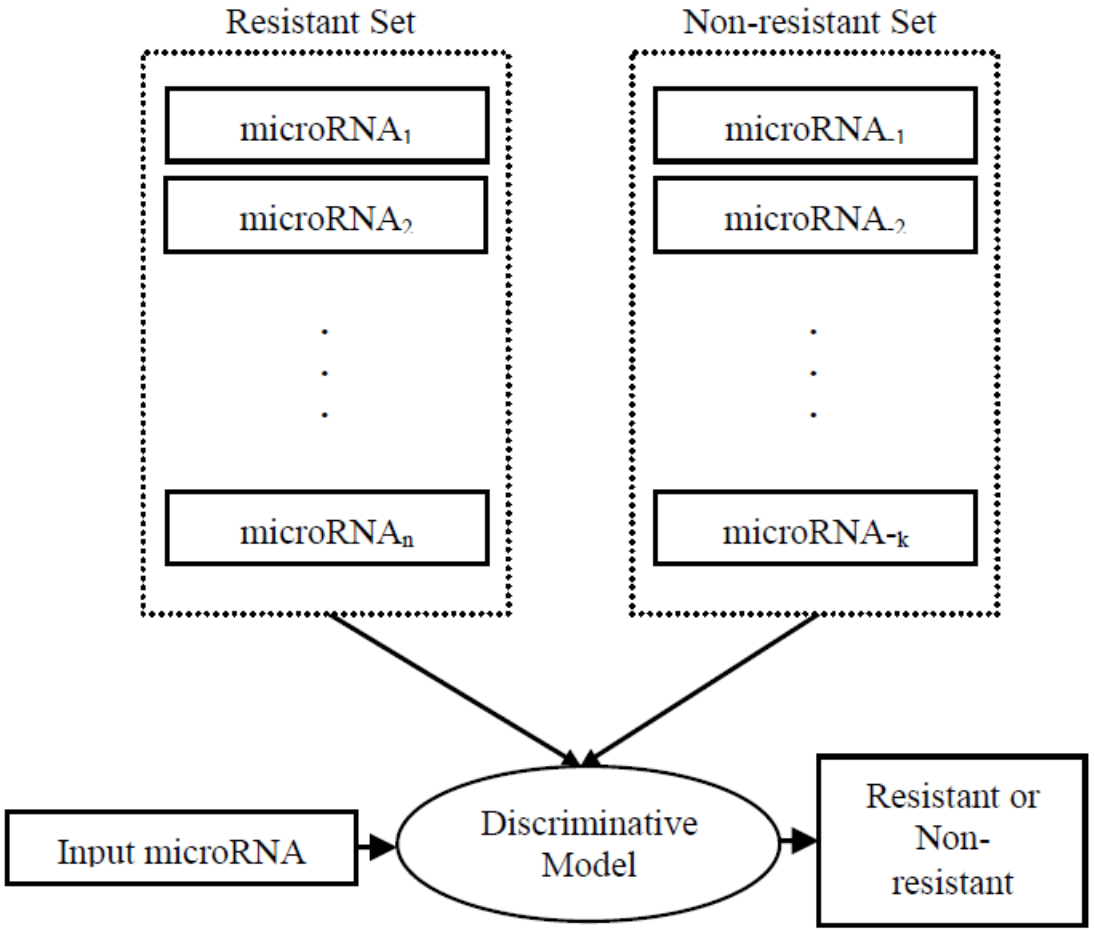

Figure 6 Discriminative model for predicting microRNA resistance to a certain chemotherapy drug

3. RESULTS

3.1. Data Sets

We used the database named CREAM – Chemotherapy Resistance Associated MiRSNP Database- that is free and online database including the microRNAs associated with 432 different chemotherapy resistances [40]. The dataset contains 1921 different microRNAs with resistance labels. Mature microRNA sequences were obtained from miRBase database [41]. Each entry in the database represents the microRNA and its corresponding mature sequences.

3.2. Experimental Setup and Evaluations

The evaluation criteria of this paper are based on ROC scores that are evaluated from ROC (Receiving Operating Characteristics) curves. ROC score was calculated from the area under the ROC curve. ROC analysis investigates and employs the relationship between sensitivity and specificity of a binary classifier. Sensitivity or true positive rate measures the proportion of positives correctly classified; specificity or true negative rate measures the proportion of negatives correctly classified. Conventionally, the true positive rate ‘tpr’ is plotted against the false positive rate ‘fpr’ [42]. All the experimental results in this paper collected as ROC scores in the [0 - 1] scale. A higher value of ROC score indicates a better prediction result where a ROC score is between 0 and 1 and the value 1 represents the perfect prediction.

We compiled the pairwise approach with dynamic- programming-based alignment, the generative model with VLMC, the discriminative model with SVM and LWL in the same experimental setup on the described data set. The methods were referred as "Pairwise", "Generative", "Discriminative (SVM)", and "Discriminative (LWL)" respectively. A ROC score was calculated from by the area that is under the ROC curve that is associated to each specific chemical.

In the pairwise method, the maximum value with the closeness to the high scored value and the maximum value with the closeness to the highest second scored value are

calculated. The average ROC value of the first scenario is found as 0.7203 and the average ROC value of the second scenario is found as 0.6581.

In the generative method, the ROC values are calculated by using the data structure type named probabilistic suffix trees, and two different scenario was applied. First scenario is using a probabilistic suffix tree with depth 3, and the second scenario is using a probabilistic suffix tree with depth 5.

In the discriminative method, support vector machines (SVM) and locally weighted learning (LWL) algorithms were used in the process. 3-mer and 2-mer forms of pyrimidine in the sequences were both used in the process.

The ROC algorithm that is depicted below as Algorithm-1 was applied to the results to calculate the ROC values of each method.

Inputs: 'L' is defined as the closest value to the set (set of tags), and the 'tag' is defined as a tag that can be 0 for false, and 1 for true.

Output: 'R' is defined as the ROC value.

Algorithm-1 Calculation of Roc value

$/*$ *assigning initialization value to* '*true positives*' $*/$

$tp -> 0$

$/*$ *assigning initialization value to* '$false$ *positives*' $*/$

$Fp -> 0$

$/*$ *assigning initialization value to the ROV value* $*/$

$R -> 0$

$/*$ $finding$ *the values of true positives and* $false$ *positives* for *each element in the set* 'L' $*/$

$for\ L\ as\ tag$

$if\ tag\ =\ 1$

$tp\ ->\ tp\ +\ 1$

$else$

$fp\ ->\ fp\ +1$

$R\ ->\ R\ +\ tp$

$end\ if$

$end\ for$

$/*\ calculation\ of\ the\ ROC\ value\ according\ to\ the\ true\ positive\ values\ and\ false\ positive\ values\ */$

$if\ tp\ =\ 0$

$R\ ->\ 0$

$else$

$if\ fp\ =\ 0$

$R\ ->\ 1$

$else$

$R\ ->\ R/tp * fp$

$end\ if$

$end\ if$

3.3. Empirical Results

All of the ROC value results are listed in the Table-5. The first column represents the names of the columns. For each chemical, the ROC values of SW-1, SW-2, PST-3, PST-5, 3-MER-SVM, 3-MER-LWL, 2-MER-SVM, 2-MER-LWL that are correspondingly stands

for SW alignment with one maximum value, SW alignment with two maximum values, probabilistic suffix tree with depth equals 3, probabilistic suffix tree with depth equals 5, support vector machine that is applied on the 3-mer structure of pyrimidine, support vector machine that is applied on the 3-mer structure of pyrimidine, support vector machine that is applied on the dimer structure of pyrimidine, locally weighted learning that is applied on the 3-mer structure of pyrimidine, locally weighted learning that is applied on the dimer structure of pyrimidine.

According to Table-4, the average values for all the methods are following:

Table 4 Average ROC values of the methods

Approach	**Minimum**	**Maximum**	**Average**
Pairwise (SW_1)	0.567	0.875	0.719
Pairwise (SW_2)	0.416	0.815	0.656
Generative (PST_3)	0.323	0.728	0.483
Generative (PST_5)	0.323	0.728	0.483
Discriminative 3-mer (LWL)	0.430	0.730	0.570
Discriminative 3-mer (SVM)	0.418	0.730	0.555
Discriminative 2-mer (LWL)	0.435	0.766	0.557
Discriminative 2-mer (SVM)	0.421	0.672	0.509

3.3.1. Pairwise method

In the pairwise method, the maximum value with the closeness to the high scored value and the maximum value with the closeness to the highest second scored value are calculated. The average ROC value of the first scenario is found as 0.7203 and the average ROC value of the second scenario is found as 0.6581.

In the Figure 7, the ROC values are demonstrated. The x-axis stands for the chemicals, each chemical has an exact number, and the y-axis stands for the ROC values that is in the range of [0, 1].

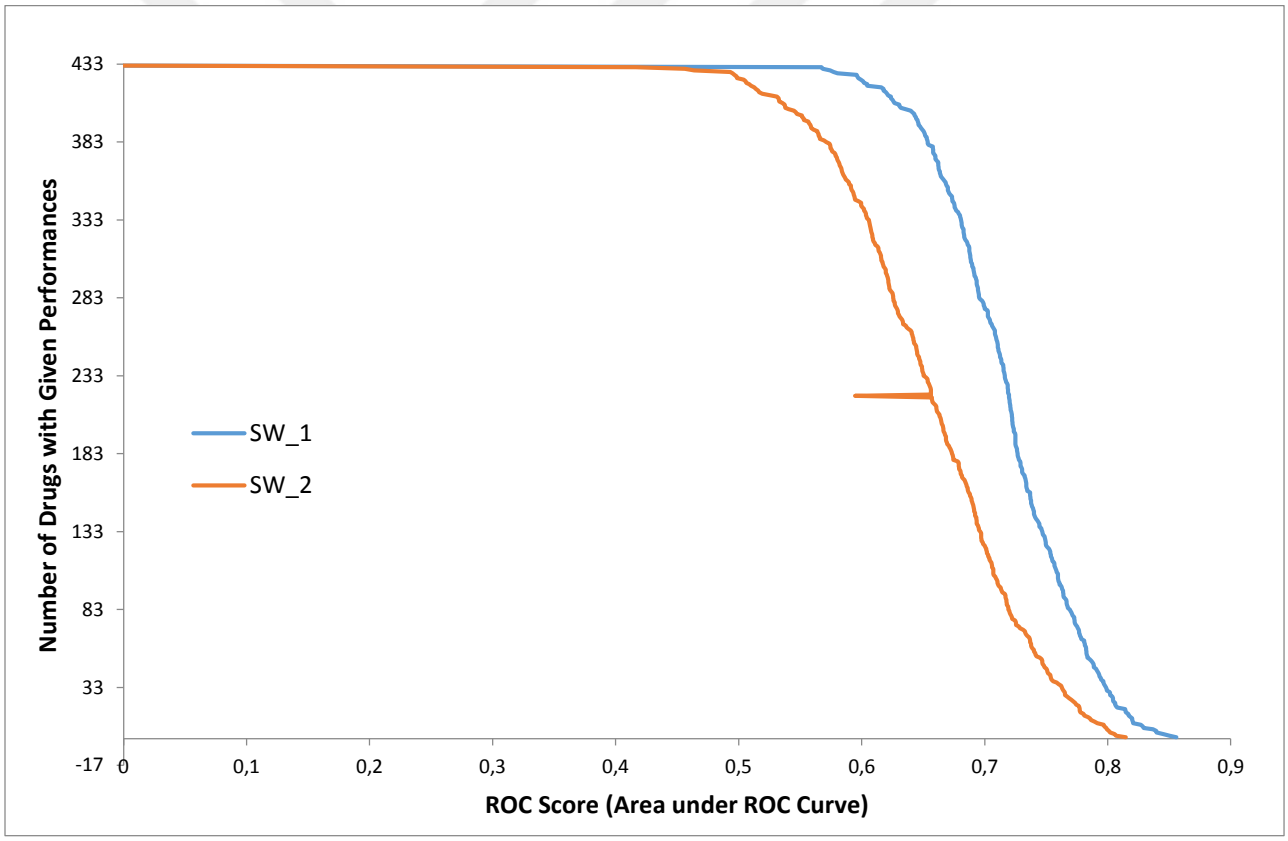

Figure 7 The ROC values of the Pairwise Methods

3.3.2. Generative method

In the generative method, the ROC values are calculated by using the data structure type named probabilistic suffix trees, and two different scenario was applied. First scenario is

using a probabilistic suffix tree with depth 3, and the second scenario is using a probabilistic suffix tree with depth 5.

