A Survey on Mathematical Modeling of Cancer Incidence Rates

Marzieh Eini Keleshteri

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Applied Mathematics and Computer Sciences

Eastern Mediterranean University
August 2011
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Applied Mathematics and Computer Sciences.

Prof. Dr. Agamirza Bashirov
Chair, Department of Applied Mathematics
and Computer Sciences

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Applied Mathematics and Computer Sciences.

Asst. Prof. Dr. Mehmet Ali Tut
Supervisor

Examining Committee

1. Prof. Dr. Nazım Mahmudov
2. Asst. Prof. Dr Mehmet Ali Tut
3. Asst. Prof. Dr Arif Akkeleş

ABSTRACT

Bioinformatics is a novel interdisciplinary field which attempts to response the biological questions by the assist of other basic sciences as well as computer sciences. Cancer modeling is a real example of such these endeavors in order to help oncologists to find new ways to cure and prevent cancer diseases or predict, estimate, and analyze this hazard in order to step forward to a better future. In this arena mathematics and statistics have played great roles and enabled biology, oncology, and epidemiology to achieve new results by applying some mathematical and statistical methods such as various graphs and tools to compare the different criteria, curve fitting as well as analyzing and predicting the future data, time series and Markov processes to model the natural phenomena and study their behaviors. However this field is still like a young sapling which requires enough patience and care of scientists to bring forth.

This thesis is mainly to make a survey generally on bioinformatics and specifically investigations on cancer cases as an application. First chapter provides a general overview about bioinformatics and its application. Then some preliminary concepts are explained including required biological information about cancer and its causes to comprehend the next concepts and implications better. Many scientists have tried to offer an applicable model for cancer incidence rate which can be acceptable and interpretable biologically. Second chapter provides some past findings about cancer incidence models as a background for readers with little biological information. In fact, the focus of this research is mainly on mathematical modeling of cancer

incidence rates. In addition, chapter three is discussing about the cancer incidence. For instance some factors affecting on the process of cancer incidence such as the place and time period of living, sex, race, and the amount of development are checked. In chapter four some curve fittings are performed by MATLAB software, and also special mathematical model which is called Furrier model has been fitted to the real cancer incidence rates data with the best goodness of fit.

Keywords: bioinformaics, biomathematics, cancer, incidence rate, mathematical modeling, curve fitting, Furrier model

ÖZ

Biyoenfoformatik, son zamanlarda gelişen çoklu disiplinlerin içinde barındırıldığı bir başlıktır. Biyolojik veri tabanları üzerindeki bilgilerin incelenmesi ve onlar üzerinde kararların verilmesi oldukça önemli bir aşamayı içermektedir.

Günümüzün en önemli sağlık vakalarından birisi durumundaki kanser olaylarının incelenmesi ve modellenmesi bu araştırma sahasının en önemli uygulamalarından biridir.

Bu tez, biyoenformatik konusunda temel tanımlamaların daha önce yapılmış çalışmaların özetlendiği ve örnek olarak kanser vakalarıyla ilgili verilerin MATLAB paketi yardımıyla modellenmesini içeren bir çalışmadır.

Anahtar kelimeler: biyoenformatik, biyomatematik, kanser, kanser vaka hızı, matematiksel modelleme, Fourier modeli

To My Family

ACKNOWLEDGMENTS

I would like to thank Asst. Prof. Dr. Mehemt Ali Tut for his continuous support and guidance in the preparation of this study. Without his invaluable supervision, all my efforts could have been short-sighted.

Assoc. Prof. Dr. Agamirza Bashirov Chairman of the Department of Applied Mathematics and Computer Sciences, Eastern Mediterranean University, helped me with various issues during the thesis and I am grateful to him. Besides, a number of friends had always been around to support me morally. I would like to thank them as well.

I owe quit a lot to my family who allowed me to travel all the way from Iran to Cyprus and supported me all throughout my studies. I would like to dedicate this study to them as an indication of their significance in this study as well as in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Bioinformatics is a new procedure which helps biologist to manage and analyze the huge amount of data which is gathered during the past decades. It is a discipline that has brought genomics, biotechnology and information technology together and involves biological data analyses, modeling the biological phenomena, the application of computer algorithms and statistics. Bioinformatics is a cross-disciplinary field that started from 1960s with the effort of some scientists like Margaret O. Dayhoff, Walter M. Fitch, Russell F. Doolittle and others. Also bioinformatics has become a strong tool for industrial researchers. They can apply the related techniques in order to produce practical drugs and therapeutic medicines, (Thampi, 2009).

This chapter will introduce some definitions related to bioinformatics which helps the readers to be familiar with it and then some applications of this new field will be mentioned.

1.1 Definition of bioinformatics

The classical point of view defines bioinformatics as to solve the biological problems by using mathematical, statistical and computing methods.

The National Center for Biotechnology Information defines bioinformatics as:

The field of science which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics by which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein management of different types of information. [NCBI2001].

1.2 Definitions of bioinformaticist and bioinformatician

A bioinformaticist is a professional who knows both how to use bioinformatics tools and how to write the programs to increase the efficiency of the tools. On the other hand a bioinformatician is a trained person who only knows how to use bioinformatic tools without a very deep vision. Hence, a bioinformaticist is at a higher level of vision and acknowledgment rather than a bioinformatician. It is similar to comparing the relation between a mechanical engineer and a car with a technician, (Thampi, 2009).

1.3 Bioinformatics research areas

1.3.1 Computational Biology

Computational biology is the study of biological, behavioral and social systems. It focuses mostly on the evolutionary, Population and the theoretical part of biology rather than the practical research on the biological concepts like cell and medicine. In this area, the powerful applied tools are stochastic statistics, mathematical modeling, computational approaches and theoretical methods. To sum up, Computational biology is not only a field, but also an approach which benefits computer and theoretical methods to study the biological phenomena, (Thampi, 2009).

1.3.2 Genomics

It is the study of the genetic components of species. Scientists researching in this field try to calculate the weight and densities of the genome sequence of various living organisms and consequently compare them with each other. In other words, every attempt to analyze inside the genome is called genomics, (Thampi, 2009).

1.3.3 Proteomics

One of the most essential parts of every living organism is protein. It is also the main part of the cell. The entire set of proteins expressed by a genome, cell, tissue or organism is called proteome and the study and research about proteome is called proteomics. Comparing with genomics this field is newer and more complicated. Since the sequence of the genome of an organism almost is constant while its proteome differs correspond to the type of the cell or the time of the research, (Thampi, 2009).

1.3.4 Pharmacogenomics (PGo)

After determining most of the living organism's genome sequence, specially the genome sequence of human being, it is time to apply this science in producing new drugs based on the patients' genome. Pharmacogenomics is the study of the effect of different genes on the patients' drug response. This field still is new and researchers are trying to produce new medicines according to genetic approaches in the laboratories, (NCBI, 2003).

1.3.5 Pharmacogenetics (PGe)

It is a subfield of pharmacogenomics which mainly focuses on the effect of inheritance characteristics, which are different according to the corresponding gene

sequences, on the patients' drug response, (NCBI, 2003).

1.3.6 Cheminformatics

Cheminformatics first was defined by Dr. Frank K. Brown in the Annual Reports of Medical chemistry. He introduces cheminformatics as the mixture of chemical databases and their resources in order to making better and faster decisions in the field of drug identification and organization. Generally it is the application of computer science in chemistry in order to produce new drugs. To do this, it uses data storing, managing, mining, retrieving and analysis, (Opera, 2004).

1.3.7 Biomedical informatics

Informatics is the science of information. Biomedical informatics is the science of information including the studying, invention and implementation medical information in order to improve human beings health. Comparing with bioinformatics, medical informatics is more dealing with structure and algorithms rather than with the data itself, (Bernstam, Smith, & Johnson, 2010).

1.4 Origin and history of bioinformatics

There are different views about the origin of bioinformatics. From one point of view the term bioinformatics was coined in the mid-1980s for the analysis of biological sequence data. (Attwood & Parry-Smith, 1999). From another point of view it is almost new term and was not appeared in the literature until 1991. Based on this opinion bioinformatics is middle aged now. (Boguski, 1998). From a general point of view bioinformatics might start over a century ago with an Austrian monk whose name was Gregor Johann Mendel. He is known as the father of the genetics. His observations on fertilizing different type of plants showed that the inheritance of traits which are different in various generations, obey particular laws. Later in 1972, the first recombinant DNA molecule using ligase was made by Paul Berg.

Simultaneously during that year other scientists produced the first DNA organism. In 1973 an approach for DNA cloning was invented. By 1977, scientists found a way for DNA sequencing and the first genetic engineering company which was named Genetech established. By 1981, 579 human genes were mapped. Later the first method for automated DNA sequencing was invented. In 1988, an international organization for human genome project, Human Genome Organization (HUGO), was established. Finally in 1989, the genome of the bacteria Haemophilus influenza mapped completely for the first time. In the next year human genome project was started. 3 years later in 1993, Genethon, a French research center, produced a physical map for human genome. It was the end of the first phase of human genome project. Bioinformatics was applied more when scientists faced to huge amount of data and decided to gather them inside databases. Different types of genome sequences were stored in various databases such as GenBank, EMBL, (Thampi, 2009).

Nowadays, with the advent of new approaches in informatics, it is very easy to manage the research results and compare the huge groups of data in a very short time. With the advent of the Internet, the management and specially the accessibility of this data increased. The first bioinformatic biological database was created after the availability of the first protein sequence in 1956. Around a decade later, in 1965, Margaret Dayhoff and colleagues gathered all available sequence data and published the first bioinformatic database in the form of a book that is “Atlas of protein sequences and structures” and later became the base for PIR; Protein Information Resource, (Attwood & Parry-Smith, 1999). Consequently, in 1972 the Protein Data Bank and SWISSPROT protein sequence database were started.

After creating databases, search tools were invented in order to find the desired data among the whole database. At the beginning simple search tools were available which only found matching keywords or a short part of a sequence of words. Later various types of algorithms for sequence database searching were written such as FASTA and Smith Waterman algorithms. About a decade ago BLAST, a very fast search algorithm, was written which was less accurate. Today, various commercial organizations such as Accelry, Genedata, Ocimum Biosolutions, Genzyme and other companies, are computing to provide better databases and much of this aim is carried out by informatics tools. Databases are still stored, organized, published and searched using flat files which are containing records that have no structured interrelationship.

One can fine a chronological history of bioinformatics in the appendix, (Thampi, 2009).

1.5 Biological database

After recognizing the structure of many proteins and specifying the sequence of the entire genome of a variety of living organisms, the next step in analyzing the sequence of information is to gather it inside a sharable source i.e. databases. Very briefly a part of a DNA sequence is similar to a sequence of letters without any meaning. It is shown in Figure 1 section a. One of the important aims of scientists is to encode this meaningless sequence to a meaning full one as shown in Figure 1 section b.

Databases make the biological data available in the computers for scientists. A simple database might be a long text file including too many records. There are different types of Bioinformatic databases, based on the type of data which is stored inside the database such as sequence data, 2D gel or 3D structure or according to the

manner of data storage like flat-files, relational databases or object-oriented databases, (Attwood & Parry-Smith, 1999). Today biological databases are vast. A few more famous databases among them are GenBank from NCBI (National Center for Biotechnology Information), SWISSPROT from the Swiss Institute of Bioinformatics and PIR from the Protein Information Resource, (Thampi, 2009).

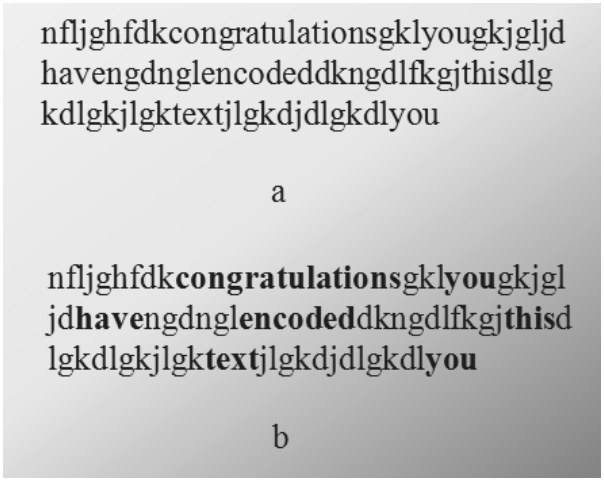

Figure 1: Briefly illustration of encoding a DNA sequence.

1.6. Bioinformatics programs and tools

Bioinformatic tools are softwares which are in the form of comprehensive packages for biological users in order to extract biological concepts from the mass data and analyze them. The most important factors to produce such these packages are the ease of the use for the end users i.e. biological scientists and the speed of the searches. Currently, various types of tools are available to serve biology related scientists. One can purchase a standard version of a product or order a customized package in order to do special researches. There are too many softwares produced for data mining which searches through the sequence data and retrieves results. The reader can find some examples of them in the following sections. Some other types are equipped with visualization tools in order to analyze proteomic databases which

are storing records including mass spectrometer; the information about protein sequences of different genes. Sequences which are related by divergence from the same ancestor are called homologous. Therefore the amount of similarity between two sequences can be interpreted corresponding to the case of their homology which can be either true or false. Homology and similarity tools can be applied to realize similarities between desired sequences and database sequences. Figure 2 shows two DNA sequences from two different species. The similar nucleotides are indicated with arrows. Also Figure 3 is a view of sequence similarity searching by PIMS (Protein Information Management System), a search tool which uses the Smith-Waterman Algorithm. This algorithm is written based on dynamic programming method which takes an arbitrary sequence and searches for an optimal sequence according to it.

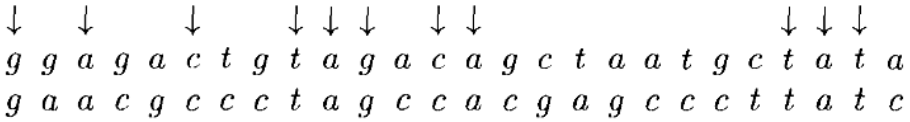

Figure 2: Searching similarities between two DNA sequences, (Ewens & Grant, 2005)

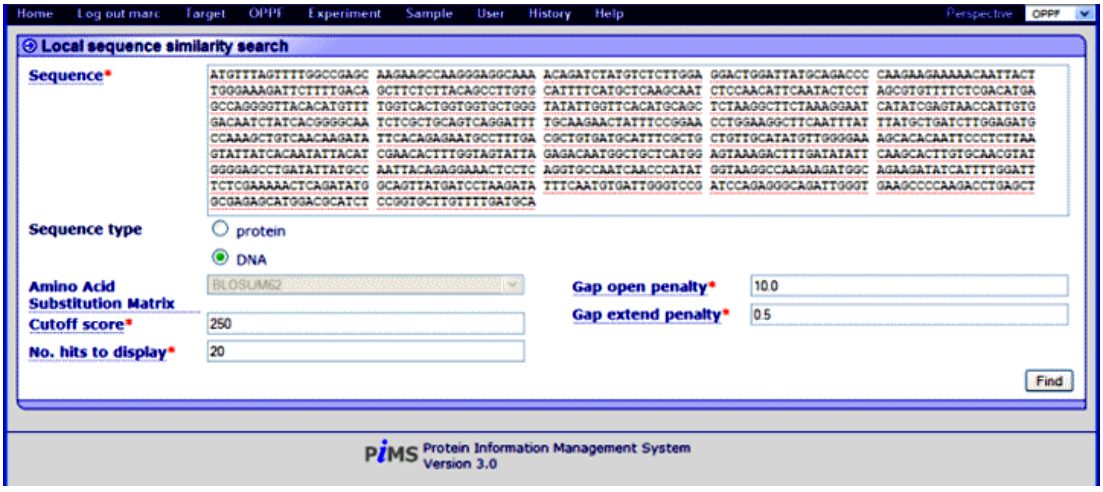

Figure 3: Sequence similarity search by PiMS. (MPSi, 2009)

Protein function analysis programs allow users to compare a special protein sequence with the protein sequence available in the database, in order to approximate the biochemical function of it. The function of a protein more than its sequence is dependent on its structure. With the structural analysis tools, comparing structures with the structures stored in the database is possible both in 2D and 3D cases. Finally sequence analysis group includes various facilities to analyze the sequences in detail and more professional, (Vizcaíno, Foster, & Martens, 2010 October). These can be categorized to homology and similarity tools, protein functional analysis tools, sequence analysis tools and miscellaneous tools. In the following section some examples of bioinformatics tools are introduced.

1.6.1 BLAST

BLAST; Basic Local Alignment Search Tool, is a sequence search program from the homology and similarity category which is designed for windows platform. The BLAST program was designed by Eugene Myers, Stephen Altschul, Warren Gish, David J. Lipman, and Webb Miller in 1990, (Myers, Altschul, Gish, Lipman, & Miller, 1990). It can be used to fast similarity searches and finding matched sequences corresponding to the entered query. BLAST contains different branches based on the type of the desired sequence to compare. Blastn is special for searching nucleotides, blastp is used for protein sequences, blastx is equipped with translated nucleotide sequences and many other various possibilities which allow biological scientists to do their custom researches, (BLAST, 2009).

1.6.2 FASTA

FASTA (pronounced FAST-AYE) stands for FAST-ALL, meaning that it searches FAST, all sequences. It is in the class of homology and similarity tools. FASTA is an alignment program produced by Pearsin and Limpman (Lipman & Pearson, 1985). This program searches at a high level of sensitivity for similarities. The

searching method of this program is that firstly it uses a fast prescreen to locate the most matching segments between the desired sequence and the data base sequences and then expands the found matching sequence to local alignments by applying more accurate algorithms like Smith-Waterman, (EBI, 2011).

1.6.3 EMBOSS

EMBOSS; Eutopean Molecular Biology Open Software Suite is a free open source sequence analyzing package. It can deal with various data formats and retrieve the results from the Web. This package includes extensive libraries and besides supports UNIX platforms, (Rice, Longden, & Bleasby, 2000).

1.6.4 Clustal

Clustal is a fully automated searcher which returns multiple sequence alignments of divergent sequences which are sequences that have similarity while having large section of divergence. Searching similarity among more than two sequences i.e. multiple sequence alignment, divergent sequences can make problem. The longer parts of divergence to search, the more difficult to find the similarities and consequently, the more error. The error which can happen is that while two sequences have similar regions, divergent parts prevent to identify them. Clustal can accurately identify theses similarities regardless of having divergence in the sequence. Also clustal can be used to predict the function and structure of proteins or identify that new protein belong to which protein families. It is available in two different versions, ClustalW which has a command line interface and ClustalX which has a graphical user interface, (Clustal, 2011), (Cates, 2007).

1.6.5 RasMol

RasMol is a 3-dimensional molecular graphics viewer program which can be used to display the structure of DNA and protein sequences. It is a powerful tool with the best and most accurate outputs. (Sayle & Milner-White, 1995)

1.6.6 PROSPECT

PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit) is a software for predicting the structure of proteins. This tool uses a computational special technique which is called protein threading in order to create a 3D model for proteins, (Thampi, 2009).

1.6.7 PatternHunter

This program is a homology and similarity tool written by Java which occupies only 40 KB of memory that is 1% of the size which BLAST does and it is while it offers a vast range of functionality and 20 times faster than BLAST. PatternHunter benefits from advanced patented algorithm and data structures, (Xia), (Thampi, 2009).

1.6.7 COPIA

COPIA (COnsensus Pattern Identification and Analysis) is a tool to search and analyze the structure of proteins. It is special for finding conserved regions or homologous regions of the sequences which are also called motifs. Motifs are useful regions to identify a new protein belongs to which family of the available protein families in the database, since similarity searches reply for sequences with more than 30% identity, (Krishnan, Li, & Issac, 2004). Moreover, COPIA has useful features for studying evolution history of the sequences and it can predict the secondary and tertiary structure and function of the proteins, (Liang, 2011).

1.7 Role of programs in bioinformatics

1.7.1 Java and its role in bioinformatics

There are many organization, research centers and scientists that are researching academic or private. Besides, there are different operating systems and hardware which are being used by researchers. Therefore, Java plays an important role in order to join the research results together. For instance, PatternHunter is a good

example of the application of Java in bioinformatics. Another example is BioJava project which is allocated to supply a framework based on Java programming in order to manage and analyze biological data. This open source project includes powerful statistical methods, special tools for file parsing, features for constructing 3D sequences and so on, (Holland R. C., et al., 2008), (Thampi, 2009).

1.7.2 Perl and its role in bioinformatics

Perl with lots of advantages such as text processing, sequence manipulation, ease of programming, file parsing which extracts and reformats sequence from its file according to defined conditions and allow users to convert multiline records into a single line record and vice versa, data format inter conversion and so on, meets the needs of the biological crowd. Although there is not any standard module designed special for bioinformatics, scientists themselves have written several modules which some of them have become known and widely used among the biologists. This population is managed by the BioPerl Project, (Stajich, 2002), (Thampi, 2009), (O'Reilly & Associates, 2001).

1.7.3 R-Statistics and its role in bioinformatics

R is a complete statistical programming language which has attracted the attention of many programmers to apply it in order to provide some addition packages useful for biological purposes. For instance BioConductor project is an open source project that produces bioinformatics tools such as sequence and genome analysis tools, data mining tools, visualization tool and so on using R, (Girke & Riverside, 2011), (Bioconductor, 2003).

1.7.4 Python and its role in bioinformatics

Python is another powerful tool in bioinformatics for performing tasks such as file parsing including the support of various file formats, strong tools for sequence

processing such as translation and transcription and integration with other useful softewares such as BioPerl and BioJava, (Kinser, 2008).

1.8 Bioinformatics Careers

There is a wide demand of bioinformatics graduates with a good specialty in computer science and software engineering. The careers in bioinformatics can be categorized into two parts, one writing and improving softwares and another applying them. To do this, they can develop new algorithms, implement softwares and apply bioinformatic tools to analyze and catch the results, construct databases special for biological data and participate in analyzing the data. Outstanding bioinformatics graduates can be employee in national or private research centers. Also because of expanding use of the Internet and IT in this field, there is an increasing demand for individuals who can manage the storages and data analyzing and retrieving over the world.

Despite the large amount of vacancies in bioinformatics careers, the advent of open source projects which has been a problem for commercial organizations to sell their products from on hand and the potential competition among different graduates from different field who have joined to this new field, should be considered as two matters of fact that make finding good jobs hard, (Edwards, 2011).

1.8 Biomathematics

Indeed without the assist of mathematical formulas, techniques and models, going forward to the future of the sciences, even for one step is impossible. Similar to other sciences joining to biology, biomathematics is an interdisciplinary field which connects mathematical techniques to biology, in order to model biological processes and compute biological parameters. Calculus, probability, statistics, linear algebra,

abstract algebra, graph theory, combinatorics, algebraic geometry, topology, dynamic systems, differential equations and coding theory are mathematical fields which are being used to reach to this aim.

