pAin Shams University

Faculty of Science

Department of Mathematics

A Study on Bioinformatics Algorithms

A thesis submitted for the award of

Ph.D. degree in Science (Computer Science)

By:

Mohammad Hashim Ali Abd El-Rahman

B.Sc. Pure Math. & Computer Science.

M.Sc. Computer Science.

Supervised by:

Prof. Dr. Abdel Badeeh M. Salem
Prof. of Computer Science,
Department of Computer Science,
Faculty of Computer & Information
Sciences, Ain shams University

Prof. Dr. Fayed F. M. Ghaleb
Prof. of Mathematics,
Department of Mathematics,
Faculty of Science,
Ain Shams University

Prof. Dr. Aliaa A.A. Youssif
Prof. of Computer Science,
Department of Computer Science,
Faculty of Computers and Information,
Helwan University

Submitted to:

Faculty of Science,

Ain Shams University,

Cairo - Egypt.

2011.

جامعة عين شمس
كلية العلوم
قسم الرياضيات

دراسة عن خوارزميات المعلوماتية الحياتية

رسالة مقدمة للحصول علي
درجة دكتوراه الفلسفة في العلوم
(علوم الحاسب)

مقدمة من

محمد هاشم علي عبد الرحمن

مدرس مساعد بقسم الرياضيات
كلية العلوم – جامعة عين شمس

تحت إشراف

ا.د. عبد البديع محمد سالم
أستاذ علوم الحاسب
قسم علوم الحاسب – كلية الحاسبات
والمعلومات – جامعة عين شمس

ا.د. فايد فائق محمد غالب
أستاذ الرياضيات
قسم الرياضيات – كلية العلوم
جامعة عين شمس

ا.د. علياء عبد الحليم يوسف
أستاذ علوم الحاسب – قسم علوم الحاسب
كلية الحاسبات والمعلومات – جامعة حلوان

مقدمة إلي

كلية العلوم
جامعة عين شمس

2011

Abstract

Gene silencing is one of the hottest problems in bioinformatics. This problem deals with how to detect and suppress genes responsible of producing unusual harmful proteins that cause some diseases and cancers.

In the present study, new algorithms are suggested to solve the gene silencing problem. The proposed algorithms are based on the hashing technique, which makes them faster than the previously suggested algorithms and also consuming fewer memory spaces. Therefore, the proposed algorithms could be applied to Human genome and other organisms with large genome sequences.

Besides, one of the proposed algorithms has a new advantage which is not included in the previous algorithms. This advantage is its capability of testing the control of a specific gene, which gives the proposed algorithm a higher medical importance that helps in detecting treatments for specific diseases.

CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

Bioinformatics is a multi-disciplinary field, at the intersection of biology, chemistry, computer science and mathematics. Bioinformatics is a rapidly evolving field, driven by advances in high technologies that result in an ever increasing variety and volume of experimental data to be managed, integrated, and analyzed [64].

Since the determining of DNA structure in 1953 by James Watson and Francis Crick [92] which is considered as a milestone achievement; biology, especially molecular biology, has grown by leaps and bounds. The sequencing of the human genome represents one of its triumphs; the sequencing of dozens of other organisms has followed. Most of these successes would be unthinkable without computers [21].

1.1 Bioinformatics Definition

Bioinformatics is sometimes called Computational Biology or Computational Molecular Biology. Bioinformatics is a multi-disciplinary science focusing on the applications of computational methods and mathematical statistics to molecular biology. This union between the two subjects is attributed to the fact that life itself is an information technology; an organism's physiology is largely determined by its genes, which at its most basic can be viewed as digital information [62]. Computational biology is being developed by computer scientists to satisfy the needs of biologists but basically requires extensive knowledge of computer science theory.

The US National Institutes of Health defined Bioinformatics as follows:

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data. Simply, Bioinformatics is any activity that deals with biological data using computational tools.

1.2 The Need for Bioinformatics

Biologists need bioinformatics to make data easily and universally interpretable by scientists. Due to efficient biotechnological methods for gathering biological data, the amount of available data currently grows much faster than the growth in available computational power. This let Anthony Kervalage to say that an experimental laboratory can produce over 100 gigabytes of data a day with ease. Thus, to make sense of the collected data; efficient algorithmic and computational techniques are required, and bioinformatics activities are expanding rapidly in both academia and industry.

On the other hand, some computer scientists have emphasized the importance of biology and its connectedness with computer science. Their views suggest that biological problems will significantly influence future directions in computer science research, including, for example, DNA computing. Thus, Donald Knuth [55] anticipates that the number of radically new results in pure computer science is likely to decrease, while scientists will continue working on biological challenges for the next 500 years. Also, Leonard Adleman [2] has argued that biological life can be equated with computation.

1.3 Aims of Bioinformatics:

Rapid developments in biotechnology have uncovered the complete sequences of numerous genomes, and many other genomes are to be mapped out in the near future. However, knowing the sequences of the genomes is only the start. In the pro-genome era, we would like to study the activities of genes within a cell.

So, a major goal in molecular biology is functional genomics, or the study of the relationships among genes in DNA and their function. Gene function can be viewed through several prisms. A common interpretation is that function describes the role of a gene product, usually a protein, in reacting with other proteins in a metabolic or signaling pathway. However, molecular biologists know that protein interactions are dependent on protein structure or shape.

Biologists and computer scientists may conclude that the ultimate objective of functional genomics is: Given the DNA of an organism, produces a simulator for a cell of that organism. That simulator (or flowchart representing metabolic and signaling pathways) embodies all that it knows about a cell's behavior, allowing in-silico experiments that enable biologists to bypass costly and ethically sensitive in-vitro or in-vivo trials. We are far from this goal, but it is an area where computer science can provide considerable research impetus.

For inferring function from the existing data, a biologist must consider three factors:

- Genes, or substrings of DNA capable of generating proteins.
- Protein structures represented in 3D space.
- The roles of these proteins within metabolic and signaling pathways.

Since data about protein shape and pathways is often unavailable, the detective work in bioinformatics consists of deducing possible function from existing information, even if it is limited. A typical example is a human gene for which there is no known protein structural data or pathways. However, corresponding data may be available for other organisms (such as the mouse and the worm). Inferring function from them might save biologists much tedious laboratory work. In addition, the inferred data might suggest key experiments that would help formulate a conjecture.

Thus, to achieve the major goal of bioinformatics, which is the complete understanding of an organism given its genome, there are threefold aims [62]

- Organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced, e.g. the Protein Data Bank for 3D macromolecular structures.
- Develops tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences, and this needs to consider what comprises a biologically significant match. Development of such resources dictates expertise in computational theory as well as a thorough understanding of biology.
- To use these tools to analyze the data and interpret the results in a biologically meaningful manner.

1.4 Branches of Bioinformatics

Some of the important branches in bioinformatics are:

- *Structural Bioinformatics*: the complete understanding of an organism given its genome.