However, the results taken from the scenario with depth 3 and scenario with depth 5 are same. In the Figure 8, the ROC values are represented. The x-axis stands for the chemicals, each chemical has an exact number, and the y-axis stands for the ROC values that is in the range of [0, 1].

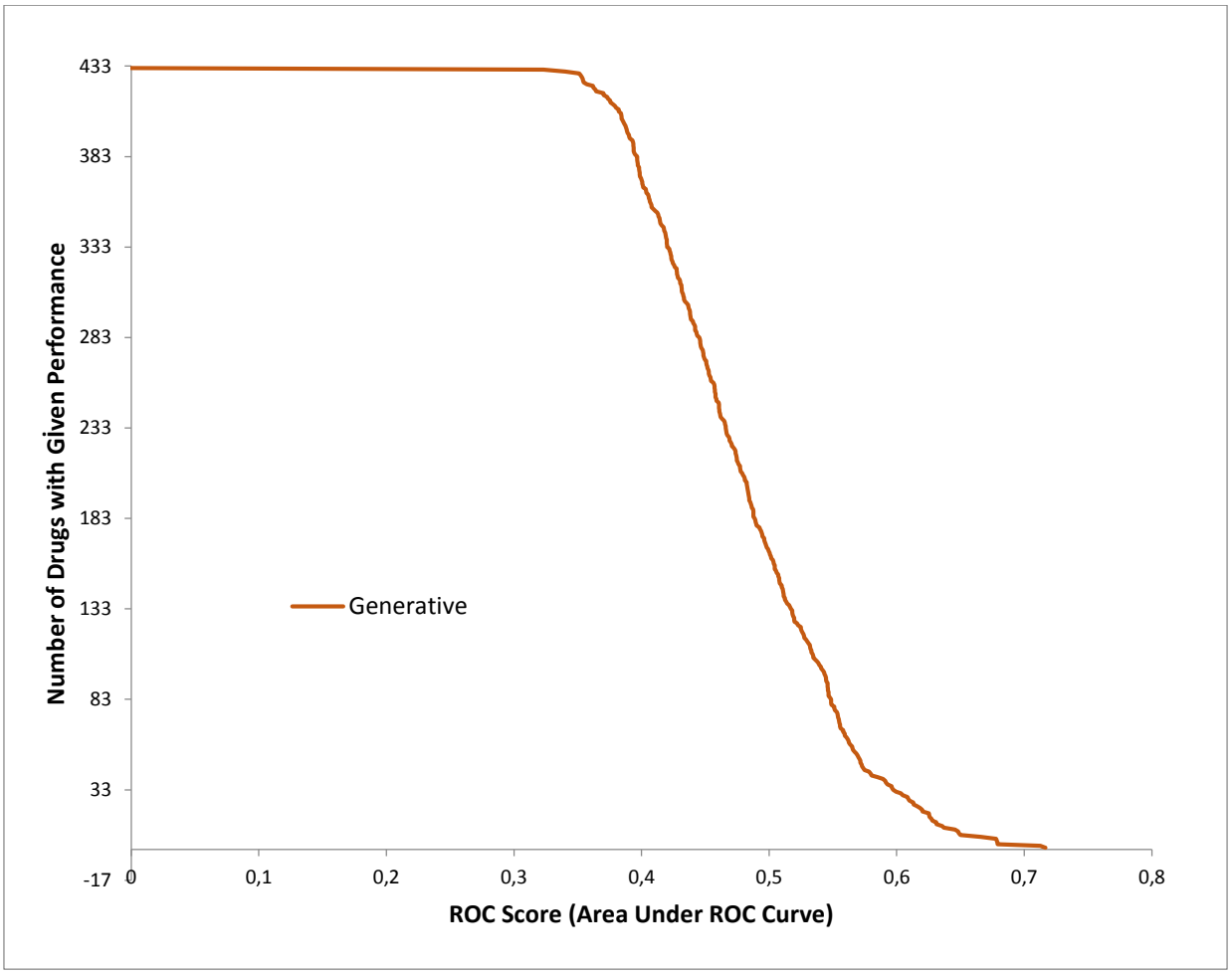

Figure 8 The ROC values of the Generative Methods

3.3.3. Discriminative method

In the discriminative method, support vector machines (SVM) and locally weighted learning (LWL) algorithms were used in the process. 3mer and 2mer representations of nucleotides in the sequences were both used in the process.

3mer SVM, 3mer LWL, 2mer SVM and 2merLWL Roc results were calculated and shown in the Figure 9.

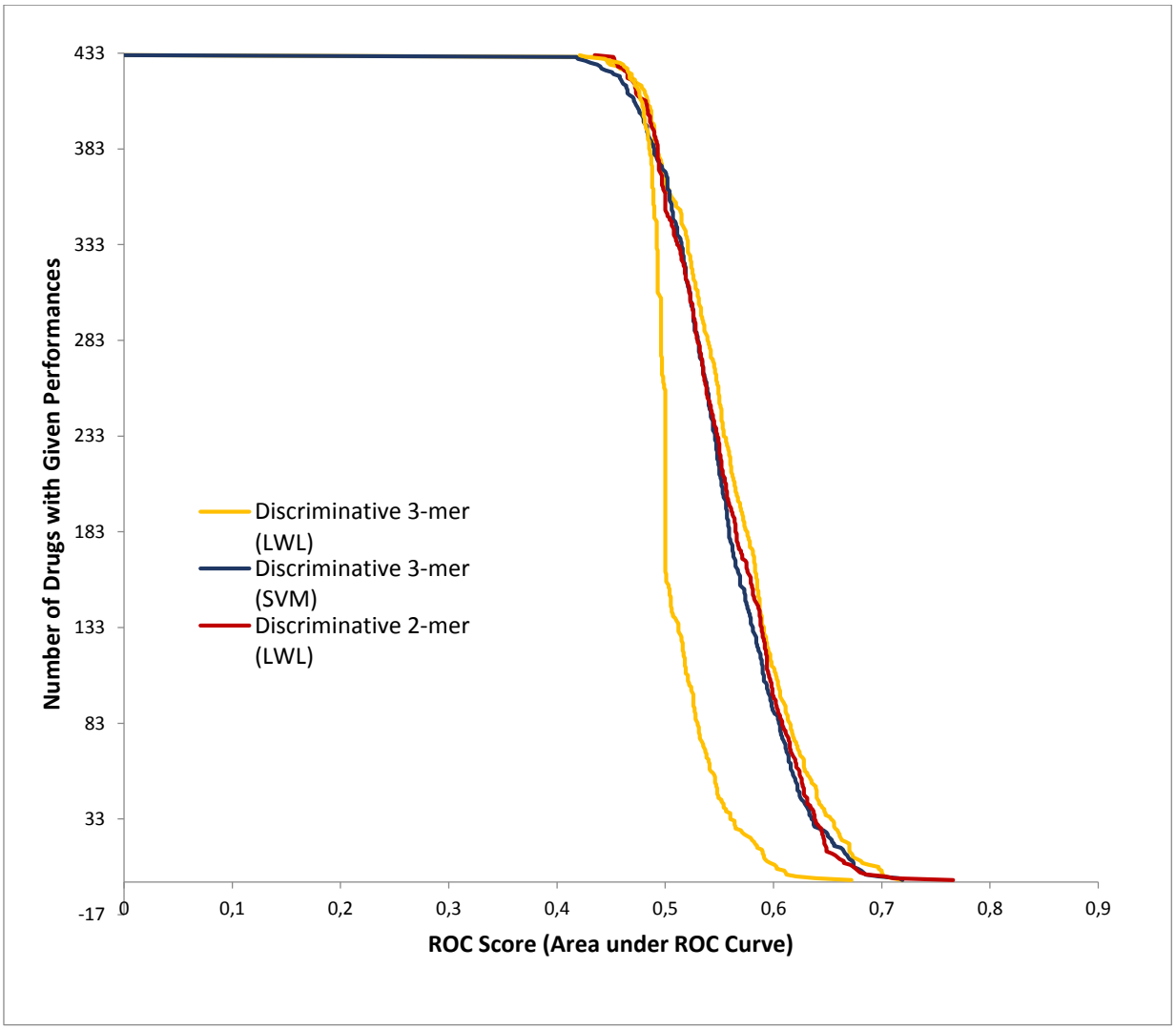

Figure 9 The ROC values of the Discriminative Methods

3.3.4. Comparison

We compiled the pairwise approach with dynamic- programming-based alignment, the generative model with VLMC, the discriminative model with SVM and LWL in the same experimental setup on the described data set. The methods were referred as "Pairwise", "Generative", "Discriminative (SVM)", and "Discriminative (LWL)" respectively. A ROC score was calculated from by the area that is under the ROC curve that is associated to each specific chemical. Figure 10 depicts the number of chemotherapy drugs with the given ROC score performance for each methodology applied.