Chapter 2

AN OVERVIEW ON CANCER

Cancer is one of the important concerns for human beings living in current century and still no certain way to cure this hazard has been found. Medical doctors and scientists researching in this field and the related fields all are trying to find some new concepts which helps to tackle this global obstacle either in some organizations or individually in private laboratories. This deathlike disease takes many victims all over the world annually.

The International Agency for Research on Cancer (IARC), estimated about 12.7 million new cancer cases and 7.6 million cancer deaths occurred in 2008 worldwide,. Moreover, according to the statistics over the world, it predicts 15 million new cancer cases 9 million cancer deaths will happen by 2015, (IARC(International Agancy for Research on Cancer), 2008). Cancer is not different from on race to another nor is from any part of the body to the others. That is, it affects all different races and all parts of body. (International Agenecy fo Research on Cancer (IARC)). Table 1 and table 2 show the incidence (the first-time diagnosis of cancer) and mortality (the number of deaths) of people diagnosed to have one type of cancers excluding non-melanoma skin cancer over the world classified according to the region of their living in five continents respectively.

Table 1: The incidence of cancer worldwide in 2008 including all cancer except non-melanoma skin cancer (IARC(International Agancy for Research on Cancer), 2008)

Continent	Male	Female	Total
Asia	3241249	285111	3526360
Africa	302786	378308	681094
Europe	302786	1508356	1811142
Latin America and Caribbean	444842	461166	906008
Oceania	74502	61362	135864

Table 2: The mortality of cancer worldwide in 2008 including all cancer except non-melanoma skin cancer, (IARC(International Agancy for Research on Cancer), 2008)

Continent	Male	Female	Total
Asia	2353611	1718721	4072332
Africa	248109	264293	512402
Europe	956284	758956	1715240
Latin America and Caribbean	279483	262568	542051
Oceania	30779	24293	55072

Therefore, the terrible statistics of the huge number of people suffering from various types of cancer has encouraged a large group of scientists, medical doctors, biologist as well as mathematicians and statistics to research for the cause and the behavior of this disease. That is the reason for defining a variety of models to describe the process of cancer by mathematicians. Indeed a good knowledge about biological process of cancer is required in order to define applicable models which give accurate outcomes and result in helping to find a cure for this global illness.

2.1 Various types of cancer and their causes

Cancer is a disease where a group of cells in patient's body grow and divide uncontrollably and abnormally without normal response to other processes related to

them. It may spread into other parts through the lymph or blood. More than 200 types of cancer have been recognized so far and it can spread in over 60 parts of body, (Cancer reasearch UK). The NCI (National Cancer Institute) has classified cancer into five major categories: carcinoma, sarcoma, leukemia, lymphoma and myeloma, and central nervous system cancers. The specific type of cancer then will be named based on the organ which is the origin of cancer in the body.

Carcinoma is the cancer of epithelial tissues that begins in the lining of an organ. Most of the cancers are carcinoma. Colon, lung, prostate and breast are belonging to this type of cancer. Scientists divide the causes of cancers into two main parts: environmental and hereditary factors. Carcinoma can be caused by both of them.

Sarcoma, which occurs rarely, is a type of cancer where originate from bone and soft tissue. Osteosarcoma is one of the cancers of this group where usually starts from the ends of the bones of the arms and legs, however it can happen anywhere in the body.

Leukemia starts from the stem cells of the bone marrow that produce blood cells. It does not construct tumors, but it makes cells fail to produce new normal cells which cause to clotting or bleeding. The reason of this cancer is not always specified, but scientists believe that it is more prevalent among the people who smoke, had radiation therapy to cure other types of cancer or have blood disorders or Down syndrome.

Lymphoma is a type of cancer where affects immune system called lymphocytes. It has about 35 different sub types itself. Non-Hodgkin's lymphoma is a prevalent example of this type. Lymphoma may occur because of chromosomal abnormalities.

Regularly in the body new plasma cells are producing to be replaced with the old ones. But in myeloma this process becomes out of control and a large number of plasma cells which are called myeloma are created abnormally. People with the hereditary factors, those who had other types of cancer like thyroid cancer and individuals who are overweight or obese are twice at risk to get this cancer rather than the others.

The last type of cancer, central nervous system cancer, attacks to the brain and spinal cord. This type of cancer is not prevalent as many as the other types. The reason of this disease yet is not specified clearly and scientists are researching to response this matter, (National Cancer Institute, 2011), (Cancer reasearch UK, 2011).

Cancer does not happen during a single process. It has a multi level treatment which can be predicted by some mathematical patterns. Most of the patterns which are defined so far are too much complicated and sophisticated such a way they are hard to apply for real examples. That is, still mathematicians are trying to find simpler models which can be more helpful to interpret the treatment of this dangerous disease, (Wodarz & Komarova, 2005).

2.2 Basic definitions and notations

In this section some required information, in the form of very brief explanations will be given to peruse the cancer concept such a way it is easily understandable for those who are not familiar with the biology sciences. Finally it should be able to estimate a high proportion of variability in the given data as well as prediction new observations with high certainty.

2.2.1 Incidence and incidence rate

Incidence is the number of new cases diagnosed to have an especial disease (here cancer) that develop in a population over time. It can be shown as the exact number of cases per year or as a rate per 100,000 persons per year. Incidence rate then is defined as the proportion of the number of new cases over the number of population at risk, (TheFreeDictionary, 2011).

2.2.1 Mortality and mortality rate

The number of deaths which happen in a given period in a specified population is called mortality. Similar to incidence data it can be published as an absolute number of deaths per year or as a rate per 100,000 persons per year, (IARC, GLOBOCAN 2008).

2.2.1 Gompertz low of mortality

This law claims that the rate of death is the addition of an aged independent part which is called Makeham and to an age dependent part which is called Gompertz function. The later part increases with age exponentially.

$$R_t = R_0 \, e^{\alpha t}$$

where R_t is the rate of mortality with age t , (Wikipedia, 2011).

2.2.2 Strehler and Mildvan's general theory of mortality

This theory states that every person has a capacity of vitality or staying to be alive. Then this vitality is indicated with $V(t)$ and defied as a linear function of age t as

$$V_t = V_0(1 - Bt)$$

where B is the slope of the vitality curve and V_0 is the original vitality, (Vapuel & Yashin, 1999).

2.2. 3 Fourier series

This interesting function firstly introduced by Joseph Fourier in order to response the heat equation which had no solution up to that time. Fourier function is the composition of multiple seines and cosines or complex exponentials, (Wikipedia, 2011).

2.2.4 Goodness of fit

After approximating data with a function it is needed to evaluate the goodness of fit. There are many formulas according to the interpolation which evaluate the error of the approximation such as residual error, goodness of fit statistics, confidence and prediction bounds.

A good fit model should have the following characteristics: first of all the model should be applicable for the data, in other words the data should be obtained from the approximated function according to the assumption of least square fitting. Secondly these coefficients of the model can be approximated with little difference from the reality, (Mathworks, 2011).

2.2.5 Stochastic multistage cancer models

This model is considered as a sequence like $C_k \rightarrow C_{k+1}$ including connected parts each of which indicating the number of mutations of a cell in the k-th part. C_k represents a cell with k mutations. In each part, a cell can divide, die or mutate, (Leonid & Wai-Yuan, 2008).

2.2.6 Markov process

A Markov process is a phenomena which changes over time randomly and during this change a special attribute holds and the conditions holding for its present , future and past are independent. This type of processes is the basic idea of the model of many natural phenomena. Some examples of such these models will be explained later in chapter 4, (Wikipedia, 2011).

2.2.7 Time series

A sequence of some observed data, which are measured repeatedly according to several time intervals, is called a time series. Applying time series model, it will be possible for the scientists to forecast the future values of a desired phenomena, (Wikipedia, 2011).

2.2.8 Mutation

Every change in the genes of a DNA sequence is called a mutation. Mutation can change the type of message carried by the genes and consequently it result in producing different proteins other than the rural one. Most of the mutations are harmful for the body while some of them are the raw factors of evolution. (David Sadava, 2006). The majority of scientists believe that cancer happens after occurrence of several mutations, (Vogelstein & Kinzler, April 1993).

2.2.9 Carcinogen

Any substance that can be lead to cancer is called carcinogen. Carcinogens are separated into two groups by IRAC: the factors which are carcinogenetic for human and the factors may be able to be carcinogenetic for human, (Quitsmokin).

2.2.10 Cancer stem and malignant cells

Cancer stem cells are able to divide and replicate in order to create the similar stem cells. They are able to increase all types of the sample cancer cells found in an individual's body. In other words cancer stem cells are able to form tumor. From the other hand, malignant cells are the cells which tend to be worse and they may lead the individual to die. Moreover malignant cells are the forming components of the malignant tumors, (Wikipedia).

2.2.11 Cellular differentiation

During the process of improvement of a cell a cell repeatedly divides and creates some new cells. For example normally cells turn over or in case that there is any injury the cells start to repair the tissue of that region. Cell differentiation happens during such these events and causes to change the attributes of the cell such as its shape and size, (Wikipedia).

2.2.12 Risk factor

Risk factors may not themselves cause to create a disease but they help to provide the conditions of appearance an illness in body. They may help to increase the disease severity.

2.2.13 Metastasis

Metastasis has a Greek origin with the meaning of displacement and it is defines as the disease transition from one part to another non-contiguous parts of the body, (Wikipedia, 2011).

2.2.14 Mutagen

Mutagen is a factor which causes more mutations in the DNA sequence. It is able to change the genetic materials such as DNA, (Wikipedia, 2011).

Chapter 3

The DETTERMINISTIC RISK FACTORS AFFECTING ON CANCER INCIDENCE

There are many reasons that people ascribe them to cancer. Scientists have proved that some of them really affect on cancer initiation while there are some other reasons which has been rejected as cause of cancer. Moreover, there are some reasons which have not been recognized yet and still they are unknown for human, (Cancer research UK, 2009).

Hereditary, bacteria, viruses, bad diet, obesity and having not enough exercise, environmental or radiation exposure nor tobacco, none can cause cancer. Cancer is the result of a series of DNA mutations. Although mutations can exist initially in an individual' DNA or can be acquired after a while of his/her birth, all the factors mentioned above can increase the chances of mutations to be enough to conduct stem cells to malignant cells, (The scientific basis of vegeteriansism, 1999).

In order to model cancer and its incidence it is important for a mathematician to recognize which parameters are important in the formulas to take account. Hereupon, in this chapter some cancer risk factors are provided before starting cancer modeling in the following chapters. It should be noted that the factors mentioned here are not the whole recognized reasons for carcinogenesis.

3.1 Genetic factors

As it is mentioned before, cancer is a multistage process including reposition of numbers of mutations which happen inside the stem cells, (Loeb & Loeb, 2000). Some carcinogenesis etiologists believe that cancer is a per se endogenous disease which can be created by inheritance and genetic reasons, (Hahn & Weinberg, 2002), (Hoyer, Gerdes, T., F., & H. B., 2002). But carcinogenesis is a set of diseases and it should be diagnosed by investigating its symptoms. The fact is, due to gene-environment interaction that other than genetic causation for cancer, there are other environmental and behavioral reasons which affect on it and lead to alter the rate of its progress, (Mucci, S., M., Trichopoulos, & Adami, 2001). Moreover there are increasing evidences that non genetic cancer causes predominate rather than genetic factors, (Lichtenstein, Holm, Verkasalo, Iliadou, Kaprio, & Koskenvua, 2000). Therefore, to model the rate of cancer incidence it is worthy to consider all types of cancer reasons. Also, to investigate more about cancer etiology it is important to reply the question that whether the progress of cancer incidence is because of the gene susceptibility to create inherited mutations or acquired somatic susceptibility existing within the population.

3.2 Lifestyle risk factors

The factors related to the style of every one's life may not be directly the reason for cancer, but they are risk factors which affect on the stages of the process of this disease through the individual's habits and exposures. In this section some more important factors are mentioned.

3.2.1 Smoking

In fact one the most important factors which leads cancer to progress is tobacco smoking. Tobacco smoke and tar contain many chemical compounds which are

equivalent to what is known as carcinogen, (Löfroth, 1988). Figure 4 shows the percentage of some common cancers among men and women tobacco smokers. As it is visible mostly it affects on lungs. To achieve a better understanding of the matter one can see the comparison between the smoking prevalence and cancer incidence in figure 5.

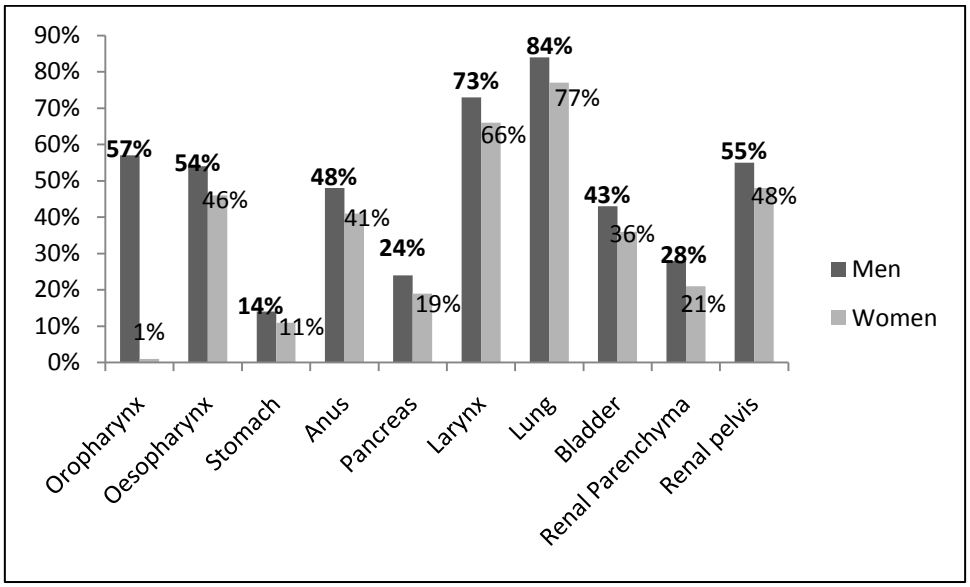

Figure 4: The percentage of cancers prevalent among tobacco smokers. Data is chosen from (English, Holman, Milne, Winter, & Hulse, 1995)

3.2.2 Alcohol consumption

In contrast with the carcinogenetic constitutive compound of tobacco smoke and tar, alcohol is not mutagenic on its own, but it promotes the effect of a carcinogen which finally ends to cancer and consequently it is classified alcohol a carcinogen, (Poschl & Seitz, 2004), (IARC(International Agenecy for Reseach on Cancer), 1988). Figure 6 shows the percentage of cancers prevalent among alcohol consumers.

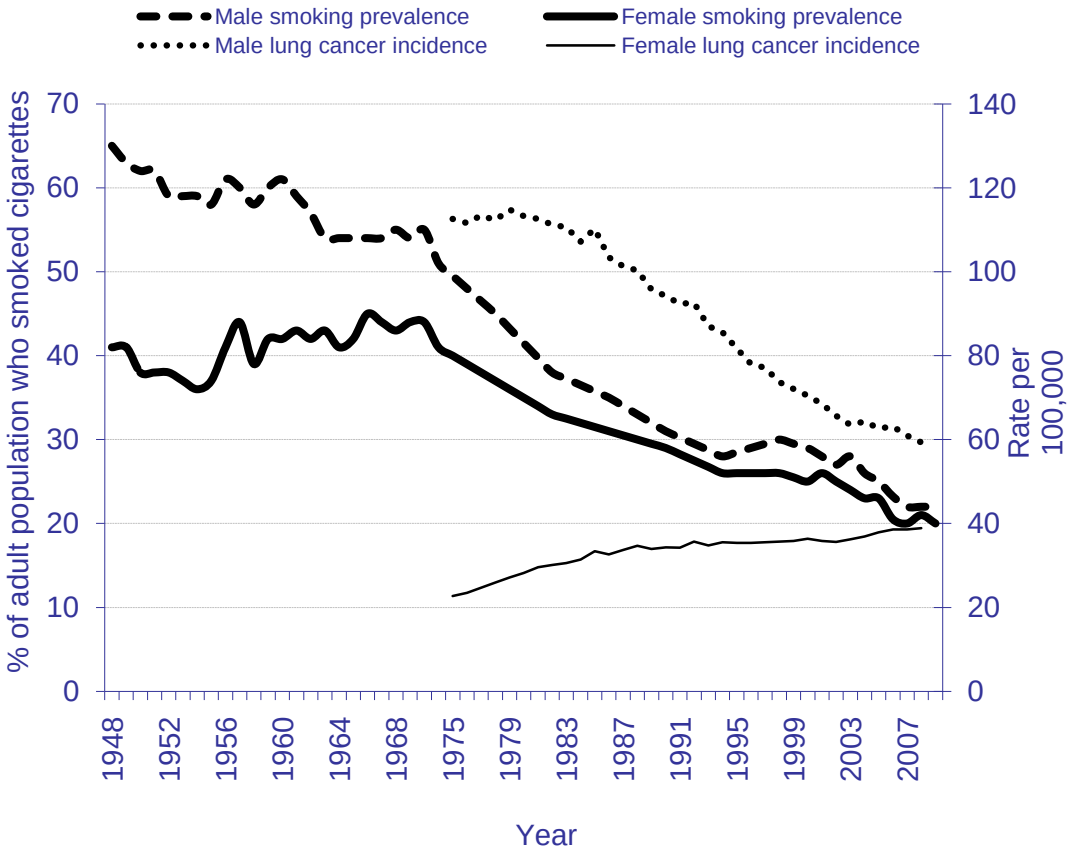

Figure 5: Comparing lung cancer incidence with the smoking prevalence in Britain during 1948 to 2007, (Cancer research UK, 2009)

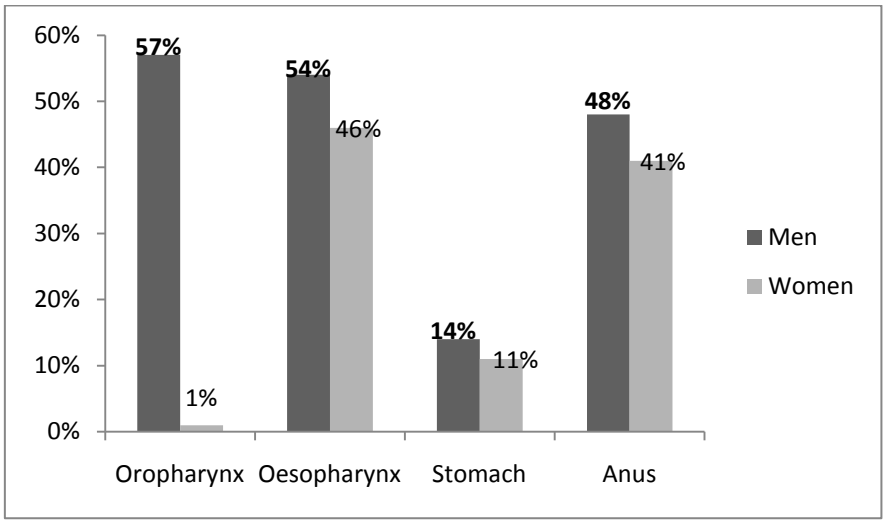

Figure 6: The percentage of cancers prevalent among alcohol consumers. Data is chosen from (English, Holman, Milne, Winter, & Hulse, 1995)

3.2.3 Diet

There are many investigations which indicate the importance of daily using fiber, fresh vegetables, and white meat and its effect on the decrease of the cancer hazard.

Diets containing high amount of fiber and low amount of calories and animal fat decrease the chance of occurring some cancers like breast, prostate, colon and endometrium. In other words, replacing adequate serves of fresh fruits and vegetables and eliminating instead of consuming processed and red meat can help to prevent the cancer incidence, (Block, Patterson, & Sabur, 1992), (Weisburger, 2002). Figure 7 illustrates the comparison between low and high fat diets and its relation with mammary tumor incidence in mice. One can see the large amount of difference between the two diets.

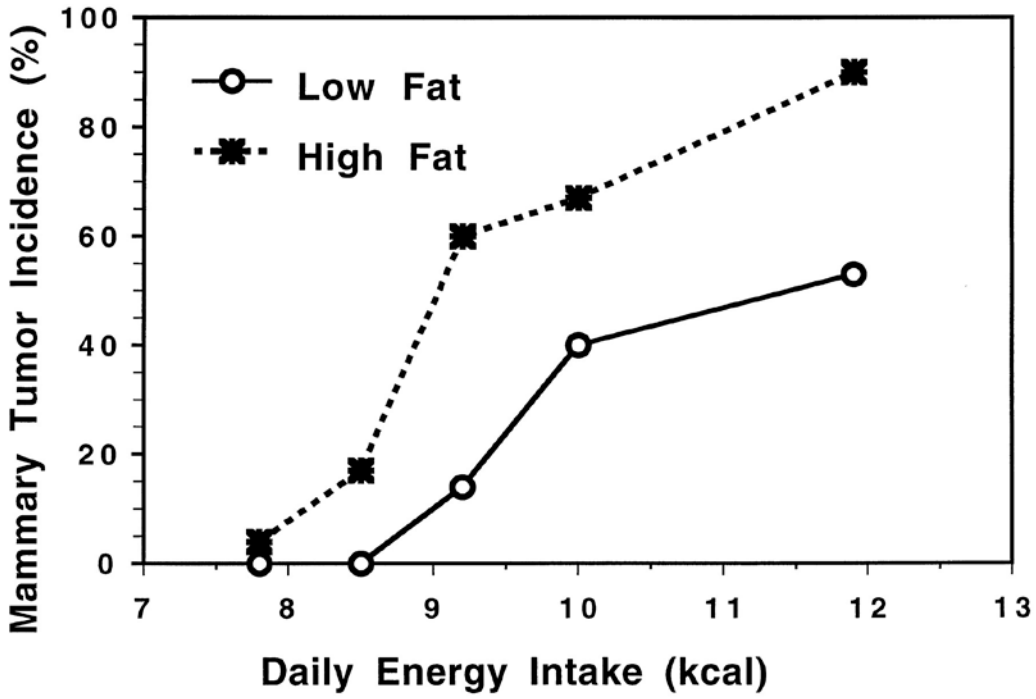

Figure 7: The effect of low/high fat at various level of caloric intake on spontaneous mammary tumorigenesis in C3H female mice, (Kufe, Pollock, R., & et al., 2003)

3.2.4 Overweight and obesity

There are many evidences which show obesity and sedentariness are factor of risk for most types of cancer including breast, colon and prostate, (Simopoulos, 1990), (IARC (International Agency for Research on Cancer), 2002). Although obesity is related to the increase of the rate of many other diseases such as diabetes and

cardiovascular problems and subsequently can increase the overall mortality, it per se has been found as the reason of the preceding cancer types except lymphoma and child cancers, (Rodriguez, A.V., Calle, Jacobs, Chao, & Thun, 2001), (Petrelli, Calle, Rodriguez, & Thun, 2002), (Willett, et al., 2005), (Uauy & Solomons, 2005).

