Sudan University Of Science And Technology

College Of Engineering

Biomedical Engineering Department

Cancer Detection Using Protein Sequence Analysis

A project submitted In Partial Fulfillment Of The Requirement For The B.Sc(Honors) Degree In Biomedical Engineering

Represented By:

1-Dania Abd Allah Salheein

2-Reem Hassan Hussain

3-Wafa Mohammed Elmodther

Supervisor by :

Mr. Alkhatim Mohammed Ahmed

October 2015

الآيــة

وَقُلِ اعْمَلُواْ فَسَيَرَى ّالله عَمَلَكُمْ وَرَسُولُهُ
وَالْمُؤْمِنُونَ وَسَتُرَدُّونَ إِلَى عَالِمِ الْغَيْبِ وَالشَّهَادَةِ
فَيُنَبِّئُكُم بِمَا كُنتُمْ تَعْمَلُونَ) صدق الله العظيم

سورة التوبة الآية (٠٥)

Dedication

To…

Our Lovely beloved father

To…

Our sweetest beloved mother

To…

Our sisters and friends

We dedicate this project

Acknowledgment

Thanks to Allah in the first, and would like to express gratitude to supervisor T. Alkhatim Mohammed for all his help guidance ,his patience and valuable encouragement during this research. Also would like to express thanks to the Staff members in the Department of biomedical engineer- of-Sudan University Of Science And Technology.

Abstract

The protein is a single ,un branched chain of amino acids ;it has 20 different types of amino acids and these amino acid sequence determine the three dimensional structure of the protein in turn determine it's function because any change in the arrangement of the amino acid sequence can affect the protein's function. This change in the function could be the beginning of infecting of cancer but we cannot detect the cancer by the medical imaging devices in this level if it does not spread in many number of cells.

The objective of this research is to analyze the protein sequence of normal and cancerous organs using a database of DNA sequence of these organs and detect the difference between them using statistical analysis from bioinformatics tools and proof the results of this analysis by using sequence dot plot and global alignment .

After using all these analysis tools we find that the average of the difference that be detected is 81.857%.

المستخلص

البروتين هو سلسلة احادية غير متفرعة من الاحماض الامينية وهو يتكون من ٢٠ حمض اميني وهذه السلسلة وترتيب الاحماض الامينية بها هي التي تحدد التركيب ثلاثي الابعاد للبروتين وبالتالي تحدد وظيفته ،لأن أي تغيير يحدث في ترتيب هذه الاحماض الامينية يؤثر على وظيفة البروتين.

هذا التغير الوظيفي قد يكون بداية للإصابة بالسرطان ولكن اجهزة التشخيص الطبي لا يمكنها كشف السرطان في هذه المرحلة المبكرة مالم ينتشر المرض في عدد من الخلايا .

لذلك فالهدف من هذا البحث هو تحليل سلاسل البروتين لأعضاء طبيعية واخرى مسرطنة و ذلك باستخدام قاعدة بيانات تحتوي على سلاسل الحمض النووي لهذه الاعضاء ومعرفة الفرق بين العضو الطبيعي والمسرطن ذلك باستخدام التحليل الاحصائي و اثبات نتائج هذا التحليل باستخدام الرسم النقطي للسلاسل و الانتظام الشامل .

وبعد استخدام ادوات التحليل تلك واثبات نتائجها وجدنا ان متوسط الفرق او الاختلاف الذي تم كشفه هو ٨١.٨٥٧% .

Table Of Contents

Contents:

Chapter three

Chapter four

Chapter five

Chapter Six

List Of Figures

List Of Tables

List Of Abbreviations

MRI	Magnetic Resonance Imaging
PET	Positron Emission Tomography
CT	Computerized Tomography
MATLAB	Matrix Laboratory
DNA	Deoxyribononucleic Acid
RNA	Ribonucleic Acid
m-RNA	Messenger Ribonucleic Acid
t- RNA	Transfer ribonucleic Acid
pH	Potential of Hydrogen
CAG	Cysteine,Alanine, Glycine
nsSNP	Nonsynonymous single nucleotide polymorphism
SNP	single nucleotide polymorphism

Introduction

1.1 General view:-

Proteins are single, un branched chains of amino acid monomers ,it has 20 different amino acids The amino acid side chains in a peptide can become modified, extending the functional repertoire of amino acids to more than hundred different amino acids and these amino acid sequence determines its three dimensional structure .In turn, a protein's structure determines the function of that protein[1].
The cancer is caused in the cell to become abnormal and lose it 's restraints on growth and divide uncontrollably and eventually forms anew growth know as a(tumor) surrounding tissue and spreads to other parts of the body. Cancer is not just one disease , but a large group of almost 100 disease .it's two main characteristics are uncontrolled growth of the cells in the human body and the ability of these cell[2].

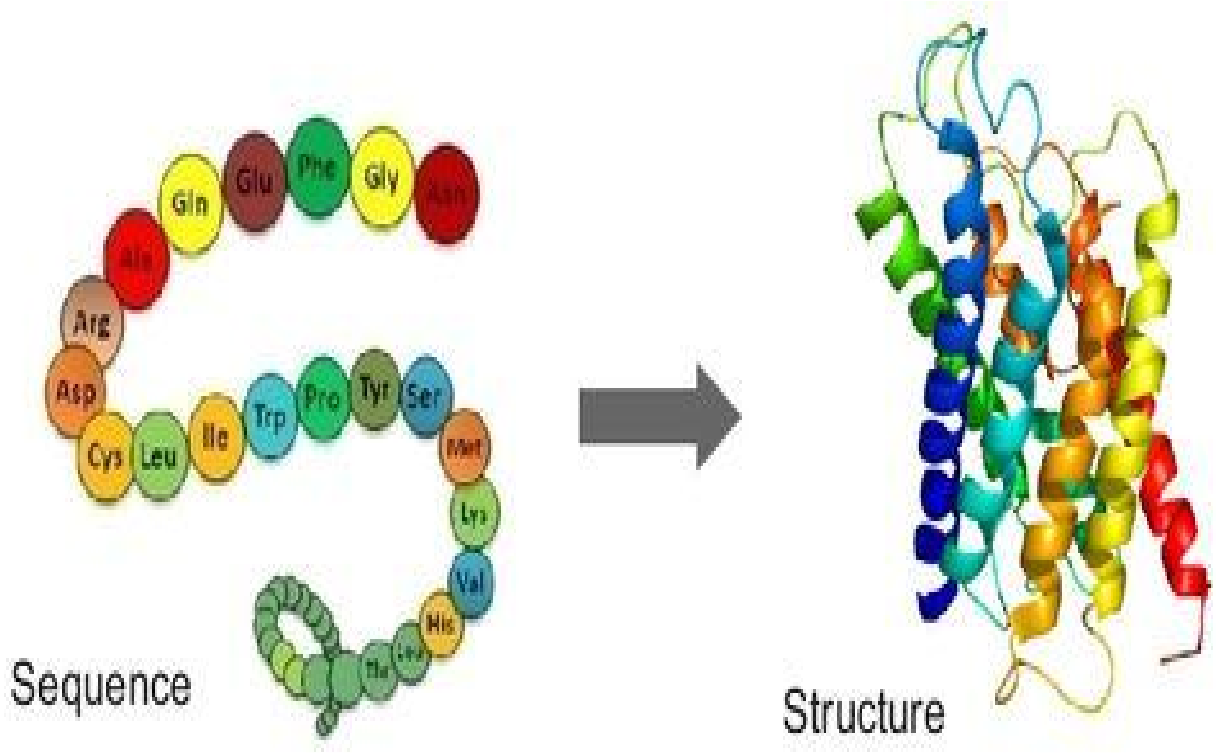


Figure 1.1: show protein sequence and structure

1.2 Problem statement:-

We can detect the cancer by many methods of medical imaging, such like MRI,PET,X-RAY but these methods cannot detect the cancer if it does not spread in many number of cells and the detection in this level would be difficult to treatment

1.3 Objective:-

1.3.1General objective:

To understand and detect the changing in amino acids and it's effects in the protein function.

1.3.2Specific objective:

-Learn MAT LAB Program and using function.

-Increase the skills in MATLAB program by learning more in bioinformatics tools.

-Design software program to analyze protein sequence.

-brings more knowledge about protein sequence.

- To compare between normal and abnormal protein to detect cancer.

1.4 Thesis Layout:

We have six chapters in this project. Each chapter speaks about different subjects. Chapter one: speaks about the introduction of the project which includes; general view, problem statement, objectives and thesis layout ,chapter two: speaks about the theoretical background which include: introduction to bioinformatics, amino acid, protein, chapter three: speaks about the literature review which include; old papers about the project, chapter four: speaks about the methodology which include; explanation of the flow chart and description of the data we collect, chapter five: speaks about the results and discussions, chapter six: speaks about conclusion and recommendations. and then we have the references.

Fundamental Background

2.1 Introduction to bioinformatics:-

Bioinformatics is conceptualizing biology in terms of molecules and applying informatics techniques derived from disciplines such as applied math's, computer science and statistics to understand and organize the information organize the information associated with these molecules.

In short, bioinformatics is a management information system for molecular biology and has many practical applications.
The aims of bioinformatics are: it is simple to organize data in a way that allows researchers to access existing information and to submit new entries as they are produced, The second aim is to develop tools and resources that aid in the analysis of data.
The third aim is to use these tools to analyze the data and interpret the results in a biologically meaningful manner[3].

2.2 Amino acid:-

Amino acids are the building blocks of proteins. They band together in chains to form the stuff from which life is born, This is a two-step process: first, they get together and form peptides or polypeptides, and it is from these groupings that proteins are made.

Twenty percent of the human body is made up of protein. Protein plays a crucial role in almost all biological processes and amino acids are the building blocks of it.

A large proportion of our cells, muscles and tissue is made up of amino acids, meaning they carry out many important bodily functions, such as giving cells their structure. They also play a key role in the transport and the storage of nutrients. Amino acids have an influence on the function of organs, glands, tendons and arteries. They are furthermore essential for healing wounds and repairing tissue, especially in the muscles, bones, skin and hair as well as for the removal of all kinds of waste deposits produced in connection with the metabolism.

All amino acids are made of the same fundamental elements: carbon, hydrogen, oxygen, nitrogen and sometimes sulfur. The basic atomic structure of an amino acid is shown in the picture above. There are nearly 80 amino acids that exist in

nature, but only 20 of these amino acids are used by the human body to make proteins.

When proteins are digested or broken down, amino acids are left. The human body uses amino acids to make proteins to help the body's Break down food, Grow , Repair body tissue , Perform many other body functions[4].

2.2.1 Function:

All proteins are composed of chains of amino acids. The sequence and type of amino acids in a protein determine a protein's function. Proteins are vital for every biological function and therefore amino acids are equally important. Proteins are made continuously in the body, and the body has a constant need for amino acids to create these proteins. Amino acids are also found free in the body, functioning mostly as chemical messengers[5]**.**

2.2.2Structure:

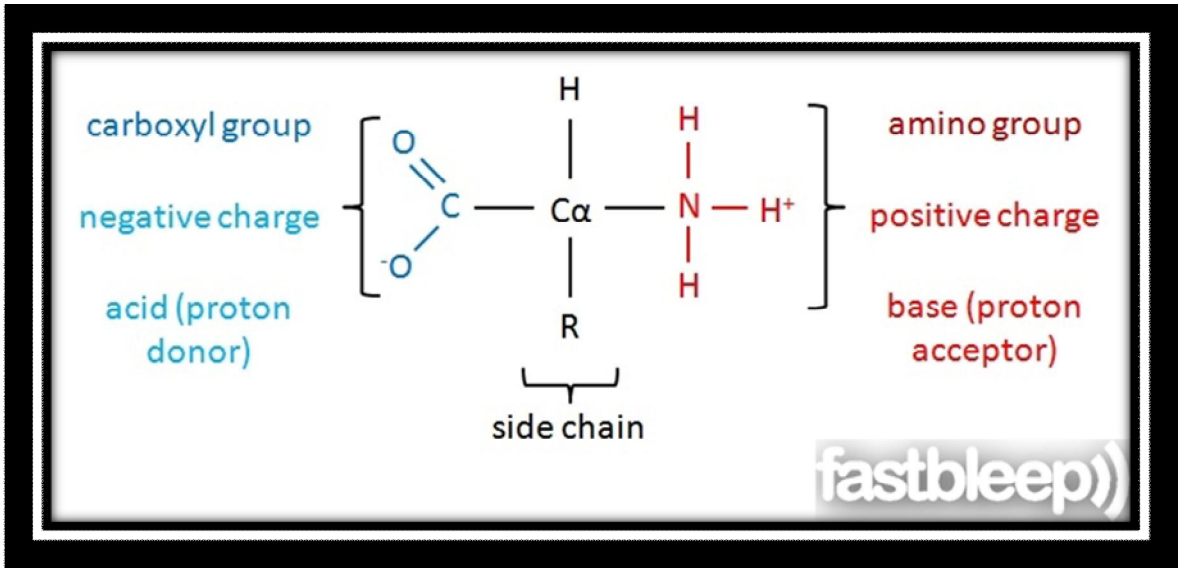


Figure 2.1:show amino acids structure

The structure of an amino acid is a carbon atom bound to a hydrogen atom, an amino group (NH2), a carboxyl group (COOH) and any variety of side chains, termed an R-group. This R-group can vary greatly between amino acids. It is this R-group portion of the amino acid that determines the chemical properties of an amino acid. This R-group can be as simple as the hydrogen atom in glycine or as complex as the aromatic ring found in tyrosine.

The amino acid sequence of protein is determined by the information found in cellular genetic code ,the genetic code is the sequence of nucleotide bases in nucleic acid (DNA and RNA),that code for amino acid the gene codes not only determine the order amino acid in a protein but they also deter mine a protein's structure and function.

Amino acids have a two-carbon bond, one of the carbons is part of a group called the carboxyl group (COO^-). A carboxyl group is made up of one carbon (C) and two oxygen (O) atoms. That carboxyl group has a negative charge, since it is a carboxylic acid (-COOH) that has lost its hydrogen (H) atom. What is left the carboxyl group is called a conjugate base. The second carbon is connected to the amino group. Amino means there is an NH_2 group bonded to the carbon atom. In the image, you see a "+" and a "-". Those positive and negative signs are there because, in amino acids, one hydrogen atom moves to the other end of the molecule. An extra "H" gives you a positive charge[5].

2.2.3 Amino Acid Classification :

While amino acids are necessary for life, not all of them can be produced naturally in the body. Of the 20 amino acids, 10 can be produced naturally .These amino acids are(alanine, proline, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, serine, glycine, and tyrosine)..0The amino acids that can not be produced naturally are called essential amino acids. They are arginine (essential for children), histidine, threonine, isoleucine, methionine, leucine, lysine, phenylalanine, tryptophan, and valine. The essential amino acids must be acquired through diet. Unlike humans, plants are capable of synthesizing all 20 amino acids[5].

Table2.1:show amino acid codes, integers, abbreviations, names and codons

Code	Integer	Abbreviation	Amino Acid Name	Codons
A	1	Ala	Alanine	GCU GCC GCA GCG
R	2	Arg	Arginine	CGU CGC CGA CGG AGA AGG
N	3	Asn	Asparagine	AAU AAC
D	4	Asp	Aspartic acid (Aspartate)	GAU GAC
C	5	Cys	Cysteine	UGU UGC
Q	6	Gln	Glutamine	CAA CAG
E	7	Glu	Glutamic acid (Glutamate)	GAA GAG
G	8	Gly	Glycine	GGU GGC GGA GGG
H	9	His	Histidine	CAU CAC
I	10	Ile	Isoleucine	AUU AUC AUA
L	11	Leu	Leucine	UUA UUG CUU CUC CUA CUG
K	12	Lys	Lysine	AAA AAG
M	13	Met	Methionine	AUG
F	14	Phe	Phenylalanine	UUU UUC
P	15	Pro	Proline	CCU CCC CCA CCG
S	16	Ser	Serine	UCU UCC UCA UCG AGU AGC
T	17	Thr	Threonine	ACU ACC ACA ACG
W	18	Trp	Tryptophan	UGG
Y	19	Tyr	Tyrosine	UAU UAC
V	20	Val	Valine	GUU GUC GUA GUG
B	21	Asx	Asparagine or Aspartic acid (Aspartate)	AAU AAC GAU GAC
Z	22	Glx	Glutamine or Glutamic acid (Glutamate)	CAA CAG GAA GAG
X	23	Xaa	Any amino acid	All codons
*	24	END	Termination codon (translation stop)	UAA UAG UGA
-	25	GAP	Gap of unknown length	NA

2.3 Protein Synthesis:-

Protein is any of a group of complex organic macromolecules that contain carbon, hydrogen, oxygen, nitrogen, and usually sulfur and are composed of one or more chains of amino acids ,The synthesis is formation of a compound from simpler compounds or elements. The term protein synthesis refers to the process of making proteins, which involves transcription of DNA and translation of RNA.

The first step of protein synthesis is the copying of three *Nucleotides* bases to a gene. During this step the *DNA* helix must unwind and unzip onto a *RNA* copy of the DNA. This first process is called *Transcription and* new RNA also known as the messenger RNA (*m-RNA*) will now move out of the nuclear membrane into the *Cytoplasm* where the *Ribosome's* are located. The m-RNA now attaches itself to a ribosome. Now a new process begins called *Translation*. Translation is the process of using coded m-RNA instructions for a sequence of amino acids to produce a polypeptide. A new type of RNA called transfer RNA (t-RNA) brings amino acids to the ribosome with the m-RNA. The dissolved *amino acids* in the *Cytoplasm* are brought by the t-RNA. From this point the t-RNA fixes its amino acids to join the polypeptide chain. The t-RNA and m-RNA form a temporary bond at the ribosome. As you may or may not ribosomes can only handle two t-RNA molecules at a time. The outgoing t-RNA will set loose its amino acid and then detaches from the m-RNA. The process of translation ends when the ribosome approaches the stop codon. The ribosome will reach the stop codon at the end of the m-RNA. It will then release the polypeptide which will eventually go through the folding process, and then finally become a protein[6].

2.3.2 Protein folding:-

Protein folding is one of the central questions in biochemistry, Protein folding is the continual and universal process whereby the long, coiled strings of amino acids that make up proteins in all living things fold into more complex three-dimensional structures. By understanding how proteins fold, and what structures they are likely to assume in their final form, researchers are then able to move closer to predicting their function[7].

This is important because incorrectly folded proteins in humans result in such devastating diseases as Alzheimer's, Parkinson's, Huntington's, emphysema and cystic fibrosis. Developing better modeling techniques for protein folding is crucial

to creating more effective pharmaceutical treatments for these and other diseases[7].

Protein folding from an unstructured amino acid sequence occurs as a complicated process of forming a special interaction among many different possible interactions on the amino acid chain.

Really the phenomenon is a probabilistic one - namely, a protein sequence won't always fold to the folded state, but it will fold with a probability that depends on the sequence itself, the solution's pH (that could in turn change the stability of different interactions in proteins), and the solution's temperature. This is why studying protein folding has become so closely related to its atomistic study[7].

All proteins in nature are made up of chains of molecules called "amino acids". Cells create proteins by "transcribing" them from RNA sequences (themselves being created from DNA sequences). When proteins are transcribed from RNA they start out as linear sequences of amino acids. Because the amino acids that make up a protein have various electrostatic and mechanical properties, the protein doesn't stay in this denatured form for long and begins to fold up into a three-dimensional structure. It is this three-dimensional structure (as well as the mechanical and electrostatic properties of the amino acid sequence) that gives the protein its functionality[7].

For example, two proteins might fold themselves in such a way that one protein presents a "lock" binding site to the other protein's corresponding "key". Fitting the key into the lock produces an electrochemical reaction that performs some essential cellular function.

The transcribed sequence of amino acids that form a protein are called the protein's "primary structure". The folded form of the protein in three-space is called the protein's "secondary structure". The secondary structure of a protein is determined in large part by the mechanical and electrostatic effects of neighboring amino acids. Proteins also have "tertiary" and "quartenary" structures. The tertiary structure refers to the overall folding path of a protein. For example, a protein might have a helical secondary structure whereas its tertiary structure might fold the overall protein into a "supercoil" where the helical protein coils around itself. The mechanics of how a protein can fold, determine a protein's structure. Tertiary structure prediction is the rough part and the focus of my thesis project, although to

predict an overall fold, all constraints from local to global folding must be considered occurred outside of the protein's innate structure or expression.

A good The quartenary structure of a protein refers to an assemblage of multiple protein strings along with the so-called "post-translational modifications" to the protein strings. "Post-translational modification" means folding or alterations of the protein string that have example of a post-translational modification is the addition of a "heme group" to hemoglobin molecules. Without this heme group, red corpuscles would be unable to carry oxygen.

One feature of proteins in nature that seems to be very constant is that when they do fold, they fold into the most energy-conservative structure possible, that is to say that the amino acids are at total rest and the protein is expending no energy to maintain its structure. This fact provides us with a key to reliably predicting a protein's structure. Theoretically, all we have to do is find the optimal conformation among all the possible conformations a protein can take.

In practice, however, this is an impractical solution. The amount of time required to test all possible conformations that a decent size protein can take on is far greater than the age of the universe, even for the fastest compute[7]

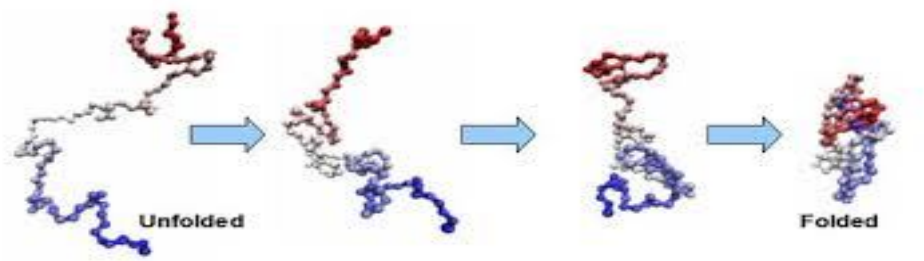


Figure2.2 : show protein folding

2.3.3Protein misfolding:-

Under favorable conditions, most proteins have no problem quickly folding to their native structures. However, there are some proteins which appear unable to fold without the presence of other helper proteins, called chaperones. In the absence of chaperones, these proteins will fail to achieve their native state and instead may associate with other unfolded polypeptide chains to form large aggregate structures (or *in vivo*, this may result in inclusion body formation). A similar scenario can occur when a protein acquires a mutation, a genetic change resulting in the alteration of one of the amino acid sites on the chain.

Protein misfolding is an intrinsic aspect of normal folding within the complex cellular environment, and its effects are minimized in living systems by the action of a range of protective mechanisms including molecular chaperones and quality control systems. Unfolded and misfolded proteins have a tendency to aggregate to form a variety of species including the highly organized and kinetically stable amyloid fi brils. The latter species represent a generic form of structure resulting from the inherent polymer properties of polypeptide chains, and their formation is associated with a wide range of debilitating human diseases. Amyloid fi brils and their precursors appear to have similar adverse effects on cellular function regardless of the sequence of the component peptide or protein. Our increasing knowledge of the interplay between different forms of protein structure and their generic characteristics provides a platform for rational therapeutic intervention designed to prevent or treat this whole family of diseases.it can also occur when a protein is subjected to unfavorable conditions, such as extremes of heat or pH. Clearly, understanding the physical basis for why folding has failed in these cases can provide clues as to the interactions vital for successful protein folding.

