Palestine Polytechnic University

College of Engineering & Technology
Electrical and Computer Engineering Department

Software Project

"Twin Gene"

Project Team

Ala Abu Al_Reesh
Gharam Al-Skafi
Nida Al-Amlh

Project Supervisor
Dr.Mahmood Al-saheb

Hebron – Palestine Polytechnic University

June, 2004

Palestine Polytechnic University

College of Engineering & Technology
Electrical and Computer Engineering Department

Software Project

"Twin Gene"

Project Team

Ala Abu Al_Reesh
Gharam Al-Skafi
Nida Al-Amlh

Project Supervisor
Dr.Mahmood Al-saheb

Hebron – Palestine Polytechnic University

June, 2004

Palestine Polytechnic University

Hebron – Palestine

"TwinGene"

By

Ala Abu-Reesh

Gharam al-Skafi

Nida Al-Amlh

According to the directions of the project supervisor and by agreement all of the committee members, this project was submitted to Department of Electrical and Computer Engineering in College of Engineering and Technology to partially fulfill to the Bachelor requirements for the department.

Supervisor Signature

Name:……..

Dept. Head Signature

Name:…….

Name:……..

Committee Members

1. Name:…….
2. Name:…….

ABSTRACT

Computer science has the characteristic that it is related to all fields of our life, one of these fields is the medical field; computer and medical fields were collected in the last few five years into bioinformatics science, so bioinformatics is the science that studies the relation of computer and medical fields. From here we decided to invest our knowledge in the bioinformatics science, so we decided to work in the field of genes. The main important problem that faces experts in medical field is that they sometimes discover that two genes are the same gene or that one gene has another similar gene that scientists think it is the same gene, so we worked in this project on the alignment of genes that is bringing similar genes for the query gene then writing each gene in one line based on the alignment rules and then dividing the genes into groups based on the correlation in differences with the query gene.

Dedication

To any one appreciates knowledge

To every one helped us

To our families and friends

Ala'

Gharam

Nida'

Acknowledgement

Our sincere thanks and appreciation to the person who gave us his guidance and support to complete this project, specially our instructor Dr. Mahmood Al-saheb and Dr Yaqub Al_Ashab.

Thanks to everyone shared in the success of the project.

Table of Contents

CHAPTER ONE

INTRODUCTION

1.1 Overview

1.2 Project Importance

1.3 What is Twin Gene

1.4 System Objectives

1.5 The Report Outline

1.1 Overview

Computer science is related to all sciences and to all aspects of our life; its relation to biology science is strengthened by Bioinformatics which is defined as conceptualizing biology in terms of molecules and then applying informatics techniques (derived from disciplines such as applied math, computer science and statistics) to understand and organize information associated with these molecules [1].

1.2 Project Importance

As the number of the projects that deal with the organization of biological information have grown explosively in the recent years the Bioinformatics research appeared which is a newly emerging interdisciplinary research area and can be defined as the interface between biological and computational sciences. Thus, the people working in this field in most cases either have a training in biology or computer science, and they learned about the other field by dealing with problems or using the tools of the other one.

Twin Gene System allows the user to make some processing on a query gene such as alignment with similar genes, alignment with one gene and determining the groups of the genes contained in one database of genes.

1.3 What is Twin Gene

Twin Gene system can be defined as the system that divides the genes into groups. It accepts a huge database taken from Gene Bank and with complex comparisons, it finds all the similar EST's (Expressed Sequence Tags) for the query Gene (using public program called BLAST). The BLAST output should be converted into a multiple alignment (MA) file. A multiple alignment arranges a set of sequences in a scheme where positions believed to be homologous (similar) are written in a common column, then it process the MA file to determine the Twin Gene (divided into groups based on correlation of differences with the query gene).

1.4 Project Objectives

1. Learn how to design and implement twin gene in a system software project.
2. To connect the biological field to the computer science.
3. To explore and evaluate the phenomenon of pseudogene expression in Human Genome by checking the prevalence of paralogous EST's that are misannotated to one gene.
4. To provide solutions for designing molecular techniques that fit genes with paralogous copy with high sequence similarity.
5. The identification of new unknown copy or copies of a given gene will lead us to more accurate mapping of the different genes in the Human Genome.

1.5 The Report Outline

This report shows all important points related to the project, this documentation helps the reader to get comprehensive view about the project.

We tried to present a documentation that can be understood easily by dividing the documentation to a set of subject related parts.

The report consists of a number of chapters. Each chapter discusses a subject related to the project, and these chapters are sorted according to the logical relations of subjects.

Chapter One: "Introduction"; gives overview about the project, project importance, and Twin Gene system objectives.

Chapter Two: "System planning"; presents system risks, development requirements, cost estimation, operational Requirements and time scheduling

Chapter Three: "System Requirements"; it includes definition, analyses and specification.

Chapter Four: "System Design"; consists of the system design model which includes the data flow diagram, structural chart and the data and control analysis.

Chapter Five: "System Implementation"; presents general view; prepares the platform of the system.

Chapter six:" System Testing"; shows how the system is going to be tested.

Chapter seven: "Future Work and Conclusions"; which is in the last chapter, it shows the suggestions for the future development, and points out the conclusions.