3.2.5 Impact of new diagnostic and screening methods

Nowadays, the advent of new diagnostic and screening techniques such as mammography to diagnose breast cancer, PSA for prostate cancer, ultra sonography for thyroid cancer cervical smears for cervical cancer and so on, allow doctors to save more cancer patients by on time diagnosing, (Solomon, 2003), (Crawford, 2003), (Eden, Mahon, & Helfand, 2002), (Parkin & Fernandez, 2006). Figure 8 shows the rate of cancer incidence in developed countries.

As it is visible in figure 9, during 1975-2003 the incidence rate of some cancer types have been increased while the mortality rate of them have been decreased. Clearly one acceptable reason for this fact is because of the better facilities to early detection of the diseases and consequently increasing the chance of survival.

Moreover there were some types of cancer 30 years ago which were not possible to diagnose because of their very slow rate of growth and progress and now they can be found by applying screening facilities very soon. However, there are some types of cancers such as melanoma, malignant lymphoma, leukemia, and childhood cancers which no screening test has been improved for them over the last 20 years. Hence there may be other reasons for this increasing such as genuine phenomenon from a biological point of view, (Irigaray, Newby, Clapp, L., & Howard, 2007).

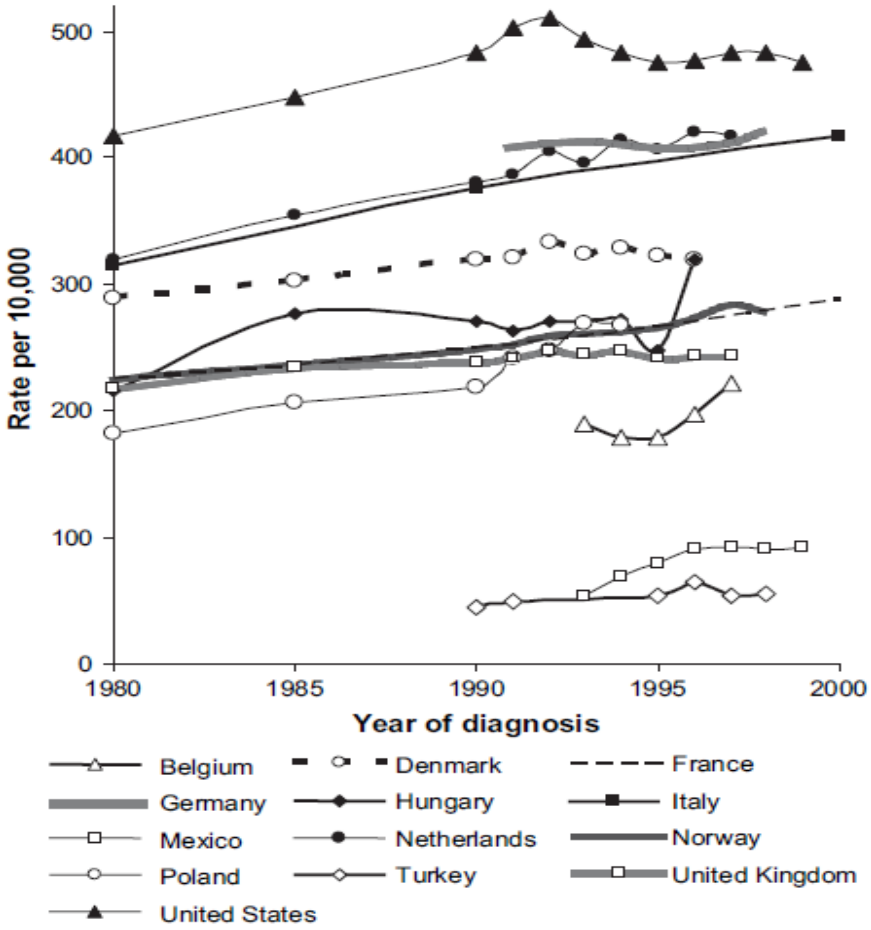

Figure 8: Comparing cancer incidence rates among developed countries, (Irigaray, Newby, Clapp, L., & Howard, 2007)

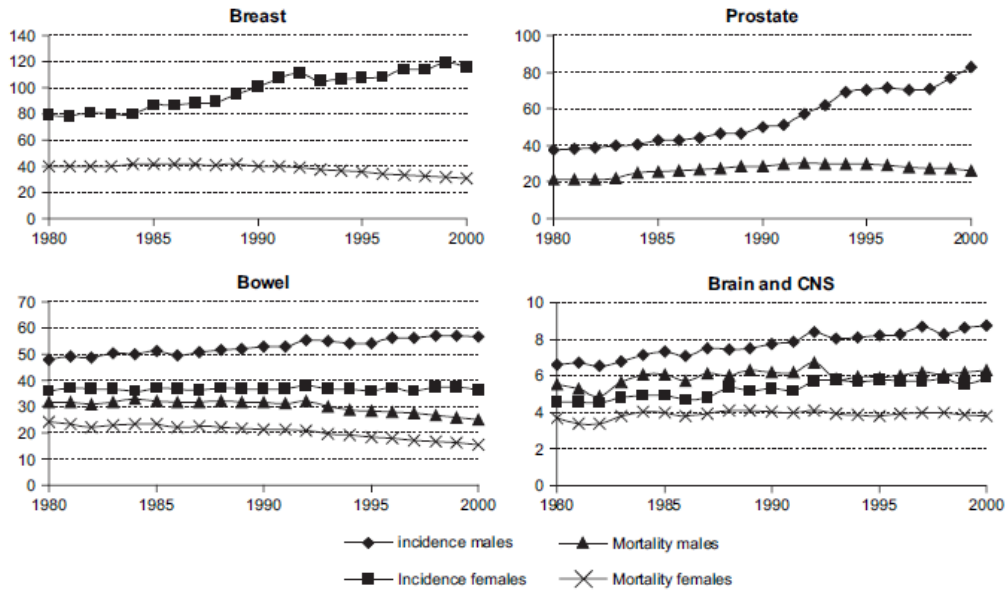

Figure 9: cancer incidence and mortality rates in UK from 1975 to 2003 for breast, prostate, bowel and brain, (Irigaray, Newby, Clapp, L., & Howard, 2007)

3.2.6 Age and increasing life expectance

This is the fact that today in most countries especially developed countries the life expectancy has been extended and therefore the age of mortality has been increased. From the other hand there is no doubt that cancer is related to age and by increasing age it increases. This increase means that there are more new cancer cases in these countries, (Ershler & Longo, 1997), (Jemal, Thomas, Murray, & Thun, 2002). However, there is another opinion about the decline of cancer incidence in oldest old ages i.e. after 85 years old, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005).

There are many age specific offered models for cancer which shows the role of ageing in the occurrence and progress of cancer. Later in chapter 4 and 5, I will discuss about some of them and the typical cancer age-pattern of the overall rate of cancer as well as a new model rating over age parameter. One opinion about the relation of the age and cancer progress is associated with the Armitage-Doll model which is mentioned in the first section of chapter 4. Armitage and Doll propose that because of the multi-stage nature of cancer people with old ages are more exposed to the danger of cancer progress via accumulating enough number of mutations inside the stem cells to create malignant cells, (Armitage & Doll, 1954). Beside this opinion, there is another fact that aging process itself can affect on cancer incidence. Because most of scientists who investigate in this field believe that the number of stem cells decrease with aging, (Knudson, 1993). From the other hand, there are biological interpretations which show that the effect of aging per se is not as considerable as the length of the exposure time to the other environmental reasons. In other words, the more aging, the more to be exposure to environmental cancer

causes and consequently the more risk to catch cancer, (Irigaray, Newby, Clapp, L., & Howard, 2007).

3.3 Environmental risk factors

From the industrial revolution so far, millions of chemical products have been produced by human beings and they have been applied in various fields such as agriculture, foods and medicines and so on. According to a European report around 100,000 chemical products have been marketed up to now and it is while there is no control on them toxicologically, (Clapp, Howe, & LeFevre, 2005). Moreover, today most countries are buckling with air pollution and traffic. Such products either directly or indirectly can act as carcinogenic compounds and therefore cause cancer. Such these cancer causes which individual exposure them during his/her life are incriminated to the environment. Here are some examples of environmental risk factors which scientists have examined their role in cancer creation or progress.

3.2.1 Radiations

Radiation is cause of some cancers such as leukemia, lymphoma, thyroid cancer, skin cancer, lung cancer, breast cancer and sarcoma. These types of cancer are stochastic results after ionizing or non-ionizing radiation, (Wakeford, 2004). Scientists who investigate the lung cancer causes have sound that about 10% of this type of cancer is caused by being exposed to low radon level existing in home environment, (Lubin & Boice, 1997) , (Darby, et al., 2005). They have found out that people mostly exposure to the products made of radon in their home or/and their workplace, (Axelson, Fredrikson, Akerblom, & Hardell, 2002).

Another source to be exposed to radiation is the application of X-rays for medical purposes. What is worthy of note is that the period of exposure is deterministic. For

instance, for a girl in her puberty age the risk of catching breast cancer will be increased if she is exposed to chest radiation during this period, (Ronckers, Erdmann, & Land, 2005). There are also some reports about increasing the incidence of total malignancies after some terrible events such as the report from Sweden after Chernobyl radioactive fallout, (Tondel, Lindgren, Hjalmarsson, Hardell, & Persson, 2006).

Also Ultraviolet (UV) rays have been recognized as definite factor of skin cancer, (IARC (International Agency for Research on Cancer) , 1992).

Recently, daily prolonged use of mobile phones for a period of 10 years has been proposed as a risk factor of brain cancer, (Hardell, Carlberg, Soderqvist, Hansson, & Morgan, 2007). There also other examples of radiation exposure as the reason of cancer which are beyond the scope of this thesis, (Feychting, Forssen, & Floderus, 1997), (Hardell & Hansson, Mobile phone use and risk of acoustic neuroma: results of the interphone case-control study in five north European countries, 2006).

3.2.2 Occupational cancers

Sir Percival Pott incriminated the scrotum as the first occupational cancer in 1775. Occupational cancers today are estimated to be around 2-10% of all types of cancer which may be really 15-20% in men, (Landrigan, Markowitz, Nicholson, & Baker, 1995). Up to 1981 only 16 agents had been recognizes as cancer causes in workplaces, (Irigaray, Newby, Clapp, L., & Howard, 2007), while today 28 factors are recognized as imperative, 27 as probable and 113 as possible cancer causes, (Siemiatycki, et al., 2004).

3.2.3 Outdoor air pollution

The smoke of factories, vehicle exhaust as well as environmental tobacco smoke (ETS) produces particles suspending through the air which are including polycyclic aromatic hydrocarbons (PHA). Indeed PHA which is the outcome of combusting organic substances is carcinogenic, (IARC (International Agency for Research on Cancer), 1989). Breathing through the air including these particles for adults increases the rate of mortality by cause of lung cancer by 8%, (Dockery, et al., 1993), (Pope, et al., 2002), (Cohen A. G., 2003). Moreover recent observations in some European countries indicate that air pollutions and exposure to tobacco smoke for never-smokers and ex-smokers is approximately around 5-7% and 16-24% respectively, (Vineis, et al., 2007). In addition nitrogen dioxide (NO2) and some existing in the air pollution and traffic exhaust can be cause of lung cancer especially in children, (Ichinose, Fujii, & Sagai, 1991), (Richters & Kuraitis, 1983), (Wertheimer N, 1979).

3.2.4 Indoor air pollution

Indoor air can hold carcinogenic particles inside it such as carbon compounds, ETS, biocides, formaldehyde, and also volatile organic compounds (VOC) such as benzene and consequently lead to lung cancer by passing time, (IARC (International Agency for Research on Cancer), 1995). Children are the most group at risk to exposure indoor air pollution. After children ex- smokers and then never smokers who are at risk at workplace more than home, (Vineis, et al., 2007).

3.2.5 Other factors

It is worthy of note that medicines and products of personal care such as cosmetics, viruses and other microorganisms, food contaminants and food additives, biocides and pesticides and, metals and metalloids all are environmental risk factors of cancer

diseases which discussion about them is beyond the scope of this thesis, (Irigaray, Newby, Clapp, L., & Howard, 2007).

3.4 Conclusion

Study to find cancer causes and its risk factors is still continuing. There are many parameters which scientists have doubt about their effect on cancer. Also there have been many investigations on many parameters which I did not mention here such as the geographical distance of countries, sexuality, and the amount of individual's income which can be observed in various populations. But most noteworthy for mathematical modeling is the notion that there are some cancer risk factors which are more important than the others. In contrast while considering effecting parameters in modeling the incidence of cancer some risk factors can be neglected. Moreover all risk factors are not associated with all cancer types. For instance while modeling lung cancer incidence there is no doubt to take account air pollution and tobacco risk factors. Maybe this is one reason for the scientists who choose some specific factors to adjust in cancer model and introduce site-specific cancers. For example there are a group of scientists who believe that approximately all cancer types rate over age and time which I discuss about them in chapter four, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005).

Chapter 4

MATHEMATICAL MODELING AND CANCER INCIDENCE RATE

To quote Joel E. Cohen, a famous mathematical biologist: “Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better” (Cohen J. E., 2004).

The attempt to find a proper pattern which is able to explain the cancer incidence behavior has a long history. There is no doubt that mathematics is a great tool in this realm. During these endeavors, the need of applying mathematical techniques and formulas in the interpretation of phenomena in the life sciences has been felt more and more. Therefore, entitling mathematical modeling of the phenomena in the life sciences as the revolution of the current century may not be so far from the fact, (BELLOMO, LI, & MAINI, 2008).

Modeling the cancer incidence rates is definitely one of the challenging frontiers of applied mathematics which could have a great effect both on the quality of human’s life and the development of mathematical sciences. This is the fact that this issue cannot be solved solely by use of mathematics. However, applied mathematics may be able to define an outline in which empirical data can be interpreted and analyzed. Moreover, mathematical modeling not only helps scientists to describe the behavior

of this phenomenon, but also estimates and predicts the future of cancer hazard among various populations.

There are two distinct categories of mathematical models interpreting cancer incidence data. Statistical models are based on mathematical formulas, rules, and techniques such as linear and logistic regression which indicate the relationship between various parameters and cancer incidence. Biomathematical models which are the translation of biological relations and hypotheses for cancer into mathematical formulas, (Kaldor & Day, 1996).

The aim of this chapter is to provide a brief background containing some famous examples of cancer incidence rates models specially age specific models beside the explanation of the features of the typical age-pattern.

4.1 General models for cancer incidence

The proposed cancer incidence rate models are vast. Among all scientists who have offered various models, there are some outstanding observers whose their models determine new classes for this epidemiological quantity. Here are some chosen samples of their great research. It is worthy of note that the goal of this section is only to introduce these models and the application of the models to the real data is beyond the scope of this research.

4.1.1 Armitage-Doll (AD) carcinogenesis model

In 1954, Armitage and Doll introduced an age distribution model for cancer based on the fact that cancer is a multi staged disease. To define this model they used two hypotheses. Firstly, they assumed that the death rate intensifies proportionally with the sixth power of age. Secondly, they accepted that a healthy cell needs to be the end result of seven successful mutations in order to convert to a cancerous cell.

Then they defined the incidence rate of cancer at age t as bellow

$$rate = kp_1p_2p_3p_4p_5p_6p_7t^6,$$

where k is a constant and p_i is the probability of occurrence the *i*-th mutation per unit time. Although Armitage- Doll's model was justified only mathematically, it gave a good fitness especially for epithelial cancer types such as colon, rectum, stomach and pancreas, (Armitage & Doll, 1954), (Wai-Yuan & Leonid, 2008).

After the advent of this model many scientists tried to revise it in order to get better fittings. With the inspiration of this model, they introduced various cancer incidence models in the form of C.[age]$^{\beta}$ for some constants C and β, (Nordling, 1953), (Doll, 1971). They improved the primary model with observation on the log-log changes of cancer incidence rates with age. For defining a multistage model of this form, one should assume that a person at age *t* has a group *X(t)* of normal cells (biologically they are called stem cells) which they reach to one mutation at a rate of *M(0)(t)*. Then the cells which have obtained one mutation in the first stage, gain the second mutation at a rate *M(1)(t)*, and similarly the cells in each stage gain a new mutation in the next stage until the *(k-1)-th* stage which the cells have gained k-1 mutations. At the *k-th* stage the primary normal cells have converted to malignant cells completely. It leads to tumor metastasis and transfer the infection to the other tissues of the individual's body. Figure 10 illustrates the process of a multi stage Armitage-Doll model from the first stage to the last, (Wai-Yuan & Leonid, 2008) , (Little, 2010).

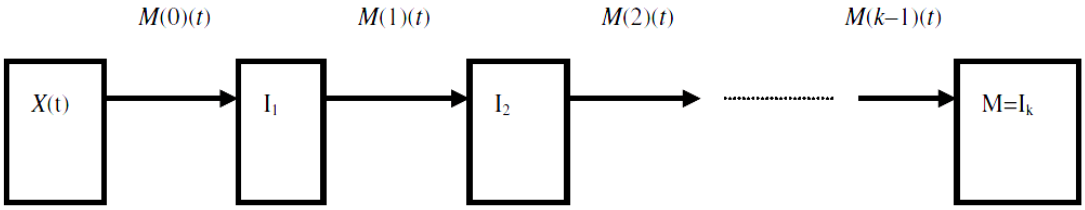

Figure 10: A diagram for Armitage-Doll multi stage model, (Wai-Yuan & Leonid, 2008)

Therefore, it can be expressed mathematically as C.[age]k-1 which C is equal to X·M(0)·M(1)·...·M(k-1)/(k-1)!, (Moolgavkar s. , 1978).

As mentioned above, although this model fits well for epithelial cancers and shows that cancer is a multi stage process, the main problem with this model is that to have a good fitting the number of stages should be between 5 and 7 which is high and acquires more number of mutations. Besides, in this model there is no possibility for the muted cell to die or include any randomness. Later, in order to improve their model, Armitage and Doll defined a model requiring two stages for mutation, but this model gave less accuracy and only better fitting for adults rather than children (Little, 2010), (Armitage & Doll, 1954).

4.1.2 The Moolgavkar, Venzon and Knudson (MVK) model for cancer

In 1971, Alfred G. Knudson improved the two stage cancer model to interpret the incidence of a type of eye cancer common among children which is named retinoblastoma. To do this, he assumed that there are two mutation rates for a stem cell; µg and µs which are corresponded to germinal and somatic mutations respectively. Also he considered that the probability for the second event to happen is m/n. where n is the total number of cells in the two retinas that have the potential for tumor formation and m is the value most compatible with the tumor distributions for unaffected, unilaterally affected, and bilaterally affected carriers of the mutant

gene.Therefore, he offered the following model for the incidence of the hereditary cases

$$f_h.i = 2q(1 - e^{-m}),$$

where i is the total incidence of retinoblastoma, f_h is the fraction of the hereditary cases and q is the population frequency of the germinal mutant gene, in this model, Knudson focused on the normal tissues only and did not consider cell mortality, (Alfred G. Jr., 1971).

Subsequently in 1979, Suresh H. Moolgavkar and David J. Venzon developed Knudson's two stage model by taking account into the dynamics of cell proliferation at all rates and also differential growth in the both normal and intermediate cells (i.e pre malignant cells which have not been completely malignant). Finally at 1981, Moolgavcar cooperating with Venzon and Knudson offered a common two stage model which is called MVK, (Moolgavkar & Venzon, 1979).

The MVK model is a Markov chain which is defined time independently. This age dependent model assumes that at age t the number of stem cells prone to get cancer is X(t). These stem cells can mute at an intermediate rate$M(0)(t)$. Moreover, the intermediate cells can divide at stage$G(1)(t)$, die or differentiate at a rate $D(1)(t)$, or transform to malignant cells at a rate $M(1)(t)$. Figure 11 illustrates the interpretation above for the MVK two stage model, (Leonid & Wai-Yuan, 2008), (Moolgavkar s. , 1978).

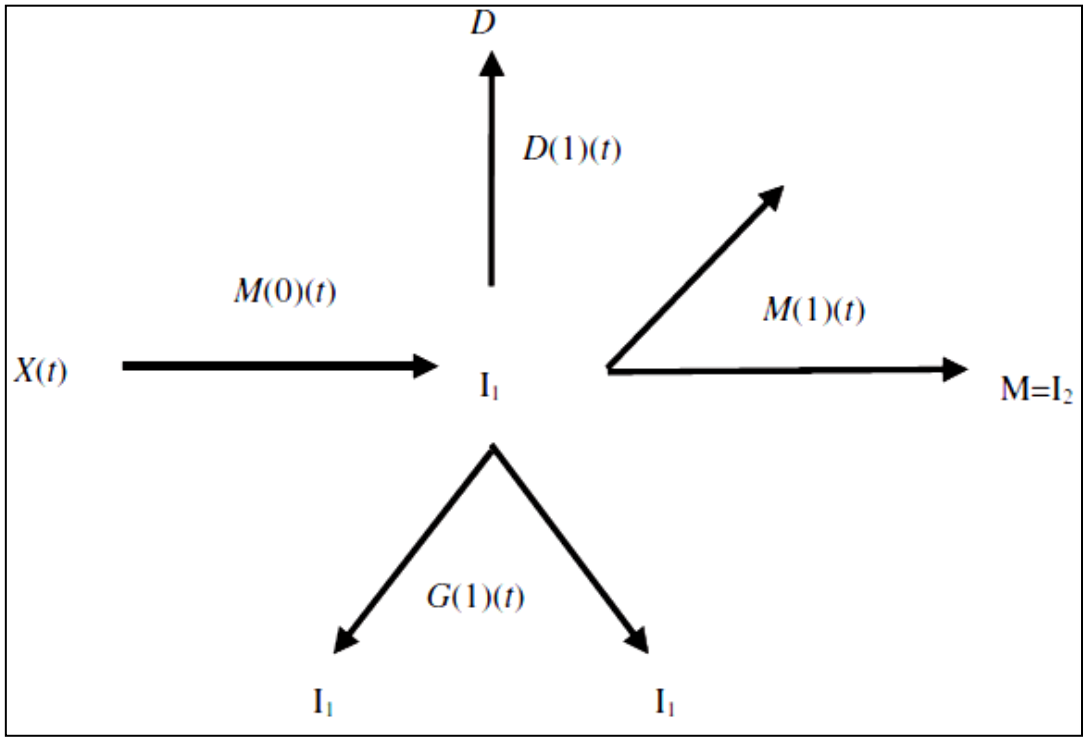

Figure 11: A diagram for two mutations MVK cancer model (Leonid & Wai-Yuan, 2008)

This model is more complicated rather than AD cancer model and is approximately applicable for all human cancer incidence data well. Also, it justifies the fact that cancer is a multi step process. But it is not the only deterministic factor to initiate and progress of cancer. Cancer can be caused by many other different reasons in different people. Therefore, the MVK model can not fit all these reasons completely.