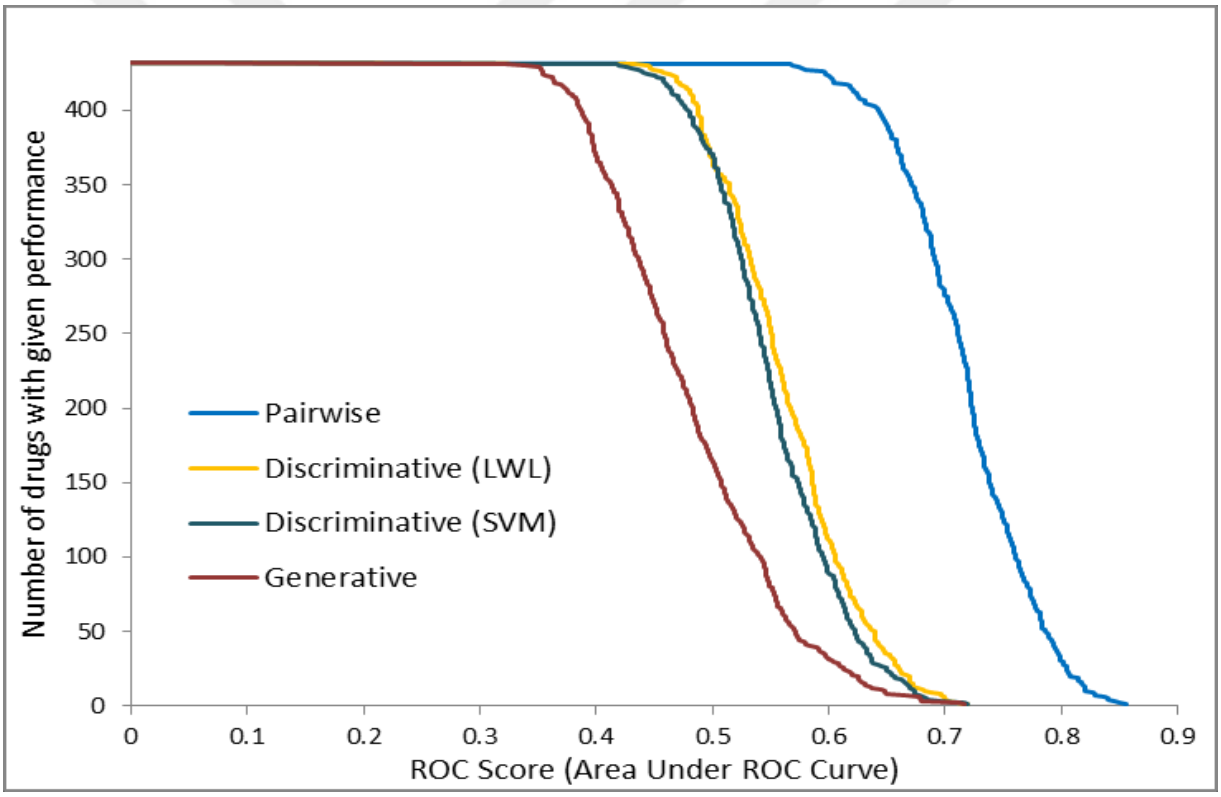

Figure 10 Comparison of methods shown by number of drug labels with given Area under ROC performance

As shown, the best accuracy was achieved by pairwise model. The highest ROC score is 0.875 with this method. In the worst case, this method attained a ROC score of 0.567. For the generative approach, the minimum and maximum ROC scores are 0.323 and 0.728 respectively.

4. CONCLUSION

We study the problem of predicting the resistance of a microRNA to a chemotherapy treatment through a certain chemical drug. To this end, we consider solely mature sequence to assign given microRNA to resistant or non- resistant class for a chemotherapy chemical under consideration. Among three different computational approaches we considered, we obtained the best results with the pairwise approach where generative and discriminative models produced almost random solutions on an experimentally validated dataset. These results determine that the microRNAs which are commonly resistant to a specific chemotherapy chemical do not necessarily have similar sequential characteristics in their mature parts. On the other hand, a high similarity between any two microRNA mature sequences may imply a similar behavior in their resistance to certain chemotherapies. The lower accuracies achieved with generative and discriminative approach can be attributed to the fact that *2-mer* or *3-mer* encoding schemes cannot be biologically representative for microRNAs to be resistant to a certain chemotherapy or not.

This is the first study that evaluates the predictability of specific chemotherapy resistance of microRNA using only sequence information. The results promote the use of pairwise techniques as a complementary tool in association studies for microRNAs and drugs in clinical environments. Future work includes the consideration of problem-specific sequence similarity measures instead of using general-purpose dynamic programming algorithms for alignment.

REFERENCES

[1] H. Li and B. B. Yang, "Friend or foe: the role of microRNA in chemotherapy resistance.," *Acta Pharmacologica Sinica 34.7,* pp. 870-879, 2013.

[2] C. E. Jansen, C. Miaskowski, M. Dodd, G. Dowling and J. Kramer, "A metaanalysis of studies of the effects of cancer chemotherapy on various domains of cognitive function.," *Cancer 104.10,* pp. 2222-2233, 2005.

[3] D. He, F. Gu, F. Gao, J. Hao, D. Gong, X. Gu, A. Mao, J. Jin, L. Fu and X. Ma, "Genome-wide profiles of methylation, microRNAs, and gene expression in chemoresistant breast cancer."," *Scientific Reports 6,* p. 24706, 2016.

[4] I. Alvarez-Garcia and E. A. Miska, "MicroRNA functions in animal development and human disease.," *Development 132.21,* pp. 4653-4662, 2005.

[5] Z. Liu, A. Sall and D. Yang, "MicroRNA: an emerging therapeutic target and intervention tool.," *International journal of molecular sciences 9.6,* p. 978–999, 2008.

[6] A. Carè, "MicroRNA-133 controls cardiac hypertrophy.," *Nature Medicine 13,* pp. 613-618, 2007.

[7] E. A. Wiemer, "The role of microRNAs in cancer: no small matter.," *Eur. J. Cancer. 43(10),* pp. 1529-1544, 2007.

[8] C. S. Sullivan, A. T. Grundhoff, S. Tevethia, J. M. Pipas and D. Ganem, "SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells.," *Nature 435(7042),* pp. 682-686, 2005.

[9] J. Krützfeldt and M. Stoffel, "MicroRNAs: a new class of regulatory genes affecting metabolism.," *Cell Metab. 4(1),* pp. 9-12, 2006.

[10] P. T. Nelson and J. N. Keller, "RNA in brain disease: no longer just "the messenger in the middle".," *J. Neuropathol Exp. Neurol. 66(6),* pp. 461-468, 2007.

[11] G. A. Calin, "MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias.," *Proc. Natl. Acad. Sci. USA, 101(32),* pp. 11755-11760, 2004.

[12] D. M. Mount , "Sequence and Genome Analysis (2nd ed.).," *Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY. ISBN 0-87969-608-7,* 2004.

[13] R. M. Karp., "Some combinatorial problems arising in molecular biology.," *ACM Symp. on Theory of Computing,* p. 278–285, 1993.

[14] L. Wang and T. Jiang , "On the complexity of multiple sequence alignment," *J Comput Biol 1 (4),* p. 337–48, 1994.

[15] I. Elias, "Settling the intractability of multiple alignment," *J Comput Biol 13 (7),* p. 1323–1339, 2006.

[16] W. Pearson and T. Wood, "Statistical Significance in Biological Sequence Comparison," *Handbook of Statistical Genetics, 2nd ed,* vol. 1, 2003.

[17] R. Durbin, S. Eddy, A. Krogh and G. Mitchison, "Biological sequence analysis - Probabilistic models of proteins and nucleic acids.," *Cambridge University Press, Cambridge, UK,* 1998.

[18] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins.," *J Mol Biol 48,* pp. 443-53, 1970.

[19] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences.," *Mol Biol 147,* pp. 195-7, 1981.

[20] A. B. Abecasis, A. M. Vandamme and P. Lemey, "equence alignment in HIV computational analysis.," *HIV sequence compendium 2007,* pp. 2-16, 2006.

[21] M. McNaughton, P. Lu, J. Schaeffer and D. Szafron, " Memory efficient A* heuristics for multiple sequence alignment," *18th National Conference on Artificial Intelligence,* 2002.

[22] D. F. Feng and R. F. Doolittle, "Progressive sequence alignment as a prerequisite to correct phylogenetic trees.," *J Mol Evol 25,* pp. 351-60, 1987.

[23] J. D. Thompson, D. G. Higgins and T. J. Gibson, "improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-

specific gap penalties and weight matrix choice.," *Nucleic Acids Res 22,* pp. 4673-80, 1994.

[24] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, "The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.," *Nucleic Acids Res 25,* pp. 4876-82, 1997.

[25] C. Notredame, D. G. Higgins and J. Heringa, "A novel method for fast and accurate multiple sequence alignment.," *J Mol Biol 302,* pp. 205-17, 2000.

[26] A. N. Arslan and Ö. Eğecioğlu, "Dynamic programming based approximation algorithms for sequence alignment with constraints.," *INFORMS Journal on Computing 16.4,* pp. 441-458, 2004.

[27] G. Bejerano, "Algorithms for variable length Markov chain modeling.," *Bioinformatics 20.5 ,* pp. 788-789, 2004.

[28] A. Galata, N. Johnson and D. Hogg, "Learning variable-length Markov models of behavior.," *Computer Vision and Image Understanding 81.3 ,* pp. 398-413, 2001.

[29] M. Mächler and B. Bühlmann, "Variable length Markov chains: methodology, computing, and software.," *Journal of Computational and Graphical Statistics,* 2012.

[30] H. Oğul, S. S. Umu, Y. Y. Tuncel and M. S. Akkaya, "A probabilistic approach to microRNA-target binding.," *Biochemical and Biophysical Research Communications 413,* pp. 111-115, 2011.

[31] D. Ron, Y. Singer and N. Tishby, "The power of amnesia: learning probabilistic automata with variable memory length.," *Machine Learning, 25 ,* pp. 117-149, 1996.

[32] Z. Sun and J. S. Deogun, "Local prediction approach for protein classification using probabilistic suffix trees.," *In Proc. Second Asia-Pacific Bioinformatics Conference (APBC2004), Dunedin, New Zealand,* 2004.

[33] G. Bejerano and G. Yona , "Variations on probabilistic suffix trees: statistical modeling," *Bioinformatics,* pp. 23-43, 2001 .

[34] G. Bejerano and G. Yona, "Modeling protein families using probabilistic suffix trees.," *Proceedings of the third annual international conference on Computational molecular biology. ACM,* 1999..

[35] C. G. Atkeson, A. W. Moore and S. Schaal, "Locally Weighted Learning," *Artificial IntelligenceReview, 11,* pp. 11-73, 1997.

[36] C. Largeron-Leténo, "Prediction suffix trees for supervised classification of sequences.," *Pattern Recognition Letters 24.16,* pp. 3153-3164, 2003.

[37] I. Guyon and A. Elisseeff, "An introduction to feature extraction.," *Springer Berlin Heidelberg,* pp. 1-25, 2006.

[38] V. N. Vapnik and A. Y. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations.," *Theory of Probability & Its Applications 26.3 ,* pp. 532-553, 1982.

[39] P. Englert, "Locally Weighted Learning," *Seminar Class on Autonomous Learning Systems,* 2012.

[40] Y. Dai, Y. Lv, F. Meng, X. Yu, Y. Zhang, S. Wang, X. Liu, D. Liu, J. Wang, X. Li and W. Jiang, "CREAM: a database for chemotherapy resistance-associated miRSNP," *Cell Death and Disease 5,* p. e1272, 2014.

[41] A. Kozomara and S. Griffiths-Jones, "miRBase: annotating high confidence microRNAs using deep sequencing data.," *Nucleic Acid Res.,* pp. D68-D73, 2015.

