Chapter 1

Introduction

In this thesis an analysis of algorithms for the set covering problem are presented. The main motivation is DNA typing, a problem from the field of bioinformatics. A mathematical formalization of DNA genotyping will be shown. It is analyzed with more detail in [7]. Some definitions will be given in order to precise some points of the problem, and how it can be seen as an NP complete problem.

1.1 Bioinformatics context

The DNA sequence analysis is a way to differentiate among species. DNA typing is of world wide importance. Techniques have appeared to unravel the information contained on it. Bioinformatics is a relatively new branch of biology whose initial intention was to acquire, store and organize the biological information contained on DNA molecule. Nowadays, its purpose extends to the analysis and interpretation of data. It involves complex problems solution, using tools from computation and informatics technology. Diagnosis and sickness treatment are some of the most important applications inside this area [7].

The DNA is a molecule discovered since XIX century. However, it was not until

1940 when the relationship between it and well known “genes” was demonstrated. The physical structure of the genetic entities was discovered 12 years later. [1]

DNA is a relatively long molecule, with a repetitive structure. It is built of nucleotides linearly linked together. Each nucleotid contains a phosphate group, a deoxyribose sugar and one of four nitrogen bases: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). It is well known that these bases are complementary to each other forming pairs, A-T and C-G. This fact allows DNA replication; this is, from one strand, the other can be synthesized.

This molecule can be seen as a chain formed from an alphabet of four letters: $\{A, G, C, T\}$. The order of the bases is unique for each species or individual. Therefore, DNA identification is very useful for the diagnosis of diseases caused by infectious agents. The typing may rely either on restriction patterns, a technique named Restriction Fragment Length Polymorphism (RFLP), or on a hybridization profile using an array of probes.

An example of how important is the typing of related sequences may be found in cervical cancer. This disease was the first human cancer where a virus was identified as cytologic cancer. Because of the Cytologic analysis, the progression from virus infection to invasive cancer could be diminished, and the early diagnosis has been improved. However, there is still a high percentage of women that are affected by this infection, and it is important to assign resources towards special women groups. Special risk groups are advanced-age women, ethnic minorities and generally women coming from low resource status. Additional independent risk factors are smoking, frequently changed sexual partners and taking one contraceptive [4].

Papilloma Virus (HPV) has been found to be a determinant factor in cervical cancer. Actually HPV is a family of viruses, each of them with different DNA sequences. They can be grouped into types and subtypes. Depending on the type there is a higher or lower risk of carcinogenic change in the cells of the uterine neck. For example, the HPV-16 and HPV-18 are high risk viruses, whereas HPV-6 and HPV-11 are low risk types, and of others like HPV-92 the risk has not yet been determined. This fact shows the importance of determining the viral type. Not only does it help in the initial diagnosis but also in the therapeutic treatment. [4]

The definition of the viral genotype in the case of HPV is a function of similarity percentage between patterns. The output is whether both viruses belong to the same type or subtype. This is:

$$\%changes(a,b) = \begin{cases} \geq 10\% & a \text{ and } b \text{ are different types,} \\ \geq 2\%, < 10\% & a \text{ and } b \text{ are different subtypes,} \\ < 2\% & a \text{ and } b \text{ are the same subtype.} \end{cases}$$

Basically, the technique known as RFLP implies that DNA is “digested” by a restriction enzyme. Each enzyme recognizes a predefined sequence of nitrogen bases, for example, it can recognize the string “CGTGC”. When an enzyme recognizes a sequence, it cuts the molecule in the middle or near the recognition sequence, depending on the enzyme.

Once the DNA is “digested”, electrophoresis is used to separate the generated fragments. A difference of charge is applied, causing migration of the DNA fragments. This migration is marked in a gel. Because small bands have less difficulty to migrate to the positive pole, the size of the sequence can be derived from the distance the bands have migrated. A pattern of bands named restriction pattern is obtained. It helps to differentiate sequences from others.