4.1.3 Age-Period-Cohort (APC) models

Some epidemiologists believe that there are three important factors which are deterministic to get cancer among different people in different times; age of the people who are being studied, period of the time when is being studied, and the specific cohort with one or more common characteristics which are chosen to observe. One good reason, therefore, to offer some models depending on these three factors is to assess and approximate the effect of the difference of these three factors among various groups of people with unequal ages, different life types, diets and habits living in different periods of time, (Kupper, Janis, Karmous, & Greenberg,

1985), . The effect of the age as well as the effect of the period is indicating the changes in the rate of time. The effect of the cohort shows the change of the rate among the different successive age groups in successive periods. Using such these models, abbreviated APC models, one can observe the cohort effect among the groups either exposed to different risks such as war, radiation ray, smoking, pollution and so on or experiencing different habits like the type of diet, exercises and other environmental and hereditary factors.

Scientists, who desire to investigate APC relations among these three factors, usually summarize the information in some two-row tables including cancer incidence rates categorized by age group and time period which is illustrated in form of an example in table 3. In this table, the cohort is assumed the birth cohort which are the diagonals with the oldest cohort placed in the left bottom corner of the table i.e. people with the age ranging from 80 to 84 years old who have been studied during the time period from 1960 to 1964 of their lives and the youngest cohort placed in the right top of the table i.e. people with the age ranging from 30 t0 34 years old who have been observed during the time period from 1980 to 1985. In other words the oldest birth cohort was born during the years from 1876 to 1884 and the youngest birth cohort ranges from 1951 to 1959.

The general linear form offered for APC model claims that the logarithm of the expected incidence rate is a linear function of age group, time period and birth cohort as below:

$$\ln\left(E\left[r_{ij}\right]\right) = \ln\left(\frac{\theta_{ij}}{N_{ij}}\right) = \mu + \alpha_i + \beta_j + \Upsilon_k$$

Table 3 : A two-way table of rates which can be used in age-period-cohort modeling (Robertson, Gandini, & Boyle, 1999)

	Time period					
Age group	1960–1964	1965–1969	1970–1974	1975–1979	1980–1984	1985–1989
Rates						
30–34	36.90	37.84	39.20	41.40	40.78	41.81
35–39	92.09	98.61	102.15	107.91	106.29	108.98
40–44	137.99	150.01	161.98	171.10	168.54	172.80
45–49	194.58	191.10	210.17	231.11	227.65	233.41
50–54	256.73	248.54	246.74	276.61	283.73	290.92
55–59	304.85	312.57	305.50	309.49	323.76	345.44
60–64	350.44	359.31	372.22	371.19	350.44	381.68
65–69	393.59	403.55	418.05	441.60	410.41	403.55
70–74	434.39	445.39	461.39	487.38	480.08	464.37
75–79	472.94	484.91	502.34	530.63	522.68	535.91
80–84	509.32	522.21	540.98	571.45	562.89	577.14
Populations						
30–34	839,500	786,800	759,500	824,000	878,800	877,500
35–39	866,000	808,900	761,200	749,800	816,700	867,600
40–44	865,400	837,200	786,300	746,600	747,700	809,900
45–49	851,000	836,600	812,200	769,400	741,400	737,900
50–54	886,400	817,900	806,100	789,500	757,000	727,400
55–59	834,500	845,500	779,000	773,200	760,900	731,900
60–64	743,900	782,800	793,200	730,200	727,700	719,200
65–69	616,100	669,700	710,700	719,700	664,700	669,500
70–74	475,700	514,800	570,000	610,200	620,200	578,100
75–79	332,300	357,000	391,300	444,200	481,000	498,500
80–84	185,000	210,300	231,100	256,500	298,900	337,600

where μ is the mean effect, α_i, β_j and γ_k are the effect of the age group i, time period j and the birth cohort k respectively. In this model, y_{ij} which is a realization of Poisson random variable denotes the number of diagnosed cases in the age group i at the period j of the time with mean θ_{ij}, where $i = 1, \dots, m$ and $j = 1, \dots, n$. Besides, it is assumed that the number of persons in the age group i at time period j who are at risk to get cancer, i.e. N_{ij} is a fixed known value, (Robertson, Gandini, & Boyle, 1999). Currently, many formulas in the form of APC model are offered by different groups and they are widely applied to represent the treatment of cancer data. However APC models have their own problems and they are not adequate for all types of cancer data, (Coleman, Esteve, Damiecki, Arslan, & Renard, 1993), (Robertson, Gandini, & Boyle, 1999).

4.1.4 Models in heterogeneous populations

Despite some scientists who apply APC models to their data, there are some others who believe that every individual even in the same age group, time period and

environmental conditions with the others has his/her own susceptibility to get an especial type of cancer, (Vapuel & Yashin, 1999). These differences among different persons have been studied for many years. Although it has been proved that genetic differences play a very important role to create the differential susceptibility among individuals, (Carins, Lyon, & Skolnick, 1980), (Knudsun, 1977), (H. & Weber, 1985), there are some other deterministic risk factors which affect on individuals to react distinctly against cancer. Cigarette smoking, being exposed to radiation, sunlight, toxic gases, and asbestos, having especial diets including vegetables and sea foods and so on are some examples of these types of reasons.

In recent decades several models have been introduced which realize heterogeneity among people in susceptibility or frailty parameters, (Cook, Doll, & Fellingham, 1969), (Manton & Stallard, 1980), (Manton & Stallard, 1982), (Manton, Stallard, & Vapuel, 1986).

Prototypical susceptibility model, for instance, applies heterogeneity hypothesis in order to define the incidence rates. Scientists assume that in a chosen cohort some individuals are prone to some special types of cancer diseases while some other are not who are called immune people. This difference which causes various range of susceptibility or immunity could be because of various risk factors such as behavioral, environmental or hereditary factors, as mentioned above. Consider that π_0 is the proportion of the population who are susceptible to get a type of cancer. Assume that to get cancer the cells of the body need some time to expose environmental risk factors, so clearly the starting age zero for this proportion of the people necessarily does not need to be zero and they may get into the two sub cohorts due to environmental and behavioral risk factors, later after their birth years.

For these sub cohorts of the whole population, the force of mortality from the cancer at age x is denoted by $\mu_c(x)$. In case that the mortality among population has any other reason, for both immune and susceptible persons the force of mortality is indicated by $\mu_0(x)$. Then, the observed force of mortality or incidence by cause of cancer for a whole considered population i.e. $\overline{\mu_c}(x)$ is defined as

$$\overline{\mu_c}(x) = \pi(x)\mu_c(x)$$

where $\pi(x)$ is the susceptible proportion of the population who are alive at age *x* and it is defined by the formula below

$$\pi(x) = \frac{\pi(0)\exp\left(-\int_0^x(\mu_c(t)+\mu_0(t))dt\right)}{\pi(0)\exp\left(-\int_0^x(\mu_c(t)+\mu_0(t))dt\right)+(1-\pi(0))\exp\left(-\int_0^x\mu_0(t)dt\right)}$$

by canceling $\pi(0)\exp\left(-\int_0^x\mu_0(t)dt\right)$, we have

$$\pi(x) = \frac{\exp\left(-\int_0^x\mu_c(t)dt\right)}{\exp\left(-\int_0^x\mu_c(t)dt\right)+\left(\frac{1}{\pi(0)}\right)-1}$$

subsequently we gain

$$\pi(x) = \frac{\exp\left(-\int_0^x\mu_c(t)dt\right)}{\exp\left(-\int_0^x\mu_c(t)dt\right)+\frac{1-\pi(0)}{\pi(0)}}$$

then by inversing both sides of the equality we have

$$\frac{1}{\pi(x)} = 1 + \frac{1-\pi(0)}{\pi(0)} exp\left(-\int_0^x \mu_c(t)\, dt\right)^{-1}$$

which is equal to

$$\pi(x) = \left(1 + \frac{1-\pi(0)}{\pi(0)} \exp\left(\int_0^x \mu_c(t)\, dt\right)\right)^{-1}$$

As you see, in the final formula the force of mortality caused by the reasons except cancer does not exist. This means that, in this model cancer and other causes are independent risks, (Vapuel & Yashin, 1999).

The prototypical model can be generalized to a different model which is called prototypical frailty model. In this model instead of having only one force of mortality from causes other than cancer, we assume to have two subpopulations P and P′ with the force of mortality by cause of some other reasons than cancer μ_0 and μ_0' respectively. Then, in case that $\mu_0 > \mu_0'$, then we can conclude that the individuals belonging to P are frail. It means that they are susceptible to get cancer and also have greater chance to die because of the reasons other than cancer. Therefore for generalizing formula given for prototypical model, we can assume that π depends on both μ_c and μ_0:

$$\pi(x) = \left(1 + \frac{1-\pi(0)}{\pi(0)} exp\left(\int_0^x (\mu_c(t) + \mu_0(t) - \mu_0'(t))dt\right)\right)^{-1}$$

where

$$\overline{\mu_0}(x) = \pi(x)\mu_0(x) + (1-\pi(x))\mu_0'(x)$$

and as before we assume that

$$\overline{\mu_c}(x) = \pi(x)\mu_c(x)$$

Note that now $\overline{\mu_c}(x)$ and $\overline{\mu_0}(x)$ are related via the common relation with $\pi(x)$ and they are not independent anymore, (Vapuel & Yashin, 1999).

Prototypical frailty model also can be generalized to some other heterogeneity models, (Vapuel & Yashin, 1999). Although heterogeneity models are adequate in order to indicate the difference in a heterogeneous population, they do not describe the internal biological that leads to observe in the dynamics. (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005).

4.1.5 An explanation for application of Game theory and ODE in modeling

One of the interesting novel fields in the approach of cancer incidence modeling is the development of some preliminary ideas to apply some fundamental paradigms such as ODEs, Game theory, phase logic, and so on in the research field under consideration. Several researchers have applied the mentioned mathematical approaches to relate cancer data to logical formulas such as Nowak and Sigmund, (Nowak & Sigmund, 2004).

4.2. Age-specific modeling for cancer incidence

In 1975, the incidence rates were observed over age for various cancer types such as bronchus, stomach, colon, rectum, pancreas, skin, male bladder, and female breast and ovary. Moreover, there were some similar observations throughout the world from 1960 to 1975. The result was interesting. The alteration of cancer incidence rate with age for all cancer types had similar behavior. Another result achieved by these observations was that cancer is dependent on the tissue where it occurs, (Dix,

1989). Recently there have been many other studies during various periods of times and among different cohorts, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005). Because of the different mechanism corresponding to a special site of the human body, age-specific incidence rates act basically different for different cancer types. For example during climacteric ages for women age- specific patterns for hormone dependent cancers such as ovarian or endometrium cancers have a wave-like shape. This behavior is because of the instability of patient's hormones which leads to morbidity and subsequently decreasing the immune balance. Despite this fact, the new observations confirms the old gained results and similarly claim that there are some prevalent cancer types such as lung, stomach, and colon which treat similarly for both male and female regardless of the place where they live and the time period when They have been diagnosed with cancer.

Although site-specific cancer studying and analyzing are very important for scientists and give much more detailed results concerning the mechanism of each site, this issue should not detract from the importance of studying on the overall cancer incidence rates. Hence from here onwards I would like to direct the readers to focus on the overall cancer incidence rates. Figures 12, 13 and 14 show the relation between age and cancer incidence rate in different periods of time in Japan.

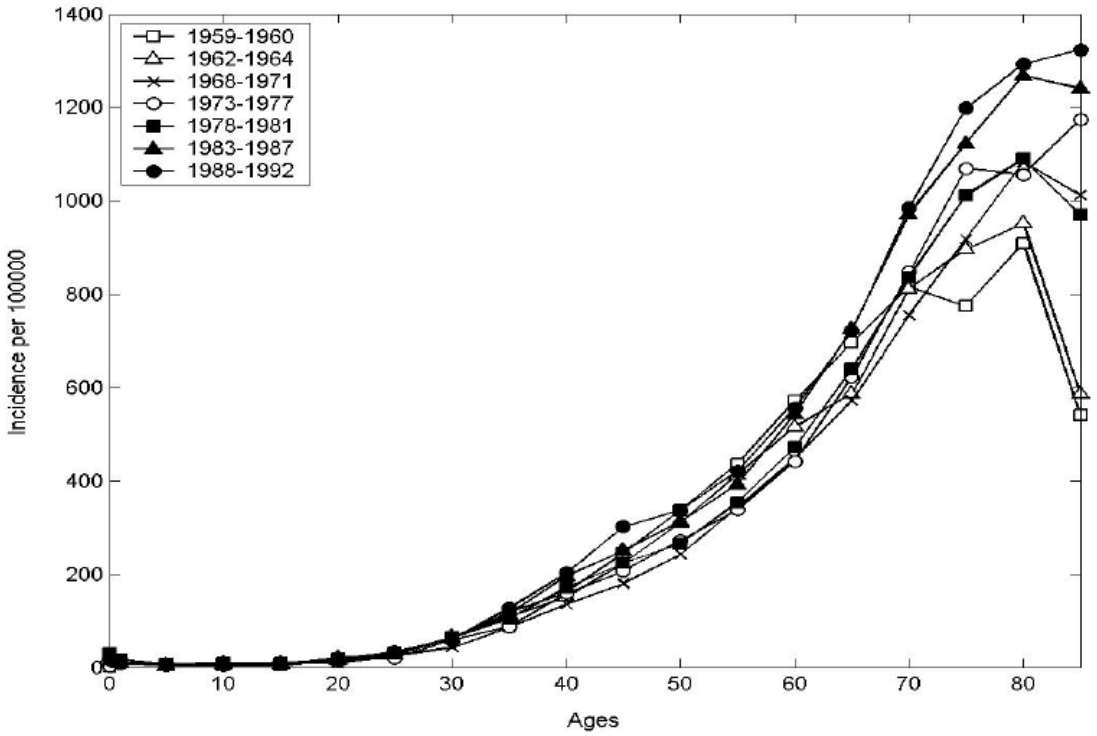

Figure 12: Cancer incidence rates over age for females in Japan (Miyagi Prefecture), (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005)

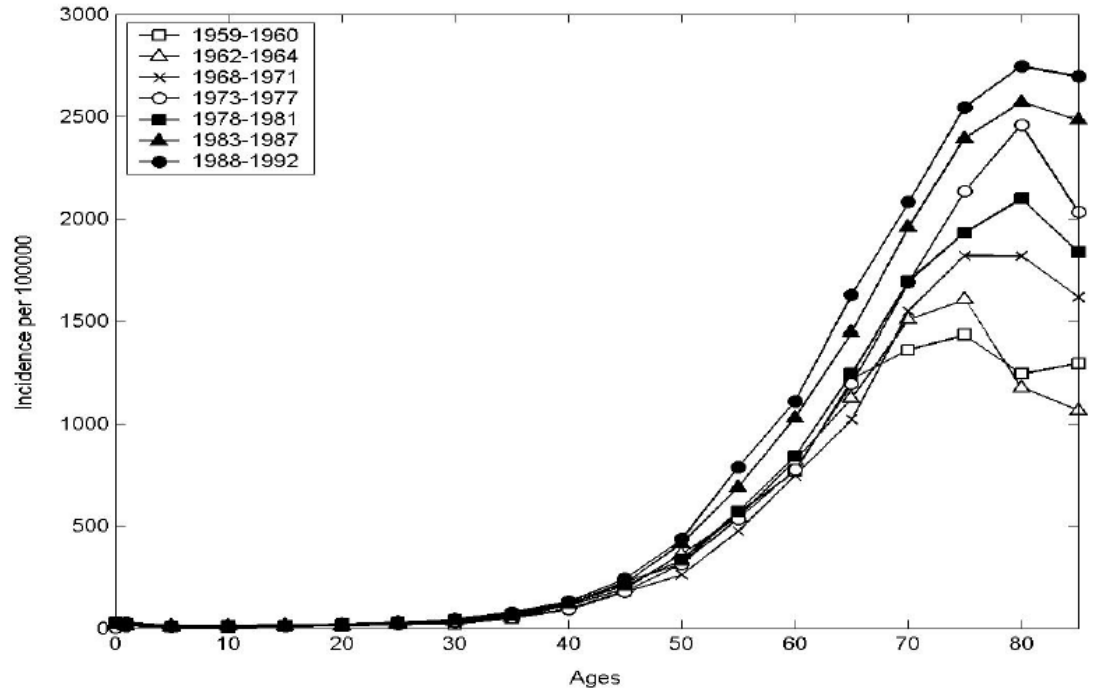

Figure 13: Cancer incidence rates over age for males in Japan (Miyagi Prefecture), (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005)

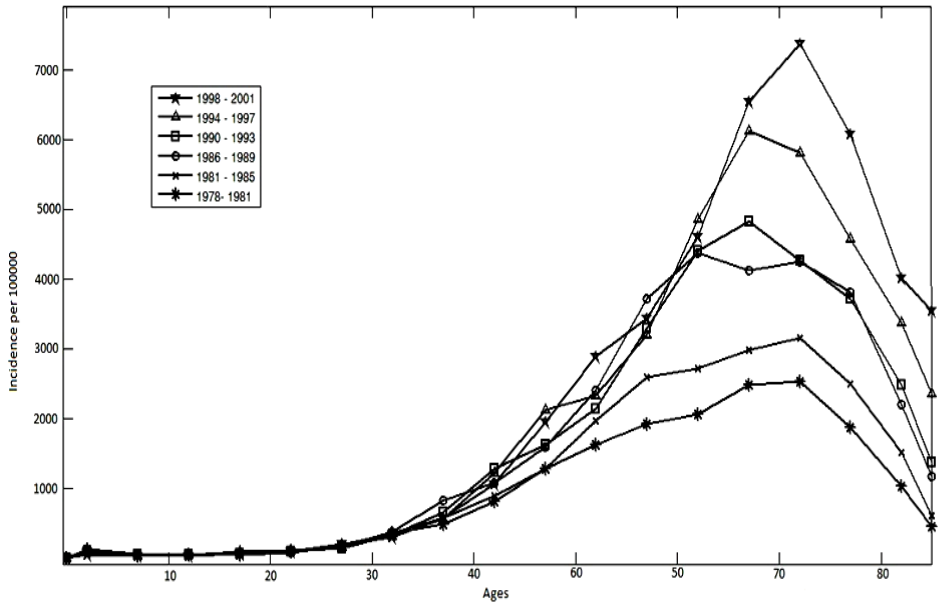

Figure 14: Overall cancer incidence rates over in Japan (Miyagi Prefecture)

4.2.1 Age pattern of the cancer incidence rate

The increase and decrease of cancer trajectory has various interpretations. The behavior of cancer incidence is somehow strange and queer. As can be seen in figures 15 and 16 cancer incidence rates level off or even decline at very old ages. It can be considered as a mortality risk factor between ages 50 and 60 while its danger declines during younger ages. Cancer incidence treatments are different for different countries whereas the overall cancer incidence curves for different countries are similar. Countries with the high rate of age- specific cancer incidence have low rates of age specific mortalities. The mortality statistics for some cancer types such as lung cancer shows an increase in recent years while the empirical mortality data for some other types of cancer such as stomach cancer indicates a decline, (Vapuel & Yashin, 1999).

Typical age pattern of the overall cancer incidence rates contains a peak during early childhood and then a low rate during youth. Then it increases during adolescence.

And finally it declines at old ages, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005).

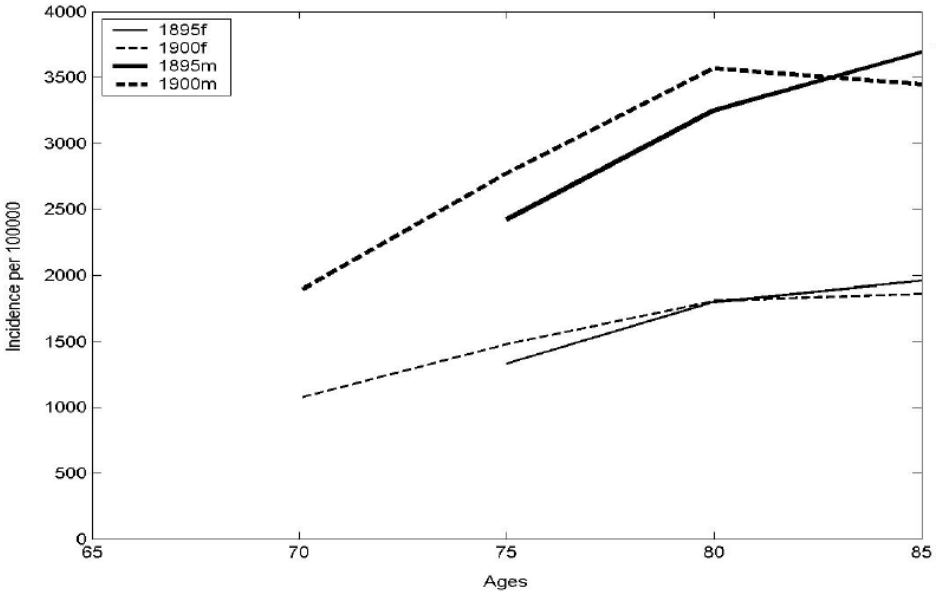

Figure 15: The decrease of cohort cancer incidence rate in the oldest old ages. Females are show with thin lines and males with thick lines in New York, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005)

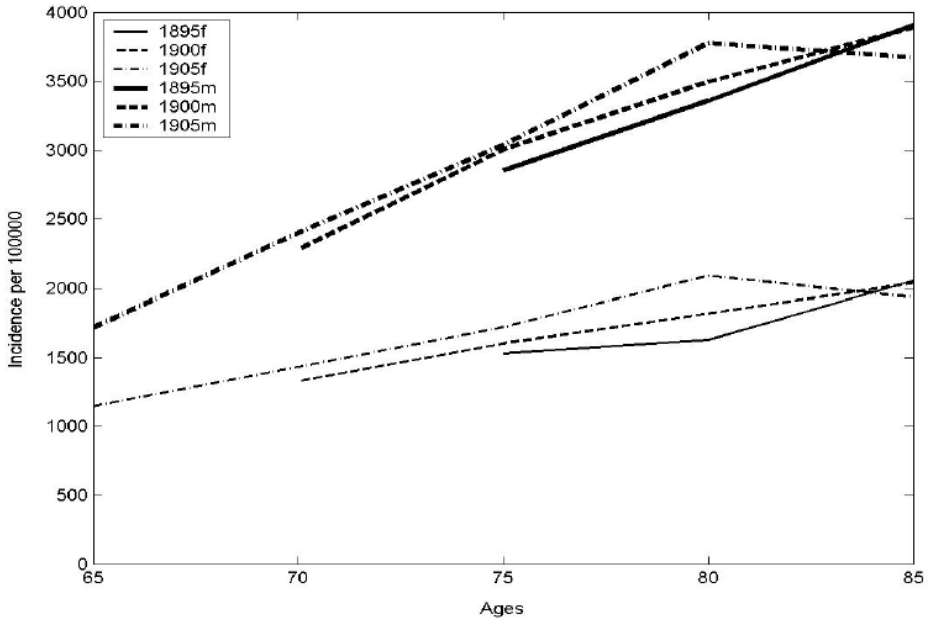

Figure 16: The decrease of cohort cancer incidence rate in the oldest old ages. Females are shown with thin lines and males with thick lines in San Francisco, (Konstantin, and et al. 2005).