[42] T. Fawcett, "An introduction to ROC analysis.," *Pattern Recognition Letters 27.8,* pp. 861-874, 2006.

APPENDIX

Table 5 ROC Values of Each Method

CHEMICAL NAME	SW_1	SW_2	PST_3	PST_5	3MER-SVM	3MER-LWL	2MER-SVM	2MER-LWL
(-)-dactylolide	0,68	0,6526	0,397	0,397	0,418	0,43	0,421	0,435
(-)-Dibenzylnortrachelogenin	0,725	0,6733	0,4228	0,4228	0,419	0,45	0,427	0,452
(-)-sesamin	0,6625	0,6606	0,5393	0,5393	0,426	0,45	0,444	0,453
(1R)-6-endo-Benzyloxycamphor	0,7291	0,6847	0,4334	0,4334	0,431	0,45	0,448	0,453
(1S,4R,5R,12R,13S)-9-[(2-{[(1S,4R,5	0,7175	0,6256	0,4143	0,4143	0,437	0,45	0,458	0,455
(1S,4R,5R,8R,12R,13S)-1,5-dimeth	0,7401	0,6538	0,4805	0,4805	0,44	0,46	0,461	0,456
(2S,3S)-2-Hydroxy-2-phenyl-3-p-to	0,6045	0,4637	0,4809	0,4809	0,441	0,46	0,462	0,456
(6R,6S)-2,4-Diamino-6-(2,5-dimetl	0,7728	0,7687	0,4574	0,4574	0,445	0,46	0,465	0,461
(6R,6S)-2,4-Diamino-6-(3,4,5-trim	0,7444	0,7401	0,5682	0,5682	0,451	0,47	0,465	0,462
(E)Methyl-3-(4-methoxyphenyl)-2-	0,72	0,7048	0,5517	0,5517	0,452	0,47	0,466	0,465
(Thiopen-2-ylmethylene-hydrazin	0,7377	0,7039	0,3873	0,3873	0,458	0,47	0,467	0,465
(Z)-3-iodo-1,2-diphenyl-2-propen-	0,7825	0,7379	0,5197	0,5197	0,458	0,47	0,467	0,465
1, 1'[Pentane-1,5-diyl)dioxy]bis(1	0,7292	0,656	0,5099	0,5099	0,459	0,47	0,467	0,465
1,2-Diphenylseleno-o-carborane	0,6701	0,6696	0,4708	0,4708	0,46	0,47	0,469	0,469
1,2-Dithiolan-3-yl-N-(3'-methoxye	0,7639	0,7061	0,4187	0,4187	0,461	0,48	0,471	0,47
1,2-Naphthoquinone, 3,7-dimethy	0,746	0,6822	0,4518	0,4518	0,464	0,48	0,472	0,472
1,3,6-Triphenyl-oxazolo(5,4-d)pyr	0,6599	0,5975	0,4109	0,4109	0,464	0,48	0,472	0,472
1,3-dimethyl-2,4-dioxo(1H,3H)pyr	0,6936	0,6563	0,4874	0,4874	0,465	0,48	0,474	0,472
1,3-Dimethyl-5'-(2-chloro-phenyla	0,8067	0,7482	0,4484	0,4484	0,465	0,48	0,475	0,473
1,4,7,10,13,16-Hexaoxacyclooctad	0,6502	0,5158	0,5546	0,5546	0,465	0,48	0,476	0,473
1,4,7,10-Tetraazacyclododecane-1	0,6399	0,5713	0,4609	0,4609	0,467	0,48	0,476	0,473
1,4-BIS(DIETHYLAMINO)-5,8-DIHYD	0,6579	0,5851	0,5615	0,5615	0,47	0,48	0,476	0,474
1,4-Butanediamine, N,N'-bis[3-[(p	0,6492	0,5695	0,4123	0,4123	0,471	0,48	0,477	0,476
1,4-Dimethyl-14-oxa-3-dioxothia-	0,771	0,7422	0,5485	0,5485	0,471	0,48	0,477	0,479
1,4-Naphthalenedione, 2-chloro-3	0,7101	0,5837	0,4856	0,4856	0,472	0,48	0,478	0,482
1,6-bis-(2,6-difluorobenzyl)-2-(2,6	0,661	0,6176	0,7126	0,7126	0,473	0,48	0,479	0,482
1,7-bis(6-methoxybenzothizol-2-y	0,6785	0,628	0,434	0,434	0,474	0,49	0,479	0,483
1,7-Heptanediamine, N,N'-bis(2-m	0,6544	0,5472	0,4058	0,4058	0,475	0,49	0,479	0,483
1,8-Octanediamine, N,N'-di-9-acri	0,6828	0,6662	0,4217	0,4217	0,476	0,49	0,48	0,483
1-(3-diethylarsino-3-thio-2-methy	0,7467	0,7287	0,4063	0,4063	0,476	0,49	0,48	0,484
1-(3-Methyl-1,4-dioxy-quinoxalin	0,7268	0,6668	0,4744	0,4744	0,478	0,49	0,48	0,484
1-(4-chlorophenyl)-3,3-diphenyl-1	0,6876	0,5788	0,3965	0,3965	0,479	0,49	0,48	0,484
1-Acetyl-3-(4H-3,1-benzoxazin-2-y	0,8144	0,7411	0,3878	0,3878	0,48	0,49	0,48	0,486
1-Acetylthio-1-phenyl-2-nitroetha	0,7783	0,6333	0,4275	0,4275	0,48	0,49	0,481	0,486
1-Azabicyclo[2.2.2]octan-3-one, 2	0,7734	0,7543	0,5219	0,5219	0,48	0,49	0,481	0,486
1-Naphthalenecarboxamide, N,N'-	0,6916	0,6336	0,4662	0,4662	0,481	0,49	0,481	0,486
1-O-Octadecyl-[(2.fwdarw.5')-3'-C	0,7785	0,7069	0,4313	0,4313	0,483	0,49	0,481	0,487
1-Piperazinecarboselenoic acid, 4	0,7263	0,6935	0,5796	0,5796	0,483	0,49	0,482	0,487
1-Piperazinecarbothioic acid, 4-(2	0,7662	0,6136	0,4876	0,4876	0,483	0,49	0,482	0,488
1-Piperidineethanol, .alpha.-[p-(p	0,8042	0,7448	0,493	0,493	0,483	0,49	0,483	0,489
1-Propanone, 1-chloro-2-methyl-,	0,7026	0,5996	0,3989	0,3989	0,484	0,49	0,484	0,49
1-Propanone, 3-(dimethylamino)-	0,7757	0,6803	0,4275	0,4275	0,484	0,49	0,484	0,49
1-[m-methyl anilino malonyl]-3-m	0,6475	0,6217	0,4429	0,4429	0,485	0,49	0,484	0,49
10H-Phenothiazine, 1,3,4-trifluor	0,5959	0,5167	0,5013	0,5013	0,486	0,49	0,484	0,491
10H-Phenothiazine, 10-(4-chlorob	0,7221	0,6337	0,4134	0,4134	0,487	0,49	0,484	0,491
11-(5'-Bromopentylidine)-5,6-dihy	0,7249	0,6293	0,4285	0,4285	0,487	0,49	0,485	0,492
11H-Indeno[1,2-c]isoquinolinium	0,7886	0,7018	0,5894	0,5894	0,489	0,49	0,485	0,492
11H-[1]Benzothiopyrano[2,3-e]im	0,6682	0,5859	0,3753	0,3753	0,489	0,49	0,485	0,493

12-Benzyl-5,12-dihydro-indeno[2'	0,6993	0,6489	0,4479	0,4479	0,49	0,49	0,485	0,493
14-FLUORO-4-DEMETHOXYDAUNO	0,7237	0,7224	0,7279	0,7279	0,49	0,49	0,485	0,493
16.beta.-N-methylpiperazino-5-an	0,6953	0,5946	0,3571	0,3571	0,49	0,49	0,486	0,493
17.beta.-Acetoxy-2-methoxyestra-	0,5993	0,4164	0,4973	0,4973	0,49	0,49	0,486	0,493
1H,3H-Benzo[c]pyrrolo[3,4-a]carb	0,7486	0,7368	0,5594	0,5594	0,492	0,49	0,486	0,493
1H-1,2,4-Triazole-5(4H)-one, 4-[[(	0,6769	0,6042	0,4312	0,4312	0,492	0,50	0,487	0,493
1H-1,4,7-Triazonine-1,4,7-triaceti	0,6887	0,6108	0,3531	0,3531	0,493	0,50	0,487	0,493
1H-1,4,7-Triazonine-1-acetic acid	0,7107	0,5995	0,3231	0,3231	0,494	0,50	0,487	0,494
1H-Benz[g]indazole-3-carboxamic	0,7183	0,6196	0,3859	0,3859	0,495	0,50	0,487	0,494
1H-Dibenz[de,h]isoquinoline-1,3-(	0,6462	0,5805	0,4459	0,4459	0,497	0,50	0,487	0,494
1H-Dipyrazolo[1,5-a:4',3'-e]pyrim	0,7211	0,6204	0,531	0,531	0,497	0,50	0,488	0,494
1H-Indol-2-one, 5-bromo-2,3-dihy	0,7084	0,6538	0,4772	0,4772	0,498	0,50	0,488	0,494
1H-Indole, 3,3'-(2,5-piperazinediy	0,7931	0,6879	0,3933	0,3933	0,5	0,50	0,488	0,494
1H-Indole, 5-chloro-2,3-dihydro-1	0,7338	0,6684	0,4173	0,4173	0,501	0,50	0,488	0,495
1H-Pyrano[3',4':6,7]indolizino[1,2	0,8147	0,7892	0,5716	0,5716	0,501	0,50	0,488	0,496
1H-Pyrano[3',4':6,7]indolizino[1,2	0,7964	0,7627	0,5556	0,5556	0,502	0,50	0,488	0,497
1H-Pyrano[3',4':6,7]indolizino[1,2	0,8205	0,7773	0,5627	0,5627	0,502	0,50	0,488	0,497
1H-Pyrano[3',4':6,7]indolizino[1,2	0,8286	0,8067	0,609	0,609	0,502	0,50	0,488	0,497
1H-Pyrano[3',4':6,7]indolizino[1,2	0,8293	0,801	0,6202	0,6202	0,502	0,50	0,488	0,497
1H-Pyrazole-1-ethanol, 4,4'-azoxy	0,7298	0,6947	0,4809	0,4809	0,502	0,50	0,488	0,497
1H-Pyrazole-5-carboxylic acid, 4,'	0,781	0,7863	0,5748	0,5748	0,502	0,50	0,488	0,497
1H-Pyrrolo[1,2-a]benzimidazole-5	0,6827	0,5743	0,4191	0,4191	0,502	0,50	0,488	0,498
2,2',5,5'-TETRAHYDROXY-4,4'-DIMI	0,7885	0,7516	0,4175	0,4175	0,504	0,50	0,489	0,498
2,2'-(1,4-phenylene)bis(1,3-dimeth	0,6874	0,5944	0,3965	0,3965	0,504	0,50	0,489	0,499
2,4-Pyrimidinediamine, 5-(4-chlo	0,6873	0,6815	0,5489	0,5489	0,504	0,51	0,489	0,5
2,6-Pyridinedicarbonitrile, 3,4,5-t	0,7532	0,765	0,5929	0,5929	0,504	0,51	0,489	0,5
2,6-Pyrimidinediamine, 4-chloro-	0,7927	0,767	0,5549	0,5549	0,504	0,51	0,489	0,5
2-(2-(Dimethylamino)ethylamino)	0,7369	0,7222	0,6084	0,6084	0,504	0,51	0,489	0,5
2-(4-chloro-2-methylphenoxy)-N-(	0,625	0,5405	0,4527	0,4527	0,505	0,51	0,489	0,5
2-(4-morpholinophenylamino)-6-(	0,6581	0,6272	0,4947	0,4947	0,506	0,51	0,489	0,5
2-(7-methyl-6-(2-phenylhydrazinyl	0,7621	0,6972	0,4611	0,4611	0,506	0,51	0,489	0,5
2-Acetylimidazo[4,5-b]pyridin 4 p	0,6207	0,5063	0,4508	0,4508	0,506	0,51	0,49	0,5
2-Azabicyclo[16.3.1]docosa-1(21)	0,6889	0,649	0,5188	0,5188	0,506	0,51	0,49	0,5
2-Azabicyclo[16.3.1]docosane, ge	0,741	0,6861	0,3937	0,3937	0,507	0,51	0,49	0,5
2-Benzylthio-4-chloro-N-[imino(3,	0,7836	0,7654	0,5375	0,5375	0,507	0,52	0,49	0,501
2-bromo-8-methoxy-6-methyl-1,4-	0,7313	0,6093	0,4365	0,4365	0,507	0,52	0,49	0,502
2-Butenoic acid, 4,4-dicyano-4-[[(	0,7107	0,6029	0,4333	0,4333	0,507	0,52	0,49	0,502
2-Butenoic acid, 4-[[4,5-dichloro-	0,7568	0,6664	0,3883	0,3883	0,507	0,52	0,49	0,504
2-carboxy-(2'-furyl)-7-chloro-3-tri	0,7191	0,6986	0,5055	0,5055	0,509	0,52	0,491	0,504
2-Ethylaminoestradiol	0,7934	0,7857	0,5005	0,5005	0,51	0,52	0,492	0,506
2-fluorphenylhydrazone alpha-la	0,6804	0,6412	0,5726	0,5726	0,51	0,52	0,492	0,506
2-Imidazoline-1-ethanol, .alpha.,4	0,7571	0,7469	0,4494	0,4494	0,511	0,52	0,492	0,506
2-Iodoestradiol	0,7377	0,6964	0,5115	0,5115	0,511	0,52	0,492	0,507
2-Methyl-4-N,N-bis-2'-cyanoethyl	0,7035	0,6911	0,5716	0,5716	0,511	0,52	0,492	0,508
2-Naphthacenecarboxamide, N-[((	0,7236	0,6727	0,4936	0,4936	0,511	0,52	0,492	0,508
2-Phenyl-oxazolo(4,5-c)quinolin-4	0,8017	0,7639	0,6311	0,6311	0,511	0,52	0,492	0,508
2-Propen-1-one, 1-(4-chlorophen	0,7767	0,7366	0,4775	0,4775	0,513	0,52	0,492	0,508
2-Propen-1-one, 3-(l,3-benzodioxo	0,7305	0,6648	0,4509	0,4509	0,514	0,52	0,492	0,509
2-Propenone, 1,1'-(2,5-pyridinedi	0,6637	0,5328	0,405	0,405	0,514	0,52	0,492	0,51