CHAPTER TWO

SYSTEM PALNNING

2.1 Overview

2.2 System Risks

2.3 Development Requirements

2.4 Cost Estimation

2.5 Operational Requirements

2.6 Time Scheduling

2.1 Overview

This chapter talks about system risks, hardware and software requirements and time scheduling.

2.2 System Risks

1. This is the first project in this field so we have not previous projects to help us.
2. Because this project contains biomedical concepts, so we need specialists in this field to help and teach us many concepts.
3. The size of the output file is so large because the database is so large and may reach to nine compact CD's or higher, as a result we may display we display samples of the output.

2.3 Development Requirements

Development Requirements includes hardware resources, software resources and the human resources.

2.3.1 Hardware Development Resources

As system needs complex searching and comparisons to find a Twin Gene, we need a professional languge to help our for the complex comparisons that we need for our project and in this project we use Visual Basic.NET , so we need a computer that can implement this system; this computer should have the following:

Requirement	Required	Recommended
Processor	PC with a Pentium II-class processor, 450 MHz	Pentium III-class, 600MHz
Ram	Windows XP Professional: 160 MB	192 MB for XP Professional
Hard disk	2 Giga	10Giga
Video card	800 x 600, 256 colors	High Color 16-bit

Table 2.1 Hardware development resources [2].

2.3.2 Software Development Resources

The PC computer should contain the following operating system and software:

Operating system	Windows XP professional
Programs	Visual studio.net (2002)

Table 2.2 Software development resources [2].

2.4 Cost Estimation

This part contains the estimation of the development costs and the implementation costs.

2.4.1 Development Cost

Hardware Costs:

Hardware Component	Cost
Petium3_600Mhz	4000$

Table (2.3): Development Hardware Costs

Software Costs:

Software Component	Cost
Windows XP	299 $
Visual Basic .NET	1,794$
Total	2093

Table (2.4): Development Software Costs

Human Resources Costs:

Number of Developers	Work Hours/day (120 days)	Cost per hour for one Developer	Total
3	5	20$	36000$

Table (2.5): Development Human Costs

Hardware Dev Costs	Software Costs	Human Dev. costs	Total
400$	2093$	36000$	38493$

Table (2.6): Total development Costs

2.5 Operational Requirements:

Required Processor	Recommended Processor	Required RAM	Recommended RAM	Programs	OS
Pentium 90 MHz	Pentium 90 MHz or faster	32 MB	96 MB or higher	Microsoft Internet Explorer 5.01 .Net framework	Win 98, win me or higher

Table 2.7 Operational Requirements [2]

2.6 Time Scheduling

Our project has to be developed during 17 weeks . figure 1.1 shows the time plan of the project .

Task / week	1	2	3	٤	٥	٦	٧	٨	٩	١٠	١١	١٢	١٣	١٤	١٥	١٦
System definition & planning.																
Requirements , specification,& analysis																
System Design																
Implementation																
Testing																
Documentation																

Table 2.8: Time Scheduling

INDEIECES

INDEX A

Theoritical Background

List of Tables

List of Figures

DNA (deoxyribonucleic acid) and proteins are biological macromolecules built as long linear chains of chemical components. In the case of DNA these components are the so-called nucleotides, of which there are four different ones, each denoted by one of the letters A, C, G and T.

DNA	adenine	guanine	cytosine	thymine
	A	**G**	**C**	**T/U**
RNA	adenine	guanine	cytosine	uracil

So the nucleotides can be made up of 20 different amino acids (or "residues") which are denoted by 20 different letters of the alphabet.

	One-letter code	Three-letter-code	Name
1	A	Ala	Alanine
2	C	Cys	Cysteine
3	D	Asp	Aspartic Acid
4	E	Glu	Glutamic Acid
5	F	Phe	Phenylalanine
6	G	Gly	Glycine
7	H	His	Histidine
8	I	Ile	Isoleucine
9	K	Lys	Lysine
10	L	Leu	Leucine
11	M	Met	Methionine
12	N	Asn	Asparagine
13	P	Pro	Proline
14	Q	Gln	Glutamine
15	R	Arg	Arginine
16	S	Ser	Serine
17	T	Thr	Threonine
18	V	Val	Valine
19	W	Trp	Tryptophan
20	Y	Tyr	Tyrosine

To compare between different nucleotides they insertions into the chain and deletions from the chain. The elementary operations allowed in the definition of sequence similarity are chosen to correspond to these events. To visualize the relationship between two similar sequences they are represented in the form of an alignment:

```
V-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DL    HAHU
VHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDL    HBHU

SH-----GSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRV    HAHU
STPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHV    HBHU

DPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR      HAHU
DPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH      HBHU
```

The two amino acid sequences compared here are the alpha chain of human hemoglobin (abbreviated HAHU) and its beta chain (HBHU). With the sequences being approximately 150 amino acids long, each block of lines contains part of the first sequence in the upper and of the second sequence in the lower line. Residues on top of each other in one block are equivalenced. Some residues are conserved (the amino acids in the column are identical), some have been exchanged and part of the chain has been deleted from the one sequence or (equivalently) inserted in the other. Insertions or deletions are indicated by a letter paired with a dash, the gap-character. An alignment can also be interpreted as representing the operations necessary to transform a sequence into another one using the same operations as evolution does.