4.2.2 Strehler and Mildvan model

This model is inspired by the Sterehler and Mildvan theory of mortality with the hypothesis that an organism has a capacity to remain healthy at age x. This capacity is called vitality and is indicated with $\mathrm{V(x)}$ and is defined as below

$$V(x) = V_0(1 - Bx),$$

where B is the slope of the vitality curve. V_0B is interpreted as the rate of physiological aging.

Assume that the intensity of events related to external stress which is indicated with K does not depend on age. Let $\mathcal{E}_{\mathrm{D}}$ be an average magnitude of stress. According to the mentioned assumptions the observed cancer incidence rates are

$$\mu(x) = Ke^{-\frac{V(x)}{\mathcal{E}_D}},$$

Strehler and Mildvan model can be related to Gompertz law of mortality if we assume that $a = Ke^{-\frac{V_0}{\mathcal{E}_D}}$ and $b = \frac{V_0B}{\mathcal{E}_D}$ then

$$Ke^{-\frac{V(x)}{\mathcal{E}_D}} = ae^{bx},$$

also there is a relationship between Gompertz parameters a and b

$$\ln a = \ln K - \frac{b}{B}.$$

An immediate result which can be concluded from this model after affecting on empirical cancer data is that it produces negative values for oldest ages. It means that this model meets the attributes of typical age pattern as mentioned before, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005).

4.2.3 Revised Mildvan and Strehler model

In this model according to the empirical data behavior it is assumed that the vitality function which was defined as a linear function in Mildvan and Strehler model, rates exponentially. In other words the hypothesis which enables to revise the in Mildvan and Strehler is that there is an age related decline in the individual vitality with age. Therefore based on this assumption we have, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005)

$$V(x) = V_0 e^{-Bx} ,$$

and the respective rate of individual aging, $r(x)$, is defined as

$$r(x) = -\frac{dV(x)}{dx} = V_0 B e^{-Bx} ,$$

the vital difference between this definition and the later offered model is in the rates of aging which is here changing with individual aging progress while in the later definition it was constant i.e. $V_0 B$.

Another difference is in the definition of parameter B, which in Mildvan and Strehler model was defined as the slope of the vitality curve while here it can be considered as the logarithmic rate of aging since

$$r_{log}(x) = -\frac{d(logV(x))}{dx} = -\frac{dV(x)}{dx}\,\frac{1}{V(x)} = \frac{r(x)}{V(x)} = B.$$

Therefore, in Revised Mildvan and Strehler model parameter B determines the slope of the logarithmic vitality curve, logV(x), and the incidence rate is defined as

$$\mu(x) = Ke^{-\frac{V_0 e^{-Bx}}{\varepsilon_D}}$$

which can be simplified as

$$\mu(x) = Ke^{-\frac{r(x)}{\varepsilon_D B}}.$$

Chapter 5

ATTEMPTS TO FIND A NEW MODEL WITH THE BEST GOODNESS OF FIT

5.1 Data

The data used here is chosen from the statistical research provided by International Agency for Research on Cancer (IARC) in nine volumes (1965-2002). In each volume one can find the information about cancer incidence both by the specific site and by overall statistics for all sites for a period of time around 3 to 5 years corresponding to various countries all over the world. For the overall cancer incidence rates here we use the data available for all cancers except non-melanoma skin cancer. Moreover the incidence data is available for both male and female corresponding to each country. The incidence rate is mentioned as a number per 100,000 in the population and it is gathered from 0-5 age group up to age 85 and above. However, for some countries the first group is divided into two groups: 0 and 0-4. The data provided in each volume is special for some countries and all countries do not appear in the all volumes except Japan which is available in all volumes. Japan has provided the longest time series especially for Miyagi prefecture. Thus it is the best region to observe cancer incidence rates over time. In addition to Japan we apply our model to some other countries all over the world such as Canada, USA and some chosen regions in Asian, African, and European countries, (IARC(International Agency for Research on Cancer), 1965), (IARC(International Agency for Research on Cancer), 1970), (IARC(International Agency for Research

on Cancer), 1976), (IARC(International Agency for Research on Cancer), 1982), (IARC(International Agency for Research on Cancer), 1976), (IARC(International Agency for Research on Cancer), 1976), (IARC(International Agency for Research on Cancer), 1976), (IARC(International Agency for Research on Cancer), 1976), (IARC(International Agency for Research on Cancer), 1976).

5.2 Goodness of fit

To compare the fitted curves some preliminary definitions are required which are explained first. For our interpolation we use four existing approaches which indicate the goodness of fit in Matlab: Goodness of fit; SSE, R-square, Adjusted R-square, and RMSE, (Mathworks, 2011).

5.2.1 The sum of squares due to error (SSE)

This statistic tool calculates the total deviation of the response values from the fitted curve to the response values. In mathematic literature it may also called as residual or sum square error. Matlab indicates it with SSE which is defined as,

$$SSE = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2 \quad ,$$

in case that SSE is close to zero, the approximation is more accurate rather than the case it is farer from zero. In other words the smaller SSE value the more adequate to predict the future values, (Mathworks, 2011).

5.2.2 R-Square

R-square is the square of the correlation between the response values and the predicted response values. It is also called the square of the multiple correlation coefficients and the coefficient of multiple determinations.

R-square is defined as the ratio of the sum of squares of the regression (SSR) and the total sum of squares (SST). SSR is defined as

$$SSE = \sum_{i=1}^{n} w_i(\hat{y}_i - \bar{y}_i)^2 \quad ,$$

SST is also called the sum of squares about the mean, and is defined as

$$SST = \sum_{i=1}^{n} w_i(y_i - \bar{y}_i)^2 \quad ,$$

where $SST = SSR + SSE$. Therefore according to the above definitions R-Square is defined as

$$\text{R-square} = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad ,$$

R-square is a number between zero and one. In case that it is close to one it means a greater proportion of variance is considered for the fitting. For instance an R-square value of 0.8751 means that the fitted curve contains the 87.51% of the total variation in the data about the average, (Mathworks, 2011).

5.2.3 Degrees of freedom adjusted R-Square

This tool applies the R-square error as defined above, and adds it according to the residual degrees of freedom. The residual degrees of freedom is defined as the number of response values n minus the number of fitted coefficients m estimated from the response values:

$$V = n - m \quad ,$$

The adjusted R-square statistic can take on any value less than or equal to 1.in case that its values close to one we get a better curve fitting. Negative values happen while the approximated function gives some values which are not helpful to estimate the new cases, (Mathworks, 2011).

5.2.4 Root Mean Squared Error (RMSE)

This approach is also called fit standard as well as the standard error of the regression. It is an estimate of the standard deviation of the random component in the data, and is defined as

$$RMSE = s = \sqrt{MSE} \quad ,$$

where MSE is the mean square error or the residual mean square

$$MSE = \frac{SSE}{v} \quad ,$$

similar to SSE, an MSE value which is closer to zero means that the fitted curve is more useful for estimation, (Mathworks, 2011).

5.3 Curve fitting of overall cancer incidence rate data

5.3.1 Attempts to find the best fit

We apply various special models on the cancer incidence rates data in different regions and time periods which were explained in the data section by using Matlab's curve fitting tool box.

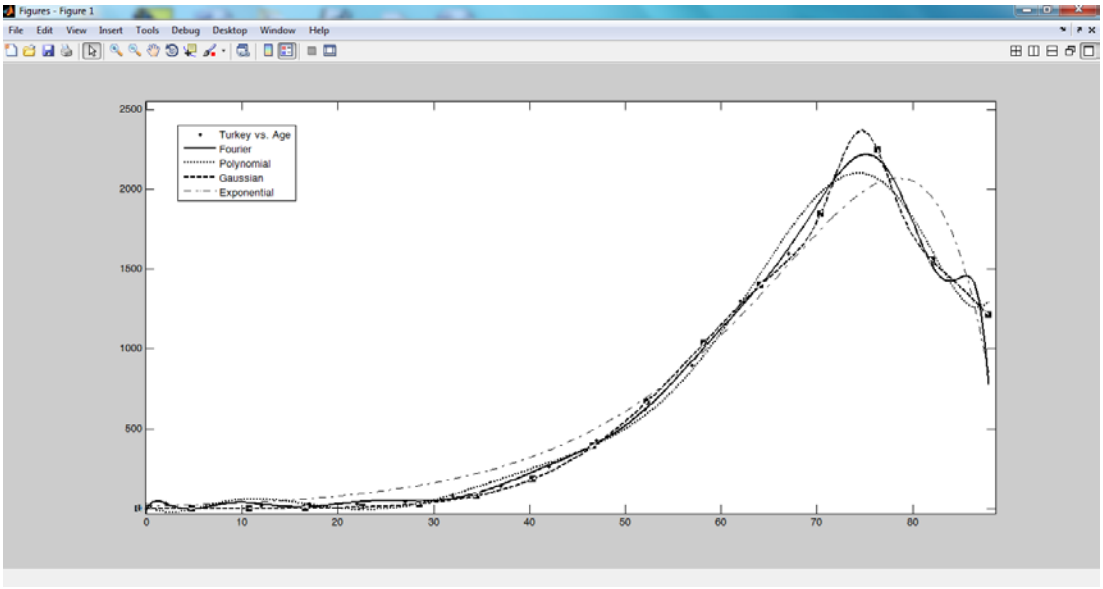

Figure 17: Four applied models to the cancer incidence rates of Turkey (Izmir) in the time period 1998-2002

The achieved functions after these fittings are as below

Fourier model:

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3*sin(3*x*w) + a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5*sin(5*x*w) + a6*cos(6*x*w) + b6*sin(6*x*w) + a7*cos(7*x*w) + b7*sin(7*x*w)

Coefficients (with 95% confidence bounds):

a0 = -6.461e+013 (-1.011e+017, 1.01e+017),

a1 = 1.03e+014 (-1.632e+017, 1.634e+017),

b1 = 4.761e+013 (-6.943e+016, 6.953e+016),

a2 = -4.957e+013 (-8.296e+016, 8.286e+016),

b2 = -5.826e+013 (-8.623e+016, 8.611e+016),

a3 = 1.047e+013 (-2.166e+016, 2.168e+016),

b3 = 3.752e+013 (-5.686e+016, 5.693e+016),

a4 = 2.325e+012 (-8.614e+014, 8.661e+014),

b4 = -1.435e+013 (-2.271e+016, 2.268e+016),

a5 = -2.1e+012 (-2.499e+015, 2.495e+015),

b5 = 3.115e+012 (-5.305e+015, 5.311e+015),

a6 = 5.147e+011 (-6.997e+014, 7.007e+014),

b6 = -3.121e+011 (-6.265e+014, 6.259e+014),

a7 = -4.483e+010 (-6.664e+013, 6.655e+013),

b7 = 5.087e+009 (-2.271e+013, 2.272e+013),

w = 0.01032 (-1.125, 1.146).

Polynomial model:

f(x) = p1*x^8 + p2*x^7 + p3*x^6 + p4*x^5 + p5*x^4 + p6*x^3 + p7*x^2 + p8*x + p9

Coefficients (with 95% confidence bounds):

p1 = 7.142e-010 (1.822e-010, 1.246e-009),

p2 = -2.328e-007 (-4.182e-007, -4.739e-008),

p3 = 3.049e-005 (4.077e-006, 5.691e-005),

p4 = -0.002061 (-0.00404, -8.236e-005),

p5 = 0.07661 (-0.006778, 0.16),

p6 = -1.525 (-3.483, 0.4323),

p7 = 14.57 (-9.044, 38.19),

p8 = -49.37 (-168.4, 69.64),

p9 = 31.61 (-122, 185.2),

Gaussian model:

f(x) = a1*exp(-((x-b1)/c1)^2) + a2*exp(-((x-b2)/c2)^2)

Coefficients (with 95% confidence bounds):

a1 = 651.2 (249.4, 1053),

b1 = 74.67 (74.22, 75.12),

c1 = 3.81 (0.457, 7.163),

a2 = 1717 (1625, 1810),

b2 = 74.42 (73.79, 75.05),

c2 = 22.86 (21.66, 24.05).

Exponential model:

f(x) = a*exp(b*x) + c*exp(d*x)

Coefficients (with 95% confidence bounds):

a = -6.75e+004 (-6.007e+013, 6.007e+013),

b = 0.08668 (-1317, 1318),

c = 6.752e+004 (-6.007e+013, 6.007e+013),

d = 0.08668 (-1317, 1318).

Table 4: Comparing the goodness of fit for the best fitted models for cancer incidence rates of Turkey (Izmir) during 1998-2002

Model	**SSE**	**R-square**	**Adjusted R-square**	**RMSE**
Fourier	68.64	1	1	4.783
Polynomial	585.4	1	1	8.065
Guassian	$4.541* 10^4$	0.9994	0.9993	53.27
Exponential	$4.931* 10^5$	0.9931	0.9918	181.3

According to the both numerical results indicated in table 4 and graphical results illustrated in figures 17 and 18 we can conclude the best fit among the four specific functions Furrier, Polynomial, Gaussian, and Exponential, the Fourier function gives the best fitting. However, Gaussian model has very small difference with the Furrier pattern for this fitting.

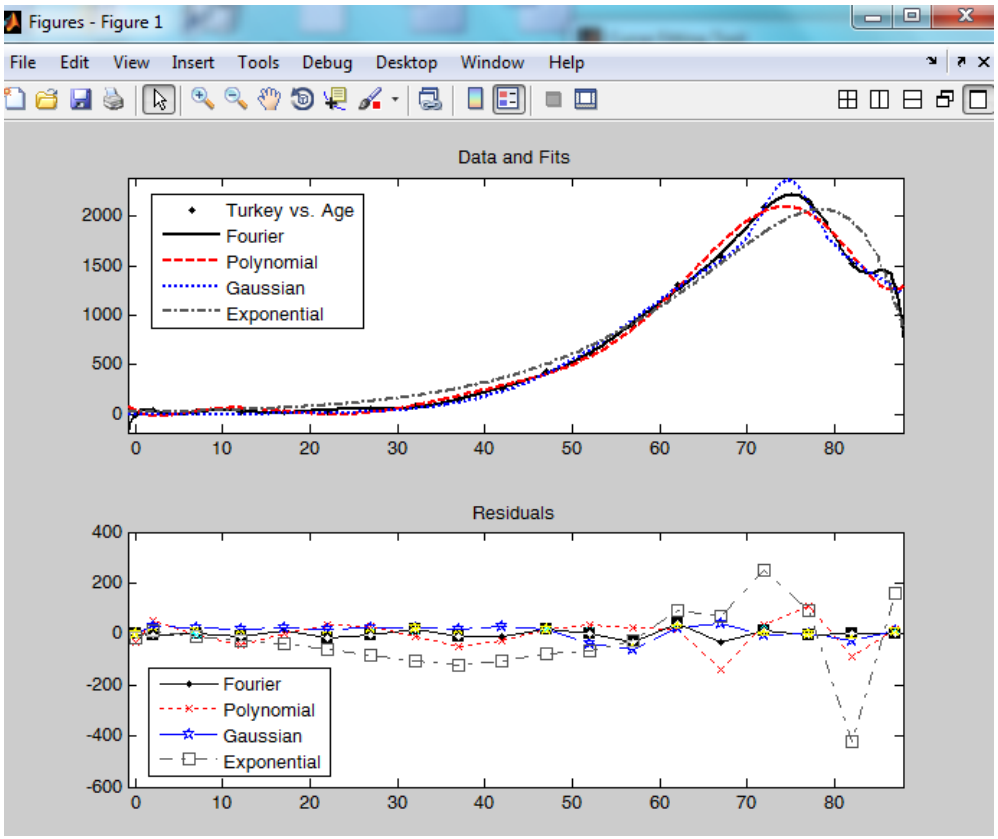

Figure 18: Comparing the residuals to determine the best fitting for the cancer incidence rates data of Turkey (Izmir) in the time period 1998-2002

Similar observation on the cancer incidence rates data of Canada (excluding Quebec, Yukon and Nunavut) during 1998 – 2002 gives the same result. It can be seen both from the numerical results existing in table 5 and graphical results shown in figure 18.

Several other observations on the data of different countries during different time periods give nearly the same result.

In the following sections some comparisons are checked according to different sexes, regions, time periods, and the amount of development in different countries.

5.3. 2 Curve fitting of the cancer incidence rates for males and females

Similar study on the separated cancer incidence data for both female and male shows that the Fourier model provides the best fitting as well as the total cancer incidence rate data.

Table 5: Comparing the goodness of fit for the best fitted models for male cancer incidence rates of Canada (excluding Quebec, Yukon and Nunavut) in 1998 – 2002

Model	**SSE**	**R-square**	**Adjusted R-square**	**RMSE**
Fourier	1.22	1	1	0.6376
Guassian	$5.087* 10^4$	0.9983	0.9981	56.38

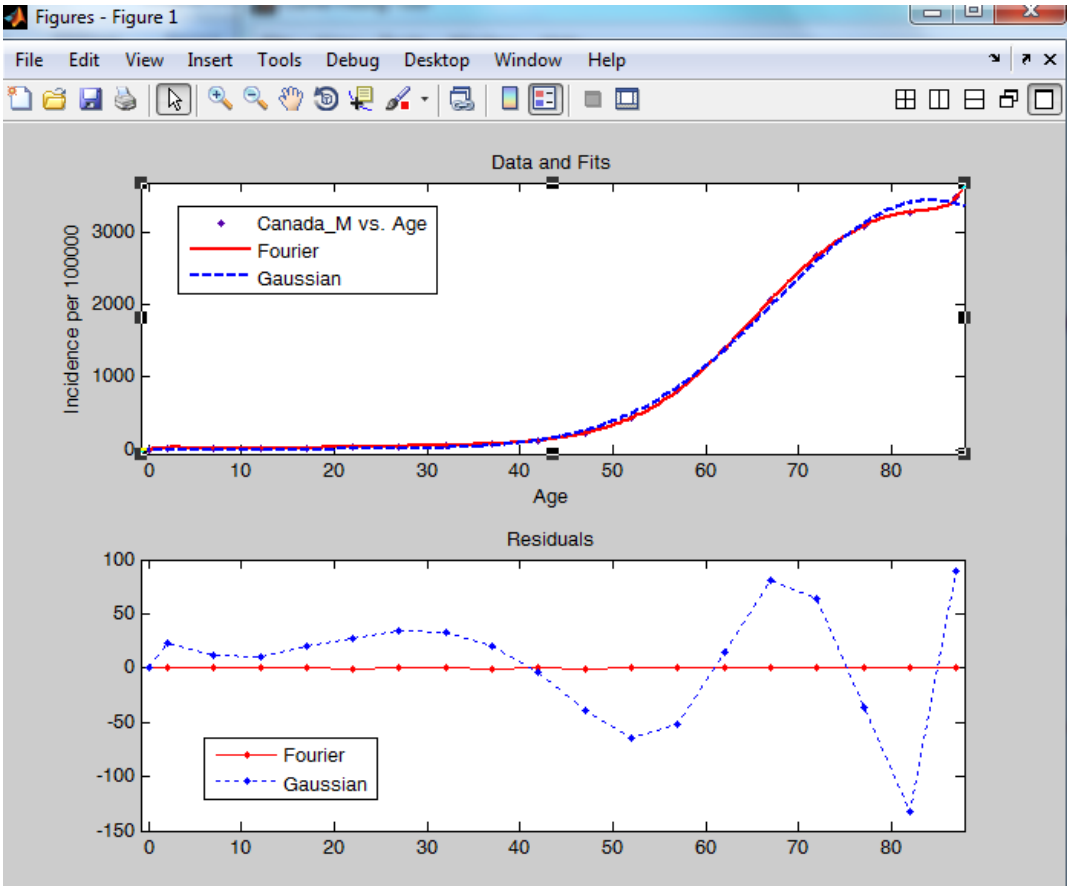

Figure 19: Comparing the residuals to determine the best fitting for the male cancer incidence rates data of Canada (excluding Quebec, Yukon and Nunavut) in 1998 – 2002

Table 6: Comparing the goodness of fit for the best fitted models for male cancer incidence rates of Canada (excluding Quebec, Yukon and Nunavut) in 1998 – 2002

Model	**SSE**	**R-square**	**Adjusted R-square**	**RMSE**
Fourier	66.97	1	1	4.725
Guassian	1568	0.9998	0.9998	9.898

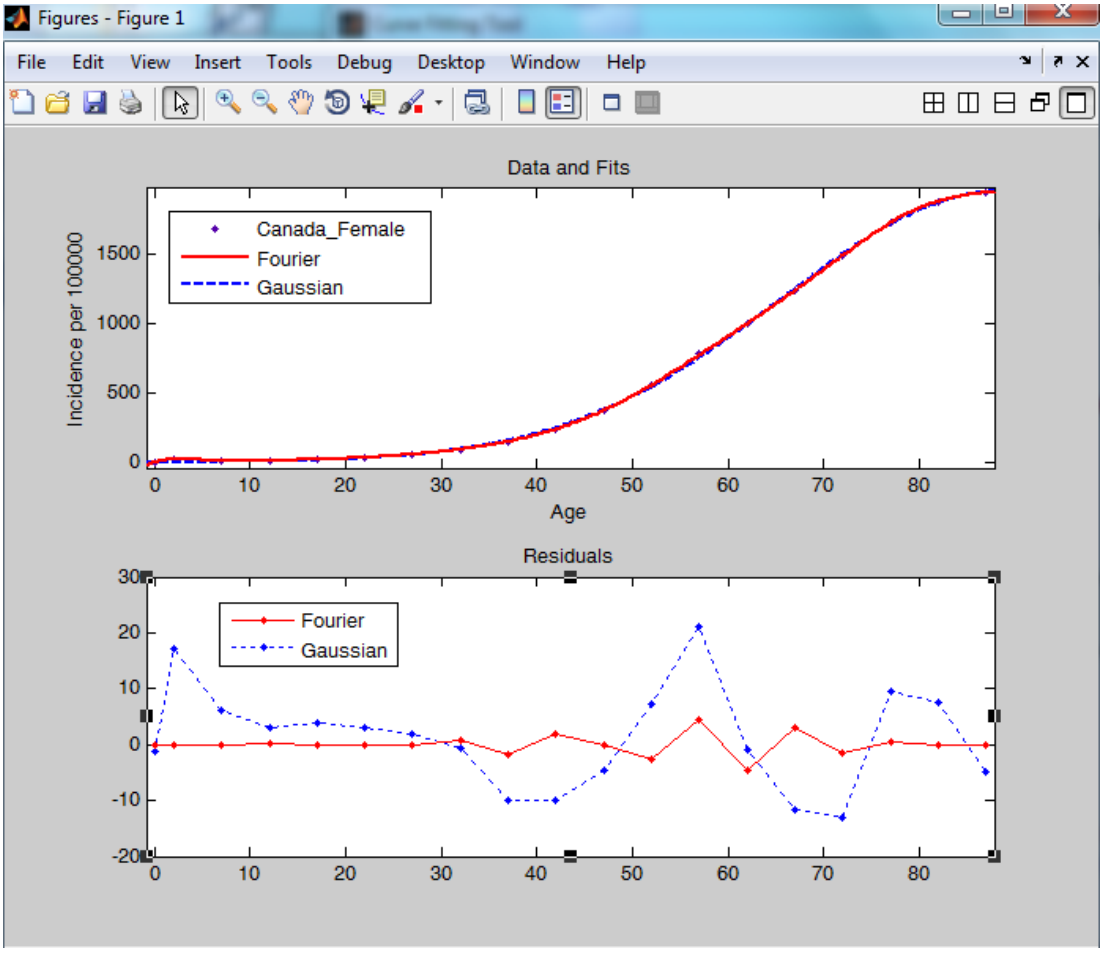

Figure 20: Comparing the residuals to determine the best fitting for the female cancer incidence rates data of Canada (excluding Quebec, Yukon and Nunavut) in 1998 – 2002