2-Quinoxalinecarboxylic acid, 3-[(	0,7163	0,6493	0,4438	0,4438	0,515	0,52	0,492	0,51
2-[(4-Methylphenyl)sulfinyl]-3-fur	0,7382	0,7144	0,4233	0,4233	0,515	0,52	0,492	0,511
2-[(Dimethylamino)-1-ethyl]-5,6-di	0,7332	0,6973	0,5079	0,5079	0,515	0,52	0,492	0,511
2-{4-[4-(4-{4-[6-(4-methyl piperazi	0,7609	0,7072	0,4377	0,4377	0,516	0,52	0,492	0,513
23,27-Epoxy-3H-pyrido[2,1-c] [1,4]	0,7987	0,6211	0,4652	0,4652	0,516	0,52	0,492	0,513
2H-1,3,5-Thiadiazine-2-thione, 3,5	0,7764	0,6137	0,483	0,483	0,516	0,52	0,493	0,514
2H-1-Benzopyran, 6-methoxy-3-nit	0,6444	0,526	0,4736	0,4736	0,516	0,52	0,493	0,514
2H-1-Benzopyran-2-one, 6-chloro-	0,6737	0,5746	0,6252	0,6252	0,517	0,52	0,493	0,514
3',4',5'-Trimethoxyflavone	0,8501	0,7972	0,5849	0,5849	0,517	0,52	0,493	0,515
3'-C-ethinylcytidine	0,7172	0,6782	0,4708	0,4708	0,517	0,52	0,493	0,515
3,5,7-Triphenyl-isoxazolo(3,4-d)py	0,7153	0,6406	0,3844	0,3844	0,517	0,52	0,493	0,515
3,8,13,18-Pentaazapentacosane pe	0,6881	0,6328	0,5804	0,5804	0,518	0,52	0,493	0,516
3-(2-PYRIDYL)-ACRYLOPHENONE	0,6854	0,5641	0,4961	0,4961	0,518	0,52	0,493	0,517
3-(5-Benzo[b]thiophen-3-yl-1H-ind	0,7086	0,6318	0,4145	0,4145	0,519	0,52	0,493	0,517
3-AZABICYCLO[3.2.2]NONANE-3-CA	0,6454	0,5661	0,5466	0,5466	0,519	0,53	0,493	0,518
3-BROMO-2,4,6-TRINITROTOLUENE	0,8171	0,7557	0,5462	0,5462	0,519	0,53	0,493	0,518
3-Deazaneplanocin, hydrochloride	0,7075	0,6146	0,3929	0,3929	0,519	0,53	0,493	0,518
3-Furancarbothioamide, [[3-(2-but	0,6705	0,6435	0,4402	0,4402	0,519	0,53	0,493	0,518
3-Hydroxy-16-(p-nitrobenzylidene)	0,8046	0,7467	0,4947	0,4947	0,519	0,53	0,493	0,519
3-Iodoacetamido-benzoylurea	0,6515	0,5934	0,4673	0,4673	0,519	0,53	0,493	0,519
3-Phenacyliden-5-brom-2-indolinc	0,7289	0,6498	0,4453	0,4453	0,52	0,53	0,493	0,519
3-Pyridinecarbonitrile, 6-[4-[(7-ch	0,7491	0,6603	0,5341	0,5341	0,52	0,53	0,493	0,52
3-Pyridinecarboxylic acid, 4-[2-cy	0,7652	0,7401	0,4516	0,4516	0,521	0,53	0,493	0,521
3-[2-pyrrolidino] ethoxy-17-butyl-	0,7186	0,6535	0,484	0,484	0,521	0,53	0,493	0,521
4'-methylpenduletin	0,6266	0,5518	0,4065	0,4065	0,522	0,53	0,493	0,522
4(3H)-Quinazolinone, 3-[[(2-nitrop	0,7378	0,6934	0,511	0,511	0,522	0,53	0,493	0,522
4,11-dihydroxy-1-[4-(4-hydroxy-1-	0,7153	0,7104	0,5405	0,5405	0,523	0,53	0,493	0,522
4,5,6-tribromo-7-(3-N-methylpiper	0,7202	0,6425	0,3736	0,3736	0,523	0,53	0,493	0,522
4,8-Ethenobenzo[1,2-c:4,5-c']dipyr	0,7204	0,6693	0,543	0,543	0,523	0,53	0,494	0,523
4-(1-Acetyloxypropen-2-yl-)-2-meth	0,7407	0,6922	0,479	0,479	0,523	0,53	0,495	0,523
4-(4-Chloro-phenyl)-7a-methyl-2,5	0,7371	0,6855	0,4824	0,4824	0,523	0,53	0,496	0,523
4-(m-Fluorobenzylidene)-2-phenyl-	0,75	0,6188	0,4878	0,4878	0,524	0,53	0,496	0,523
4-AMINO-3-PENTADECYLPHENOL	0,7496	0,6062	0,4996	0,4996	0,524	0,53	0,496	0,523
4-Benzylidene-2-phenyl-5(4H)-oxa	0,7964	0,6933	0,4845	0,4845	0,524	0,53	0,496	0,525
4-Methylthio-N-(3,4,5-trimethoxyb	0,6578	0,6295	0,3995	0,3995	0,525	0,53	0,496	0,525
4-Morpholinepentanoic acid, .alpl	0,6643	0,576	0,385	0,385	0,526	0,53	0,496	0,525
4-[(Z)-2-(5-nitro-2-furyl)ethenyl]-6-	0,7118	0,6675	0,4616	0,4616	0,526	0,53	0,496	0,525
4-[(Z)-2-phenylethenyl]-6-(3,5,5-tri	0,6935	0,6902	0,6182	0,6182	0,526	0,53	0,496	0,526
4-[4-(2-Diethylamino-ethylamino)-	0,7977	0,7778	0,4006	0,4006	0,526	0,53	0,496	0,526
4-[N-(4-amino-4-deoxy-N10-methy	0,7545	0,7471	0,527	0,527	0,526	0,53	0,496	0,526
4H,9H-Naphtho[2,3-b]furan-2-carb	0,7042	0,6608	0,556	0,556	0,526	0,53	0,496	0,526
4H-1,3,6,2-Dioxazaphosphocinium	0,7264	0,6734	0,3969	0,3969	0,526	0,54	0,496	0,527
4H-1-Benzothiopyran-4-one, 2-(me	0,7216	0,64	0,39	0,39	0,527	0,54	0,496	0,527
4H-1-Benzothiopyran-4-one, 2-(me	0,7752	0,7038	0,4458	0,4458	0,527	0,54	0,496	0,528
5,5'-bis(2-methyl-1H-indol-3-yl)-3,	0,7223	0,7012	0,4527	0,4527	0,527	0,54	0,496	0,528
5,6-Dihydro-6-[3-(2-hydroxyethyl)a	0,6812	0,6074	0,4318	0,4318	0,527	0,54	0,496	0,528
5-(p-chlorobenzylidene)-3-phenyl-	0,6622	0,6572	0,5253	0,5253	0,527	0,54	0,496	0,528
5-CHLORO-6-[[(3,4-DICHLOROPHEN	0,6945	0,6793	0,5623	0,5623	0,529	0,54	0,496	0,528
5-Hydroxy-3,7-dimethoxy-3',4'-met	0,7305	0,6641	0,4832	0,4832	0,529	0,54	0,496	0,528
5-Hydroxy-7-amino(1,2,3)thiadiaz	0,8207	0,802	0,6487	0,6487	0,53	0,54	0,496	0,529