The problem in practice is that differentiation between sequences can not be done with only one enzyme [6]. Different DNA sequences can generate the same pattern for a particular enzyme either by chance or due to the similarity between these sequences. In general, all the typing techniques use reagents to identify sequence features and these reagents have a different capacity to distinguish between pairs of sequences. Complexity of the problem is introduced at this point. Nowadays there are 90 types of HPV viruses known, but the number continues growing. There are about 220 different enzymes. The main problem can be enunciated as follows: Given a set T of viruses and a set E of enzymes, which is the minimum subset of enzymes that allow to distinguish all types of viruses in T?

1.2 Problem formalization

In this section, a formal exposition of the problem is presented. There are some definitions that help to understand the typing problem and its relation with the set covering problem. The instance RFLP of the problem will be still used to help with the explanation; however, as it was mentioned, the problem can be generalized to any sequences typing technique.

Given a set X, a metric d is the distance between two points [7]. This distance can be discretized, for example, into a binary metric. In the specific problem of similarity between restriction patterns, the metric used has been:

- Squared migration difference between virus fragment pairs.
- Absolute value of the sum of migration differences.
- Maximum value among absolute values of these differences.

In this case, a metric can not be exact because laboratory analyses are prone to precision errors. Therefore, a threshold h is defined. If the distance is less or equal than this threshold, both patterns are considered equal. Otherwise, they are considered different.

Definition 1.1. Let S be a set. A binary relation in S is any set $C \in S \times S$. It is denoted as xRy, this is, the pair $< x, y > \in R$. An equivalence relation $R \in C$ is a relation with some properties as follows:

1. Symmetrical. This is, $< x, x > \forall x \in S$.

2. Reflexive. This is, $< x, y > \in S \rightarrow < y, x > \in S$.

3. Transitive. $(< x, y > \in S \wedge < y, z > \in S) \rightarrow < x, z > \in S$. [17]

In RFLP, the relation is not transitive in principle. Let a and b be two patterns taken as equal, and let c be a third pattern found similar to b with a length below a considered threshold, and then it is equal to b. In other words, $< a, b > \in S$ and $< b, c > \in S$ but $< a, c > \notin S$.

If there is an equivalence relation, then a partition over set T of types can be gotten. From a set of data, a distance matrix for each enzyme was constructed. The binary complement of this matrix has the property of being reflexive, transitive and symmetrical.

Matrices are created for each enzyme. Rows and columns are identified as different patterns. The size of the matrix is $n \times n$ where n is the number of patterns. A value of 1 is assigned to an entry when the enzyme differentiates between the patterns, and a value of 0 is assigned otherwise; in this way we can build the complement of the matrix.

Example 1.1. Let A and B be two enzymes, which help to separate the patterns in the following ways:

Enzyme A: $\{\{a, b\}, \{c, d\}\}$ Enzyme B: $\{\{a\}, \{b, c, d\}\}$

There is an equivalence relation, including also the symmetric pairs $< x, x >$:

A: $\{< a, b >, < c, d >\}$

B: $\{< b, c >, < b, d >, < c, d >\}$

The complement of binary distance matrices for this example result as follows:

Matrix A

	a	b	c	d
a	1	1	0	0
b	1	1	0	0
c	0	0	1	1
d	0	0	1	1

Matrix B

	a	b	c	d
a	1	0	0	0
b	0	1	1	1
c	0	1	1	1
d	0	1	1	1

The objective of differentiation between types of virus can be defined as an operation of $M1 \wedge M2$, where $M1$ and $M2$ are the binary inverse of distance matrices for two different enzymes. The symbol $\wedge$ is the logic operation *and*. This operation represents a partition generated by two different enzymes and the final differentiation obtained from combining both enzymes.

The problem can be redefined as follows: To find the minimum set of matrices such that the output of the *and* operation between them is the identity matrix. In other words, this set of enzymes produces the finest partition over set T, this is, the genotypes.