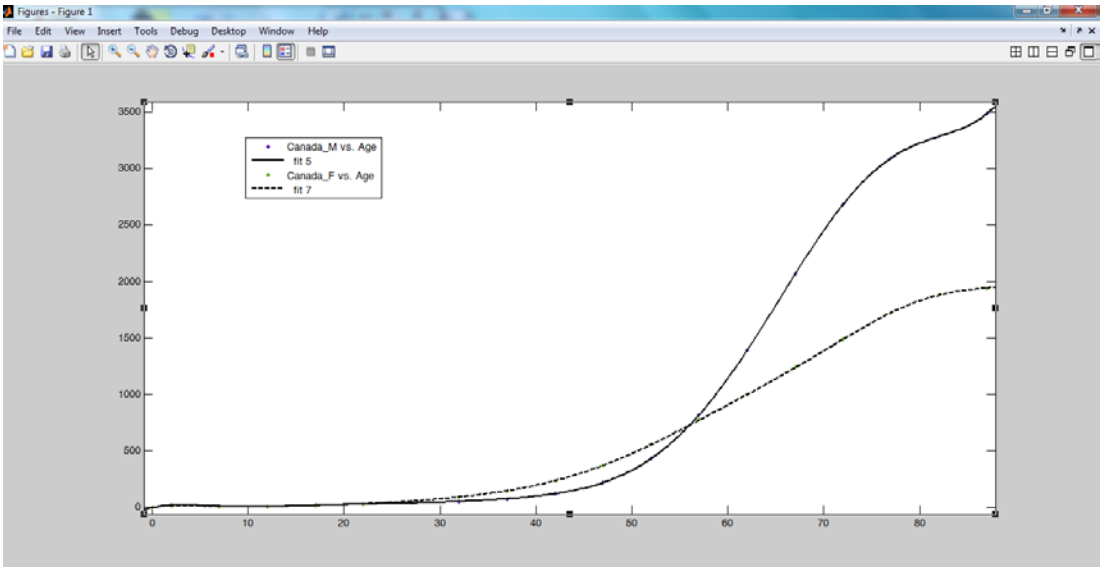

Figure 21: Comparing male and female cancer incidence rates data of Canada (excluding Quebec, Yukon and Nunavut) in 1998 – 2002

5.3. 3 Curve fitting of the cancer incidence rates of different regions

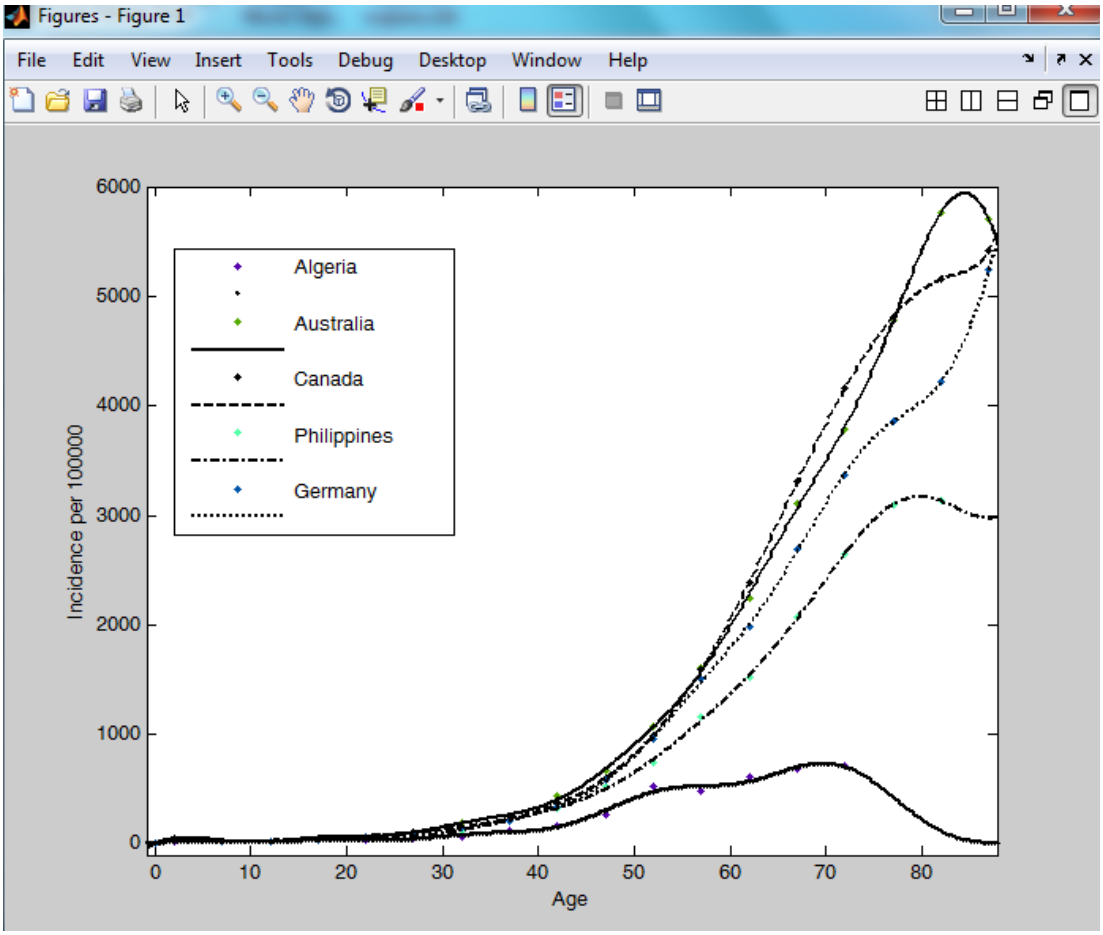

Figure 22: Cancer incidence rates curve fitting of Algeria(Setif), Philippines,(Manila), Australian capital territory, and Germany(Hamburg)

Although the incidence rates for different regions are different, their behaviors obey the same pattern. Figure 22 shows the curve fitting of different regions in five continuums which are all fitted by the same formula i.e. Furrier model. As it can be seen regardless of the amount of incidence rates which are not the same the shape of the curves are similar.

5.3. 4 Curve fitting of the cancer incidence rates of different time periods

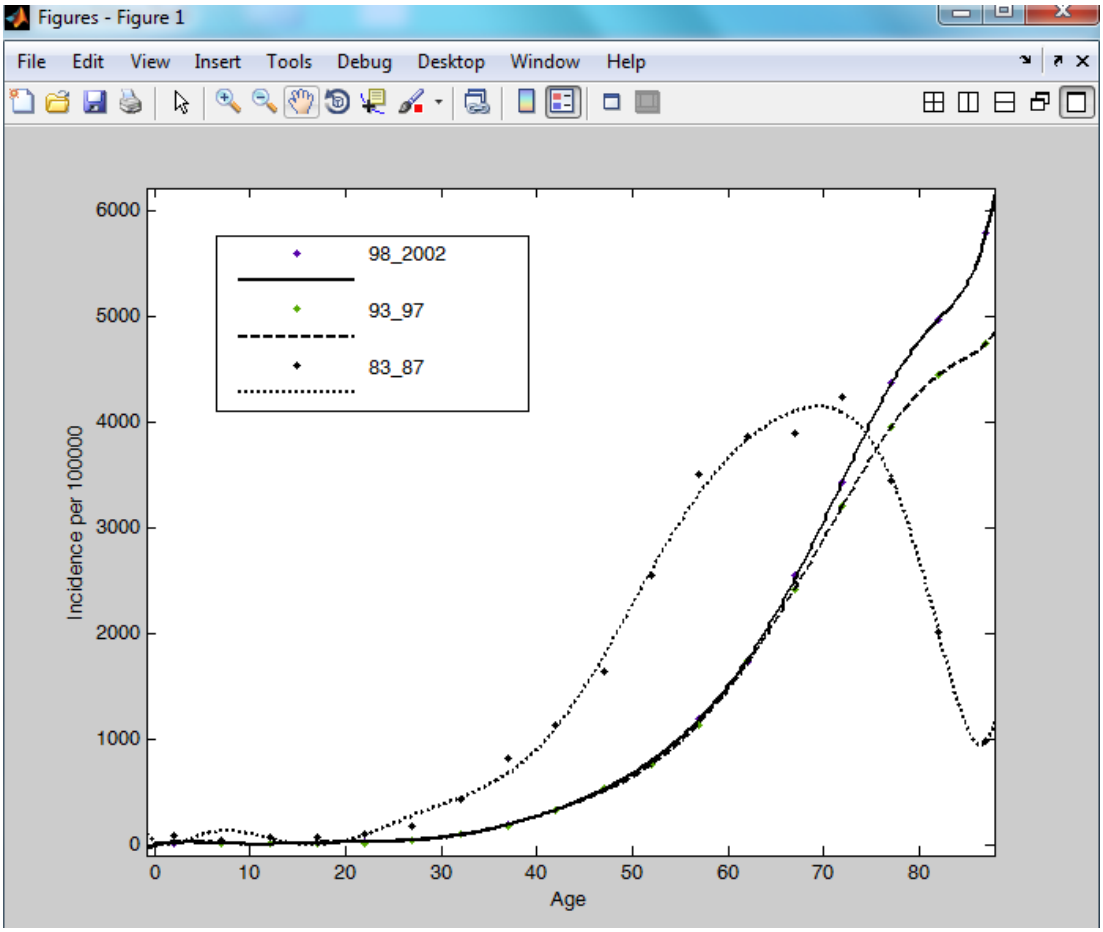

Figure 23 Comparing the cancer incidence rates of Japan (Miyagi prefecture) during 1983-1987,1993-1997, and 1998-2002

As it can be seen from the figure 23 the incidence rates curves for the two subsequent time periodss from 1993 to 1997 and from 1998 to 2002 have the same behavior, while the similar curve for the time period from 1983 to 1987 does not behave the same. It may be interpreted because of the outcomes after the second war

in 1941. However in recent years the rate of cancer incidence has been increased for older people in Japan which can be acceptable regarding to the recent increase of the number of smokers in Japan.

5.3. 5 Curve fitting of the cancer incidence rates of different races

To do this observation the USA different races including White, Black, Chinese, Japanese, and Filipino are chosen to compare.

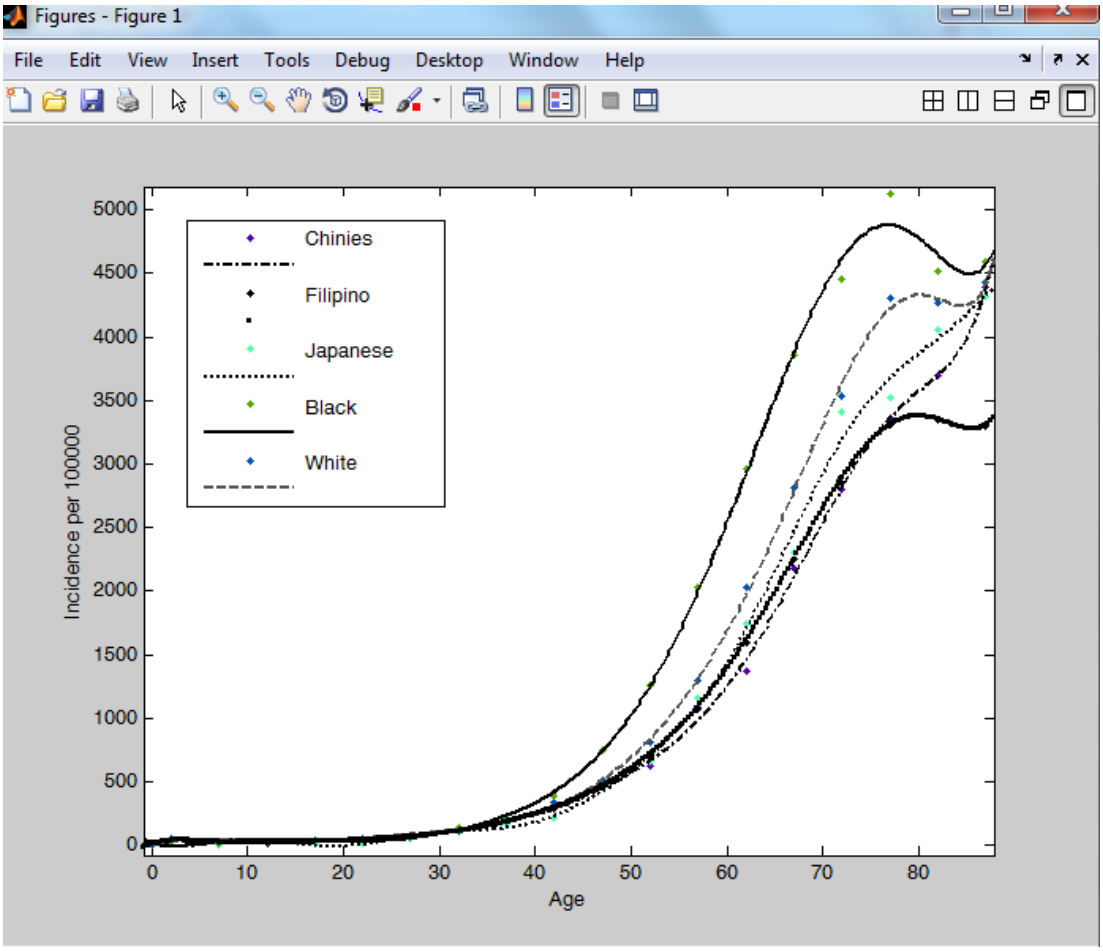

Figure 24: Comparing the cancer incidence rates among different races in USA, California, Greter San Francisco Bay Area including Chinese, Japanese, Filipino, Black, White during 1998-2002

Cancer incidence rates among American Blacks are the highest rather than the other races and this rate is the lowest for Filipinos. Regardless of the amount of incidence rates all the cancer incidence curves for all races behave the same.

5.3. 6 Curve fitting of the cancer incidence rates of different developments

We compare the cancer incidence rates of several developed countries with developing countries. The interesting result is that this rate is higher in developed countries.

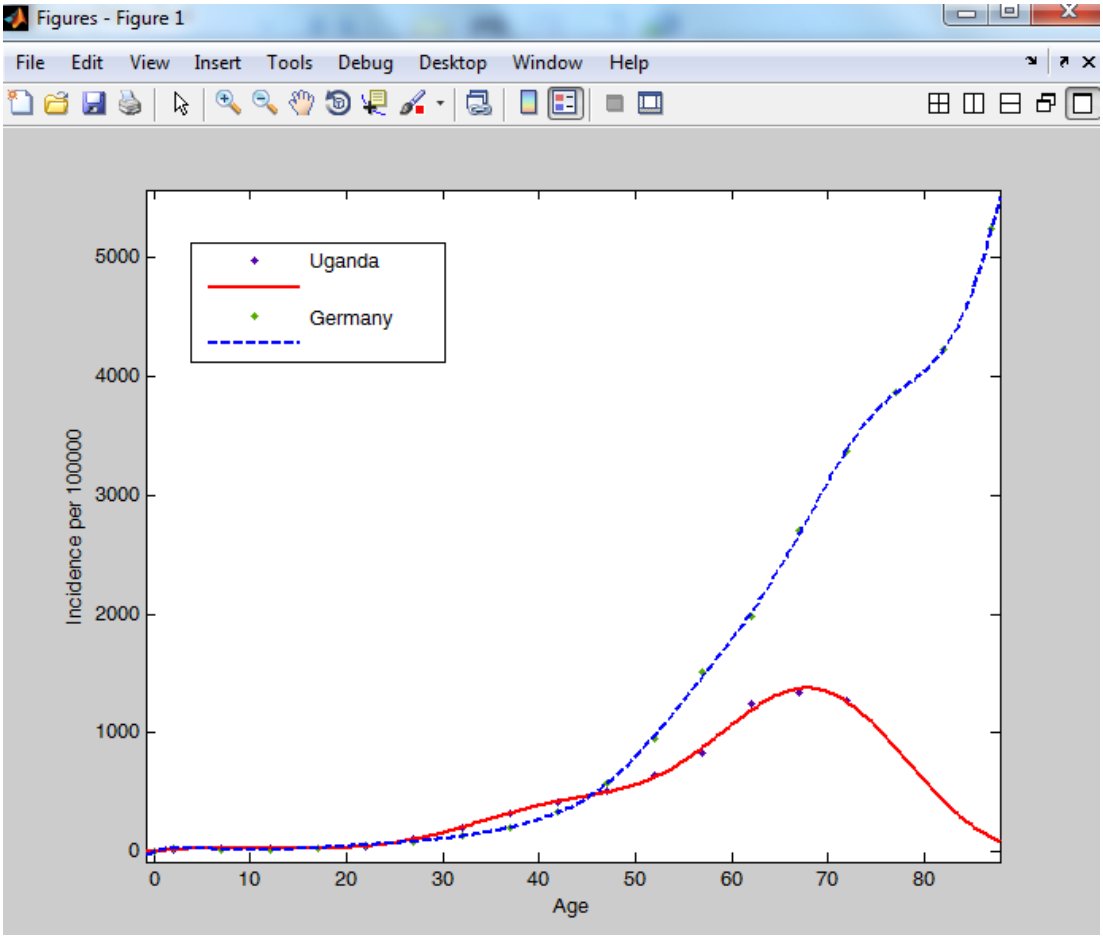

Figure 25: Comparing Uganda (Kyadondo County) with Germany (Hamburg) during 1998-2002 as a developing and a developed country respectively.

This may be interpreted because of traffic and air pollution and using medicines and chemical products which are more available in a developed country rather than a developing one. Also as the later explanations in chapter 3, the advent of new technologies such as mobile phones and the application of new facilities such as X-ray and toxic medical products, all can be the reasons of this increase in developed countries. Figure 25 is an example of this observation which illustrates this fact clearly.

5.3.5 Comparing the cancer incidence between males and females

In addition to the cancer incidence rates also we applies some models to the cancer incidence data. Figure 26 shows the best three curve fitting models which are applied to the cancer incidence data of Denmark during 1999 to 2002.

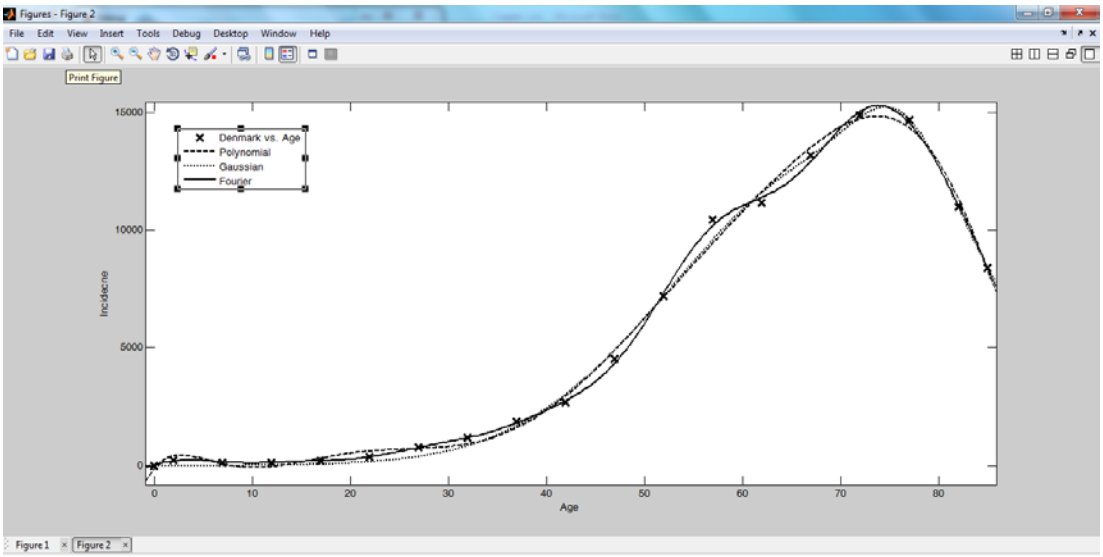

Figure 26: Three applied models to the cancer incidence of Denmark in time period 1999-2002

The goodness of fit for these three models is available in table 4.

Table 4: Comparing the goodness of fit for the best fitted models for cancer incidence of Denmark during 1999-2002

Model	**SSE**	**R-square**	**Adjusted R-square**	**RMSE**
Polynomial	$2.121 * 10^6$	0.9962	0.9924	485.5
Gaussian	$1.636 * 10^6$	0.9971	0.996	354.8
Furrier	$2.946 * 10^5$	0.9995	0.9969	313.4

The three models provide an adequate fit to the cancer incidence rate data in different countries. Comparing four available tools for checking goodness of fit, Furrier model has achieved the best results.

Figure 27 shows fitting cancer incidence data with Furrier model for Australia (New South Wales), Canada (Alberta), Denmark, Japan (Miyagi prefecture).