5-Phenyl-4-oxo-oxadiazolo(5,4-c)q	0,6161	0,5041	0,3512	0,3512	0,53	0,54	0,496	0,529
5-Pyrimidinecarboxaldehyde, 2-am	0,6305	0,5376	0,5199	0,5199	0,53	0,54	0,496	0,529
5H-Pyrido[3,2-d][2]benzazepin-6(7I	0,7564	0,7137	0,533	0,533	0,531	0,54	0,496	0,53
5H-[1,3]Dioxolo[5,6]indeno[1,2-c]is	0,7732	0,7254	0,6499	0,6499	0,531	0,54	0,496	0,53
6,7-dimethyl-1,4-di-N-oxide-3-meth	0,7209	0,6158	0,5322	0,5322	0,531	0,54	0,496	0,531
6-(10'-tert-BOC-aminodecyl)-5,6-di	0,7705	0,6864	0,5409	0,5409	0,531	0,54	0,496	0,532
6-(10'-tert-BOC-aminododecyl)-5,6-	0,7219	0,7028	0,4379	0,4379	0,531	0,54	0,496	0,532
6-(10-aminodecyl)-5,6-dihydro-5,1	0,6899	0,6445	0,4198	0,4198	0,531	0,54	0,496	0,532
6-(11-aminoundecyl)-5,6-dihydro-5	0,7067	0,6577	0,4992	0,4992	0,532	0,54	0,496	0,532
6-(3-Aminopropyl)-5,6-dihbydro-3-	0,6734	0,5913	0,447	0,447	0,532	0,54	0,496	0,533
6-(3-Aminopropyl)-5,6-dihydro-3-n	0,7047	0,6913	0,5963	0,5963	0,532	0,54	0,496	0,533
6-(3-Aminopropyl)-5,6-dihydro-8,9-	0,7992	0,7621	0,6778	0,6778	0,533	0,54	0,497	0,533
6-(3-Aminopropyl)-5,6-dihyro-9-ioc	0,7523	0,6899	0,6133	0,6133	0,534	0,54	0,497	0,534
6-(3-Aminopropyl)-9-ethoxy-5,6-dih	0,7971	0,7841	0,5531	0,5531	0,534	0,55	0,497	0,534
6-(3-Azido-propyl)-2,3-dimethoxthy	0,6625	0,6071	0,5226	0,5226	0,534	0,55	0,497	0,534
6-(3-Azidopropyl)-5,6-dihydro-9-me	0,725	0,7133	0,6661	0,6661	0,535	0,55	0,497	0,534
6-(3-Bromopropyl)-5,6-dihydro-9-m	0,7138	0,6788	0,4778	0,4778	0,535	0,55	0,497	0,535
6-(3-Chloropropyl)-5,6-dihydro-2,3	0,8394	0,8148	0,551	0,551	0,535	0,55	0,497	0,535
6-(3-fluoro-4-methoxyphenyl)-5-(3,	0,6623	0,6633	0,5156	0,5156	0,535	0,55	0,497	0,535
6-Aminotoyocamycin	0,6318	0,6058	0,5542	0,5542	0,536	0,55	0,497	0,535
6-Benzylthioinosine	0,6834	0,6034	0,4617	0,4617	0,536	0,55	0,497	0,535
6-chloro-2-(4-methyl-piperazin-1-y	0,7392	0,661	0,413	0,413	0,537	0,55	0,498	0,536
6-[(3-Thiazolylamino)-1-propyl]-5,6	0,7249	0,7192	0,5551	0,5551	0,537	0,55	0,498	0,536
6-[3-(Bis-hydroxyethylamino)-1-pro	0,7136	0,6972	0,4665	0,4665	0,538	0,55	0,498	0,536
6-[3-[2-[1,2,4]-Triazolyl-1-propyl]-5	0,7474	0,7004	0,5113	0,5113	0,538	0,55	0,498	0,537
6H-Indeno[1,2-c]isoquinoline-5,11-	0,7948	0,7631	0,5347	0,5347	0,538	0,55	0,499	0,537
6H-Pyrido[4,3-b]carbazolium, 9-hy	0,7741	0,7177	0,6272	0,6272	0,539	0,55	0,499	0,538
7-(2-Pyrrolidinoethyl)oximino-5-ar	0,7386	0,6652	0,4078	0,4078	0,539	0,55	0,499	0,538
7-chloro-3-(4-fluorophenyl)-5-meth	0,7911	0,6573	0,5911	0,5911	0,539	0,55	0,5	0,538
7-O-.beta.-D-Galactosyl-befeldin A	0,7642	0,7362	0,5085	0,5085	0,54	0,55	0,5	0,538
7-{[3-(dibutylamino)propyl]amino}	0,7223	0,6302	0,3619	0,3619	0,54	0,55	0,5	0,538
7.alpha.,12.alpha.-Diacetoxy-5.beta	0,8061	0,7654	0,4011	0,4011	0,54	0,55	0,5	0,539
7H-1,2,4-Triazolo[3,4-b][1,3,4]thiac	0,7197	0,7069	0,628	0,628	0,54	0,55	0,5	0,539
7H-Pyrrolo[2,3-d]pyrimidine, 5,6-di	0,7445	0,6684	0,398	0,398	0,54	0,55	0,5	0,54
7H-Pyrrolo[2,3-d]pyrimidine-5-carl	0,8042	0,7486	0,4632	0,4632	0,54	0,55	0,5	0,541
7H-Pyrrolo[3,2-f]quinazoline-1,3-di	0,6011	0,5086	0,4482	0,4482	0,54	0,55	0,5	0,541
7H-Thieno[3',4':3,4]cyclopent[1,2-d	0,7092	0,6444	0,5539	0,5539	0,541	0,55	0,5	0,542
7H-Thieno[3',4':3,4]cyclopent[1,2-d	0,6884	0,5839	0,4303	0,4303	0,541	0,55	0,5	0,542
8-Chloro-3-(2-hydroxyphenyl)-2-me	0,8061	0,7705	0,476	0,476	0,542	0,55	0,5	0,542
8-chloro-3-(4-methylphenylamino)-	0,7729	0,6923	0,637	0,637	0,542	0,55	0,5	0,543
9,10-Anthracenedione, 1-[(oxiranyl	0,6829	0,6225	0,5582	0,5582	0,542	0,55	0,5	0,543
9,10-o-Benzenoanthracene-1,4-dior	0,6529	0,6521	0,679	0,679	0,542	0,55	0,5	0,544
9-AC	0,7205	0,6894	0,6785	0,6785	0,543	0,55	0,5	0,544
9-AMINO-20-CAMPTOTHECIN	0,8398	0,7993	0,7165	0,7165	0,544	0,55	0,5	0,544
9-METHOXYCAMPTOTHECIN	0,7828	0,7532	0,6791	0,6791	0,544	0,55	0,5	0,545
Acetamide, 2-[[[2-(1,2-dihydro-2-ox	0,7345	0,6935	0,4525	0,4525	0,544	0,55	0,5	0,545
Acetamide, N,N'-(1,5-pentanediyl)bi	0,7188	0,5464	0,4279	0,4279	0,544	0,55	0,5	0,545
Acetic acid, [1,4,7,10-tetraazacyclo	0,644	0,5659	0,4835	0,4835	0,544	0,55	0,5	0,546
Adenine, 9-[1,2-bis(ethoxycarbonyl	0,695	0,6214	0,4419	0,4419	0,544	0,55	0,5	0,547
AI 3-63984	0,68	0,6218	0,4486	0,4486	0,545	0,55	0,5	0,547
Amquinate	0,6945	0,5942	0,5485	0,5485	0,546	0,55	0,5	0,547

ANTINEOPLASTIC-328785	0,5804	0,4983	0,5727	0,5727	0,546	0,55	0,5	0,548
ANTINEOPLASTIC-376265	0,6978	0,6502	0,5462	0,5462	0,546	0,56	0,5	0,548
ANTINEOPLASTIC-625543	0,7325	0,6835	0,5581	0,5581	0,546	0,56	0,5	0,549
ANTINEOPLASTIC-654379	0,729	0,6562	0,4903	0,4903	0,547	0,56	0,5	0,549
ANTINEOPLASTIC-655905	0,7231	0,717	0,6309	0,6309	0,547	0,56	0,5	0,549
Arteanuine B	0,7994	0,7772	0,4588	0,4588	0,547	0,56	0,5	0,55
ARTEMISININ DIMER HEMISUCCINA	0,6993	0,6467	0,4572	0,4572	0,547	0,56	0,5	0,55
as-Triazine-3,5(2H,4H)-dione, 2-.be	0,6899	0,6006	0,4491	0,4491	0,547	0,56	0,5	0,55
Aspidospermin	0,6752	0,5813	0,4325	0,4325	0,548	0,56	0,5	0,55
AT 116, benzylamine salt	0,7428	0,7387	0,5387	0,5387	0,548	0,56	0,5	0,55
Austocystin D	0,725	0,7172	0,4844	0,4844	0,548	0,56	0,5	0,55
Avarol	0,7025	0,6309	0,4608	0,4608	0,548	0,56	0,5	0,551
AZETIDINE CARBOTHIOIC ACID, [1-(2	0,6624	0,6452	0,4684	0,4684	0,548	0,56	0,5	0,551
AZG	0,7498	0,7177	0,5486	0,5486	0,549	0,56	0,5	0,551
Azoxydapsone {1,2-bis-[4-(4'-amino	0,6938	0,6255	0,5037	0,5037	0,549	0,56	0,5	0,552
BAS100 Monohydrate	0,6178	0,5102	0,5996	0,5996	0,549	0,56	0,5	0,552
Bassic acid	0,7539	0,7169	0,5974	0,5974	0,55	0,56	0,5	0,552
Benzamide, N-[2-(diethylamino)eth	0,7641	0,7585	0,5029	0,5029	0,55	0,56	0,5	0,552
Benzimidazo[1,2-a][1,8]naphthyrid	0,6789	0,6659	0,4925	0,4925	0,55	0,56	0,5	0,552
Benzoic acid, 2,4-dichloro-, 6-acety	0,7025	0,6689	0,5606	0,5606	0,55	0,56	0,5	0,553
Benzoic acid, 2-(2,2-diphenylcyclop	0,6482	0,6169	0,3866	0,3866	0,55	0,56	0,5	0,553
Benzoic acid, 2-hydroxy-, (2,6-pyric	0,7044	0,7056	0,6099	0,6099	0,551	0,56	0,5	0,553
Benzoic acid, 2-[(4,6-dimethyl-1-ox	0,7596	0,7346	0,5262	0,5262	0,551	0,56	0,5	0,554
Benzoxazole, 5-chloro-2-(chlorome	0,8185	0,7312	0,5071	0,5071	0,552	0,56	0,5	0,555
Benzo[1,2-b:5,4-b']dithiophene-4,8-	0,7584	0,688	0,4466	0,4466	0,552	0,56	0,5	0,555
Benzo[1,2-b:5,4-b']dithiophene-4,8-	0,7814	0,7278	0,3943	0,3943	0,552	0,56	0,5	0,555
Benzo[1,2-c:4,5-c']dipyrrole-1,3,5,7	0,6538	0,5639	0,5347	0,5347	0,552	0,56	0,5	0,556
Benzo[g]pteridine-2,4-dione, 8-chlo	0,7813	0,7463	0,5512	0,5512	0,553	0,57	0,5	0,556
Bis(indole)dibenzyltindichloride	0,6179	0,5193	0,5099	0,5099	0,553	0,57	0,5	0,556
Bis-{(5,6-dihydro-5,11-diketo-11H-i	0,6671	0,5355	0,4877	0,4877	0,553	0,57	0,5	0,556
Bisarylpurine	0,6926	0,6077	0,4508	0,4508	0,553	0,57	0,5	0,557
Bis[1-(4-methylpiperidino)-4-(2-pyr	0,6812	0,6206	0,5107	0,5107	0,554	0,57	0,5	0,557
Bis[1-azepanyl-4-(2-pyridyl)-2,3-di	0,7582	0,6893	0,4387	0,4387	0,554	0,57	0,5	0,557
BKF	0,7677	0,7508	0,5557	0,5557	0,555	0,57	0,5	0,557
Boronic acid, [3-methyl-1-[[1-oxo-4	0,6916	0,625	0,4721	0,4721	0,556	0,57	0,5	0,558
Bromocriptine methanesulfonate	0,7188	0,6792	0,5112	0,5112	0,556	0,57	0,5	0,558
Buclizine Hydrochloride	0,7528	0,6644	0,4576	0,4576	0,556	0,57	0,5	0,559
Butanamide, N-[2-[(4-fluorophenyl	0,6737	0,6041	0,5142	0,5142	0,556	0,57	0,5	0,559
Butanedioic acid, 2-bromo-3-fluoro	0,8019	0,733	0,4755	0,4755	0,556	0,57	0,5	0,56
Butanimidoyl chloride, N-[[[(3-fluo	0,669	0,5781	0,4186	0,4186	0,557	0,57	0,5	0,561
C.I. 37559	0,7166	0,707	0,5447	0,5447	0,557	0,57	0,5	0,561
C.I. 42600	0,7509	0,7413	0,5059	0,5059	0,557	0,57	0,5	0,562
C.I. 65020	0,742	0,7086	0,4079	0,4079	0,557	0,57	0,5	0,562
Camptothecin ethylglycinate ester l	0,7775	0,6887	0,6458	0,6458	0,557	0,57	0,5	0,563
Camptothecin hemisuccinate sodiu	0,8143	0,7029	0,596	0,596	0,558	0,57	0,5	0,564
Camptothecin lysinate HCl	0,8451	0,7916	0,613	0,613	0,558	0,57	0,5	0,564
CAP 1 hydrochloride	0,7638	0,6664	0,4738	0,4738	0,558	0,57	0,5	0,564
Carbonimidodithioic acid, [5-(4-nit	0,6749	0,6139	0,4282	0,4282	0,558	0,57	0,5	0,565
Carbonotrithioic acid, 2,3,5,6-tetra	0,763	0,7179	0,4315	0,4315	0,559	0,57	0,5	0,565
Carquniostatin B	0,6954	0,5835	0,4149	0,4149	0,559	0,57	0,5	0,565