This problem is NP-complete kind. To explain this, there will be defined P and NP problems. Problems are named P when solution time is a polynomial function in a deterministic machine, depending on the size of the input. NP problems are those whose solution can be verified in polynomial time. Observe that P problems are also NP. The inverse implication has not been proved yet, it is an open problem nowadays [21]. More detailed definitions related to P and NP can also be found in [21].

NP complete problems are those whose solution in less than exponential time has not been found in a deterministic machine. It is suspected that there is no such solution. Polynomial approaches towards this kind of problems are very important. Although there is a lost in precision of the solution, reasonable running times are obtained.

The decision problem associated to the "set covering", and the generalization to the "weighted set covering", are NP-complete problems too. A characteristic of this kind of problems is that they can be reduced one to each other in a polynomial time. The problem of minimum multiplication of matrices can be reduced to set covering and vice versa, as it is shown in [6]. This is very useful because of the amount of analysis and information about set covering. There are a lot of approaches and exact algorithms, whether they have a reasonably bounded error or several reductions and prunes that improve time in the average case.

In the sequence typing problem, there is an unexplored aspect, this is, the case that there is no interest to distinguish virus from others. For example, between subtypes of virus (this is, when the change percentage is $> 2\%$ and $< 10\%$). Another case is when the viruses have the same degree of risk. This specialization can be approached with generalized set covering or weighted set covering. However, this work will focus on simple set covering applied to distinguish among the whole set of types.

Up today, the algorithms analyze the problem from different angles:

- There are algorithms using a solution vector and getting the cost of weight for each subset. Originally the subsets are or are not in the solution. In this approach constraint relaxation is used, and the subset can take any real value in the range [0, 1].

- Greedy algorithms. These algorithms look for the best solution in each iteration. In this case, the subset covering the largest number of elements with the minimum cost is selected.

- Neural networks. There are units representing the rows of inequalities and others containing the variables or costs, these are the columns. This neural network is connected as a bipartite graph representing incidental relations between rows and columns in a given matrix A.

Moreover, there is a "greedy" algorithm, applied directly to the problem of matrix multiplication. It uses Entropy concept, from Information theory, to select the best partition each time.

Generalization of set covering to weighted set covering is also applicable to get quality in the selected partitions. In this case, the weight associated to the subsets is the minimum. When the weight has a value of 1 is the simplest form of the problem.

1.3 Algorithms for set covering

As the set covering algorithm is a very old and studied problem, it will be useful to solve typing sequences problem mapping matrices to an instance of set covering. Design and test algorithms for set covering on instances of DNA typing will be helpful to optimize the solution. On the other side, there can be found some properties of these algorithms to compare them with analyzed polynomial solutions to set covering problem.

Therefore, the problem becomes a problem of examining algorithms to solve set covering. Two paths can be followed: an exact algorithm and polynomial algorithms. The exact algorithm is interesting because some properties about pruning and time can be found. The main disadvantage is that it does not give a different perspective. In the worst case, it becomes an exhaustive search. However, an exact algorithm contributes with results to the theory of NP-completeness.

On the other hand, polynomial algorithms can be analyzed. In this case, an exact algorithm can be used to compare solutions given by these algorithms, but only with small inputs. Polynomial algorithms are the realistic way to solve a real NP complete problem. They are also interesting because there is an ample range of approaches to be explored. Moreover, some properties can be found in order to learn more about when some approach or another can be applied. In this text, an exact algorithm and two polynomial algorithms are presented, as well as some results from the implementation. The main objective is to report results about how all the algorithms work, in order to apply them to the genotyping of DNA sequences problem.

In the next chapter, Chapter 2, are presented main concepts concerning to algorithms and complexity. The set covering problem will be inserted in a framework next to some related research about it. Chapter 3 explains the algorithms with more detail, as well as the implementation. Chapter 4 presents results. Results are divided in statistics and specific results for the algorithms, like some examples where an algorithm is better than others. Finally, Chapter 5 presents a summary of the work and some ideas about interesting paths to explore in the set covering problem.