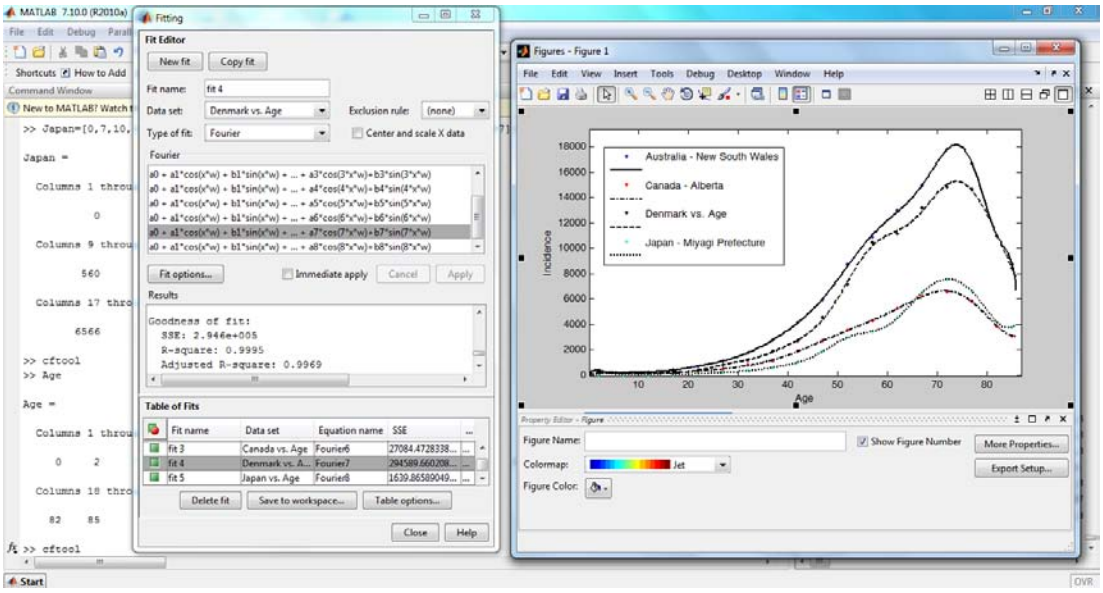

Figure 27: Cancer incidence data fitting for Australia (New South Wales), Canada (Alberta), Denmark, Japan (Miyagi prefecture) in time period 1999-2002

The result after applying the Fourier formula to the mentioned data is explained in detail in the following

Furrier model:

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3*sin(3*x*w) + a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5*sin(5*x*w) + a6*cos(6*x*w) + b6*sin(6*x*w) + a7*cos(7*x*w) + b7*sin(7*x*w)

Coefficients (with 95% confidence bounds):

a0 = -2.812e+014 (-1.309e+017, 1.304e+017),

a1 = 4.495e+014 (-2.112e+017, 2.121e+017),

b1 = 2.046e+014 (-8.846e+016, 8.887e+016),

a2 = -2.187e+014 (-1.087e+017, 1.083e+017),

b2 = -2.511e+014 (-1.105e+017, 1.1e+017),

a3 = 4.843e+013 (-2.923e+016, 2.933e+016),

b3 = 1.625e+014 (-7.293e+016, 7.326e+016),

a4 = 8.646e+012 (-4.366e+014, 4.539e+014),

b4 = -6.269e+013 (-2.943e+016, 2.931e+016),

a5 = -8.736e+012 (-3.039e+015, 3.022e+015),

b5 = 1.383e+013 (-6.943e+015, 6.971e+015),

a6 = 2.191e+012 (-8.746e+014, 8.79e+014),

b6 = -1.44e+012 (-8.43e+014, 8.401e+014),

a7 = -1.942e+011 (-8.506e+013, 8.467e+013),

b7 = 3.059e+010 (-3.293e+013, 3.299e+013),

w = 0.01056 (-0.3364, 0.3576).

With the goodness of fit:

SSE: 8.273e+004

R-square: 0.9999

Adjusted R-square: 0.9993

RMSE: 166.1

5.2.6.2 Canada (Alberta)

Fourier model:

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3*sin(3*x*w) + a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5*sin(5*x*w) + a6*cos(6*x*w) + b6*sin(6*x*w)

Coefficients (with 95% confidence bounds):

a0 = -7.666e+005 (-8.916e+007, 8.763e+007),

a1 = 2.961e+005 (-5.197e+007, 5.256e+007),

b1 = 1.347e+006 (-1.481e+008, 1.508e+008),

a2 = 8.909e+005 (-8.688e+007, 8.866e+007),

b2 = -4.241e+005 (-7.142e+007, 7.057e+007),

a3 = -3.479e+005 (-5.435e+007, 5.366e+007),

b3 = -4.266e+005 (-3.214e+007, 3.129e+007),

a4 = -1.318e+005 (-3.769e+006, 3.506e+006),

b4 = 1.865e+005 (-2.522e+007, 2.56e+007),

a5 = 6.021e+004 (-6.6e+006, 6.721e+006),

b5 = 1.823e+004 (-1.977e+006, 2.014e+006),

a6 = -889 (-6.626e+005, 6.608e+005),

b6 = -8623 (-6.695e+005, 6.523e+005),

w = 0.03055 (-0.2038, 0.2649),

with the goodness of fit:

SSE: 2.708e+004

R-square: 0.9997

Adjusted R-square: 0.999

RMSE: 73.6

5.2.6.3 Denmark

Fourier model:

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3*sin(3*x*w) + a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5*sin(5*x*w) + a6*cos(6*x*w) + b6*sin(6*x*w) + a7*cos(7*x*w) + b7*sin(7*x*w)

Coefficients (with 95% confidence bounds):

a0 = 4920 (-852.9, 1.069e+004),

a1 = -3990 (-2.478e+004, 1.68e+004),

b1 = -5641 (-2.138e+004, 1.01e+004),

a2 = -1888 (-1.24e+004, 8620),

b2 = 2053 (-1.622e+004, 2.033e+004),

a3 = 203.1 (-1.028e+004, 1.068e+004),

b3 = 891.7 (-1159, 2943),

a4 = 609.9 (-3728, 4948),

b4 = 311.9 (-5546, 6169),

a5 = -22.6 (-2536, 2491),

b5 = -209.1 (-1461, 1042),

a6 = 71.87 (-4243, 4387),

b6 = 205.7 (-437.9, 849.2),

a7 = 119.5 (-1009, 1248),

b7 = -88.78 (-2652, 2474),

w = 0.06038 (0.0147, 0.1061).

Goodness of fit:

SSE: 2.946e+005

R-square: 0.9995

Adjusted R-square: 0.9969

RMSE: 313.4

5.2.6.4 Japan (Miyagi prefecture)

Fourier model:

f(x) = a0 + a1*cos(x*w) + b1*sin(x*w) + a2*cos(2*x*w) + b2*sin(2*x*w) + a3*cos(3*x*w) + b3*sin(3*x*w) + a4*cos(4*x*w) + b4*sin(4*x*w) + a5*cos(5*x*w) + b5*sin(5*x*w) + a6*cos(6*x*w) + b6*sin(6*x*w) + a7*cos(7*x*w) + b7*sin(7*x*w) + a8*cos(8*x*w) + b8*sin(8*x*w)

Coefficients (with 95% confidence bounds):

a0 = 3483 (-2.127e+004, 2.824e+004),

a1 = -376.2 (-2.225e+004, 2.15e+004),

b1 = -4204 (-3.847e+004, 3.006e+004),

a2 = -625.4 (-9291, 8041),

b2 = -1153 (-4.087e+004, 3.856e+004),

a3 = -1348 (-4e+004, 3.731e+004),

b3 = -1152 (-1.366e+004, 1.135e+004),

a4 = -1740 (-2.38e+004, 2.033e+004),

b4 = 392.7 (-2.548e+004, 2.626e+004),

a5 = -287.1 (-1.238e+004, 1.18e+004),

b5 = 1353 (-1.708e+004, 1.978e+004),

a6 = 482.8 (-9090, 1.006e+004),

b6 = 430.8 (-2087, 2948),

a7 = 327 (-3724, 4378),

b7 = 45.08 (-4301, 4391),

a8 = 84.19 (-1761, 1929),

b8 = -131.1 (-2160, 1898),

w = 0.05596 (0.004395, 0.1075).

with the goodness of fit:

SSE: 1640

R-square: 1

Adjusted R-square: 0.9998

RMSE: 40.5

As it is clear from the observation above, although there are some other formulas which can interpret the cancer incidence data, the best goodness of fit similar to the data special for cancer incidence rates belongs to Furrier model. In the following section a model based on this observation has been offered such a way it can cover biological part of fitting as well as mathematical part.

5.3 Analyzing Fourier Model as a differential solution of cancer incidence trend

In Fourier model we assume that the magnitude of stress ε_D, which is defined in Mildvan and Strehler model, (Konstantin, Svetlana V., Lyubov, & Anatoli, 2005), is not a real number. In other words it can be written as $a + bi$, where a and b belong to real numbers and $i = \sqrt{-1}$. In this case if we replace it by such this complex number we approach to the new formula as follow

$$\mu(x) = K\cos\left(\frac{v(x)}{\varepsilon_D}\right) + Ki\sin\left(\frac{v(x)}{\varepsilon_D}\right), \qquad (1)$$

But it may raise the question that what will be the interpretation of this complex number for a cognitive parameter for the readers. Replying this question, we assume that $\mu(x)$ is the combination of the terms in form of (1) where K does not agree with the assumption in the Sterehler and Mildvan model. Instead, since we have assumed that ε_D is complex, the external stress K is complex as well. Note that we also assume that K is still depending on age and is not a constant value. To do this we consider K=K(x) as a linear combination of the subsections of a conditional function which is defined corresponding to the altering of the values of x. This concept leads us to imagine a Fourier model to finalize our discussion as bellow:

$$\mu(x) = \sum_{n=1}^{\infty} A_n \cos(nx) - B_n \sin(nx), \qquad (2)$$

where A_n and B_n are the finalized values after the total alteration of external stress coefficients and n is zero when it is not close to the value of $\frac{v(x)}{\varepsilon_D}$. For the cases other than this we assume that n is the closest positive integer value to $\frac{v(x)}{\varepsilon_D}$.

The goodness of fit of this model is much better than the other offered models. Regarding to this new estimation we can consider that μ(x) is a solution of a second order linear differential equation with constant coefficients with the characteristic equation whose roots are complex numbers.

5.5 Conclusion

In this survey we checked several very specific mathematical models to find the best fitting for the available data. To do this, we observed through various cohort cancer data in different time periods for different regions. The best model as we found before is the furrier model. In the future, our aim is to find convincing reasons and proofs for this claim. After analyzing the model we found out that

1) The model agrees with the attributes which are expected to be existed in general age pattern for the overall cancer incidence which is explained in chapter four, section 4.2.1.

2) There is difference between male and female cancer incidence data. At older ages, a greater amount of men are diagnosed to have cancer rather than the opposite sex. However, they can be fitted by Furrier model adequately as well.

3) Although different regions have different cancer incidence rates, all of them treat the same and this behavior obeys the Fourier pattern.

4) Cancer incidence curve for the periods of time with the similar situations have nearly the same behavior, but for the special periods of time which an event has affected the cohort data we may expect to have different behaviors.

5) Cancer incidence rates may alter among different races, but they have the same behaviors and all of them can be fitted by Fourier model well.

6) Cancer incidence rates in more developed countries are higher than the lower developed ones. It may be able to be interpreted because of the advent of new facilities, products, and technologies as well as traffic and air pollution in more developed countries.

7) This interesting pattern needs further argument, from both mathematical and biological points of view. Provided to have enough reasoning, this model can cover other models and improve them.

8) Here we only examined the rate of cancer incidence over altering age, but clearly still there many unanswered questions in this area such as the rate of overall cancer incidence over time or some other biological parameters which all of them worth to study and check.

REFERENCE

Alfred G. Jr., K. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA , 68: 820-823.

Armitage, P., & Doll, R. (1954, march). The age distributing of cancer and a multi-stage theory of carcinogenesis. pp. VIII, 1.

Attwood, T. K., & Parry-Smith, D. J. (1999). Introduction to bioinformatics. Delhi: Pearson Education.

Axelson, O., Fredrikson, M., Akerblom, G., & Hardell, L. (2002). Leukemia in childhood and adolescence and exposure to ionizing radiation in homes built from uranium-containing alum shale concrete. Epidemiology , 13:146-50.

BELLOMO, N., LI, N. K., & MAINI, P. K. (2008). ON THE FOUNDATIONS OF CANCER MODELLING: SELECTED TOPICS, SPECULATIONS, AND PERSPECTIVES. Mathematical Models and Methods in Applied Sciences , Vol. 18, No. 4, 593–646.

Bernstam, E. V., Smith, J. W., & Johnson, T. R. (2010). What is biomedical informatics? Journal of Biomedical Informatics , 43 (1), 104-110.

Bioconductor. (2003). Bioconductor. Retrieved from http://www.bioconductor.org/

Biology Direct. (n.d.). Retrieved from http://www.biology-direct.com/content/5/1/19/figure/F1?highres=y

BLAST. (2009, October 28). BLAST. Retrieved from http://blast.ncbi.nlm.nih.gov/Blast.cgi

Block, G., Patterson, B., & Sabur, A. (1992). Fruit, vegetables, and cancer prevention:. Nutr Cancer , 18:1-29.

Boguski, M. S. (1998). Trends guide to bioinformatics. trends supplement , p1.

Cancer research UK. (2009). Cancer research UK. Retrieved from http://info.cancerresearchuk.org/cancerandresearch/all-about-cancer/what-is-cancer/what-causes-cancer/

Carins, J., Lyon, J. L., & Skolnick, M. (1980). Cancer incidence in defined populations. Bandury report; 4. Cold spring Harbor, NY: Cold Spring Harbor Laboratory.

Cates, S. (2007, September 12). Multiple Sequence Alignment. Retrieved from http://cnx.org/content/m11036/latest/

Clapp, R., Howe, G., & LeFevre, M. (2005). Environmental and occupational causes of cancer: review of recent scientific literature. The Lowell Center for Sustainable Production, University of Massachusetts Lowell .

Clustal. (2011, June 27). Clustal. Retrieved from http://www.clustal.org/

Cohen, A. G. (2003). Air pollution and lung cancer: what more do we need to know? Thorax , 58:1010-2.

Cohen, J. E. (2004). Mathematics Is Biology's Next Microscope, Only Better; Biology Is Mathematics' Next Physics, Only Better. PLoS Biol , 2(12): e439. doi:10.1371/journal.pbio.0020439.

Coleman, M., Esteve, J., Damiecki, P., Arslan, A., & Renard, H. (1993). Trends in cancer incidence and mortality. Lyon: IARC.

Cook, P. J., Doll, R., & Fellingham, S. A. (1969). A mathematical model for the age distribution of cancer in man. International journal of cancer , 4(1):93-112.

Crawford, E. D. (2003). Epidemiology of prostate cancer. Urology , 62:3e12.

Darby, S., Hill, D., Auvinen, A., Barros-Dios, J. M., Baysson, H., Bochicchio, F., et al. (2005). Radon in homes and risk of lung cancer: collaborative analysis of individual data from 13 European case-control studies. BMJ , 330:223-8.

David Sadava, H. C. (2006). Life the science of biology. USA: The courier companies.

Dix, D. (1989). The Role of Aging in Cancer Incidence: An Epidemiological Study. J Gerontol , 44(6): 10-18 .

Dockery, D., Pope, C., Xu, X., Spengler, J. D., Ware, J. H., Fay, M., et al. (1993). An association between air pollution and mortality in six U.S. cities. N Engl J Med , 1753-9.

Doll, R. (1971). The age distribution of cancer: Implications for models of carcinogenesis. journal of roy statist soc series A , 134: 133-166.

EBI. (2011). EBI. Retrieved from http://www.ebi.ac.uk/2can/tutorials/nucleotide/fasta.html

Eden, K., Mahon, S., & Helfand, M. (2002). Screening high-risk populations for thyroid cancer. Med Pediatr Oncol , 36:583e91.

Edwards, J. (2011). Bioinformatics Careers. Retrieved from Graduating Engineer & Computer Careers : http://www.graduatingengineer.com/articles/20091031/Bioinformatics-Careers

English, D. R., Holman, C. D., Milne, E., Winter, M. G., & Hulse, G. K. (1995). the qualification of drug caused morbidity and mortality in Australia. Commonwealth Department of human service and health: AGPS.

Ershler, W. B., & Longo, D. L. (1997). The biology of aging: the current research agenda. Cancer , 80:1284-93.

Ewens, W. J., & Grant, G. R. (2005). Statistical Methods in Bioinformatics: An Introduction (Statistics for Biology and Health). New York: Springer.

Feychting, M., Forssen, U., & Floderus, B. (1997). Occupational and residential magnetic field exposure and leukemia and central nervous system tumors. Epidemiology , 8:384-9.

Girke, T., & Riverside, U. C. (2011). R & Bioconductor Manual. Retrieved from http://manuals.bioinformatics.ucr.edu/home/R_BioCondManual#biocon_affypack

H., M., & Weber, W. (1985). Familial cancer. Proceeding of the 1th international research conference on familial cancer; 1985 Sep (pp. 16-21). Basel, Schweiz. Karger.

Hahn, W. C., & Weinberg, R. (2002). Modelling the molecular circuitry of cancer. Nat Rev Cancer , 2:331-41.

Hardell, & Hansson. (2006). Mobilephone use and risk of acoustic neuoma: result of the interphone case-control study in five north. Br J Cancer , 93:1348-9.

Hardell, L., & Hansson, K. (2006). Mobile phone use and risk of acoustic neuroma: results of the interphone case-control study in five north European countries. Br J Cancer , 93:1348-9.

Hardell, L., Carlberg, M., Soderqvist, F., Hansson, M. K., & Morgan, L. L. (2007). Long-term use of cellular phones and brain tumours - increased risk associated with use for >10 years. Occup Environ Med , 64:626-32.

Holland, R. C., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., et al. (2008). BioJava: an open-source framework for bioinformatics. Bioinformatics , 24 (18): 2096-2097.

Holland, R. C., Down, T. A., Pocock, M., Prlić, A., Huen, D., James, K., et al. (2008). BioJava: an open-source framework for bioinformatics. Bioinformatics , 24, Issue: 18, Pages: 2096-7.

Hoyer, A. P., Gerdes, A. M., T., J., F., R., & H. B., H. (2002). Organochlorines, p53 mutations in relation to breast cancer risk and survival. A Danish cohort-nested case-controls study. Breast Cancer Res Treat , 71:59-65.

IARC (International Agency for Research on Cancer) . (1992). Solar and ultraviolet radiation In: IARC monographs on the evaluation of carcinogenic risk to humans vol. 55. Lyon: IARCPress.

IARC (International Agency for Research on Cancer). (1989). Occupational exposures in petroleum refining; crude oil and major petroleum fuels. In In: IARC monographs on the evaluation of carcinogenic risk to humans, vol. 45. Lyon: IARCPress.

IARC (International Agency for Research on Cancer). (2002). Weight control and phisical activity. In IARC handbooks of cancer prevention, vol. Lyon: IARCPress.

IARC (International Agency for Research on Cancer). (1995). Wood dust and formaldehyde. In IARC monographs on the evaluation of carcinogenic risk to humans, vol. 62. Lyon: IARCPress.

IARC. (1976). Cancer Incidence in Five Continents. Volume III. Lyon: IARC Sci Publ,15.

IARC. GLOBOCAN 2008. Retrieved from http://globocan.iarc.fr/

IARC(International Agancy for Research on Cancer). (2008). International Agancy for Research on Cancer- GLOBOCAN. Retrieved from http://globocan.iarc.fr/

IARC(International Agency for Research on Cancer). (1965). Cancer Incidence in Five Continents. Volume I. Lyon: International Agency for Research on Cancer.

IARC(International Agency for Research on Cancer). (1970). Cancer Incidence in Five Continents. Volume II. Lyon: International Agency for Research on Cancer.

IARC(International Agency for Research on Cancer). (1982). Cancer Incidence in Five Continents. Volume IV. Lyon: IARC Sci Publ, 42.

IARC(International Agency for Research on Cancer). (1976). Cancer Incidence in Five Continents. Volume IX. Lyon: IARC Sci Publ,160.

IARC(International Agency for Research on Cancer). (1976). Cancer Incidence in Five Continents. Volume V. Lyon: IARC Sci Publ,88.

IARC(International Agency for Research on Cancer). (1976). Cancer Incidence in Five Continents. Volume VI. Lyon: IARC Sci Publ,120.

IARC(International Agency for Research on Cancer). (1976). Cancer Incidence in Five Continents. Volume VIII. Lyon: IARC Sci Publ,155.

IARC(International Agenecy for Reseach on Cancer). (1988). Alcohol drinking in : IARC monographs on the evaluation of carcinogenic risk to human. Lyon: IARCPress, vol 44.

Ichinose, T., Fujii, K., & Sagai, M. (1991). Experimental studies on tumor promotion by nitrogen dioxide. Toxicology , 67:211-25.

International Agenecy fo Research on Cancer (IARC). (n.d.). Retrieved from http://www.iarc.fr/

Irigaray, P., Newby, J. A., Clapp, R., L., H., & Howard, V. (2007). Lifestyle-related factors and environmental agents causing cancer: An overview. Biomedicine & Pharmacotherapy , 61; 640-58.

Jemal, A., Thomas, A., Murray, T., & Thun, M. (2002). Cancer statistics. 2002. CA Cancer J Clin , 52:23-47.

Kaldor, J., & Day, N. (1996). Mathematical models in cancer epidemiology. In F. J. Edited by Schottenfeld D, Cancer Epidemiology (pp. 127-137). New York: Oxford University Press.

Kinser, J. M. (2008). Python For Bioinformatics. New York: Jones and Brtlett publishers.

Knudson, A. G. (1993). Antioncogenes and human cancer. Proc Natl Acad Sci U S A , 90:10914-21.

Knudsun, A. J. (1977). Gentic prediction to cancer. In Origins of human cancer. Book C: Human risk assessment. Cold spring harbor conferences on cell proliferation; 4 proceeding of a conference; 1976 Sep (p. 45). Cold Spring Harbor labortary.

Konstantin, G. A., Svetlana V., U., Lyubov, S. A., & Anatoli, I. Y. (2005). Mathematical models for human cancer incidence rates. Demographic research , vol12/10.

Krishnan, A., Li, K.-B., & Issac, P. (2004). Rapid detection of conserved regions in protein sequences using wavelets. Silico Biology , 4, Issue: 2, Pages: 133-148.

Kufe, D. W., Pollock, R. E., R., W. R., & et al., e. (2003). Holland-Frei Cancer Medicine. 6th edition. Hamilton (ON): BC Decker Inc.

Kupper, L. L., Janis, J. M., Karmous, A., & Greenberg, B. G. (1985). Statistical age-period-cohort analysis: A review and critique. J Chronic Dis , 38:811-830.

Landrigan, P. J., Markowitz, S. B., Nicholson, W. J., & Baker, D. B. (1995). Cancer prevention in the workplace. In K. B. Greenwald P, Cancer prevention and control. (pp. 39-410). New York: Marcel Dekker.

Leonid, H., & Wai-Yuan, T. (2008). Hand book of cancer models with applications. Singapore: World Scientific.