CCI-779	0,7325	0,6268	0,5274	0,5274	0,559	0,58	0,5	0,565
CHAETOCHROMIN	0,7082	0,6157	0,4893	0,4893	0,559	0,58	0,5	0,565
CHAETOGLOBOSIN A	0,7834	0,7809	0,4353	0,4353	0,559	0,58	0,5	0,566
CHALCONE ANALOG	0,6808	0,5913	0,4419	0,4419	0,559	0,58	0,5	0,566
Chlorazin	0,7906	0,711	0,3701	0,3701	0,56	0,58	0,5	0,566
Chlorodestruxin	0,7668	0,6908	0,4381	0,4381	0,56	0,58	0,5	0,566
Choline, hydroxide, 2-methoxy- 3-[r	0,6591	0,5311	0,4036	0,4036	0,561	0,58	0,5	0,566
Choline, hydroxide, 3-methoxy- 2-[r	0,6861	0,6545	0,3773	0,3773	0,562	0,58	0,5	0,567
Chrysanthin	0,6819	0,5883	0,3982	0,3982	0,562	0,58	0,5	0,567
CI-941	0,7191	0,6739	0,5016	0,5016	0,562	0,58	0,5	0,568
cis-Dichlorobis(4-methoxypheneth	0,7236	0,6471	0,4539	0,4539	0,562	0,58	0,5	0,568
Clonixin	0,7304	0,6905	0,4233	0,4233	0,563	0,58	0,5	0,57
Cromolyn sodium	0,7023	0,624	0,4201	0,4201	0,563	0,58	0,5	0,57
Crotonosid	0,7666	0,7217	0,4852	0,4852	0,563	0,58	0,5	0,571
Curromycin A.B mixture	0,7917	0,6944	0,565	0,565	0,564	0,58	0,5	0,571
Cyclo-C	0,7341	0,6816	0,5783	0,5783	0,565	0,58	0,5	0,571
Cycloalkannin	0,7221	0,5802	0,4462	0,4462	0,565	0,58	0,5	0,574
Cyclobutanemethanol, 1-[(2-amino	0,6427	0,6018	0,6481	0,6481	0,565	0,58	0,5	0,575
Cyclohepta[cd]benzofuran	0,7223	0,6109	0,4773	0,4773	0,565	0,58	0,5	0,575
CYCLOMETHYLENOMYCIN A	0,6756	0,6364	0,5299	0,5299	0,566	0,58	0,5	0,575
Cyclopent-4-ene-1,3-dione, 2-[[4-hy	0,7136	0,5997	0,4198	0,4198	0,567	0,58	0,5	0,575
Cyclopentanone, 2-[(4-morpholinyl	0,6908	0,6296	0,4661	0,4661	0,567	0,58	0,5	0,576
Cyclopentanone, 2-[(dimethylamin	0,6948	0,5818	0,3938	0,3938	0,568	0,58	0,5	0,576
Cyclopentanone, 2-[(dimethylamin	0,7026	0,6479	0,4005	0,4005	0,569	0,58	0,501	0,577
Cyclopentanone, 2-[(dimethylamin	0,7667	0,7091	0,4621	0,4621	0,569	0,58	0,501	0,578
Cyclopentanone, 2-[(dimethylamin	0,6685	0,5872	0,382	0,382	0,569	0,59	0,501	0,579
Cyclopentanone, 2-[(dimethylamin	0,6237	0,5373	0,3838	0,3838	0,569	0,59	0,501	0,579
Cyclopentanone, 2-[(dimethylamin	0,733	0,6742	0,4177	0,4177	0,569	0,59	0,501	0,579
Cyclopentanone, 2-[2-methyl-2-[4-(	0,681	0,5756	0,3886	0,3886	0,569	0,59	0,502	0,579
Cyclopentanone, 5-[(4-morpholinyl	0,6963	0,6065	0,3792	0,3792	0,571	0,59	0,503	0,58
CYTOVARICIN	0,7444	0,7055	0,5441	0,5441	0,573	0,59	0,503	0,58
D-Amethopterin	0,7093	0,6624	0,5659	0,5659	0,573	0,59	0,504	0,581
Dasatinib	0,7124	0,6287	0,4038	0,4038	0,573	0,59	0,504	0,581
dechlorinated rebeccamycin	0,7596	0,684	0,4981	0,4981	0,574	0,59	0,504	0,581
Dichlorobis[1-(4-nitrophenyl)-3-ph	0,7276	0,602	0,4328	0,4328	0,574	0,59	0,504	0,581
Diketocoriolin B	0,7059	0,6069	0,4404	0,4404	0,574	0,59	0,505	0,582
Dipyrazolo[3,4-f:3',4'-g]-2-pyrindin	0,7227	0,7254	0,508	0,508	0,574	0,59	0,505	0,582
Disulfide, bis[2-amino-4-(4-methox	0,7151	0,642	0,3643	0,3643	0,575	0,59	0,505	0,583
ERIOFLORIN	0,7225	0,6229	0,4386	0,4386	0,575	0,59	0,505	0,584
Ethanamine, N,N-dimethyl-2-[4-(1,1	0,6908	0,5896	0,3967	0,3967	0,576	0,59	0,505	0,585
Ethanedioic acid, mono(6-nitroben	0,7209	0,6744	0,4794	0,4794	0,576	0,59	0,505	0,586
Ethanesulfonic acid, compd. with 2	0,6935	0,6467	0,5196	0,5196	0,577	0,59	0,506	0,586
Ethanesulfonic acid, compd. with 2	0,7252	0,7114	0,4825	0,4825	0,577	0,59	0,506	0,587
Ethanesulfonic acid, compd. with 4	0,6599	0,6221	0,5034	0,5034	0,578	0,59	0,506	0,587
Ethanesulfonic acid, compd. with 4	0,6429	0,6179	0,4699	0,4699	0,578	0,59	0,507	0,587
Ethanesulfonic acid, compd. with 5	0,7372	0,6913	0,5276	0,5276	0,579	0,59	0,507	0,587
Ethanol, 2-[(4,8,12-trimethyl-3,7,11	0,7153	0,6455	0,3903	0,3903	0,579	0,59	0,509	0,588

Ethyl 4-[6-(2-furyl)-2-(methylsulfan	0,7339	0,7059	0,4074	0,4074	0,579	0,59	0,51	0,588
Ethylenediaminetetraacetic acid bi	0,7768	0,7369	0,536	0,536	0,579	0,59	0,511	0,588
EUPACUNIN	0,664	0,5379	0,4236	0,4236	0,58	0,59	0,512	0,588
Everolimus	0,5964	0,5531	0,4743	0,4743	0,581	0,59	0,512	0,588
FCDR	0,7277	0,6999	0,5006	0,5006	0,581	0,59	0,512	0,589
Fluorene, 9-bromo-2,4,7-trichloro-	0,7257	0,7105	0,3987	0,3987	0,581	0,59	0,512	0,589
Furan-2-yl-(1-oxy-3,6-bis-trifluoro	0,7123	0,6563	0,4827	0,4827	0,582	0,59	0,512	0,589
gamma-glutamyl-alpha-amino-beta	0,7692	0,7538	0,6255	0,6255	0,583	0,59	0,513	0,59
Gardenin	0,7158	0,6746	0,4978	0,4978	0,584	0,59	0,514	0,59
GEL-I-195-1	0,6722	0,5829	0,5467	0,5467	0,584	0,59	0,515	0,59
Geldanamicin	0,737	0,6921	0,3542	0,3542	0,584	0,59	0,515	0,591
Guanidine, 1-(4-chloro-.alpha.,.alp	0,6534	0,5449	0,4648	0,4648	0,584	0,59	0,515	0,591
Helenin	0,7436	0,6589	0,4157	0,4157	0,585	0,59	0,516	0,592
Herveline O	0,7568	0,6612	0,3938	0,3938	0,585	0,59	0,516	0,592
Hexadecylphospho(N-acetyl)-L-seri	0,7107	0,6932	0,4968	0,4968	0,586	0,60	0,516	0,592
Hydrazinecarbodithioic acid, [1-(2-	0,82	0,7537	0,6032	0,6032	0,586	0,60	0,516	0,592
Hydrazinecarbothioamide, 2,2'-(1,5	0,7674	0,6476	0,5071	0,5071	0,587	0,60	0,516	0,593
Hydrazinecarbothioamide, 2-[(2-py	0,5773	0,4562	0,5125	0,5125	0,588	0,60	0,517	0,593
Hydrazinecarbothioamide, 2-[(2-py	0,6514	0,5331	0,4607	0,4607	0,588	0,60	0,517	0,593
Imidazole, 1-(4-chlorophenyl)-4-(4-	0,6986	0,6036	0,4261	0,4261	0,588	0,60	0,517	0,594
Imidazo[2,1-b]thiazole-5-carboxan	0,5684	0,4967	0,5182	0,5182	0,589	0,60	0,518	0,594
Indeno[1,2-c]isoquinoline-5,11-dic	0,827	0,8057	0,6357	0,6357	0,589	0,60	0,518	0,594
Isokanugin	0,755	0,7336	0,4863	0,4863	0,589	0,60	0,518	0,594
Keenamide A	0,686	0,6442	0,5123	0,5123	0,59	0,60	0,518	0,594
Kikubasaponin	0,6831	0,5789	0,5458	0,5458	0,59	0,60	0,518	0,594
L-Aspartic acid, N-[4-[[(2-amino-4-h	0,6735	0,6493	0,5043	0,5043	0,59	0,60	0,519	0,594
L-Lysine, N-[(methylamino)carbony	0,6956	0,6091	0,4302	0,4302	0,59	0,60	0,519	0,594
LMU-5 HERZ	0,715	0,5915	0,4544	0,4544	0,59	0,60	0,519	0,595
Lobinaline, monohydrochloride	0,6883	0,5866	0,3984	0,3984	0,591	0,60	0,519	0,595
Maxima isoflavone D	0,781	0,7735	0,5331	0,5331	0,591	0,60	0,521	0,596
Megacarpidin	0,875	0,8333	0,5459	0,5459	0,591	0,60	0,521	0,596
MELAMPODIN B ACETATE	0,7161	0,6421	0,3624	0,3624	0,591	0,60	0,521	0,597
Methanaminium, N-[bis[4-(dimethy	0,8076	0,7509	0,4895	0,4895	0,593	0,60	0,521	0,597
Methanone, (tricyclo[3.3.1.1(3,7)]d	0,7132	0,6196	0,4532	0,4532	0,594	0,60	0,522	0,598
Methyl-7.alpha.,12.alpha.-diaceto	0,7626	0,7179	0,5594	0,5594	0,594	0,61	0,522	0,598
METHYLUNDECYLPIPERIDINE, TRANS	0,7551	0,6789	0,384	0,384	0,594	0,61	0,524	0,598
Mitoxantrone	0,6239	0,6129	0,5635	0,5635	0,595	0,61	0,524	0,599
MNQ	0,7198	0,6415	0,4225	0,4225	0,595	0,61	0,524	0,599
Morpholine, 4,4'-(3,4-dichloro-2,4-	0,7854	0,7803	0,4573	0,4573	0,596	0,61	0,525	0,599
Morpholine-N-dithiocarbamate	0,74	0,6381	0,3755	0,3755	0,597	0,61	0,526	0,599
Musennin	0,7494	0,7129	0,6258	0,6258	0,597	0,61	0,526	0,6
MYCOTRIENINE-LIKE	0,6457	0,5567	0,4655	0,4655	0,597	0,61	0,526	0,6
Myricitin	0,6762	0,6259	0,5737	0,5737	0,598	0,61	0,526	0,602
N,N' Bis{2-(4-phenylene)benzimida	0,7189	0,6852	0,4245	0,4245	0,598	0,61	0,526	0,602
N,N-bis(2-chloropropyl)-N-methyla	0,8559	0,7967	0,4571	0,4571	0,598	0,61	0,526	0,602
N-(1H-benzo[d]imidazol-2(3H)-ylid	0,6838	0,5821	0,4899	0,4899	0,599	0,61	0,526	0,603
N-(4-Amino-4-deoxy-N10-methypte	0,5972	0,5928	0,5694	0,5694	0,599	0,61	0,527	0,603
N-(6'-Chrysenyl)-4-(1'-piperidinyl)-	0,7105	0,6448	0,4689	0,4689	0,6	0,61	0,527	0,604
N-Ethyl-1,2,3,4-Tetrahydro-6,7-dim	0,7865	0,7988	0,382	0,382	0,601	0,61	0,527	0,604

N-[p-[N-(2,4,7-triamino-6-pteriddir	0,6878	0,6806	0,5323	0,5323	0,604	0,61	0,528	0,605
NANAOMYCIN	0,6196	0,5128	0,4315	0,4315	0,604	0,61	0,528	0,606
Naphthalene-1,2-dione, 1,2-dihydr	0,773	0,7743	0,4092	0,4092	0,604	0,61	0,528	0,606
Naphtho[2,1-b]furan-6,9-dione, 7-(	0,6462	0,5589	0,4244	0,4244	0,605	0,61	0,528	0,607
Naphtho[2,3-d]thiazole-4,9-dione,	0,7888	0,7327	0,4843	0,4843	0,605	0,62	0,529	0,608
Naphth[2,3-d]thiazole-4,9-dione, 2	0,7368	0,6971	0,5185	0,5185	0,606	0,62	0,53	0,608
Neocuproin	0,6538	0,5797	0,5653	0,5653	0,606	0,62	0,53	0,608
Nitrotolazoline	0,802	0,7193	0,5181	0,5181	0,606	0,62	0,531	0,609
Nonanamide, N-(2-aminoethyl)- N-	0,6629	0,5587	0,4	0,4	0,606	0,62	0,531	0,609
null	1	1	1	1	0,607	0,62	0,531	0,611
OLENDRIGENIN	0,7249	0,6974	0,5477	0,5477	0,607	0,62	0,531	0,611
Oligomycin A	0,6664	0,6063	0,4878	0,4878	0,608	0,62	0,532	0,612
Olomoucine	0,602	0,5057	0,354	0,354	0,608	0,62	0,532	0,613
Oltipraz	0,658	0,6571	0,4032	0,4032	0,609	0,62	0,532	0,614
Ossamycin	0,5743	0,5649	0,62	0,62	0,61	0,62	0,533	0,614
Oxalic acid, mono(4,9-dihydro-4,9	0,8368	0,7764	0,4864	0,4864	0,611	0,62	0,534	0,615
Pap H	0,5674	0,5141	0,5915	0,5915	0,611	0,62	0,535	0,615
Petriosamine A	0,6814	0,592	0,4579	0,4579	0,611	0,62	0,535	0,615
Phenol, 2,2'-methylenebis[5-chlorc	0,7056	0,6556	0,4667	0,4667	0,612	0,62	0,536	0,615
Phenol, 4,4'-methylenebis[2-[(prop	0,7698	0,7014	0,442	0,442	0,612	0,62	0,537	0,616
Phenol, 4-(5,6-dimethyl-1H-benzim	0,7	0,5738	0,4141	0,4141	0,613	0,62	0,537	0,616
Phenothiazine, 2-azido-10-[4-(4-m	0,6874	0,6152	0,3951	0,3951	0,614	0,63	0,538	0,617
Phosphinic amide, P,P-bis(1-aziric	0,7165	0,6861	0,6048	0,6048	0,614	0,63	0,538	0,618
Phosphorodiamidic acid, N,N-bis(2	0,7602	0,6984	0,57	0,57	0,614	0,63	0,538	0,619
Phosphorodiamidic acid, N,N-bis(2	0,7595	0,719	0,5422	0,5422	0,614	0,63	0,54	0,62
PHYTOENE	0,7113	0,6361	0,4942	0,4942	0,616	0,63	0,54	0,62
Pimozide	0,671	0,6078	0,5043	0,5043	0,616	0,63	0,541	0,621
Piperazine-2,2,5,5-tetracarbonitril	0,7828	0,7071	0,4198	0,4198	0,616	0,63	0,541	0,621
Piperidine, 1,1'-(3,4-dichloro-2,4-c	0,7463	0,6972	0,4661	0,4661	0,616	0,63	0,541	0,621
Piperidine, 2,6-diphenyl-1-methyl-	0,6637	0,616	0,5082	0,5082	0,617	0,63	0,541	0,623
Propanedinitrile, 1H-indazol-6-ylh	0,7126	0,6225	0,4192	0,4192	0,618	0,63	0,542	0,624
Propanimidoyl chloride, 2-methyl-	0,6843	0,5937	0,3843	0,3843	0,618	0,63	0,544	0,624
Propanimidoyl chloride, 2-methyl-	0,7489	0,6564	0,3909	0,3909	0,619	0,63	0,545	0,624
Propanimidoyl chloride, 2-methyl-	0,7596	0,6954	0,4055	0,4055	0,62	0,63	0,546	0,625
Propanimidoyl chloride, 2-methyl-	0,7271	0,6678	0,3638	0,3638	0,62	0,64	0,546	0,626
Propanimidoyl chloride, N-[[[(3,4-(	0,7831	0,7199	0,4414	0,4414	0,621	0,64	0,546	0,626
Propanimidoyl chloride, N-[[[(3-ch	0,7262	0,6504	0,3801	0,3801	0,622	0,64	0,546	0,626
Propanimidoyl chloride, N-[[[(4-br	0,7638	0,7099	0,4279	0,4279	0,622	0,64	0,547	0,627
Propanimidoyl chloride, N-[[[(4-me	0,6896	0,6087	0,3935	0,3935	0,622	0,64	0,547	0,627
Propanimidoyl chloride, N-[[[(5-ch	0,67	0,5896	0,422	0,422	0,622	0,64	0,548	0,628
Propanoic acid, 3,3,3-trifluoro- 2-[	0,6412	0,5846	0,5433	0,5433	0,624	0,64	0,548	0,628
Pyrazino[1,2-a]benzimidazole, 1,3-	0,6308	0,5609	0,4299	0,4299	0,624	0,64	0,548	0,628
Pyrazino[1,2-a]benzimidazole, 3-(4	0,6507	0,5745	0,3971	0,3971	0,624	0,64	0,548	0,628
Pyridine-3-carbonitrile, 6-(4-chlor	0,7249	0,6476	0,5163	0,5163	0,625	0,64	0,549	0,629
Pyridone-OCH3	0,7714	0,7202	0,4836	0,4836	0,626	0,64	0,549	0,631
Pyrimidin-2-amine, 4-(2-hydroxy-4	0,7269	0,6871	0,4747	0,4747	0,628	0,64	0,552	0,631
Pyrimidin-2-amine, 4-(2-hydroxy-4	0,7578	0,6686	0,4255	0,4255	0,629	0,64	0,552	0,631
Pyrimidine, 2-[1,5-bis(4-chlorophe	0,6912	0,6406	0,4013	0,4013	0,63	0,64	0,554	0,631
Quinoline-3,3,4,4-tetracarbonitrile	0,6581	0,495	0,389	0,389	0,631	0,65	0,554	0,634