Liang, C. (2011). COPIA: A New Software for Finding Consensus Patterns in Unaligned Protein Sequences. Retrieved from http://hdl.handle.net/10012/1050

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., & Koskenvua, M. (2000). Environmental and heritable factors in the causation of cancer-analyses of cohorts of twins from Sweden, Denmark and Finland. N Engl J Med , 342:78-85.

Lipman, D., & Pearson, W. (1985). Rapid and sensitive protein similarity searches. Proceedings of the National Academy of Sciences of the United States of America , 85 (8): 2444–8.

Little, M. P. (2010). Cancer models, genomic instability and somatic cellular Darwinian evolution. Biology Direct , 5:19.

Loeb, K. R., & Loeb, L. A. (2000). Significance of multiple mutations in cancer. Carcinogenesis , 21:379-85.

Löfroth, G. (1988). Environmental tobacco smoke: overview of chemical composition and genotoxic components. Mutation Research/Genetic Toxicology , Volume 222, Issue 2, Pages 73-80 .

Lubin, J. H., & Boice, J. D. (1997). Lung cancer risk from residential radon: a meta-analysis of eight epidemiologic studies. J Natl Cancer Inst , 89:49-57.

Manton, G., & Stallard, E. K. (1980). A two-disease model of female breast cancer: mortality in 1969 among white females in the United States. Journal of the National Cancer Institute , 64(1):9-16.

Manton, K. G., & Stallard, E. (1982). Bioactuarial models of national mortality time series data. Health Care Finincing Review , 3(3):89-109.

Manton, K. G., Stallard, E., & Vapuel, J. W. (1986). Alternative models for the heterogeneity of mortality risk among the aged. Journal of the American Statistical Association , 81(395)635-44.

Mathworks. (2011). Curve Fitting Toolbox. Retrieved from http://www.mathworks.com/help/toolbox/curvefit/bq_5ka6-1_1.html

Moolgavkar, S. H., & Venzon, D. (1979). Two-event models for carcinogenesis-incidence curves for childhood and adult tumors. Mth BioSci , 47: 55-77.

Moolgavkar, s. (1978). The multi stage theory of carcinogenesis and the age distribution of cancer in man. Journal of natl Cancer Inst , 61:49-52.

MPSi. (2009). MPSi. Retrieved from http://www.mpsi.ac.uk:8080/pims/help/HelpLocalSimilaritySearch.html

Mucci, L. A., S., W., M., T. R., Trichopoulos, D., & Adami, H. O. (2001). The role of geneeenvironment interaction in the aetiology of human cancer: examples from cancers of the large bowel, lung and breast. J Intern Med , 249:477-93.

Myers, E., Altschul, S., Gish, W., Lipman, D. J., & Miller, W. (1990, October). Basic local alignment search tool. Journal of Molecular Biology , pp. 215(3): 403-410.

National Cancer Institute. (2011). Retrieved from http://www.cancer.gov/

NCBI. (2003). A science primer. One size does not fit all: the promise of pharmacogenetics. Retrieved from National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/About/primer/pharm.html

Nordling, C. O. (1953). A new theory on cancer-inducing mechanism. Br J Cancer , 7: 68-72.

Nowak, M. A., & Sigmund, K. (2004). Evolutionary dynamics of biological games. Science , 303, 793–799.

Opera, T. I. (2004). cheminformatics in drug discovery. WILEY-VCH Verlag GmbH & Co.KGaA.

O'Reilly & Associates. (2001). Retrieved from Bioon: http://www.bioon.com/book/biology/Beginning%20Perl%20for%20Bioinformatics/19.htm

Parkin, D. M., & Fernandez, L. M. (2006). Use of statistics to assess the global burden. Breast J , 12:S70e80.

Petrelli, J., Calle, E., Rodriguez, C., & Thun, M. J. (2002). Body mass index, height, and postmenopausal breast cancer mortality in a prospective cohort of US women. Cancer Causes Control , 13:325e32.

Pope, C. A., Burnett, R. T., Thun, M. J., Calle, E. E., Krewski, D., Ito, K., et al. (2002). Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. JAMA , 287:1132-41.

Poschl, G., & Seitz, H. K. (2004). Alcohol and cancer. Alcohol , 39:155-65.

Quitsmokin. (n.d.). Retrieved from http://quitsmoking.about.com

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics , 16, (6) ;276-277.

Richters, A., & Kuraitis, K. (1983). Air pollutants and the facilitation of cancer metastasis. Environ Health Perspect , 52:165-8.

Robertson, C., Gandini, S., & Boyle, P. (1999). Age-Period-Cohort models: A comparitive study of available methodologies. J Clin Epidemiol , vol. 52, No. 6, pp. 569--583.

Rodriguez, C., A.V., P., Calle, E. E., Jacobs, E., Chao, A., & Thun, M. (2001). Body mass index, height, and prostate cancer mortality in two large cohorts of adults men in the Unites States. Cancer Epidemiol Biomarkers Prev , 10:345e53.

Ronckers, C. M., Erdmann, C. A., & Land, C. E. (2005). Radiation and breast cancer: a review of current evidence. Breast Cancer Res , 7:21-32.

Sayle, R. A., & Milner-White, E. J. (1995, April 16). RASMOL: biomolecular graphics for all. Trends in Biochemical Sciences , 20:374-376.

Siemiatycki, J., Richardson, L., Straif, K., Latreille, B., Lakhani, R., Campbell, S., et al. (2004). Listing occupational carcinogens. Environ Health Perspect , 112:1447-59.

Simopoulos, A. P. (1990). Energy imbalance and cancer of the breast, colon and. Med Oncol Tumor Pharmacother , 7:109e20.

Solomon, D. (2003). Chapter 14: Role of triage testing in cervical cancer screening. J Natl Cancer Inst Monogr , 31:97e101.

Stajich, J. E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. Genome research , 12: 1611-161.

Thampi, S. M. (2009, 12 22). Introduction to Bioinformatics. Retrieved from Cornel University Library: http://arxiv.org/abs/0911.4230

The scientific basis of vegeteriansism. (1999). Retrieved from Cancer and the Vegetarian Diet: http://www.vegsource.com/harris/cancer_vegdiet.htm

TheFreeDictionary. (2011). Retrieved from http://medical-dictionary.thefreedictionary.com/incidence+rate

Tondel, M., Lindgren, L., Hjalmarsson, P., Hardell, L., & Persson, B. (2006). Increased incidence of malignancies in Sweden after the Chernobyl accidentd-a promoting effect? Am J Ind Med , 49:159-68.

TR, H. (1992). Analysing the temporal effects of age, period and cohort. Stat methods med res , 1:317-337.

Uauy, R., & Solomons, N. (2005). Diet, nutrition, and the life-course approach to. J Nutr , 135:S2934e45.

Vapuel, J. W., & Yashin, A. I. (1999). Cancer rates over age, time, and place: Insights from stochastic models of heterogeneous population. MPIDR working paper WP , WP #88-01-1.

Vineis, P., Hoek, G., Krzyzanowski, M., Vignia-Taglianti, F., Vegalia, F., Airoldi, L., et al. (2007). Lung cancers attributable to environmental tobacco smoke and

air pollution in non-smokers in different European countries: a prospective study. Environ Health , 6:7-13.

Vizcaíno, J. A., Foster, J. M., & Martens, L. (2010 October). Proteomics data repositories: Providing a safe haven for your data and acting as a springboard for further research. J Proteomics. , 73 (11), 73(11):2136-46.

Vogelstein, B., & Kinzler, K. W. (April 1993). The multistep nature of cancer. Trends in genetics , Volume 9, Issue 4, 138-141.

Wai-Yuan, T., & Leonid, H. (2008). Hand book of cancer models with applications. Singapore: World Scientific Publishing Co. Pte. Ltd.

Wakeford, R. (2004). The cancer epidemiology of radiation. Oncogene , 23:6404-28.

Weisburger, J. H. (2002). Lifestyle, health and disease prevention: the underlying. Eur J Cancer Prev , 11;S1-7.

Wertheimer N, L. E. (1979). Electrical wiring configurations and childhood cancer. Am J Epidemiol , 109:273-84.

Wikipedia. (n.d.). Retrieved from http://en.wikipedia.org/wiki/Malignant_cell

Wikipedia. (n.d.). Retrieved from http://en.wikipedia.org/wiki/Cellular_differentiation

Wikipedia. (2011). Retrieved from http://en.wikipedia.org/wiki/Gompertz%E2%80%93Makeham_law_of_mortality

Wikipedia. (2011). Retrieved from http://en.wikipedia.org/wiki/Fourier_series

Wikipedia. (2011). Retrieved from http://en.wikipedia.org/wiki/Markov_process

Wikipedia. (2011). Retrieved from http://en.wikipedia.org/wiki/Time_series

Wikipedia. (2011). Retrieved from http://en.wikipedia.org/wiki/Metastasis

Wikipedia. (2011). Retrieved from http://en.wikipedia.org/wiki/Mutagen

Willett, E. V., Skibola, C. F., Adamson, P., Skibola, D. R., Morgan, G. J., Smith, M. T., et al. (2005). Non-Hodgkin's lymphoma, obesity and energy Non-Hodgkin's lymphoma, obesity and energy. Br J Cancer , 93:811e6.

Wodarz, D., & Komarova, N. L. (2005). Computational Biology of cancer. Singapore: World Scientific Publishing.

Xia, X. (2011).Retrieved from http://gchelpdesk.ualberta.ca/WebTextBook/37-38.htm

APPEMDIX

Appendix: A Chronological History of Bioinformatics

Year	Activities
1951	Pauling and Corey propose the structure for the alpha-helix and beta-sheet.
1953	Watson & Crick propose the double helix model for DNA based x-ray data obtained by Franklin & Wilkins.
1954	Perutz's group develop heavy atom methods to solve the phase problem in protein crystallography.
1955	The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger.
1958	The first integrated circuit is constructed by Jack Kilby at Texas Instruments. The Advanced Research Projects Agency (ARPA) is formed in the US
1962	Pauling's theory of molecular evolution.
1965	Margaret Dayhoff's Atlas of Protein Sequences
1968	Packet-switching network protocols are presented to ARPA
1969	The ARPANET is created by linking computers at Stanford, UCSB, The University of Utah and UCLA.
1970	The details of the Needleman- Wunsch algorithm for sequence comparison are published.
1971	Ray Tomlinson (BBN) invents the email program.
1972	The first recombinant DNA molecule is created by Paul Berg and his group.
1973	The Brookhaven Protein DataBank is announced. Robert Metcalfe receives his PhD from Harvard University. His thesis describes Ethernet.
1974	Vint Cerf and Robert Khan develop the concept of connecting networks of computers into an "internet" and develop the Transmission Control Protocol (TCP).
1975	Microsoft Corporation is founded by Bill Gates and Paul Allen. Two-dimensional electrophoresis, where separation of proteins on SDS polyacrylamide gel is combined with separation according to isoelectric points, is announced by P. H. O'Farrell
1976	The Unix-To-Unix Copy Protocol (UUCP) is developed at Bell Labs. E. M. Southern published the experimental details for the Southern Blot technique of specific sequences of DNA.
1977	The full description of the Brookhaven PDB (http://www.pdb.bnl.gov) is published (Bernstein, F.C.; Koetzle, T.F.; Williams, G.J.B.; Meyer, E.F.; Brice, M.D.; Rodgers, J.R.; Kennard, O.; Shimanouchi, T.; Tasumi, M.J.; J. Mol. Biol., 1977, 112:, 535). Allan Maxam and Walter Gilbert (Harvard) and Frederick Sanger (U.K. Medical Research Council), report methods for sequencing DNA. DNA sequencing and software to analyze it (Staden)
1978	The first Usenet connection is established between Duke and the University of North Carolina at Chapel Hill by Tom Truscott, Jim

	Ellis and Steve Bellovin.
1980	The first complete gene sequence for an organism (FX174) is published. The gene consists of 5,386 base pairs which code nine proteins. Wüthrich et. al. publish paper detailing the use of multi-dimensional NMR for protein structure determination. IntelliGenetics, Inc. founded in California. Their primary product is the IntelliGenetics Suite of programs for DNA and protein sequence analysis.
1981	The Smith-Waterman algorithm for sequence alignment is published. IBM introduces its Personal Computer to the market. The concept of a sequence motif (Doolittle).
1982	Genetics Computer Group (GCG) created as a part of the University of Wisconsin of Wisconsin Biotechnology Center. The company's primary product is The Wisconsin Suite of molecular biology tools. GenBank Release 3 made public Phage lambda genome sequenced.
1983	The Compact Disk (CD) is launched. Name servers are developed at the University of Wisconsin. Sequence database searching algorithm (Wilbur-Lipman) LANL (Los Alamos National Laboratory) and LLNL (Lawrence Livermore National Laboratory) begin production of DNA clone (cosmid) libraries representing single chromosomes. DNA analysis becomes viable with the discovery of Polymerase Chain Reaction. It allows small samples of DNA to be multiplied to produce a large enough sample to analyse.
1984	Jon Postel's Domain Name System (DNS) is placed on-line. The Macintosh is announced by Apple Computer.
1985	• The FASTP/FASTN algorithm is published. Robert Sinsheimer holds meeting on human genome sequencing at University of California, Santa Cruz. At OHER, Charles DeLisi and David A. Smith commission the first Santa Fe conference to assess the feasibility of a Human Genome Initiative 1986 - Following the Santa Fe conference, DOE OHER announces Human Genome Initiative. With $5.3 million, pilot projects begin at DOE national laboratories to develop critical resources and technologies. The term "Genomics" appeared for the first time to describe the scientific discipline of mapping, sequencing, and analyzing genes. The term was coined by Thomas Roderick as a name for the new journal. Amoco Technology Corporation acquires IntelliGenetics. The SWISS-PROT database is created by the Department of Medical Biochemistry of the University of Geneva and the European Molecular Biology Laboratory (EMBL). The PCR reaction is described by Kary Mullis and co-workers.
1987	The use of yeast artifical chromosomes (YAC) is described (David T. Burke, et. al., Science, 236: 806-812). The physical

	map of e. coli is published. Perl (Practical Extraction Report Language) is released by Larry Wall. Congressionally chartered DOE advisory committee, HERAC, recommends a 15-year, multidisciplinary, scientific, and technological undertaking to map and sequence the human genome. DOE designates multidisciplinary human genome centers. NIH NIGMS begins funding of genome projects
1988	National Center for Biotechnology Information (NCBI) created at NIH/NLM EMBnet network for database distribution The Human Genome Intiative is started (commission on Life Sciences, National Research council. Mapping and sequencing the Human Genome, National Academy Press: washington, D.C.), 1988. The FASTA algorith for sequence comparison is published by Pearson and Lupman. A new program, an Internet computer virus desined by a student, infects 6,000 military computers in the US. Reports by congressional OTA and NAS NRC committees recommend concerted genome research program. HUGO founded by scientists to coordinate efforts internationally First annual Cold Spring Harbor Laboratory meeting on human genome mapping and sequencing. DOE and NIH sign MOU outlining plans for cooperation on genome research. Telomere (chromosome end) sequence having implications for aging and cancer research is identified at LANL.
1989	The genetics Computer Group (GCG) becomes a privatae company. Oxford Molceular Group,Ltd.(OMG) founded, UK by Anthony Marchigton, David Ricketts, James Hiddleston, Anthony Rees, and W.Graham Richards. Primary products: Anaconds, Asp, Cameleon and others (molecular modeling, drug design, protein design). DNA STSs recommended to correlate diverse types of DNA clones. DOE and NIH establish Joint ELSI Working Group.
1990	The BLAST program (Altschul,et.al.) is implemented. Molecular applications group is founded in California by Michael Levitt and Chris Lee. Their primary products are Look and SegMod which are used for molecular modeling and protein deisign. InforMax is founded in Bethesda, MD. The company's products address sequence analysis, database and data management, searching, publication graphics, clone construction, mapping and primer design. DOE and NIH present joint 5-year U.S. HGP plan to Congress. The 15-year project formally begins. Projects begun to mark gene sites on chromosome maps as sites of mRNA expression. Research and development begun for efficient production of more stable, large-insert BACs
1991	The research institute in Geneva (CERN) announces the creation of the protocols which make -up the World Wide Web. The creation and use of expressed sequence tags (ESTs) is described. Incyte Pharmaceuticals, a genomics company headquartered in Palo Alto California, is formed. Myriad Genetics, Inc. is founded in Utah. The company's goal is to lead in the discovery of major

	common human disease genes and their related pathways. The company has discovered and sequenced, with its academic collaborators, the following major genes: BRCA1, BRACA1 , CHD1, MMAC1, MMSC1, MMSC2, CtIP, p16, p19 and MTS2. Human chromosome mapping data repository, GDB, established.
1992	Low-resolution genetic linkage map of entire human genome published. Guidelines for data release and resource sharing announced by DOE and NIH.
1993	Sanger Centre , Hinxton, UK . CuraGen Corporation is formed in New Haven, CT. Affymetrix begins independent operations in Santa Clara, California. International IMAGE Consortium established to coordinate efficient mapping and sequencing of gene-representing cDNAs. DOE-NIH ELSI Working Group's Task Force on Genetic and Insurance Information releases recommendations. DOE and NIH revise 5-year goals [Science 262, 43-46 (Oct. 1, 1993)] IOM releases U.S. HGP-funded report, "Assessing Genetic Risks." LBNL implements novel transposon-mediated chromosome-sequencing system. GRAIL sequence-interpretation service provides Internet access at ORNL.
1994	Netscape Communications Corporation founded and releases Naviagator, the commerical version of NCSA's Mozilla. Gene Logic is formed in Maryland. The PRINTS database of protein motifs is published by Attwood and Beck. Oxford Molecular Group acquires IntelliGenetics. EMBL European Bioinformatics Institute , Hinxton, UK. Genetic-mapping 5-year goal achieved 1 year ahead of schedule. Completion of second-generation DNA clone libraries representing each human chromosome by LLNL and LBNL.
1995	The Haemophilus influenzea genome (1.8) is sequenced. LANL and LLNL announce high-resolution physical maps of chromosome 16 and chromosome 19, respectively The Mycoplasma genitalium genome is sequenced. Moderate-resolution maps of chromosomes 3, 11, 12, and 22 maps published . Physical map with over 15,000 STS markers published. First (nonviral) whole genome sequenced (for the bacterium Haemophilus influenzae). Sequence of smallest bacterium, Mycoplasma genitalium, completed; provides a model of the minimum number of genes needed for independent existence.
1996	The genome for Saccharomyces cerevisiae (baker's yeadt, 12.1 Mb) is sequenced. The prosite database is reported by Bairoch, et.al. Methanococcus jannaschii genome sequenced; confirms existence of third major branch of life on earth. DOE initiates 6 pilot projects on BAC end sequencing. Saccharomyces cerevisiae (yeast) genome sequence completed by international consortium Affymetrix produces the first commerical DNA chips. Sequence of the human T-cell receptor region completed
1997	The genome for E.coli (4.7 Mbp) is published. Oxford Molecualr

	Group acquires the Genetics Computer Group. LION bioscience AG founded as an intergrated genomics company with strong focus on bioinformatics. The company is built from IP out of the European Molecualr Biology Laboratory (EMBL), the European Bioinformtics Institute (EBI), the GErman Cancer Research Center (DKFZ), and the University of Heidelberg. paradigm Genetics Inc., a company focussed on the application of genomic technologies to enhance worldwide food and fiber production, is founded in Research Triangle Park, NC. deCode genetics publishes a paper that described the location of the FET1 gene, which is responsible for familial essential tremor, on chromosome 13 (Nature Genetics). NIH NCHGR becomes National Human Genome Research Institute (NHGRI). Second large-scale sequencing strategy meeting held in Bermuda High-resolution physical maps of chromosomes X and 7 completed. DOE-NIH Task Force on Genetic Testing releases final report and recommendations. DOE forms Joint Genome Institute for implementing high-throughput activities at DOE human genome centers, initially in sequencing and functional genomics.
1998	The genomes for Caenorhabitis elegans and baker's yeast are published. The Swiss Institute of Bioinformatics is established as a non-profit foundation. Craig Venter forms Celera in Rockville, Maryland. PE Informatics was formed as a center of Excellence within PE Biosystems. This center brings together and leverges the complementary expertise of PE Nelson and Molecualr Informatics, to further complement the genetic instrumention expertise of Applied Biosystems. Inpharmatica, a new Genomics and Bioinformatics company, is established by University College London, the Wolfson Institute for Biomedical Research, five leading scientists from major British academic centres and Unibio Limited. GeneFormatics, a company dedicated to the analysis and predication of protein structure and function, is formed in San Diego. Molecualr Simulations Inc. is acquired by Pharmacopeia.
1999	deCode genetics maps the gene linked to pre-eclampsia as a locus on chromosome 2p13. First Human Chromosome Completely Sequenced! On December 1, researchers in the Human Genome Project announced the complete sequencing of the DNA making up human chromosome 22. Joint Genome Institute sequencing facility opens in Walnut Creek, CA. Major Drug Firms Create Public SNP Consortium. HGP advances goal for obtaining a draft sequence of the entire human genome from 2001 to 2000.
2000	The genome for Pseudomonas aeruginosa (6.3 Mbp) is published. The A.thaliana genome (100 Mb) is secquenced. The D.melanogaster genome (180 Mb) is sequenced. Pharmacopeia acquires Oxoford Molecular Group. HGP leaders and President Clinton announce the completion of a "working draft" DNA sequence of the human genome. International research consortium publishes chromosome 21 genome, the smallest human chromosome and the second to be completely sequenced. DOE researchers announce completion of chromosomes 5, 16, and 19

	draft sequence. International collaborators publish genome of fruit fly Drosophila melanogaster .
2001	The huam genome (3,000 Mbp) is published. Human Chromosome 20 Finished - Chromosome 20 is the third chromosome completely sequenced to the high quality specified by the Human Genome Project.
2002	Structural Bioinformatics and GeneFormatics merge An international sequencing consortium published the full genome sequence of the common house mouse (2.5 Gb). Whitehead Institute researcher Kerstin Lindblad-Toh is the lead author on the paper; her institution lead the project and contributed about half of the sequence. Washington University School of Medicine delivered about 30 percent of the sequence, and created the mouse BAC-based physical map. The Wellcome Trust Sanger Institute in the UK was the third major partner. Other institutes in the International Mouse Genome Sequencing Consortium included the University of California at Santa Cruz, the Institute for Systems Biology, and the University of Geneva. Mouse Genome Sequencing Consortium publishes its draft sequence of mouse genome in the December 5, 2002, issue of Nature International consortium led by the DOE Joint Genome Institute publishes draft sequence of Fugu rubripes.
2003	Human Genome Project Completion, April 2003. Human Chromosome 14 Finished - Chromosome 14 is the fourth chromosome to be completely sequenced.
2004	The draft genome sequence of the brown Norway laboratory rat, Rattus norvegicus, was completed by the Rat Genome Sequencing project Consortium.