Many genetic diseases can be caused by mutations that make proteins less stable or disrupt the normal three-dimensional folding of proteins. A few examples include lysososmal storage diseases (LSDs), Alzheimer's, Parkinson'

Literature Review

3.1 Shannon Limit in Sequence-Structure Communication:-

Cellular production of polypeptides was modeled as a serial process where over time many chains are synthesized by translational and ribosomal apparatus and Our model hypothesis was a Shannon-weaver communication channel between amino acid sequences(source or sender) ,the Source and destination are linked with three consecutive components: an encoder a noisy channel and decoder.
The source is here defined as a series of connected primary sequences $\{seq_t\}$ resulting in a stream $\mathbf{s_A}$ of letters from the amino acid alphabet A with alphabet size |A|= 20 and The encoder is a map that uses a block code of fixed length n to encode the source through a set of code words, it maps every sequences seq_t on to one single code word $X^n(seq_t)$ represented by n-vector$(x_1 \ldots\ldots, x_n)$ of integers.
The message input is defined by:

$$X^n(seq_t) = (x_1 \ldots\ldots, x_n) \qquad (3.1)$$

This input is transmitted over a noisy communication channel which outputs an n-vector $Y^n\ (str_t)$:

$$str_t = (y_1, \ldots\ldots\ldots, y_n) \qquad (3.2)$$

now representing the folded protein chain str_t. This step mirrors the physical folding process in which a geometrically unspecified sequence becomes a functionally determined 3D structure and communicational noise is interpreted as any physical interaction of the protein with its environment so that the original input X^n is randomly distorted in to an output Y^n in a last step a decoder deciphers $Y^n(str_t)$ by selecting on member in a code book that registers the completed structure. This decoding produces an output sequence S_{A^*} of structural symbol in A^* and its complete the communication process.

The channel capacity is defined as :

$$C=\max I_{P(A)}\ (A, A^*) \qquad (3.3)$$

The channel capacity gives the maximum amount of information that can be

transferred in a single use of the channel.

The code rate R is defined as:

$$R = H(A)/n \qquad (3.4)$$

where H(A) is the information entropy of the amino acid sequence (source) and n is the code block length used by the

encoder. If code rate R and channel capacity C are known, then Shannon's theorem tells us whether communication over the channel is possible, and the point where capacity C equal the rate R is the Shannon limit.

A direct application of Shannon noisy channel theorem confirmed that communication between protein amino acid sequences and native structures was achievable .If C>R the information can be transmitted through the communication channel but no information can be reliably transmitted at capacity less than the rate R according to the Shannon theorem.

Evidence has been given that protein amino acid sequences and their tertiary structures constitute the source and the destination of a digital communication channel. In direct consequence, Shannon's noisy channel theorem could be applied and a Shannon limit in the sequence-structure map quantitatively predicted[9].

3.2 Predicting the effects of amino acid substitutions on protein function:-

Non synonymous single nucleotide polymorphisms (nsSNPs) are coding variants that introduce amino acid changes in their corresponding proteins. Because nsSNPs can affect protein function, they are believed to have the largest impact on human health compared with SNPs in other regions of the genome Therefore, it is important to distinguish those nsSNPs that affect protein function from those that are functionally neutral. Here we provide an overview of amino acid substitution (AAS) prediction methods, which use sequence and/or structure to predict the effect of an AAS on protein function.

Most methods predict approximately 25–30% of humans SNPs to negatively affect protein function, and such nsSNPs tend to be rare in the population Direct comparisons between AAS prediction methods are difficult because they were trained and tested on different data sets using different versions of sequence and structural databases as resources. The performance of an AAS prediction method depends on the datasets the method is tested on. AAS prediction methods are

typically tested on two types of data sets: a non neutral set, which contains substitutions assumed to affect protein function, and a neutral set, which contains substitutions assumed to have no effect. An AAS prediction method should predict the substitutions in the non neutral set to be damaging to protein function.
The percentage of non-neutral substitutions incorrectly predicted to be tolerated
is an approximation of the false negative errorrate. The AAS prediction method should also predict the majority of the substitutions in a neutral set as having no effect on protein Function The percentage of amino acid substitutions that can be predicted by an AAS method is defined as the method's coverage.
Coverage for AAS methods that rely on protein structure only is approximately 14% (83), whereas coverage for AAS methods that use sequence can be as high as 81% (53). Notably, most methods now have a sequence-based score and analysis of structure has become an option offered by the method.
These studies suggest that AAS prediction methods can provide insights into phenotypic differences observed between species. Because of false positive error, it would be difficult to study this on an individual
gene basis. However, by grouping genes within protein families or pathways, it may be possible to identify pathways that
have undergone relaxed selection in certain species The progress that has been made over the past few years with AAS prediction methods
is promising: methodology has improved and applications have proliferated. AAS prediction methods have proven successful for Mundelein traits and may eventually play an important role in identifying complex disease variants. Because AASs are a source of fundamental changes between and within species, AAS prediction methods will continue to be of major importance in the future[10].

3.3 Nanotechnology in cancer prevention, detection and treatment:-

This paper is an overview of advances and prospects in applications of nanotechnology for cancer prevention, detection and treatment. We begin with a brief description of the underlying causes of cancer. Then we address preventive treatment, disease-time treatment, and diagnosis in the context of some of the most recent advances in nanotechnology. Nanoparticle science is also briefly addressed as the foundation upon which most nanotechnology cancer therapy is based. It is demonstrated how nanotechnology can help solve one of the most challenging and

longstanding problems in medicine, which is how to eliminate cancer without harming normal body tissue. Introduction "*Nanotechnology will change the very foundations of cancer diagnosis, treatment, and prevention*". We have already seen how nanotechnology, an extremely wide and versatile field, can affect many of its composing disciplines in amazingly innovative and unpredictable ways. In fact, nanotechnology and the ideas and methods that it encompasses can be applied to almost any problem that leading researchers face today. since everything around us is made up of atomic and molecular matter, and all of our problems are ultimately rooted in atomic and molecular arrangements. Nanotechnology has at last provided a way for us to rearrange and restructure matter on an atomic scale, Genetics of cancer we can see how even a single nucleotide base inserted (or removed) out of sequence can cause the entire subsequent chain of amino acids, which are the building blocks of proteins, to be incorrect. When this happens in the gene sequence that codes for the protein(s) responsible for apoptosis or damping cell division, the cell permanently loses the ability to carry out that particular function It must be added that many other mutations can be also cancerous Every protein is unique and specially tailored to participate in certain complex biochemical reactions. A difference in the sequence of amino acids that make up a protein results in a difference in shape, thus altering or destroying the functionality of the protein .Although in some cases a point mutation may not cause a change in the amino acid sequence, and in others a change in the sequence may not cause a change in functionality, in a vast majority of cases even the slightest change will render the protein useless Conclusion Prevention, diagnosis and treatment of cancer have always been a formidable medical challenge. In fact, cancer has long been considered an incurable disease and it is grouped with Hepatitis C and AIDS. Throughout the bulk of human history, cancer tended to be fatal in those who were unfortunate to develop it. Cancer will continue to be a big problem since it is a disease related mostly to age. As our population average age increases due to medical advances cancer will be a major disease of the aging Although there is still plenty of work to be done, some very promising new nanotechnology treatment methods are in the works. Nanotechnology treatments can be used in both the preemptive and in the disease-time approaches to dealing with cancer. As we become better able to engineer more sophisticated nano devices and to equip them with effective targeting techniques for biomolecules, we will gain the ability to treat various kinds of cancer more and more effectively. . An important aspect of

cancer treatment is its early detection. There have been significant improvements largely due to breakthroughs, both, in the bottom-up and in the top-down nanotechnology. Developments in such areas as in nanoarrays, nanosensors, liposomes, monoclonal antibodies, improved nanoparticles (dendrimers, diamondoids, gold-based nanoparticles, magnetic nanoparticles, and quantum dots) and nanoelectronics are making early detection , prevention and treatment with a high degree of accuracy and ease possible. Also other recent discoveries and inventions in nanotechnology are suggesting that a safe and effective cure for cancer is just around the corner[11].

.

Methodology

This chapter is talk about the method we follow in our research, that include the program executed and the functions used with flow chart that abbreviate the steps of the program. this chapter also include a description of the data used .

4.1 Flow Chart:-

We collect fourteen data base from internet, these data base contains a DNA sequence of seven normal organs and seven of the same organs but with cancer to compare between them. The organs are; blood, kidney, lung, breast, skin, bones and colon, and using statistical analysis to compare between the normal and cancerous one.

Firstly convert the DNA sequence to amino acid sequence because we work on protein not DNA then calculate the number of amino acid's bases in the protein sequence; how many A in the sequence, how many R in the sequence, how many N in the sequence and so on . Then filter the sequence from noise by replace a substring by another and in this program replace (*) by (F).plot density of nucleotides along sequence using. calculates the atomic composition of the protein sequence such as: the number of carbon atoms in the sequence, the number of hydrogen atoms in the sequence, the number of nitrogen atoms in the sequence, the number of oxygen atoms in the sequence and the number of sulfur atoms in the sequence ,then calculates the molecular weight of amino acid. calculate isoelectric point for amino acid sequence; isoelectric point It is the pH at which the amino acid or proteins are electrically neutral

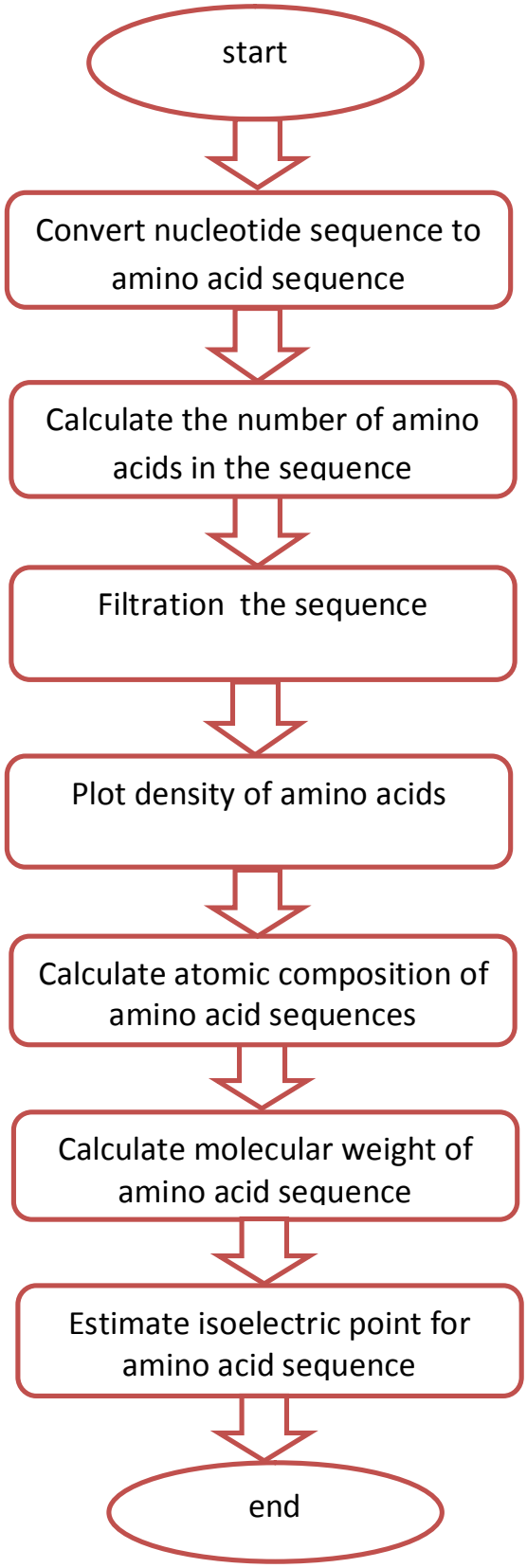


Figure 4.1: show the flow chart of the program

Results and Discussion

5.1 Results:-

5.1.1 Normal and Cancerous Blood:-

5.1.1.1 aacount Before Filtration:

Table 5.1 : show the amino acid count of normal and cancerous blood before filtration

Normal blood	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	30	37	28	18	22	32	21	27	16	38	72	33	12	50	43	99	35	18	17	36
Cancerous blood	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	34	24	18	21	27	18	45	18	41	86	34	12	54	33	58	38	16	26	36

Normal blood	Others
	29
Cancerous blood	Others
	28

Normal Blood

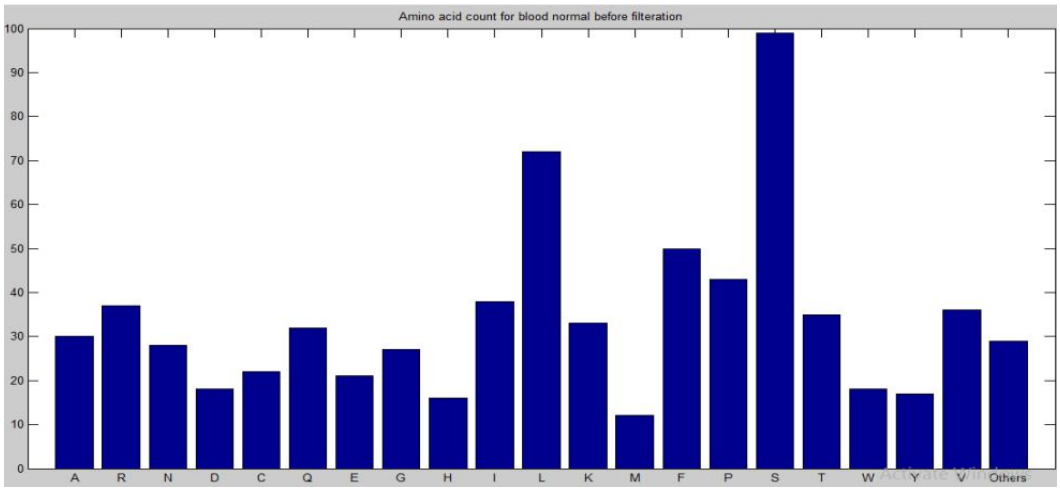


Figure 5.1 : show amino acid count of normal blood bef2ore filtration

Cancerous Blood:

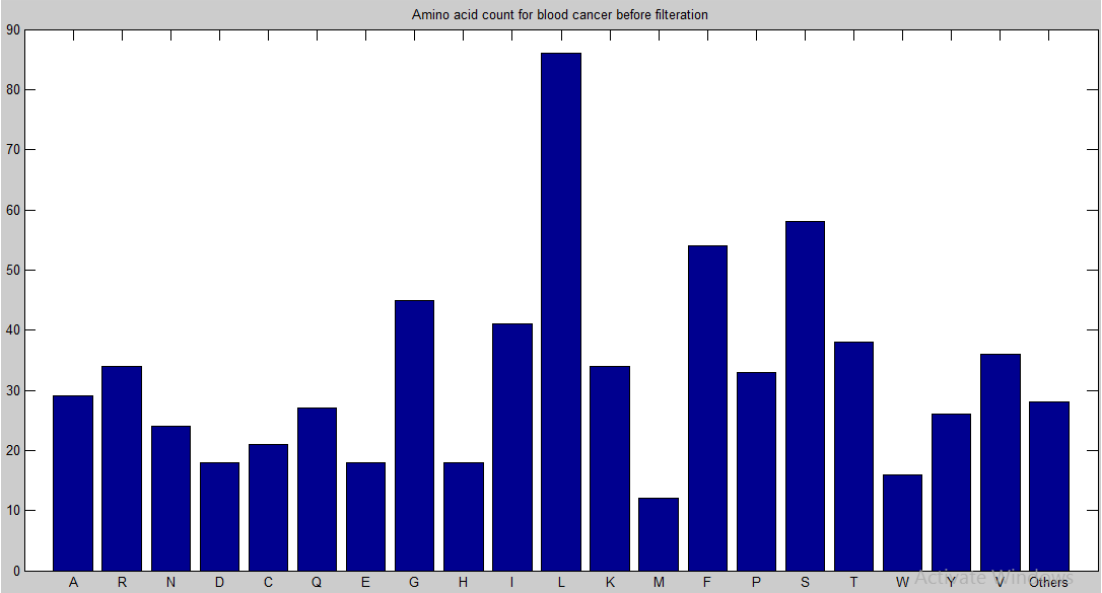


Figure 5.2 : show the amino acid count of cancerous blood before filtration

5.1.1.2 aacount After Filtration:

Table 5.2 : show the amino acid count of normal and cancerous blood after filtration

Normal blood	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	30	37	28	18	22	32	21	27	16	38	72	33	12	79	43	99	35	18	17	36
Cancerous blood	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	34	24	18	21	27	18	45	18	41	86	34	12	82	33	58	38	16	26	36

Normal Blood:

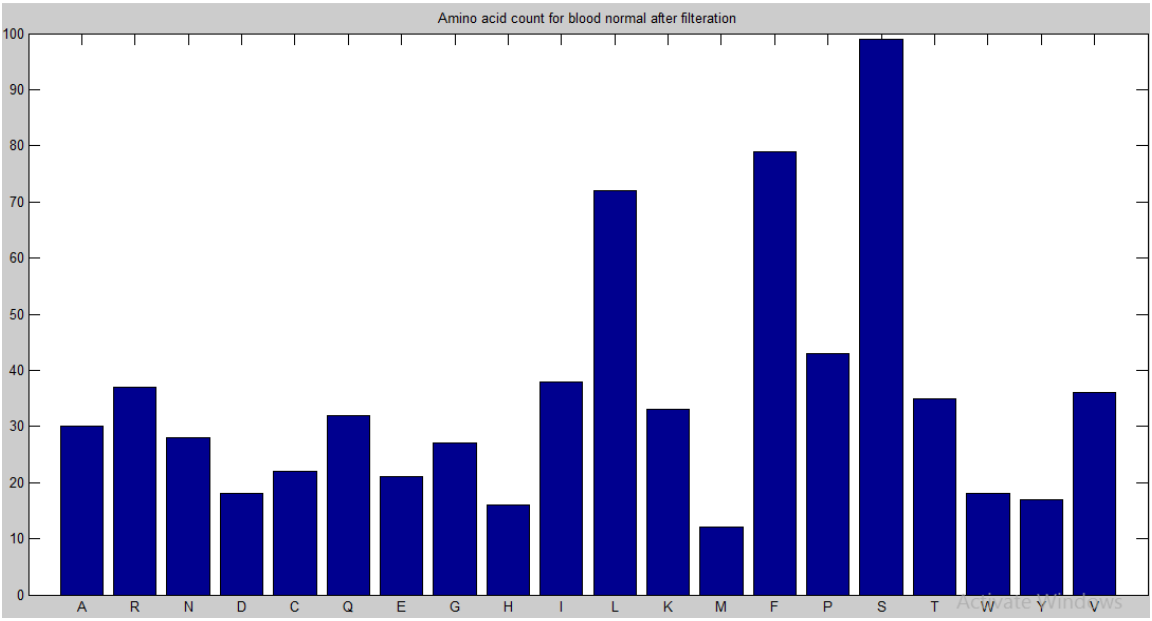


Figure 5.3 : show amino acid count of normal blood after filtration

Cancerous blood:

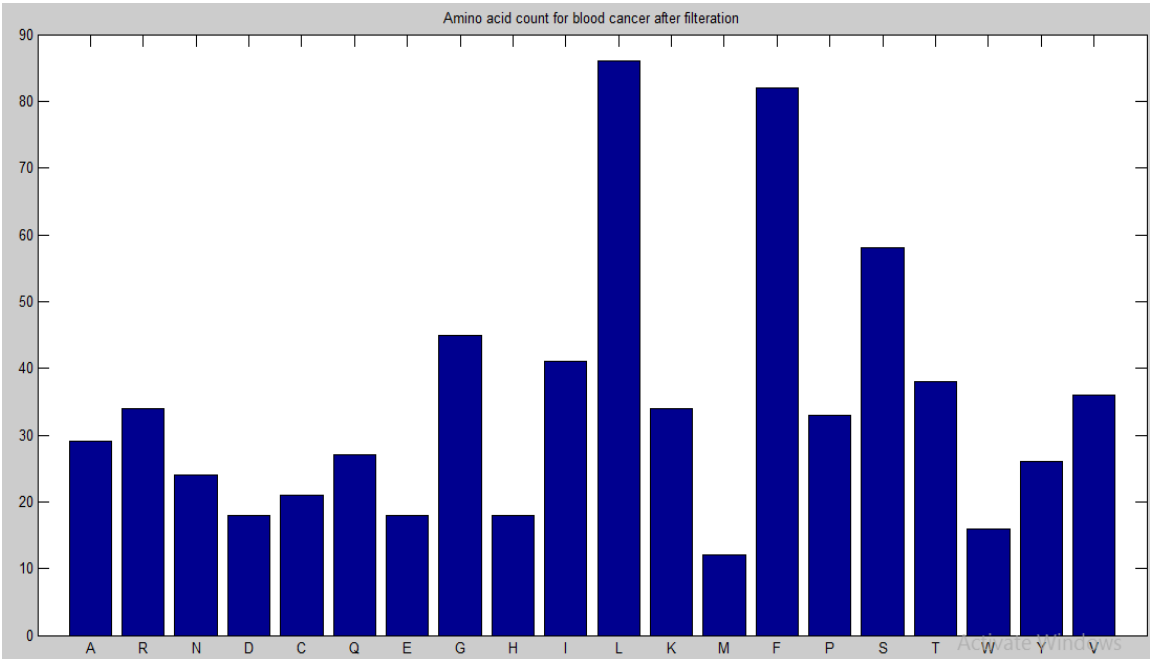


Figure 5.4 : show amino acid count of cancerous blood after filtration

5.1.1.3 Ntdensity:

Normal Blood:

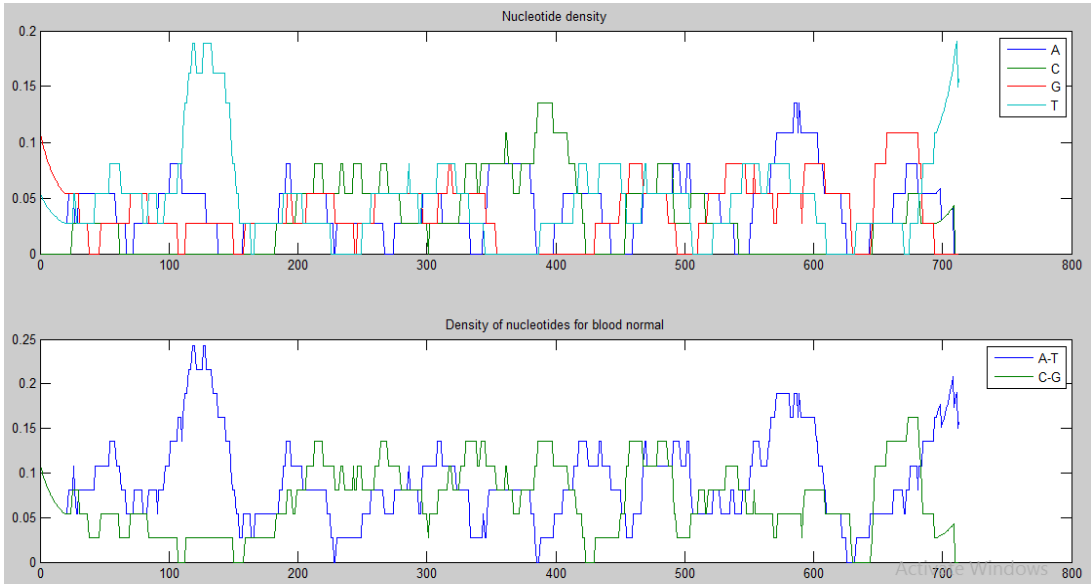


Figure 5.5 : show the density of normal blood

Cancerous Blood:

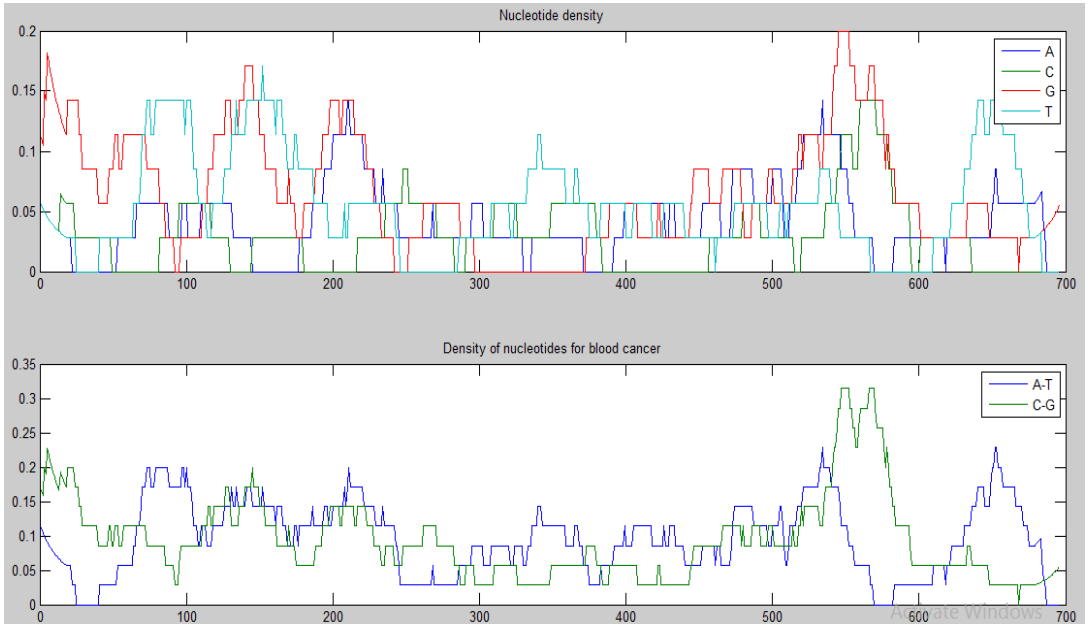


Figure 5.6 : show the density of cancerous blood

5.1.1.4 atomiccomp:

Table 5.3: show the atomic composition of normal and cancerous blood

Normal Blood	C	H	N	O	S
	3789	*5683*	*967*	*1003*	*34*
Cancerous Blood	C	H	N	O	S
	3790	*5653*	*935*	*942*	*33*

5.11.5 molweight:

Table 5.4 : show the molecular weight of normal and cancerous blood

Normal blood	8.1919×10^4
Cancerous blood	8.0444×10^4

5.1.1.6 isoelectric point:

Normal Blood:

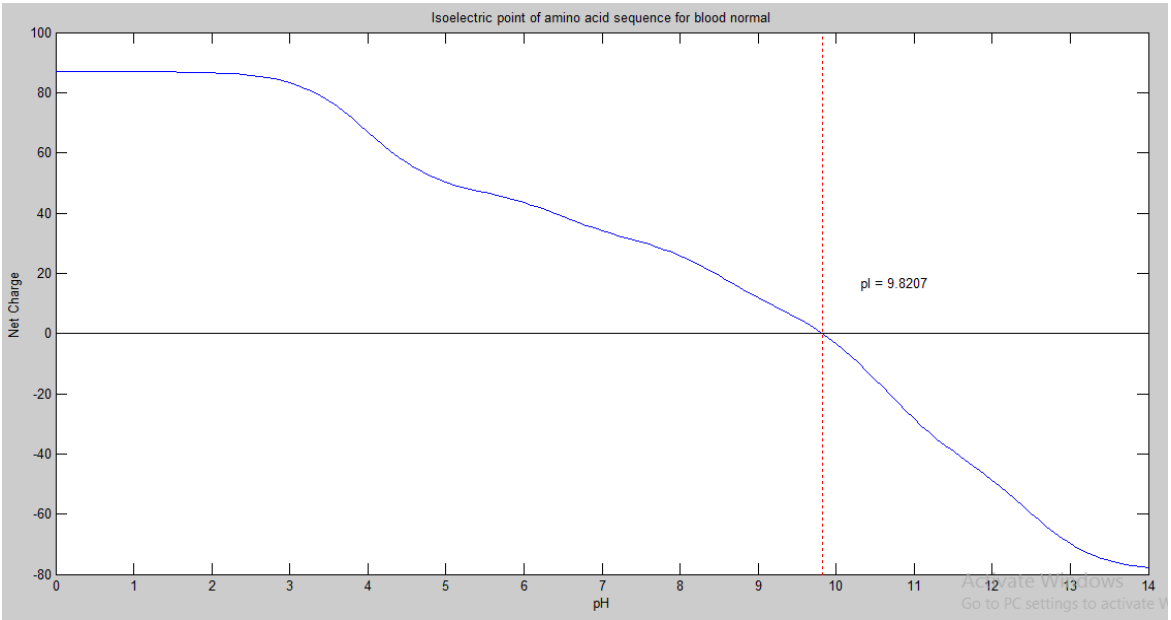


Figure 5.7: show the isoelectric point of normal blood

Cancerous Blood:

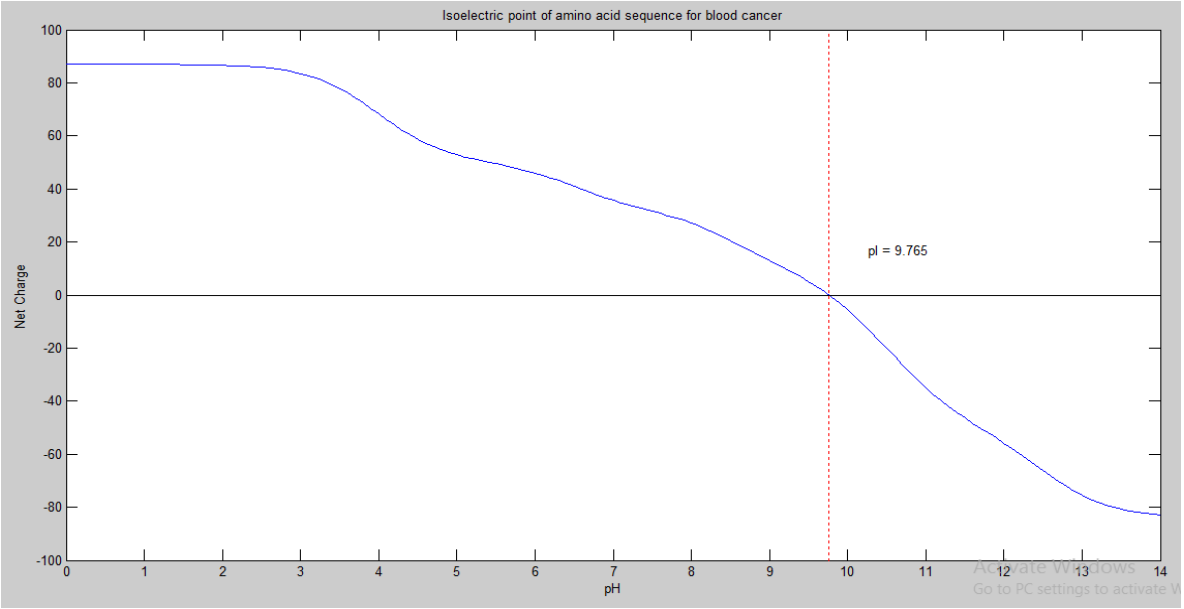


Figure 5.8: show the isoelectric point of cancerous blood

5.1.2 Normal And Cancerous Kidney:-

5.1.2.1 aacount before filtration:

Table 5.5: show the amino acid count of normal and cancerous kidney before filtration

Normal kidney	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	1370	1840	1565	843	1304	1437	1386	1653	1132	2370	4138	2302	776	2293	1510	3220	1700	596	1274	1895
Cancerous kidney	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	1383	1807	1527	854	1221	1396	1318	1752	1087	2356	4104	2281	664	2335	1692	3236	1808	609	1289	1898
Normal kidney	Others																			
	2165																			
Cancerous kidney	Others																			
	2108																			

Normal Kidney:

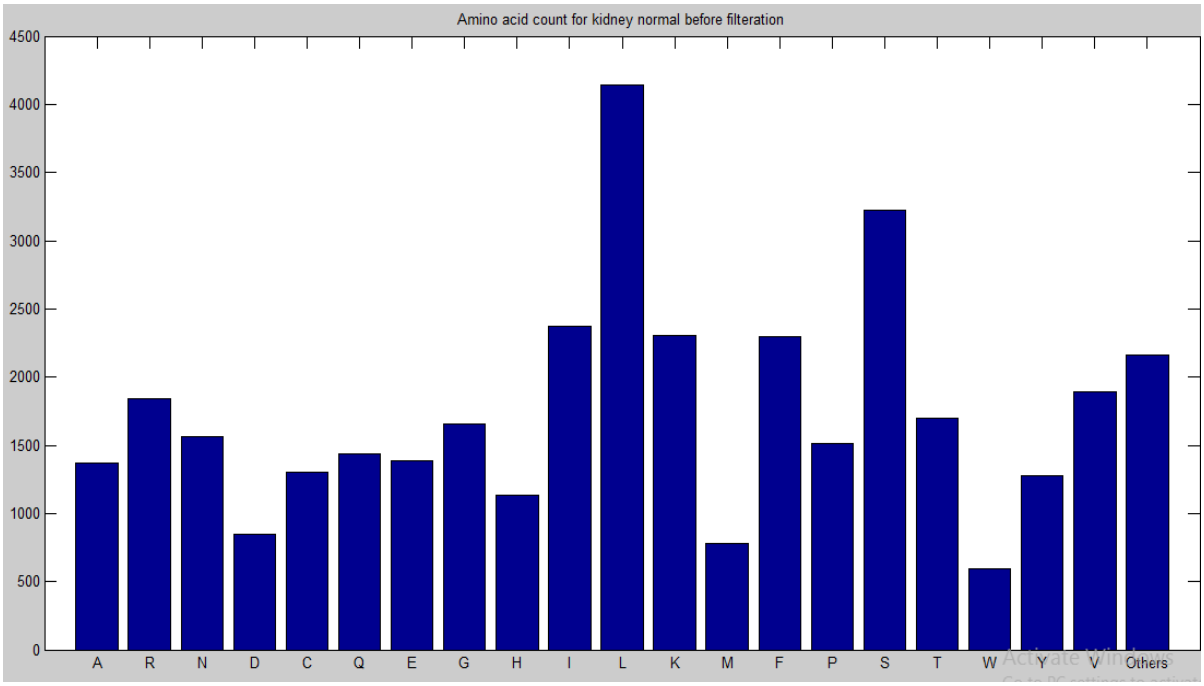


Figure 5.9 : show amino acid count of normal kidney before filtration

Cancerous Kidney:

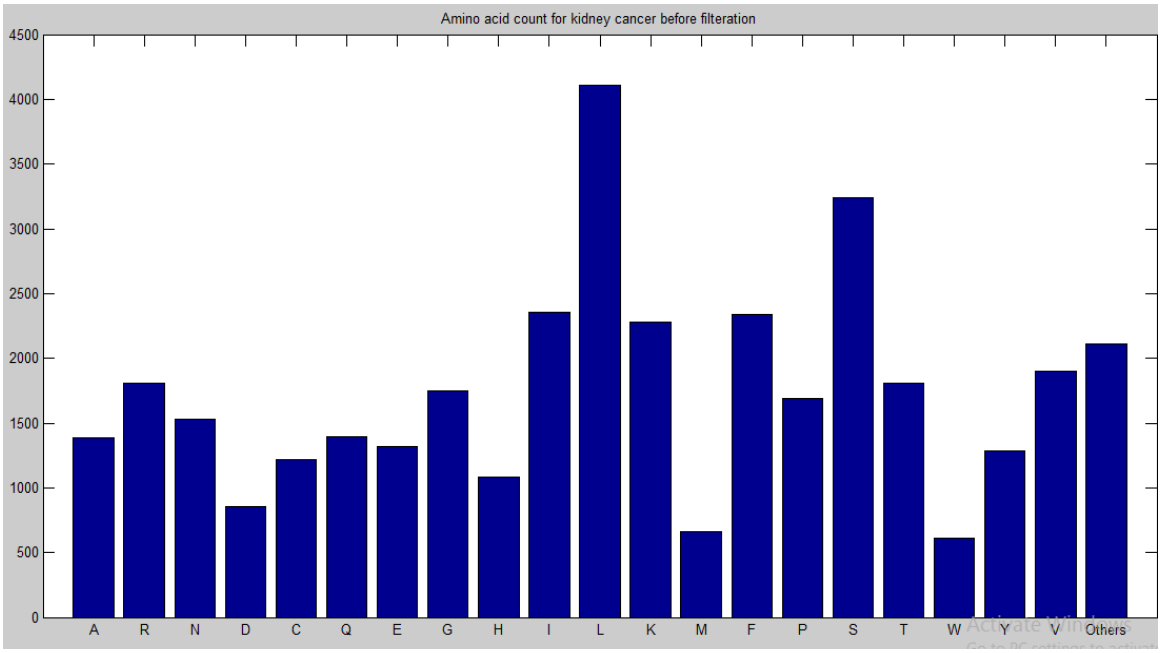


Figure 5.10 : show amino acid count of cancerous kidney before filtration

5.1.2.2 aacount after filtration:

Table 5.6 : show the amino acid count of normal and cancerous kidney after filtration

Normal kidney	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	1370	1840	1565	843	1304	1437	1386	1653	1132	2370	4138	2302	776	4458	1510	3220	1700	596	1274	1895
Cancerous kidney	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	1383	1807	1527	854	1221	1396	1318	1752	1087	2356	4104	2281	664	4443	1692	3236	1808	609	1289	1898

Normal Kidney:

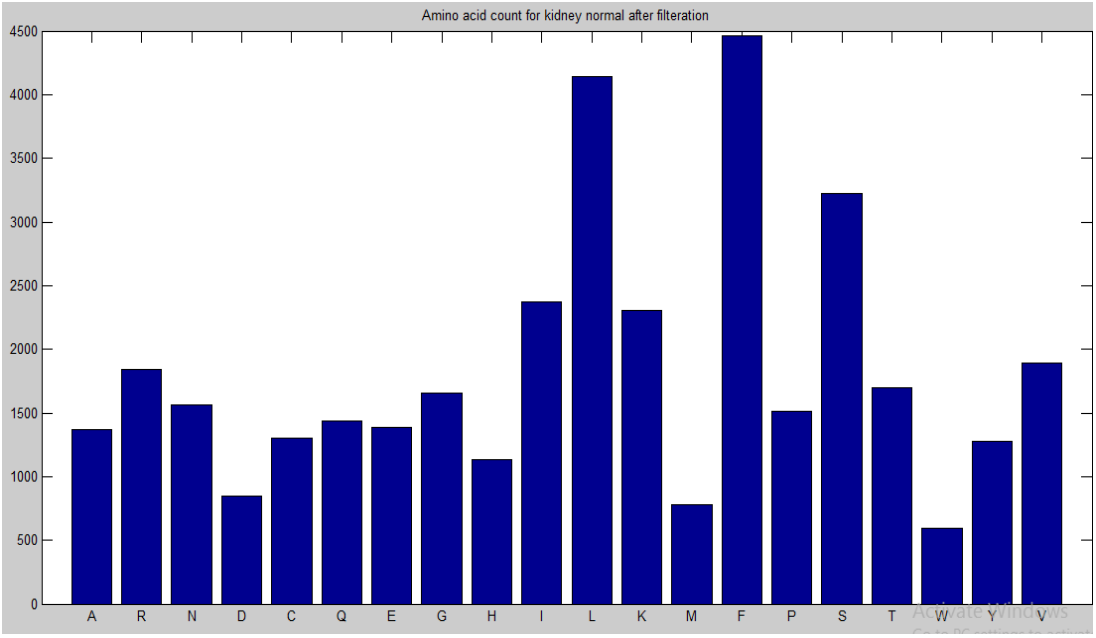


Figure 5.11 : show amino acid count of normal kidney after filtration

Cancerous Kidney:

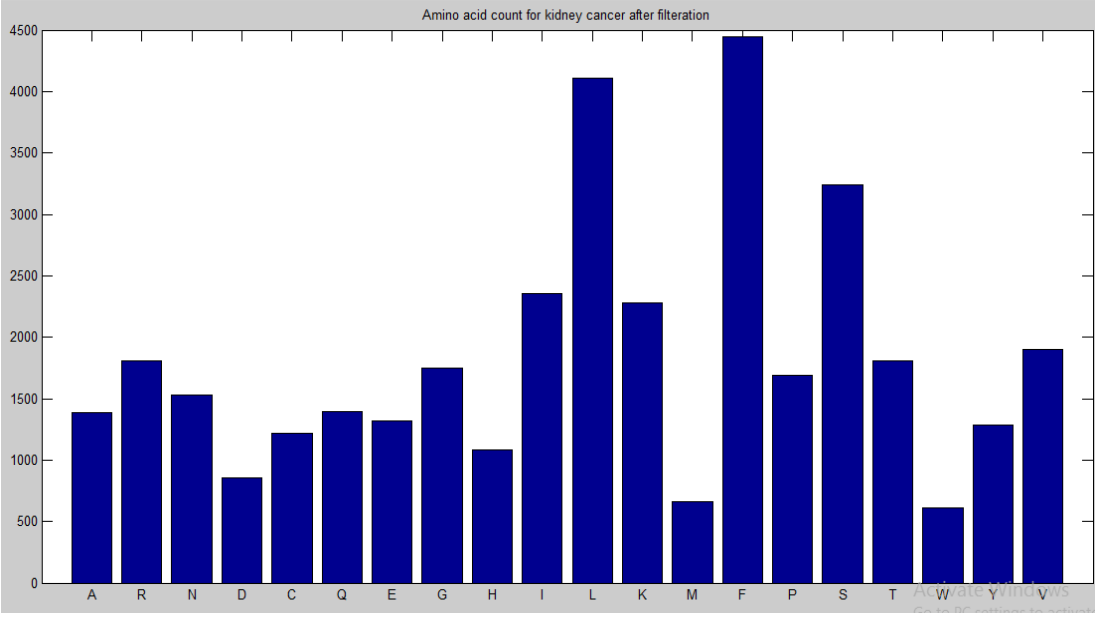


Figure 5.12 : show amino acid count of cancerous kidney after filtration

5.1.2.3 ntdensity:

Normal Kidney:

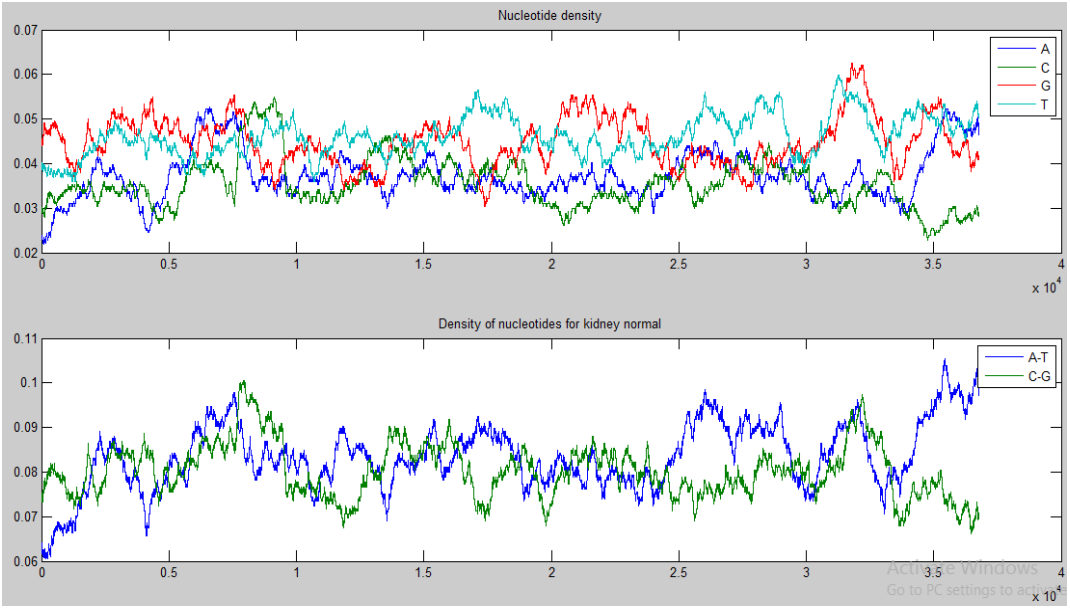


Figure 5.13 : show the density of normal kidney

Cancerous Kidney:

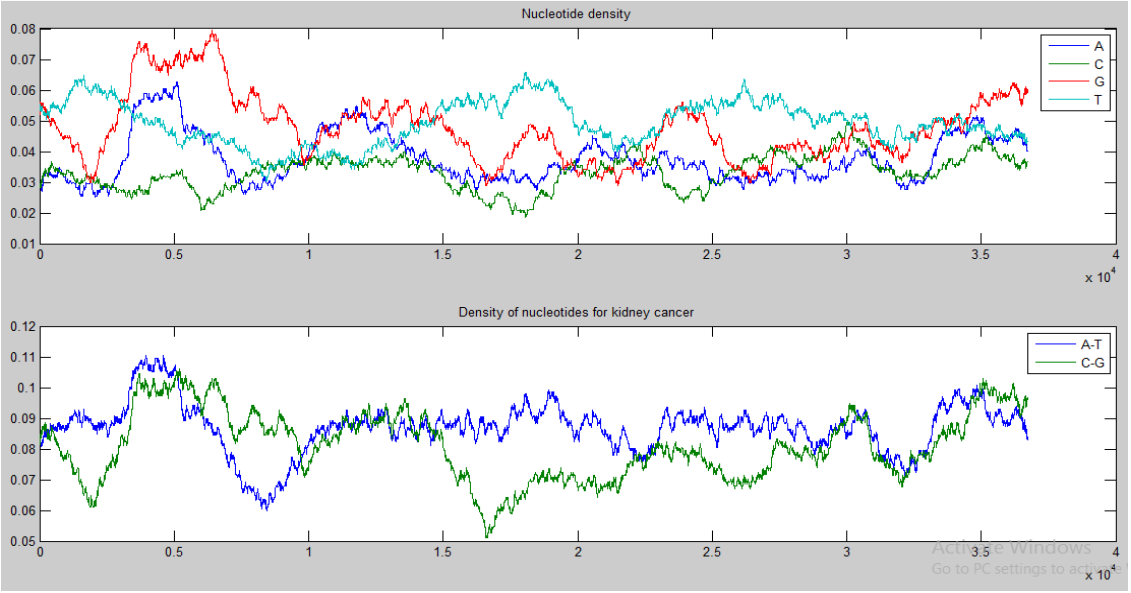


Figure 5.14 : show the density of cancerous kidney

5.1.2.4 atomiccomp:

Table 5.7 : show the atomic composition of normal and cancerous kidney

Normal Kidney	C	H	N	O	S
	201276	302507	50453	50424	2080
Cancerous Kidney	C	H	N	O	S
	200717	301245	50133	50326	1885

5.1.2.5 molweight:

Table 5.8 : show the molecular weight of normal and cancerous kidney

Normal kidney	4.3025×10^6
Cancerous kidney	4.2822×10^6

5.1.2.6 isoelectric point:

Normal Kidney:

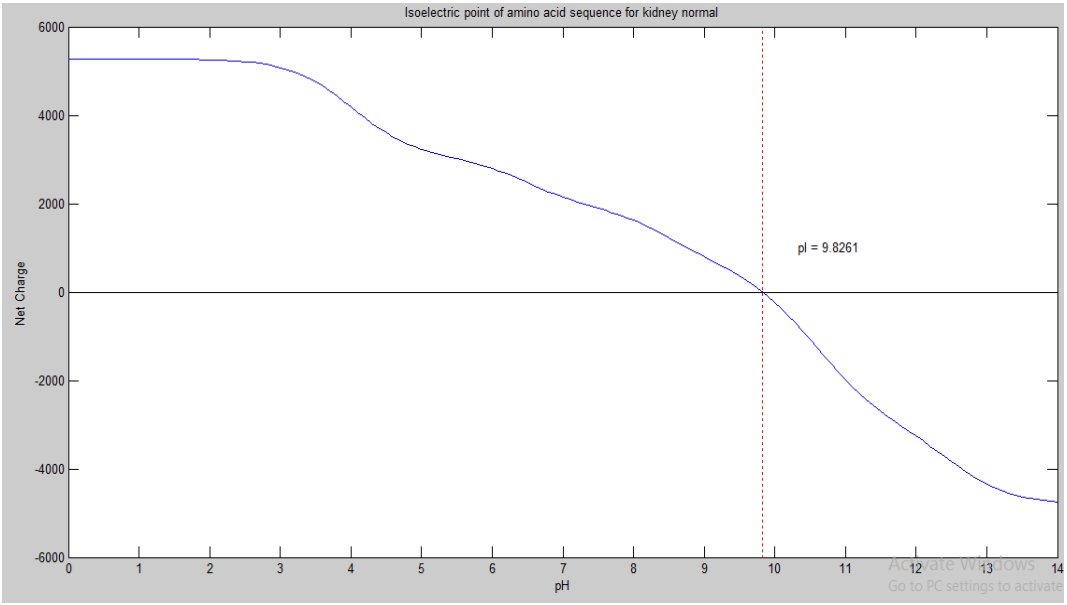


Figure 5.15 : show the isoelectric point of normal kidney

Cancerous Kidney:

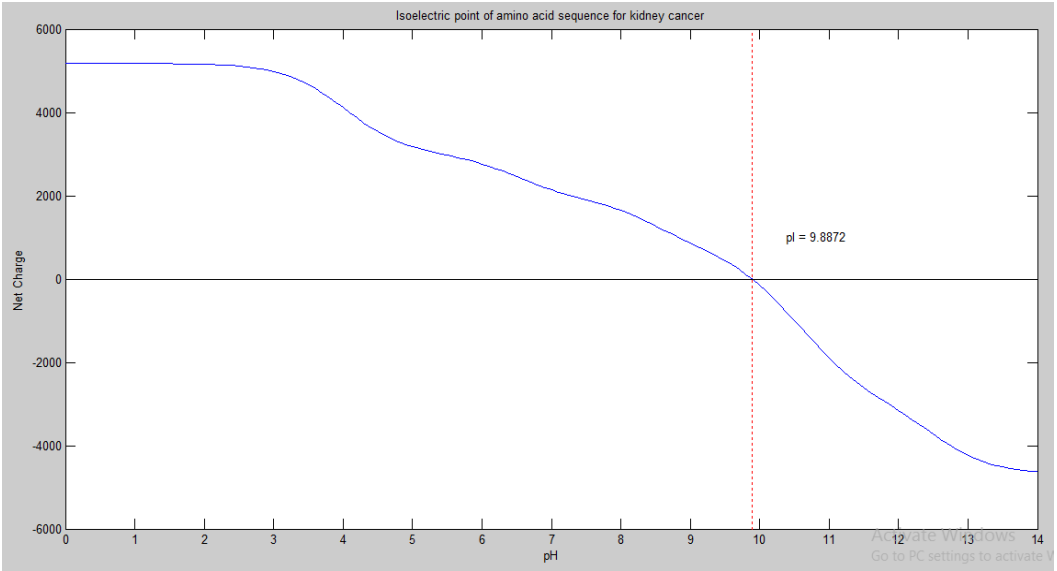


Figure 5.16 : show the isoelectric point of cancerous kidney

5.1.3 Normal And Cancerous Lung:-

5.1.3.1 aacount before filtration:

Table 5.9 : show the amino acid count of normal and cancerous lung before filtration

Normal lung	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	40	*39*	*43*	*28*	*24*	*42*	*45*	*49*	*34*	*55*	*68*	*78*	*27*	*31*	*37*	*58*	*64*	*16*	*31*	*38*
Cancerous lung	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	*34*	*24*	*18*	*21*	*27*	*18*	*45*	*18*	*41*	*86*	*34*	*12*	*54*	*33*	*58*	*38*	*16*	*26*	*36*
Normal lung	Others																			
	44																			
Cancerous lung	Others																			
	28																			

Normal Lung:

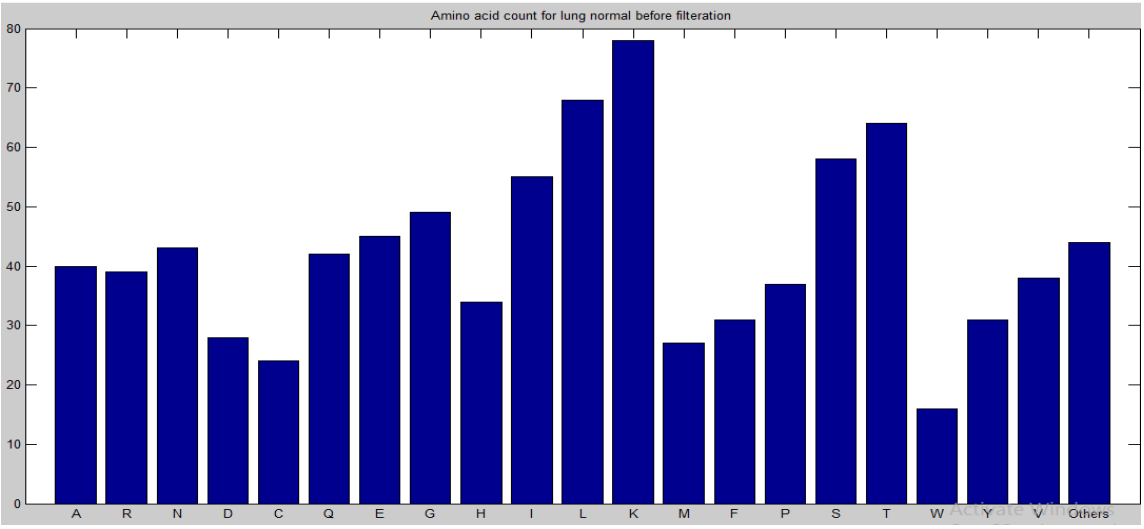


Figure 5.17 :show amino acid count of normal lung before filtration

Cancerous Lung:

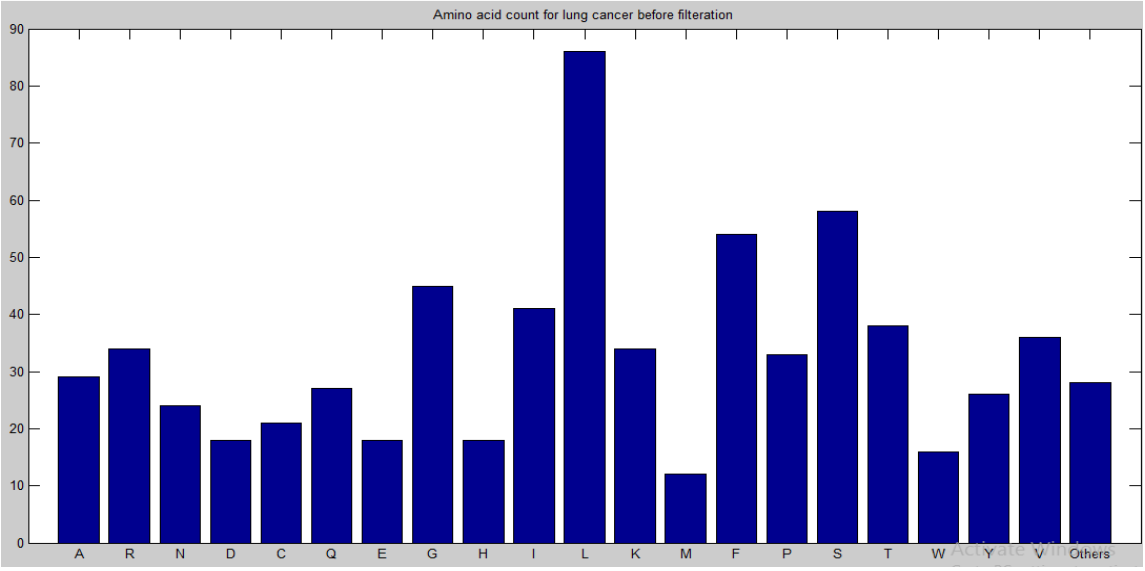


Figure 5.18 : show amino acid count of cancerous lung before filtration

5.1.3.2 aacount after filtration:

Table 5.10 : show the amino acid count of normal and cancerous lung after filtration

Normal lung	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	40	*39*	*43*	*28*	*24*	*42*	*45*	*49*	*34*	*55*	*68*	*78*	*27*	*75*	*37*	*58*	*64*	*16*	*31*	*38*
Cancerous lung	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	*34*	*24*	*18*	*21*	*27*	*18*	*45*	*18*	*41*	*86*	*34*	*12*	82	*33*	*58*	*38*	*16*	*26*	*36*

Normal Lung:

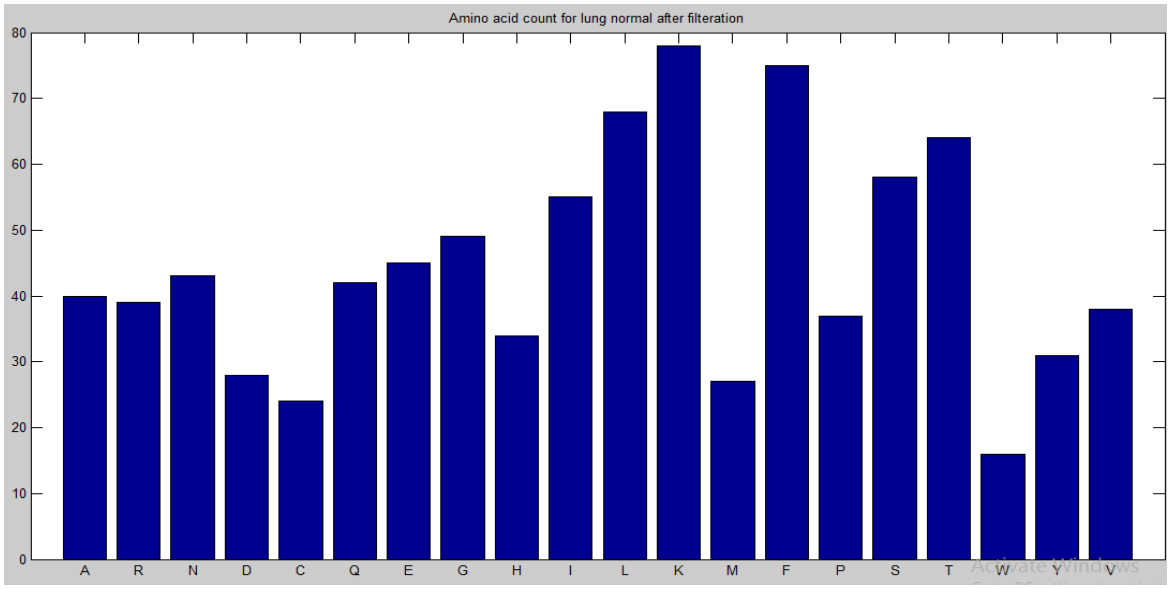


Figure 5.19 : show amino acid count of normal lung after filtration

Cancerous Lung:

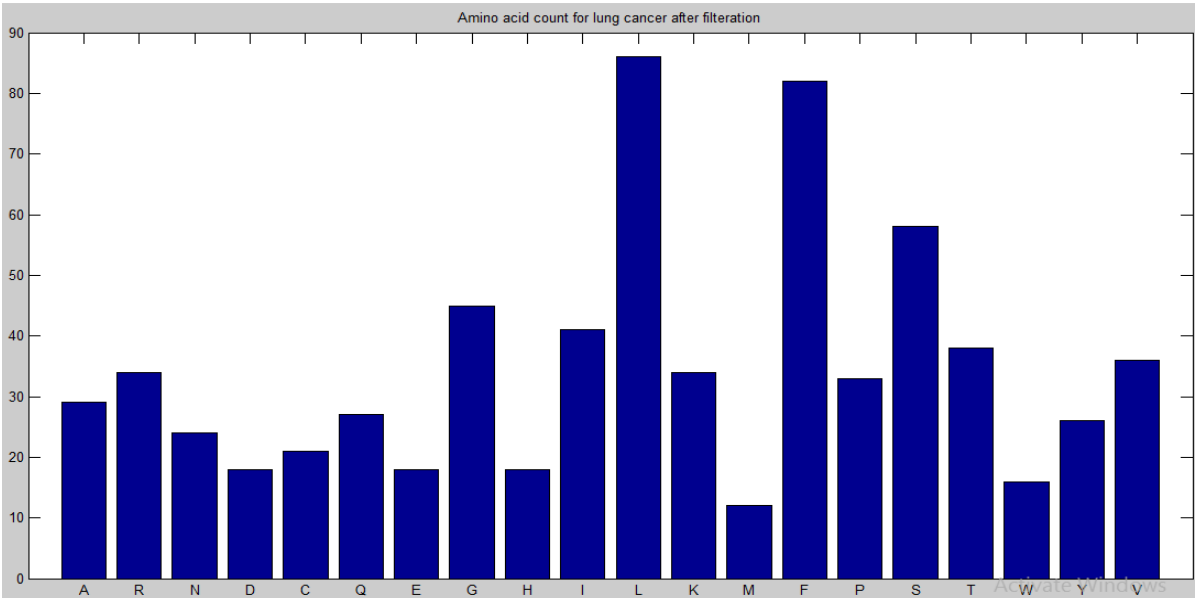


Figure 5.20 : show amino acid count of cancerous lung after filtration

5.1.2.3 ntdensity:

Normal Lung:

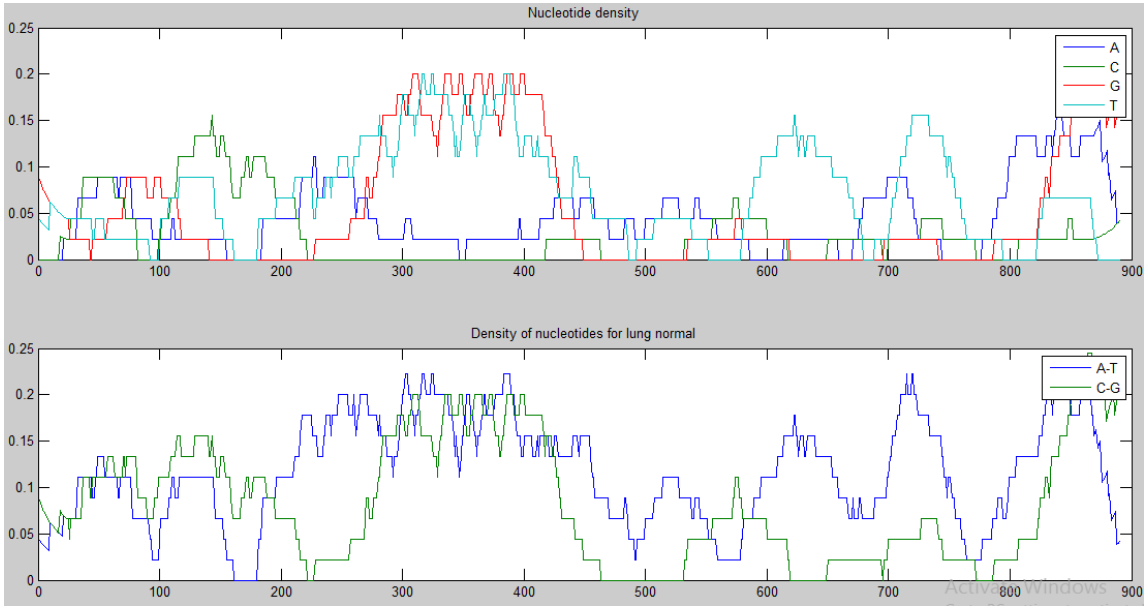


Figure 5.21 : show the density of normal lung

Cancerous Lung:

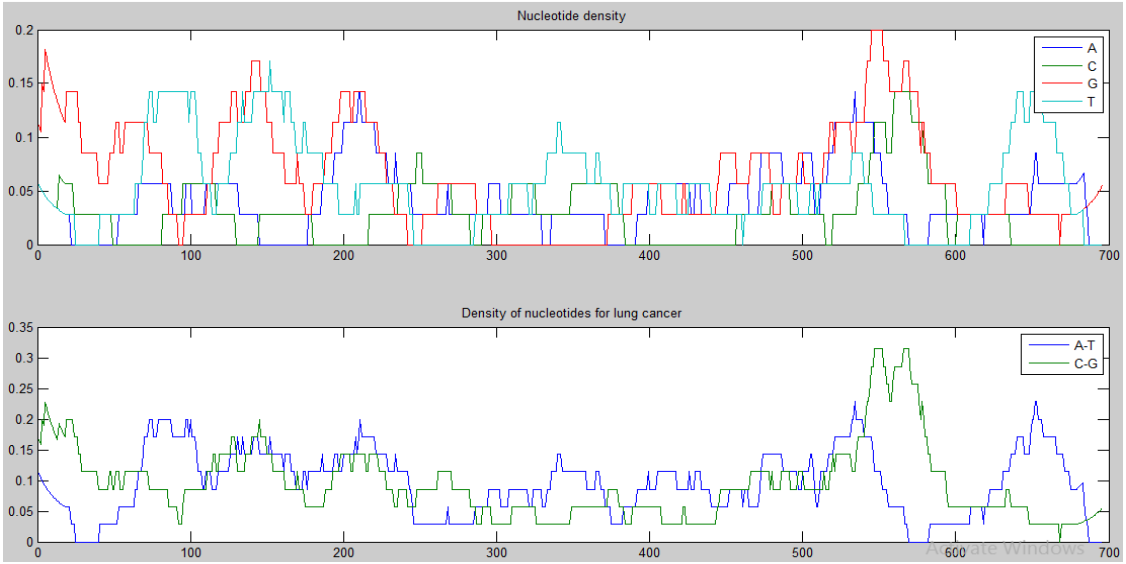


Figure 5.22 : show the density of cancerous lung

5.1.3.4 atomiccomp:

Table 5.11 : show the atomic composition of normal and cancerous lung

Normal Lung	C	H	N	O	S
	4723	*7209*	*1255*	*1276*	*51*
Cancerous Lung	C	H	N	O	S
	3790	*5653*	*935*	*942*	*33*

5.1.2.5 molweight:

Table 5.12 : show the molecular weight of normal and cancerous lung

Normal lung	*1.0362×10^5*
Cancerous lung	*8.0444×10^4*

5.1.2.6 isoelectric point:

Normal Lung:

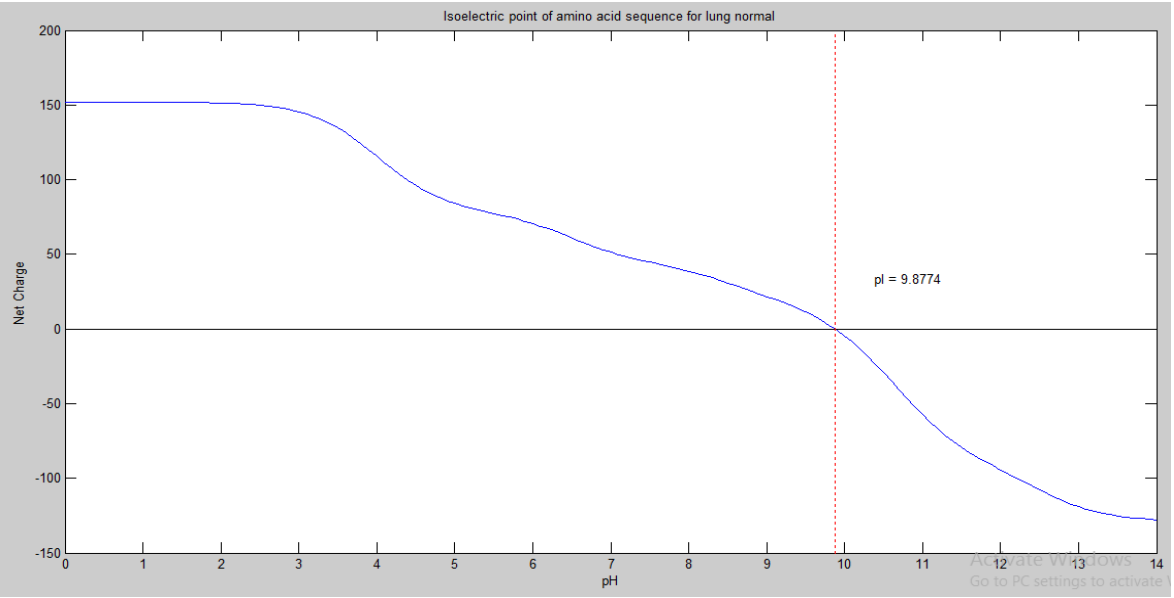


Figure 5.23 : show the isoelectric point of normal lung

Cancerous Lung:

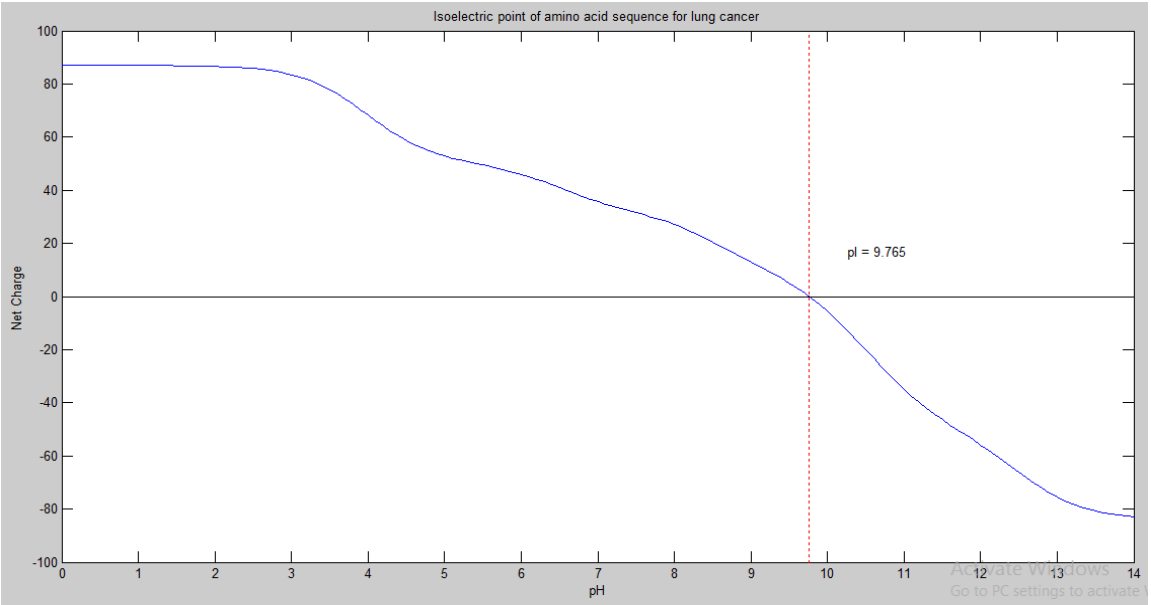


figure 5.24 : show the isoelectric point of cancerous lung

5.1.4 Normal And Cancerous Breast:-

5.1.4.1 aacount before filtration:

Table 5.13 : show the amino acid count of normal and cancerous breast before filtration

Normal breast	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	30	*51*	*39*	*21*	*22*	*50*	*32*	*58*	*40*	*58*	*67*	*74*	*26*	*34*	*37*	*70*	*57*	*10*	*25*	*33*
Cancerous breast	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	*34*	*24*	*18*	*21*	*27*	*18*	*45*	*18*	*41*	*86*	*34*	*12*	*54*	*33*	*58*	*38*	*16*	*26*	*36*

Normal breast	Others
	34
Cancerous breast	Others
	28

Normal Breast:

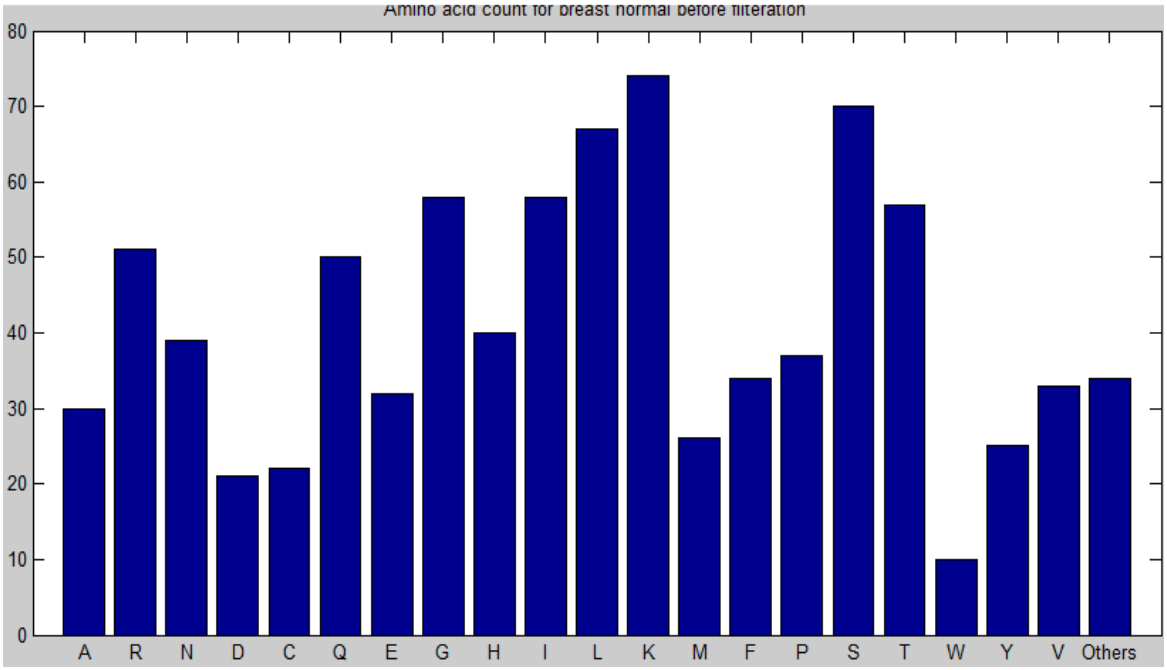


Figure 5.25 : show amino acid count of normal breast before filtration

Cancerous Breast:

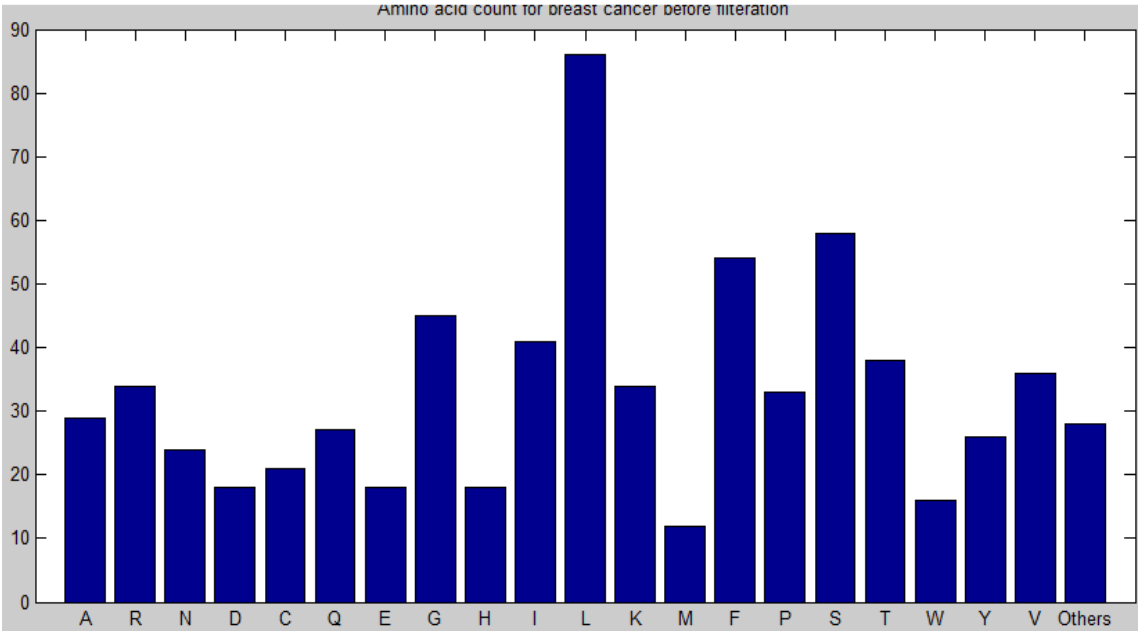


Figure 5.26 : show amino acid count of cancerous breast before filtration

5.1.4.2 aacount after filtration:

Table 5.14 : show the amino acid count of normal and cancerous breast after filtration

Normal breast	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	30	*51*	*39*	*21*	*22*	*50*	*32*	*58*	*40*	*58*	*67*	*74*	*26*	68	*37*	*70*	*57*	*10*	*25*	*33*
Cancerous breast	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	*34*	*24*	*18*	*21*	*27*	*18*	*45*	*18*	*41*	*86*	*34*	*12*	82	*33*	*58*	*38*	*16*	*26*	*36*

Normal Breast:

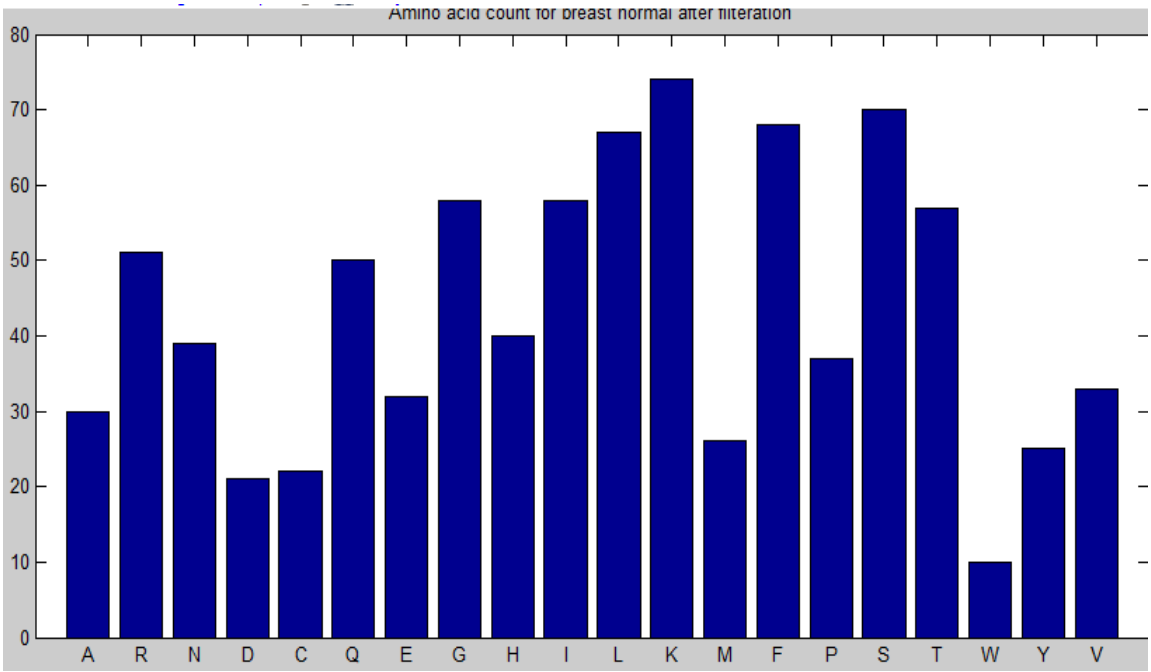


Figure 5.27 : show amino acid count of normal breast after filtration

Cancerous Breast:

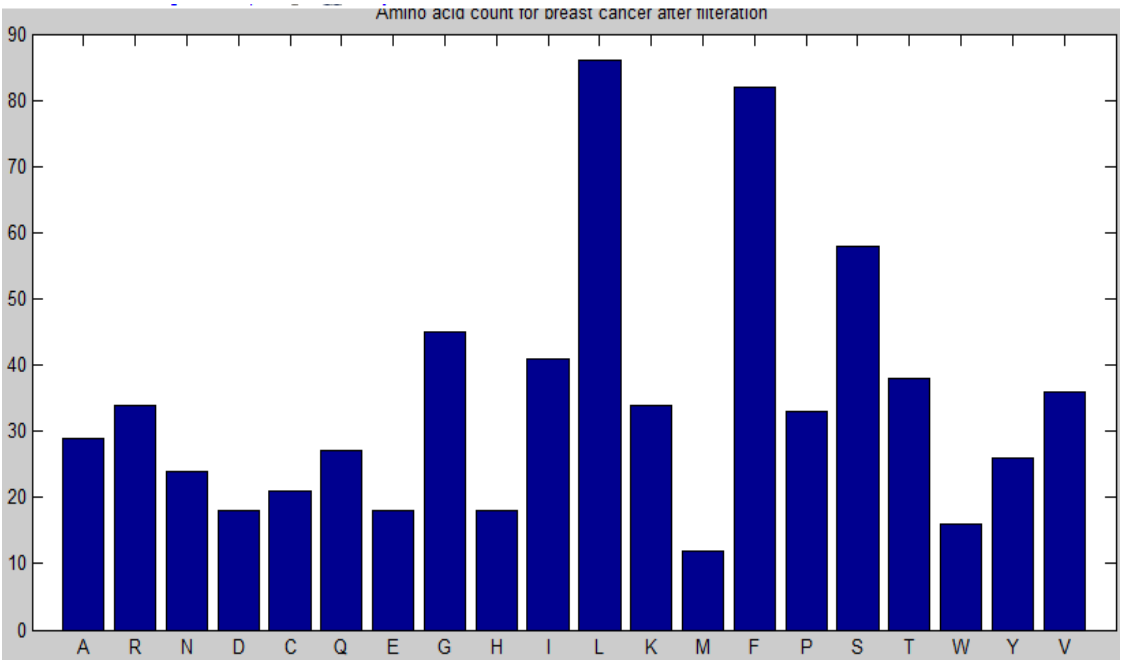


Figure 5.28 : show amino acid count of cancerous breast after filtration

5.1.4.3 Ntdensity:

Normal Breast:

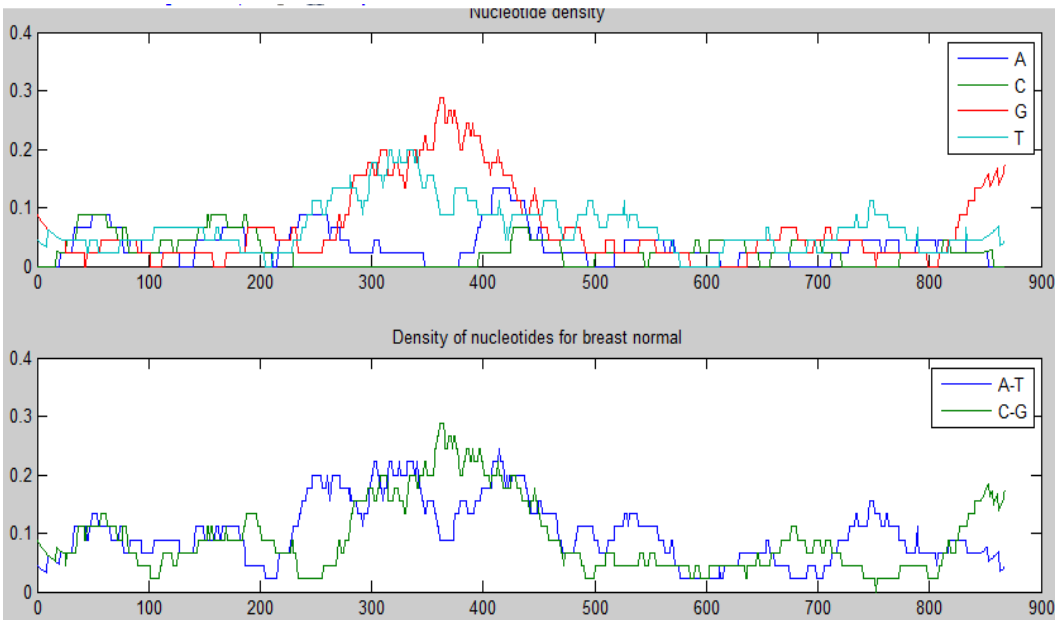


Figure 5.29 : show the density of normal breast

Cancerous Breast:

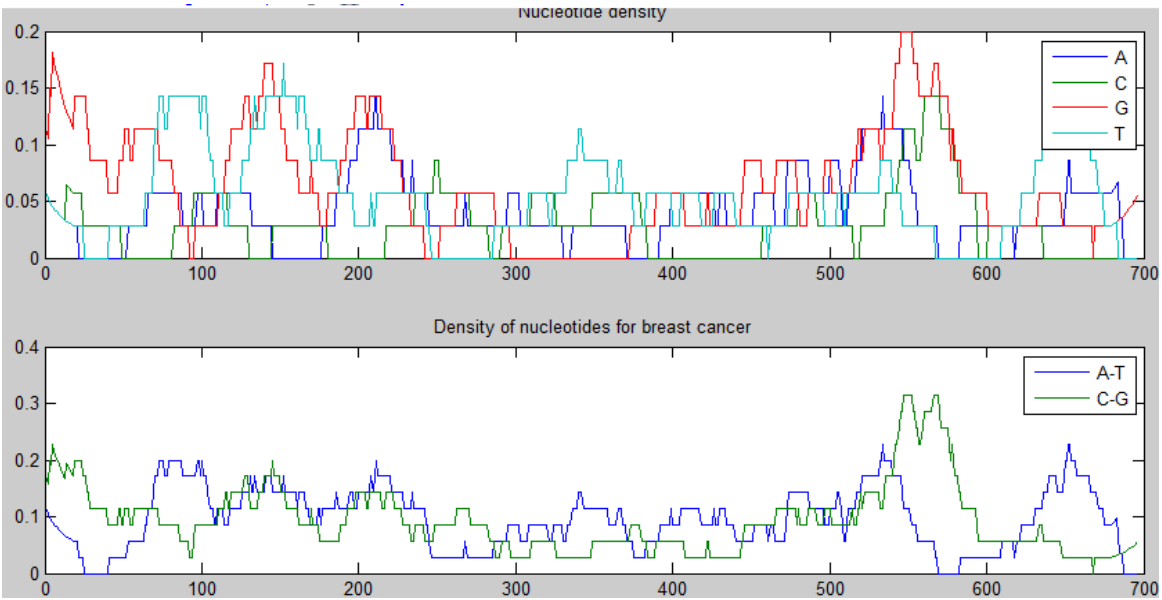


Figure 5.30 : show the density of cancerous breast

5.1.4.4 atomiccomp:

Table 5.15 : show the atomic composition of normal and cancerous breast

Normal Breast	C	H	N	O	S
	4527	*7030*	*1274*	*1216*	*48*
Cancerous Breast	C	H	N	O	S
	3790	*5653*	*935*	*942*	*33*

5.1.4.5 molweight:

Table 5.16 : show the molecular weight of normal and cancerous breast

Normal breast	1.0030×10^5
Cancerous breast	8.0444×10^4

5.1.4.6 isoelectric point:

Normal Breast:

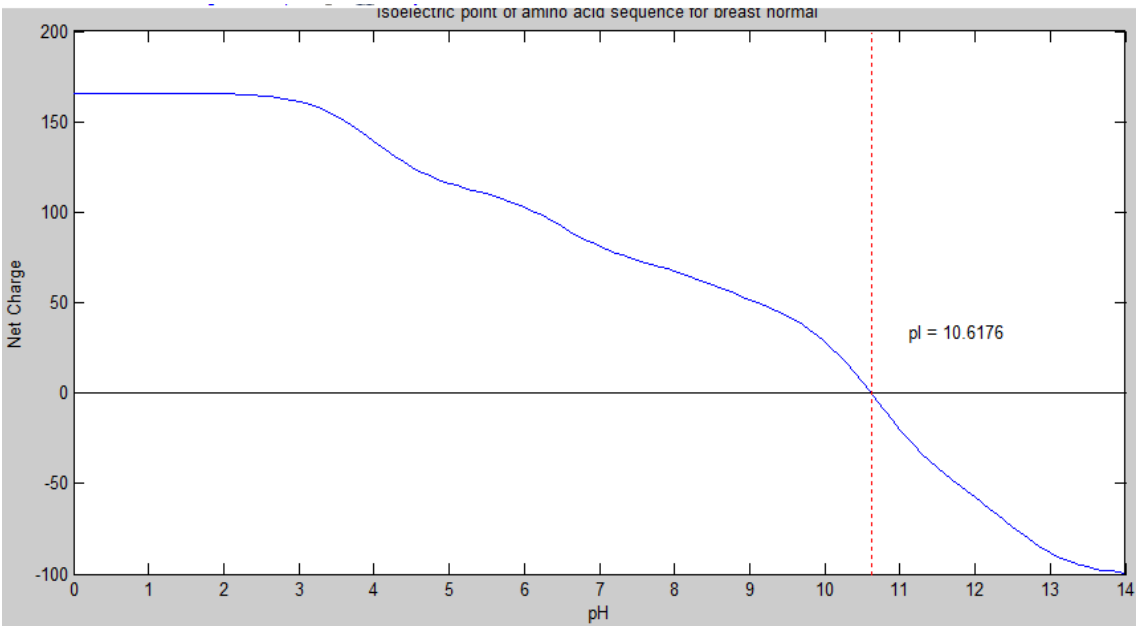
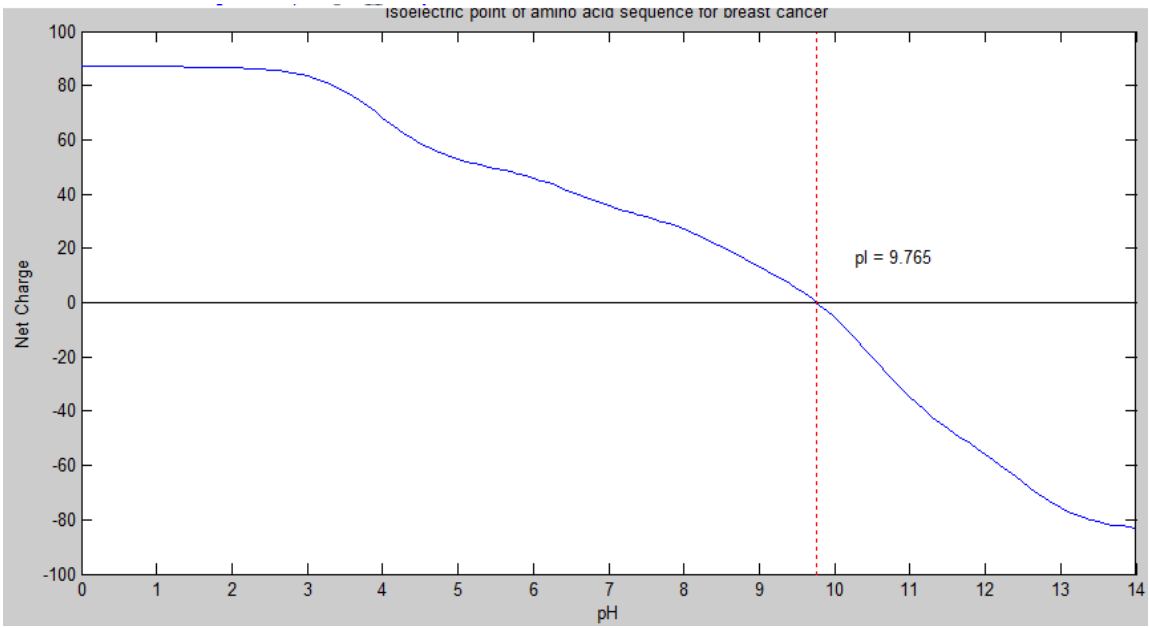


Figure 5.31 : show the isoelectric point of normal breast

Cancerous Breast:



(Figure 5.32) : show the isoelectric point of cancerous breast

5.1.5 Normal And Cancerous Skin:-

5.1.5.1 aacount before filtration:

Table 5.17 : show the amino acid count of normal and cancerous skin before filtration

Normal skin	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	6221	7120	5198	3125	4988	5787	5137	7549	5033	7836	14898	7795	2745	8449	7275	12362	6804	2634	4132	7158
Cancerous skin	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	6744	7343	4297	3142	5460	5137	5140	8331	4248	6378	15386	6873	2395	7971	7939	12480	6469	2985	3312	7987

Normal skin	Others
	6815
Cancerous skin	Others
	6668

Normal Skin:

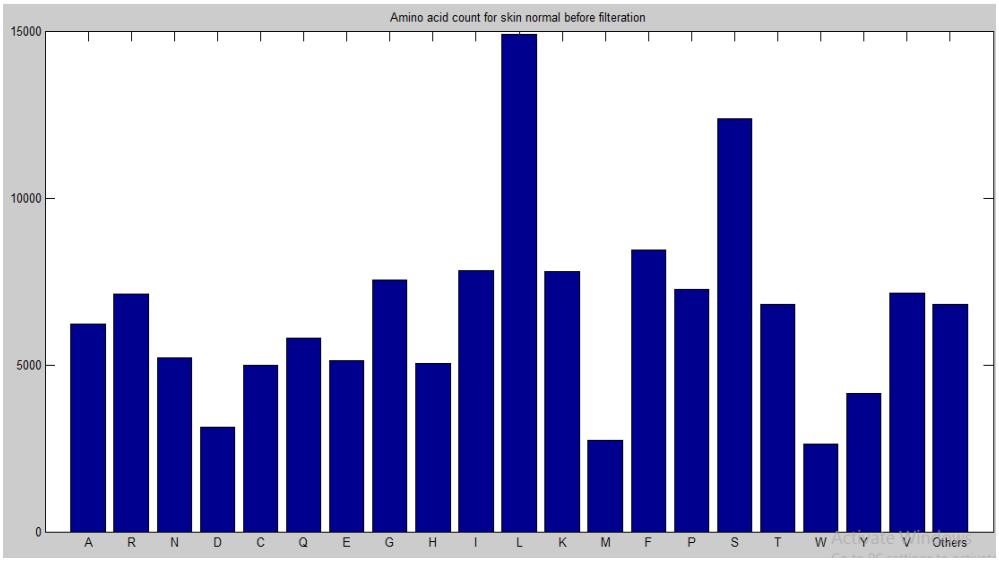


Figure 5.33 : show amino acid count of normal skin before filtration

Cancerous Skin:

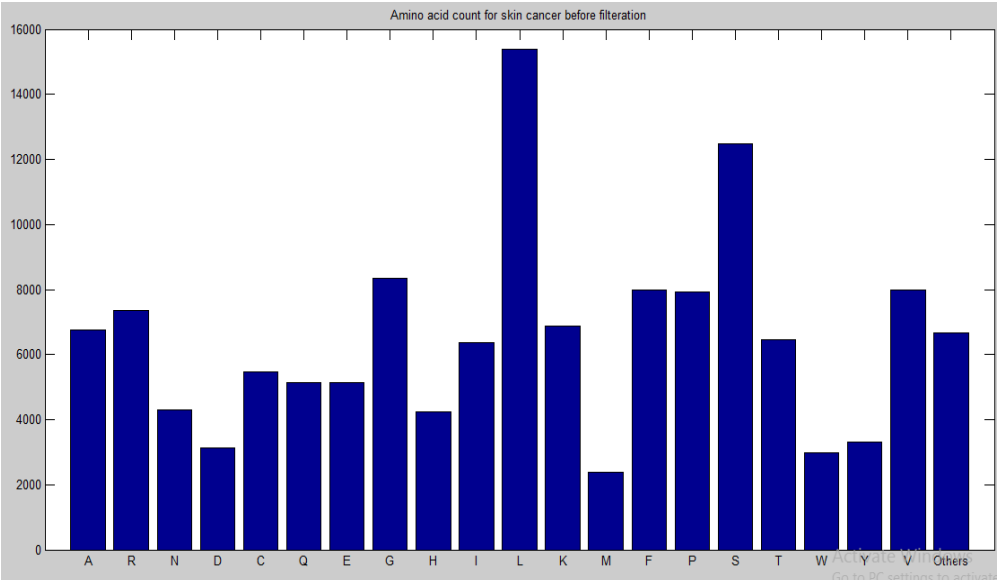


Figure 5.34 : show amino acid count of cancerous skin before filtration

5.1.5.2 aacount after filtration:

Table 5.18 : show the amino acid count of normal and cancerous skin after filtration

Normal skin	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	6221	7120	5198	3125	4988	5787	5137	7549	5033	7836	14898	7795	2745	15264	7275	12362	6804	2634	4132	7158
Cancerous skin	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	6744	7343	4297	3142	5460	5137	5140	8331	4248	6378	15386	6873	2395	14639	7939	12480	6469	2985	3312	7987

Normal Skin:

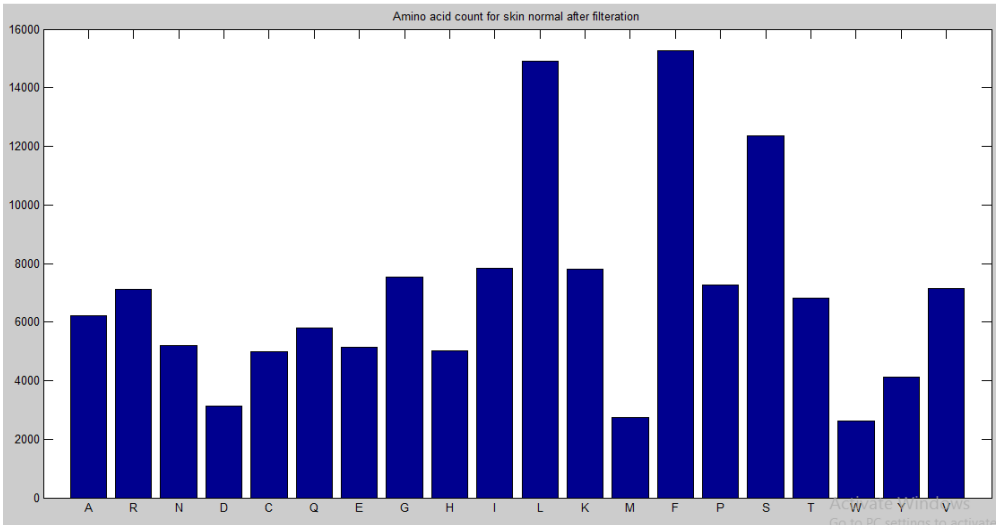


Figure 5.35 : show amino acid count of normal skin after filtration

Cancerous Skin:

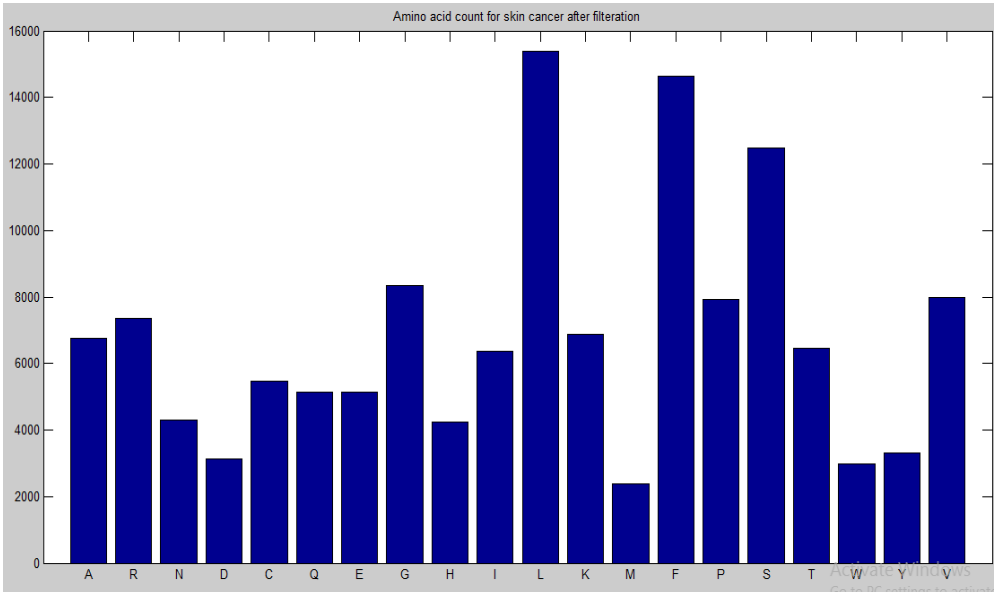


Figure 5.36 : show amino acid count of cancerous skin after filtration

5.1.5.3 ntdensity:

Normal Skin:

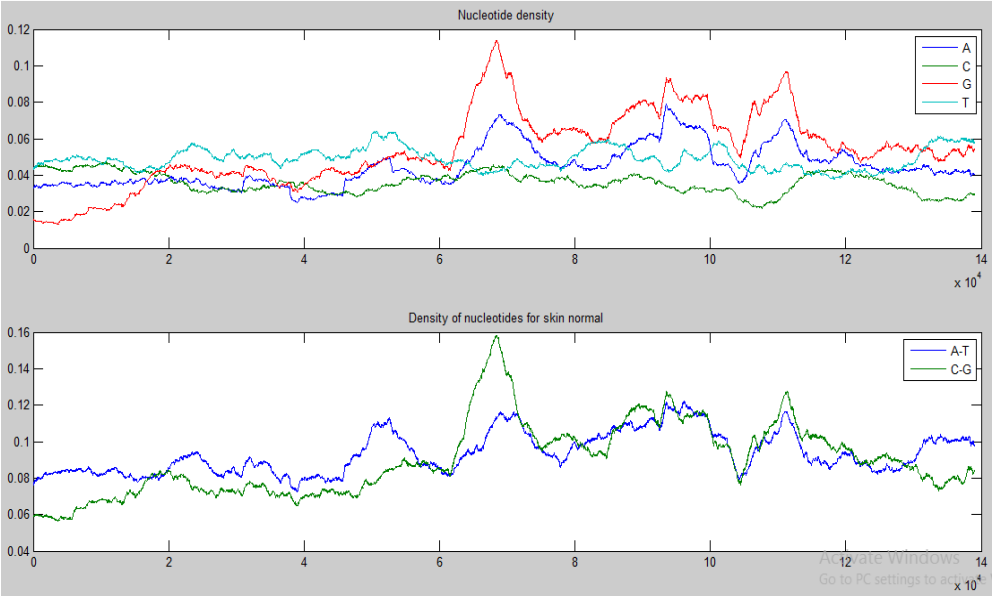


Figure 5.37 : show the density of normal skin

Cancerous Skin:

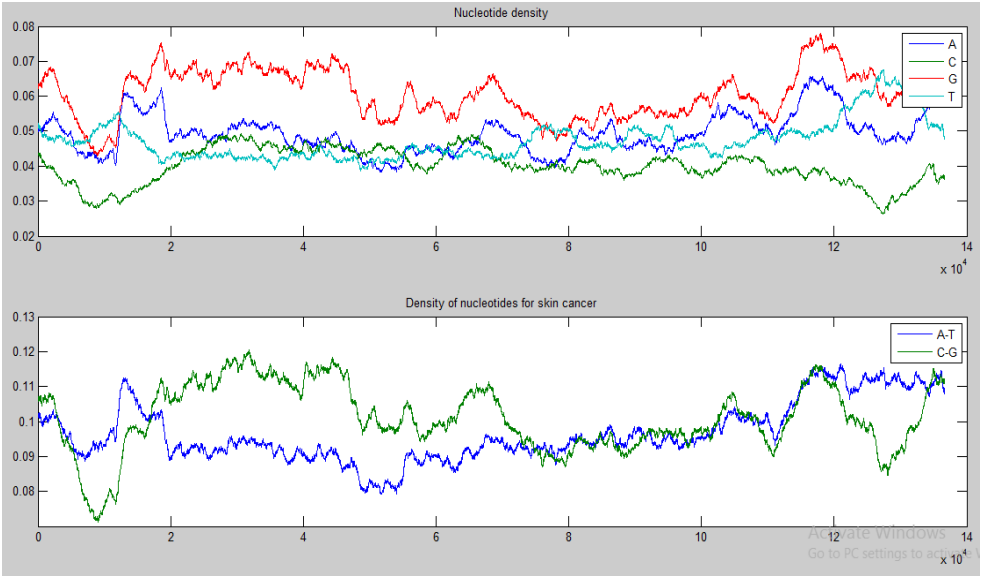


Figure 5.38 : show the density of cancerous skin

5.1.5.4 atomiccomp:

Table 5.19 : show the atomic composition of normal and cancerous skin

Normal Skin	C	H	N	O	S
	746459	1122441	191901	189869	7733
Cancerous Skin	C	H	N	O	S
	725098	1092418	186502	184945	7855

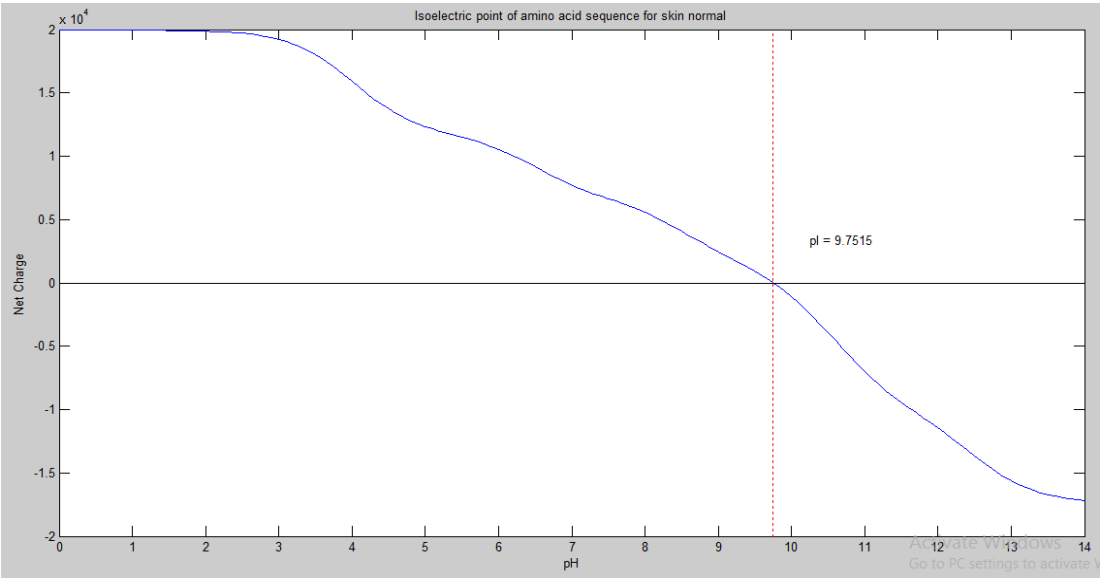
5.1.5.5 molweight:

Table 5.20 : show the molecular weight of normal and cancerous skin

Normal skin	1.6071×10^7
Cancerous skin	1.5633×10^7

5.1.5.6 isoelectric point:

Normal Skin:



(Figure 5.39) : show the isoelectric point of normal skin

Cancerous Skin:

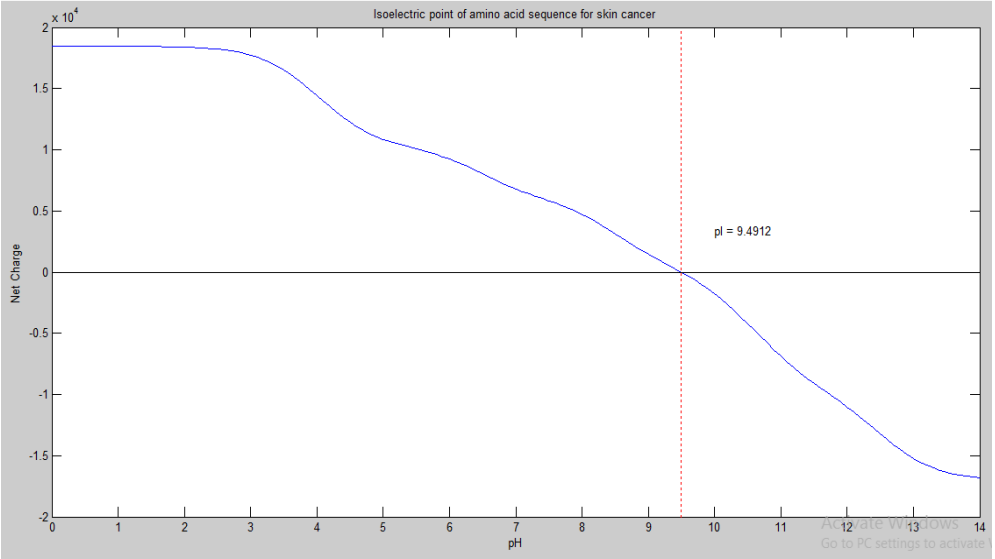


Figure 5.40 : show the isoelectric point of cancerous skin

5.1.6 Normal And Cancerous Bone:-

5.1.6.1 aacount before filtration:

Table 5.21 : show the amino acid count of normal and cancerous bone before filtration

Normal bone	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	246	*272*	*145*	*96*	*163*	*159*	*167*	*308*	*152*	*173*	*432*	*225*	*70*	*205*	*335*	*424*	*239*	*100*	*101*	*219*
Cancerous bone	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	246	*272*	*145*	*96*	*163*	*159*	*167*	*267*	*116*	*215*	*495*	*200*	*83*	*267*	*282*	*373*	*240*	*95*	*129*	*235*

Normal bone	Others	
	187	
Cancerous bone	Others	
	194	

Normal Bone:

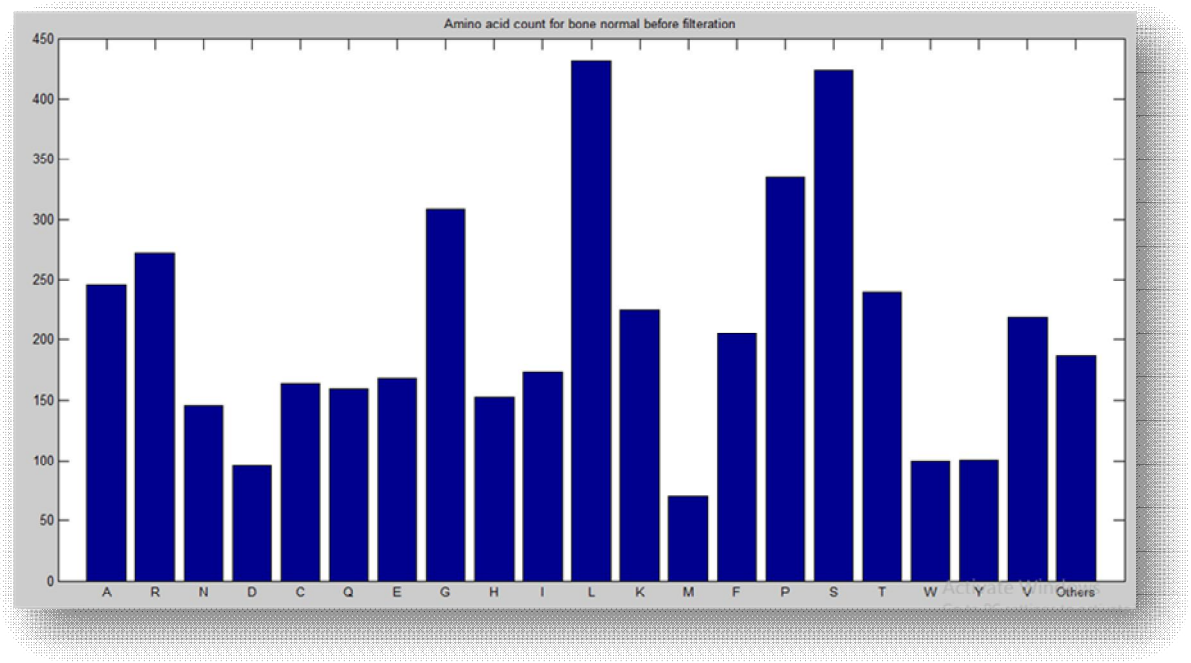


Figure 5.41 : show amino acid count of normal bone before filtration

Cancerous Bone:

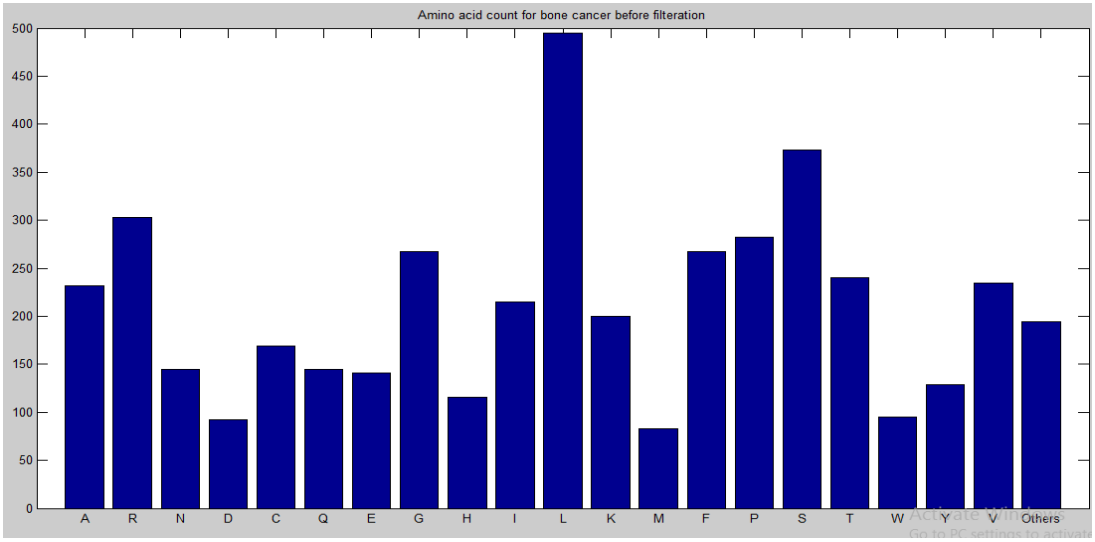


Figure 5.42 : show amino acid count of cancerous bone before filtration

5.1.6.2 aacount after filtration:

Table 5.22 : show the amino acid count of normal and cancerous bone after filtration

Normal bone	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	246	*272*	*145*	*96*	*163*	*159*	*167*	*308*	*152*	*173*	*432*	*225*	*70*	*392*	*335*	*424*	*239*	*100*	*101*	*219*
Cancerous bone	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	246	*272*	*145*	*96*	*163*	*159*	*167*	*267*	*116*	*215*	*495*	*200*	*83*	*461*	*282*	*373*	*240*	*95*	*129*	*235*

Normal Bone:

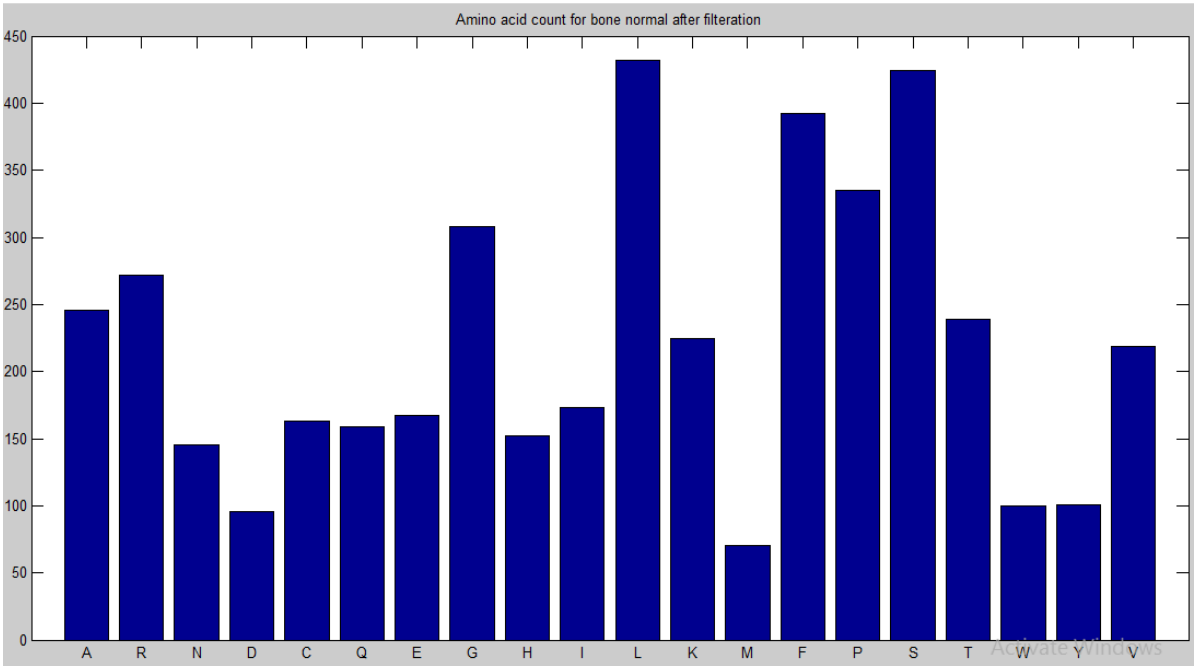


Figure 5.43 : show amino acid count of normal bone after filtration

Cancerous Bone:

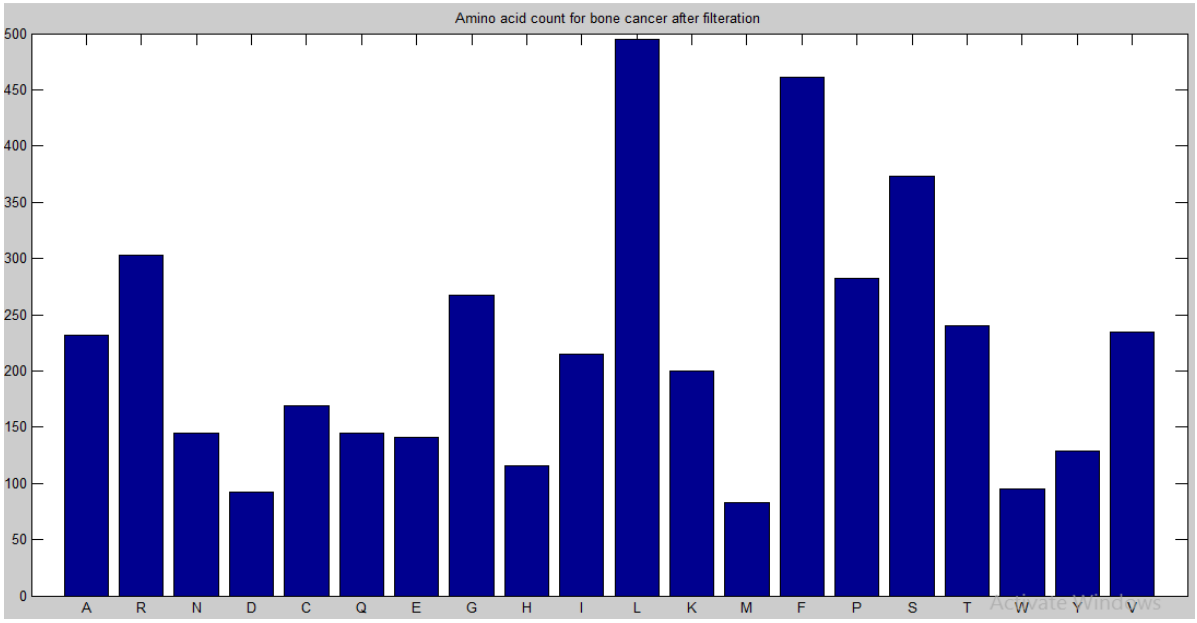
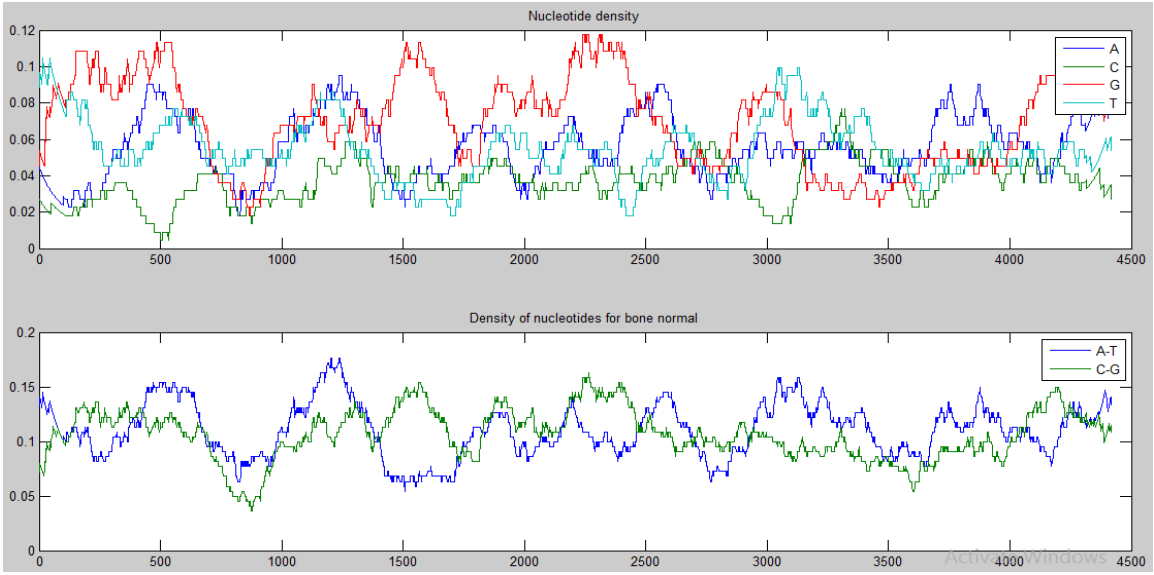


Figure 5.44 : show amino acid count of cancerous bone after filtration

5.1.6.3 ntdensity:

Normal Bone:



(Figure 5.45) : show the density of normal bone

Cancerous Bone:

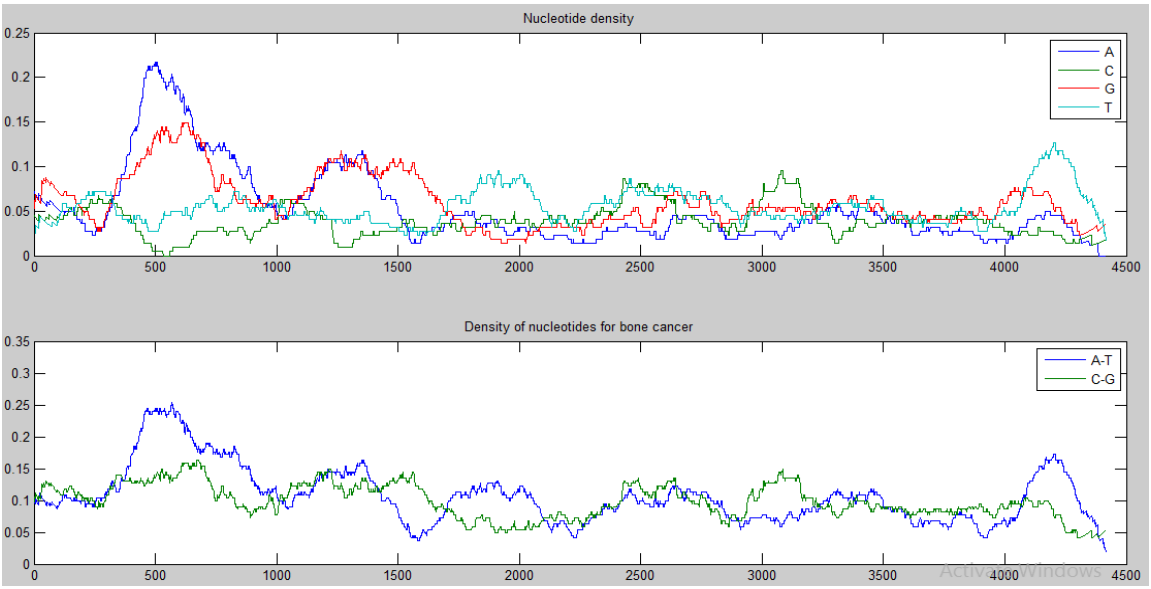


Figure 5.46 : show the density of cancerous bone

5.1.6.4 atomiccomp:

Table 5.23 : show the atomic composition of normal and cancerous bone

Normal Bone	C	H	N	O	S
	22846	*34621*	*6167*	*6013*	*233*
Cancerous Bone	C	H	N	O	S
	23523	*35584*	*6144*	*5917*	*252*

5.1.6.5 molweight:

Table 5.24 : show the molecular weight of normal and cancerous bone

Normal bone	*4.9935×10^5*
Cancerous bone	*5.072×10^5*

5.1.6.6 isoelectric point:

Normal Bone:

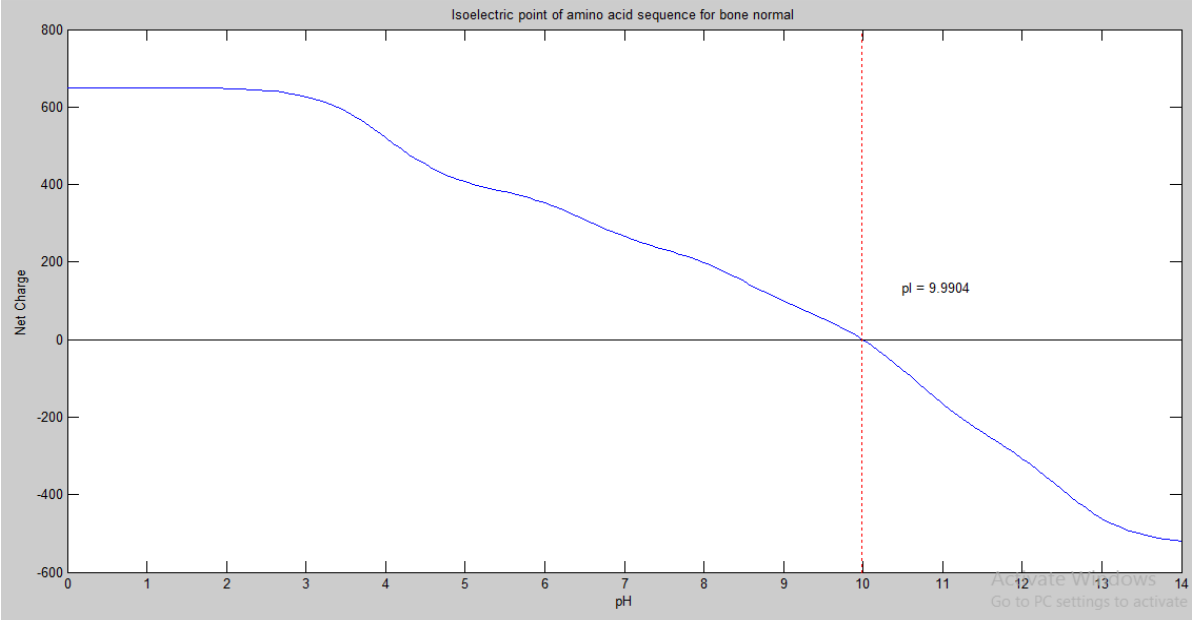


Figure 5.47 : show the isoelectric point of normal

Cancerous Bone:

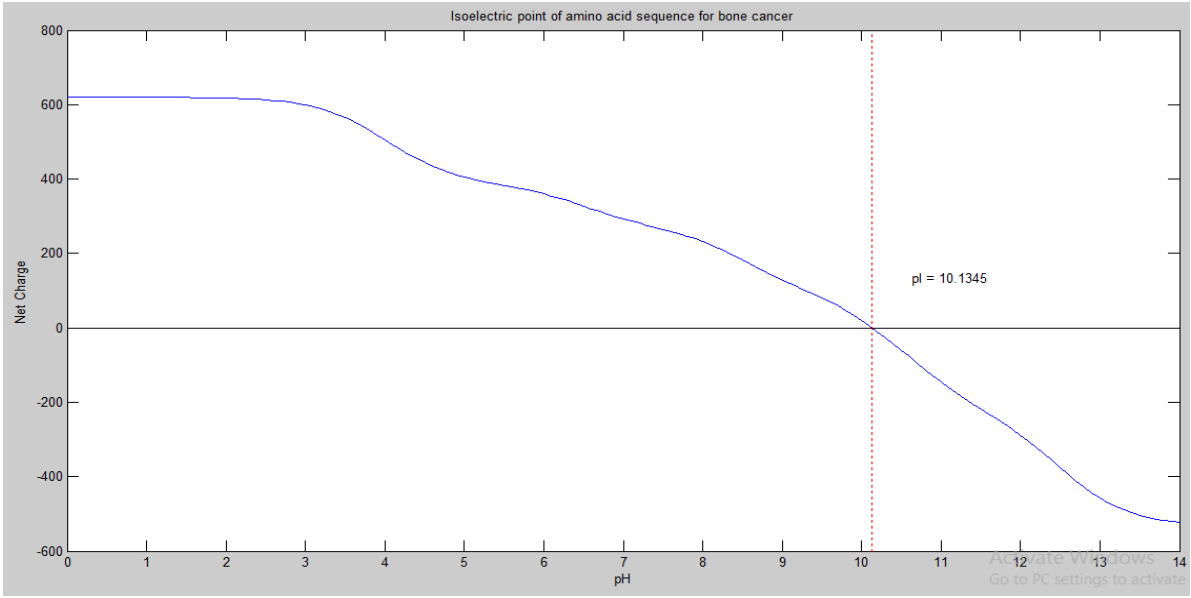


Figure 5.48 : show the isoelectric point of cancerous bone

5.1.7 Normal And Cancerous Colon:-

5.1.7.1 aacount before filtration:

Table 5.25 : show the amino acid count of normal and cancerous colon before filtration

Normal colon	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	62	*47*	*8*	*16*	*23*	*33*	*35*	*87*	*20*	*7*	*65*	*20*	*6*	*11*	*86*	*67*	*26*	*25*	*2*	*35*
Cancerous colon	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	*34*	*24*	*18*	*21*	*27*	*18*	*45*	*18*	*41*	*86*	*34*	*12*	*54*	*33*	*58*	*38*	*16*	*26*	*36*
Normal colon	Others																			
	15																			
Cancerous colon	Others																			
	28																			

Normal Colon:

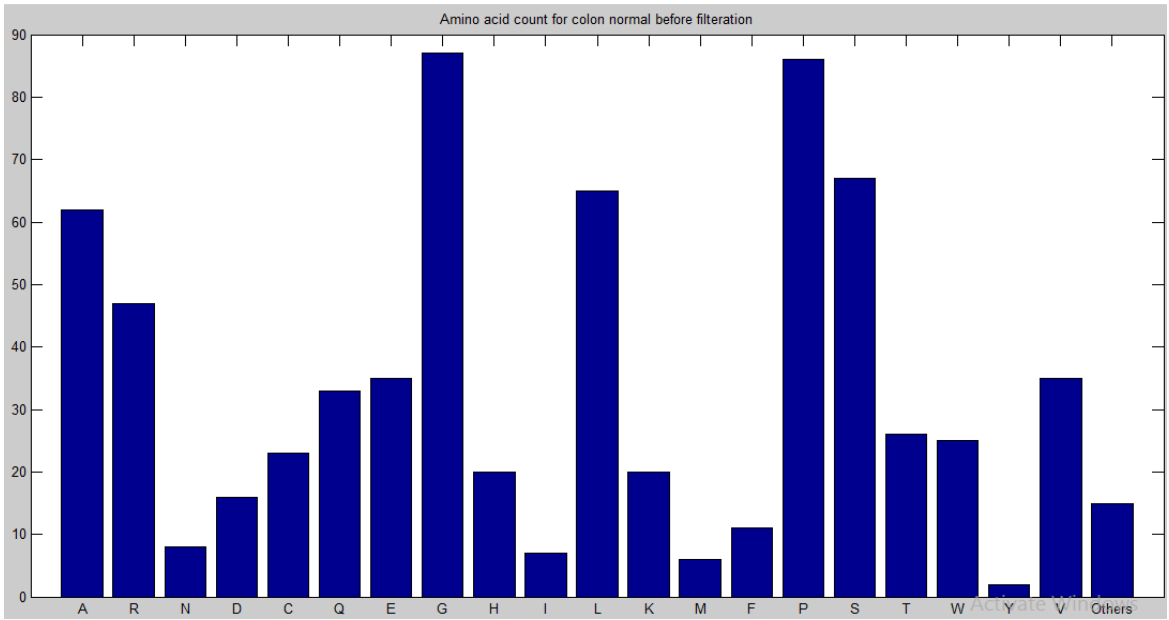


Figure 5.49 : show amino acid count of normal colon before filtration

Cancerous Colon:

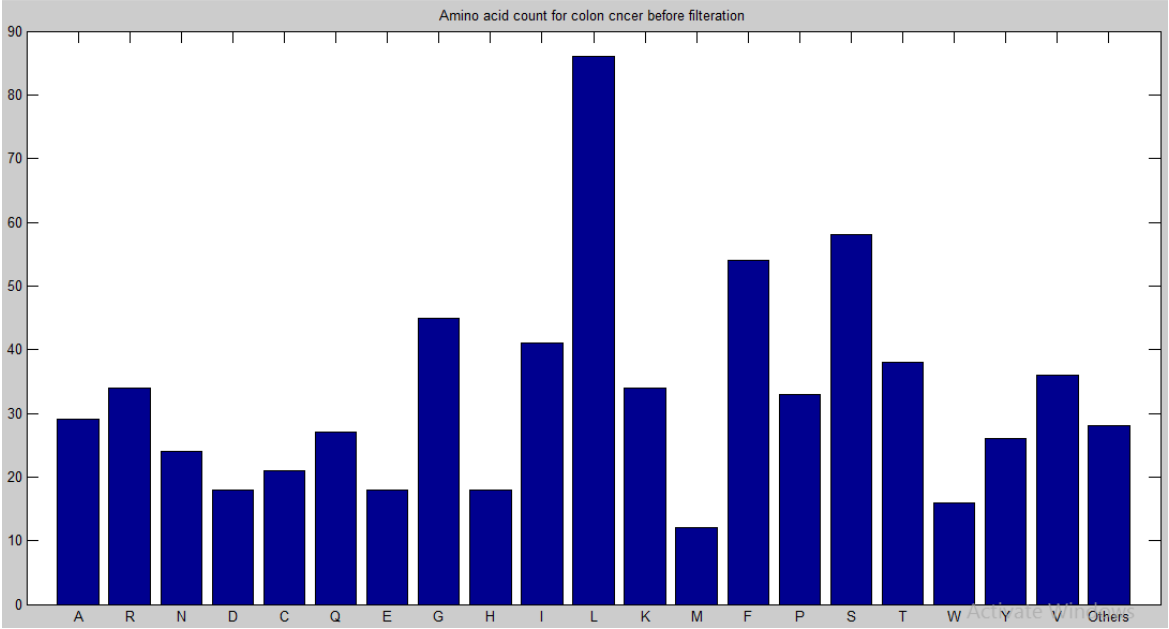


Figure 5.50 : show amino acid count of cancerous colon before filtration

5.1.7.2 aacount after filtration:

Table 5.26 : show the amino acid count of normal and cancerous colon after filtration

Normal colon	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	62	*47*	*8*	*16*	*23*	*33*	*35*	*87*	*20*	*7*	*65*	*20*	*6*	*26*	*86*	*67*	*26*	*25*	*2*	*35*
Cancerous colon	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	29	*34*	*24*	*18*	*21*	*27*	*18*	*45*	*18*	*41*	*86*	*34*	*12*	*82*	*33*	*58*	*38*	*16*	*26*	*36*

Normal Colon:

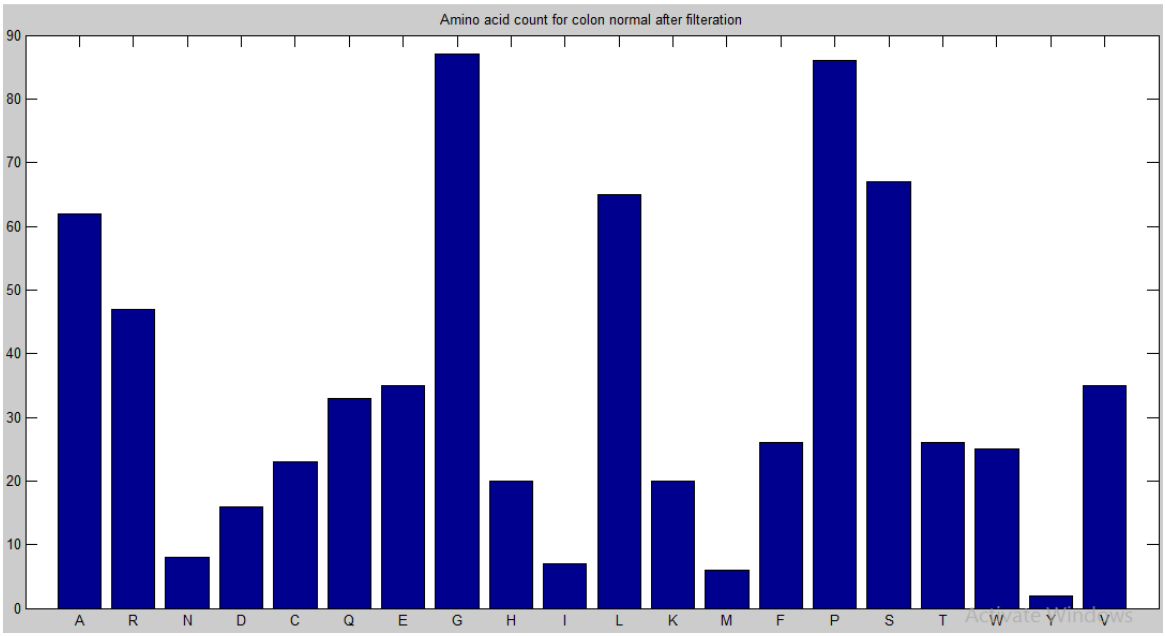


Figure 5.51 : show amino acid count of normal colon after filtration

Cancerous Colon:

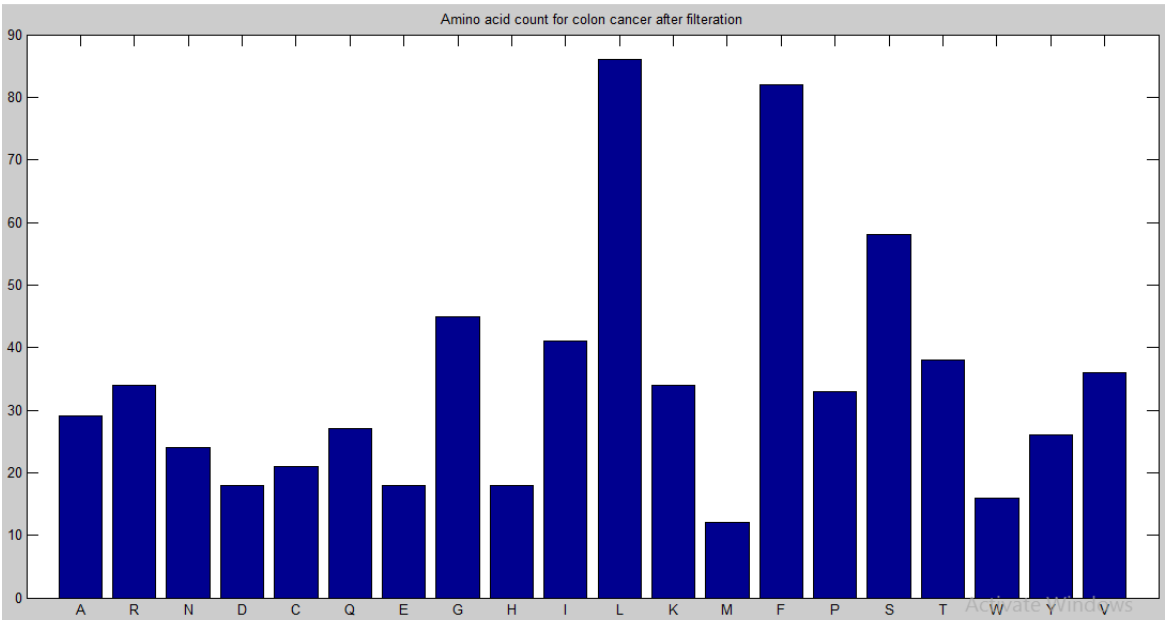


Figure 5.52 : show amino acid count of cancerous colon after filtration

5.1.7.3 ntdensity:

Normal Colon:

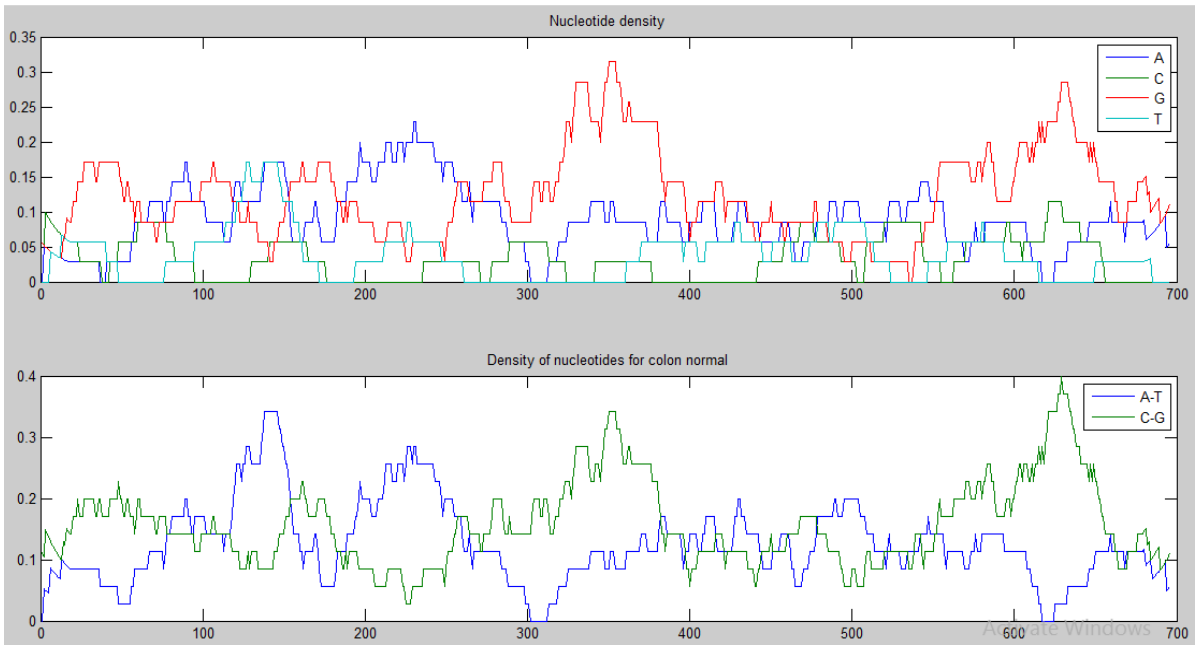


Figure 5.53 : show the density of normal colon

Cancerous Colon:

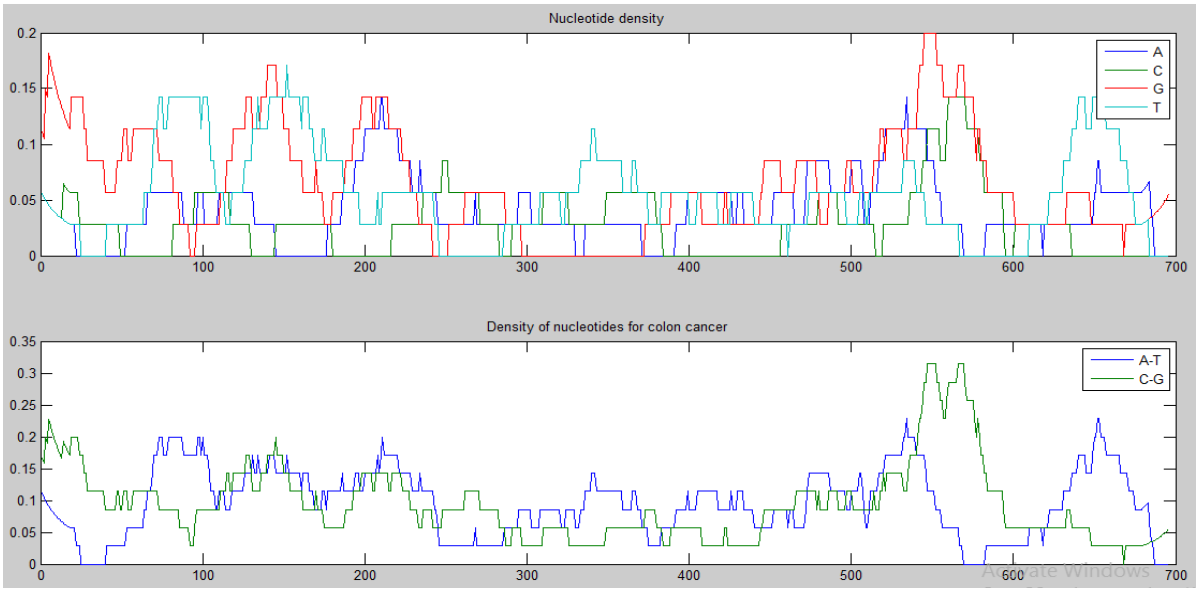


Figure 5.54 : show the density of cancerous colon

5.1.7.4 atomiccomp:

Table 5.27 : show the atomic composition of normal and cancerous colon

Normal Colon	C	H	N	O	S
	3286	*5051*	*963*	*935*	*29*
Cancerous Colon	C	H	N	O	S
	3790	*5653*	*935*	*942*	*33*

5.1.7.5 molweight:

Table 5.28 : show the molecular weight of normal and cancerous colon

Normal colon	*7.3936×10^4*
Cancerous colon	*8.0444×10^4*

5.1.7.6 isoelectric point:

Normal Colon:

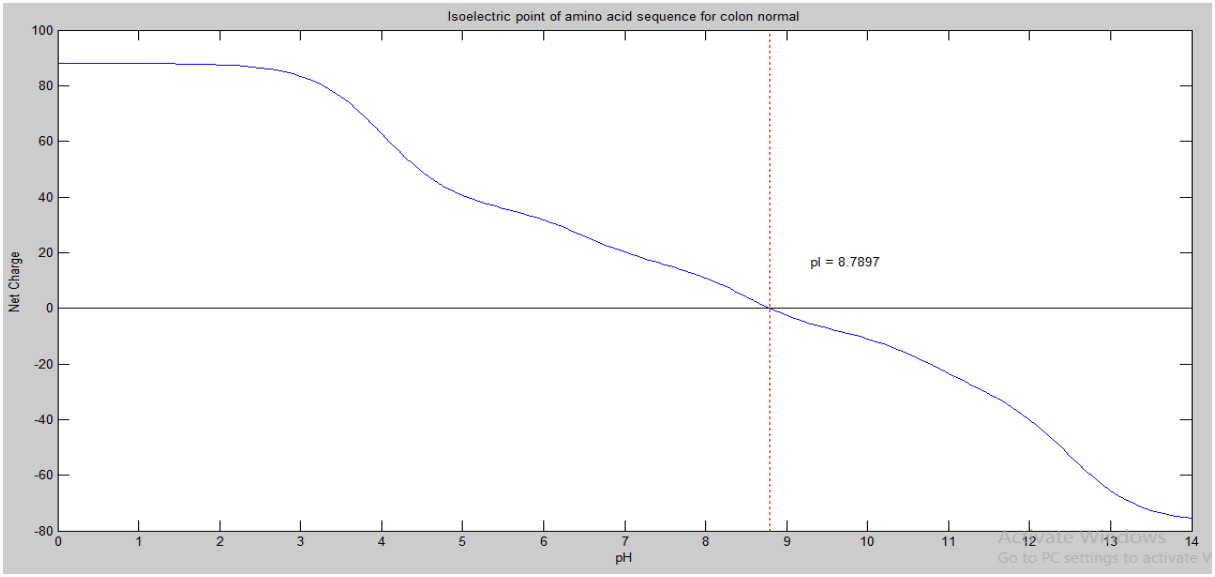


Figure 5.55 : show the isoelectric point of normal colon

Cancerous Colon:

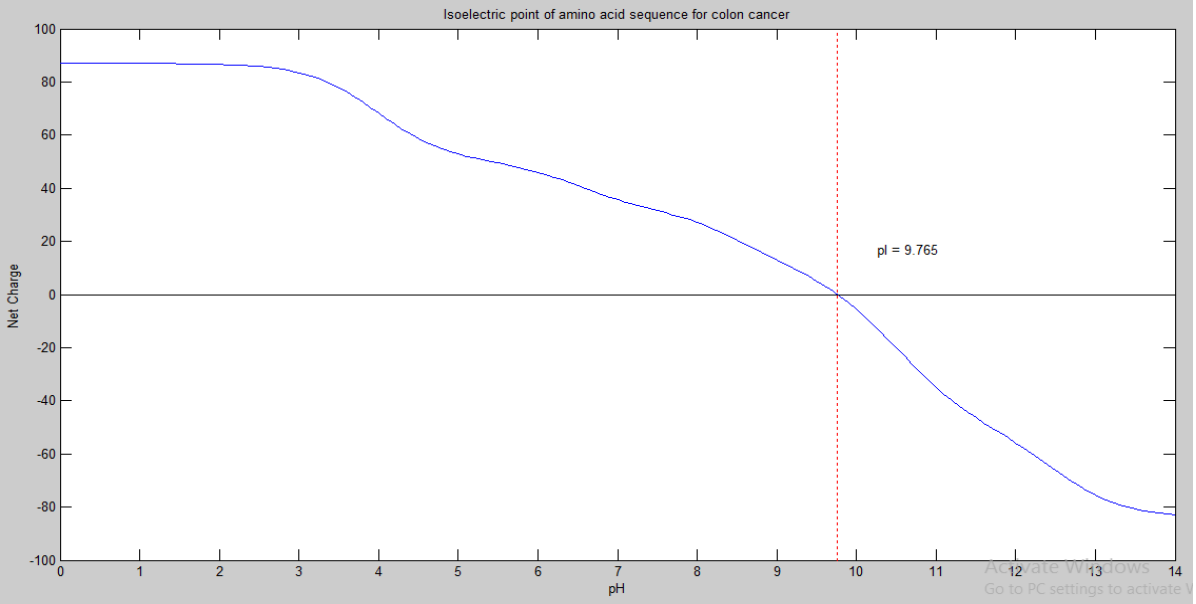


Figure 5.56 : show the isoelectric point of cancerous colon

5.2 Discussion:-

We use the seqdotplot and sequence alignment using Needleman-wunsch algorithm (nwalign) from bioinformatics tools to make compare between the normal and the cancerous organs.

5.2.1 seqdotplot:-

A dotplot is a graphical method that allows the comparison of two protein or DNA sequences and identify regions of close
similarity between them[12].

5.2.1.1 Normal And Cancerous Blood:-

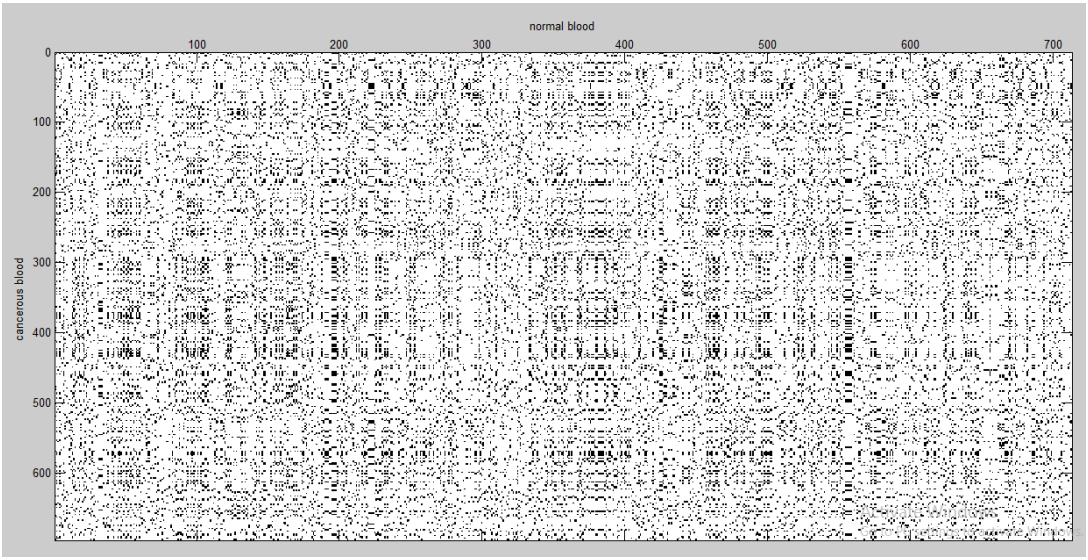


Figure 5.57 : show the dotplot of normal and cancerous blood

5.2.1.2 Normal And Cancerous Kidney:-

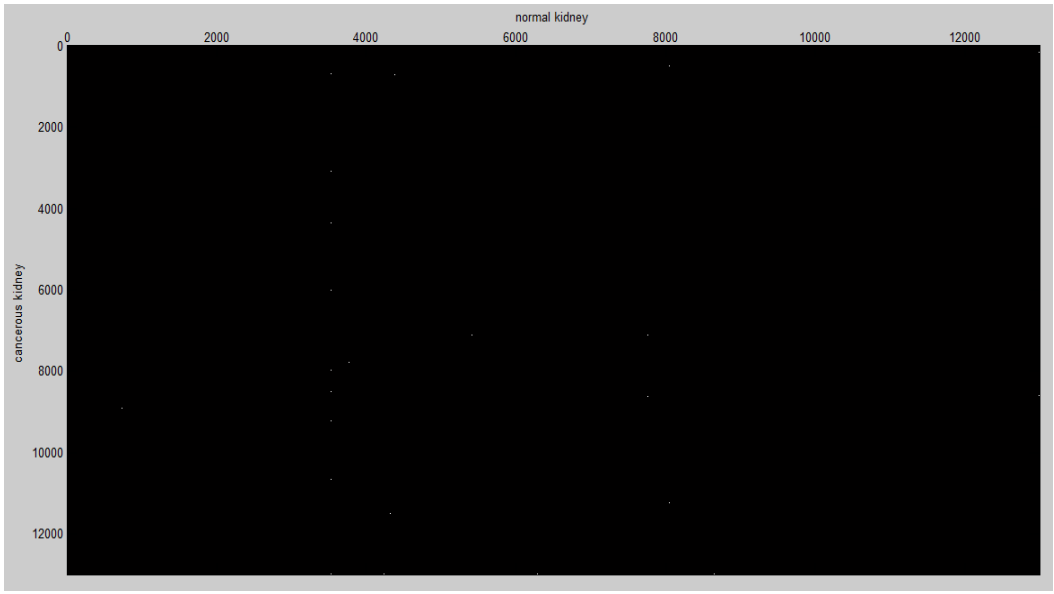


Figure 5.58 : show the dotplot of normal and cancerous kidney

5.2.1.3 Normal And Cancerous Lung:-

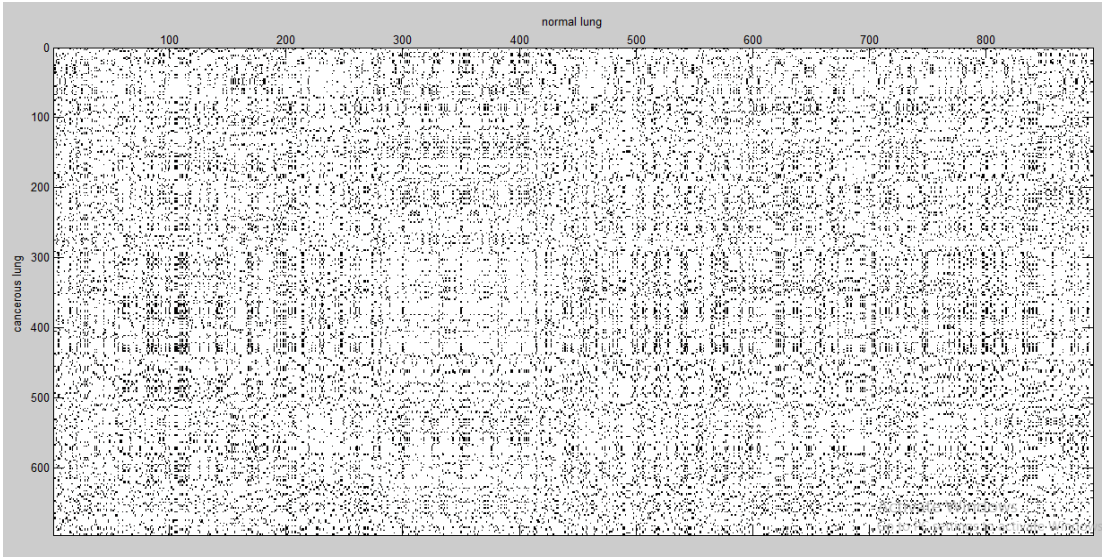


Figure 5.59 : show the dotplot of normal and cancerous lung

5.2.1.4 Normal And Cancerous Breast:-

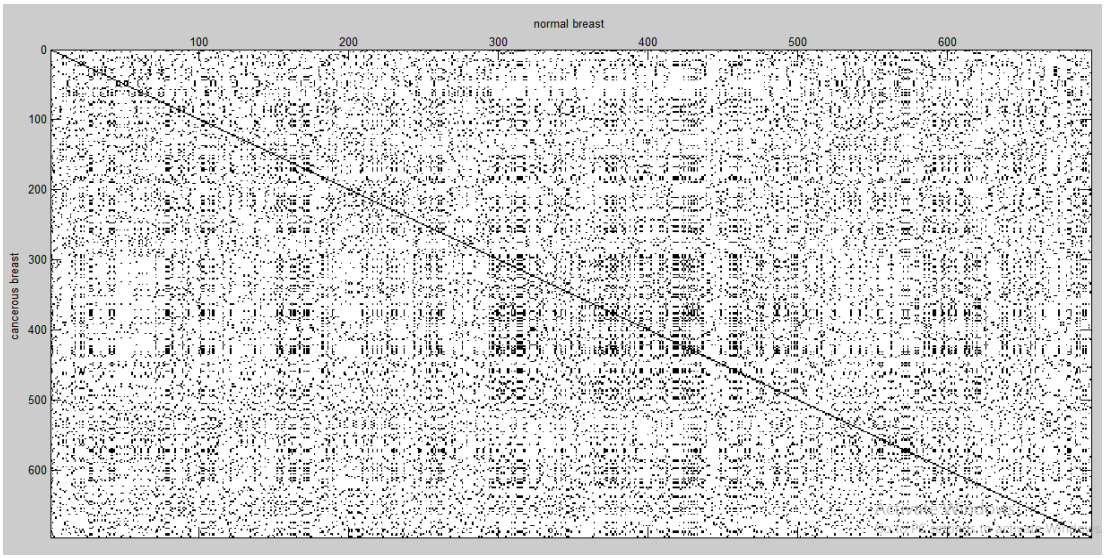


Figure 5.60 : show the dotplot of normal and cancerous breast

5.2.1.5 Normal And Cancerous Skin:-

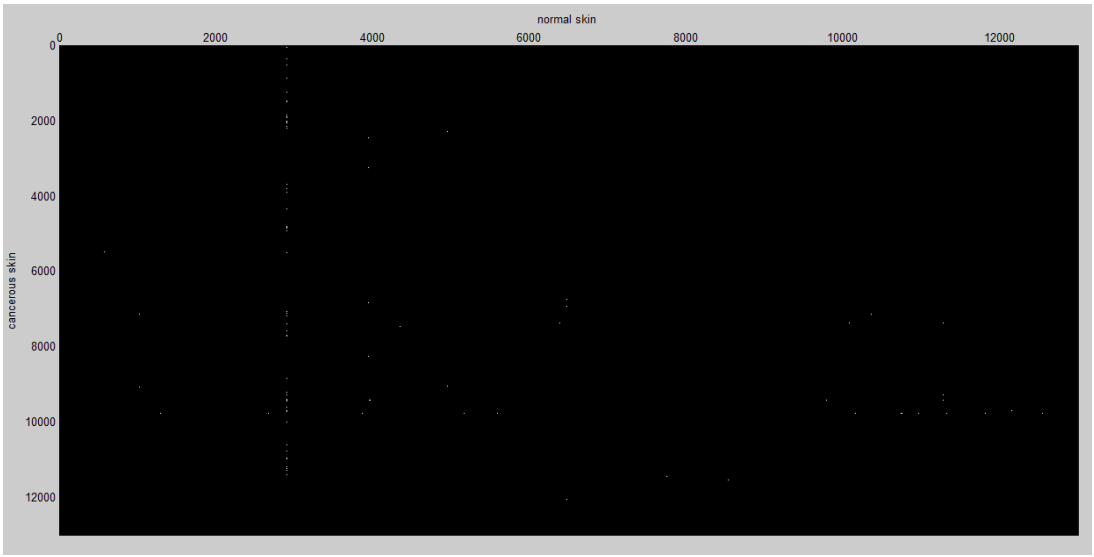


Figure 5.60 : show the dotplot of normal and cancerous skin

5.2.1.6 Normal And Cancerous Bone:-

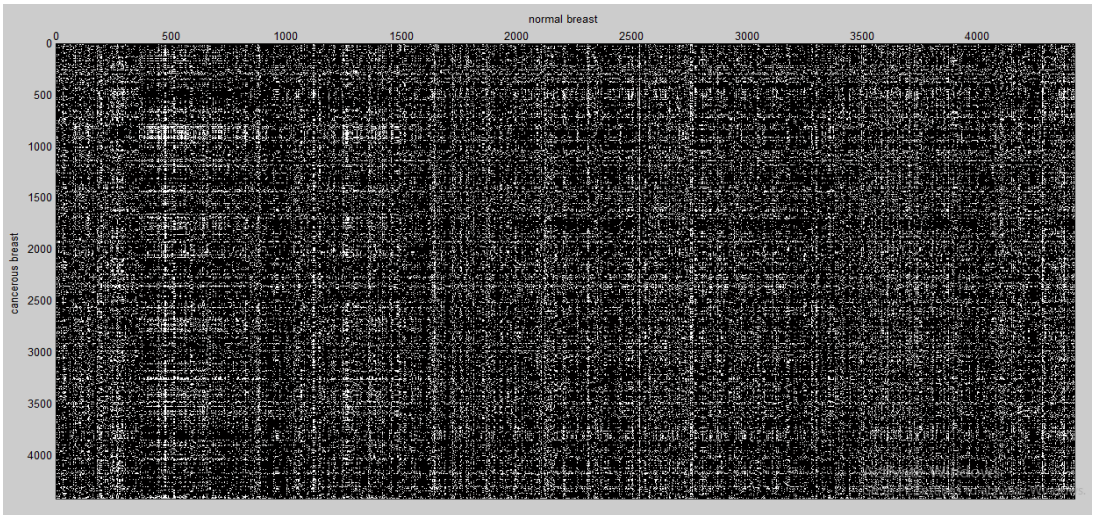


Figure 5.62 : show the dotplot of normal and cancerous bone

5.2.1.7 Normal And Cancerous Colon:-

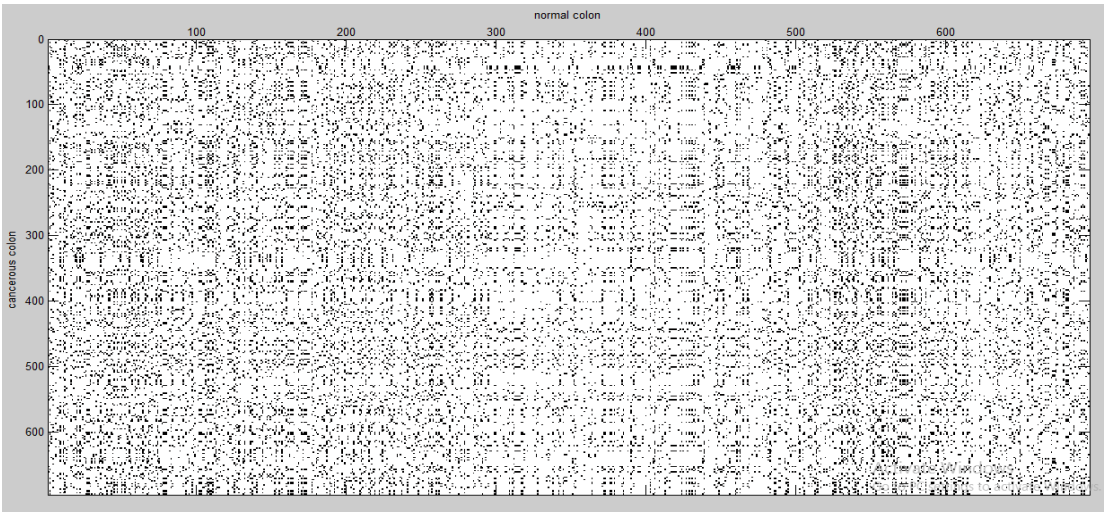


Figure 5.63 : show the dotplot of normal and cancerous colon

5.2.2 nwalign:-

Needleman-wunsch algorithm is the global alignment of full length of two sequences[13] .there is difference between identity and positives as identity refers to the exact number of matches between the two sequences and positives refer to the amino acids which are similar according to their physiochemical property[14].

5.2.2.1 Normal And Cancerous Blood:-

Identities=127/741 (17%).

Positives=350/741(47%).

5.2.2.2 Normal And Cancerous Kidney:-

Identities=2583/13684 (19%).

Positives=6469/13684 (47%).

5.2.2.3 Normal And Cancerous Lung:-

Identities=168/896 (19%).

Positives=412/896 (46%).

5.2.2.4 Normal And Cancerous Breast:-

Identities=162/871 (19%).

Positives=392/871 (45%).

5.2.2.5 Normal And Cancerous Skin:-

Identities=2415/13666 (18%).

Positives=6476/13666 (47%).

5.2.2.6 Normal And Cancerous Bone:-

Identities=837/4657 (18%).

Positives=2119/4657(46%).

5.2.2.7 Normal And Cancerous Colon:-

Identities=124/729 (17%).

Positives=311/729 (43%).

Conclusion And Recommendations

6.1 Conclusion:-

After we collect the data of normal and cancerous organs and analyzing them by sequence analysis from bioinformatics tools then implement dot plot and global alignment ; we found that there are differences between the normal and cancerous of the same organs because the average of identity between them are 18.143% and the average of differences are

81.857%.

6.2 Recommendations:-

The people who will continue the research should bring their database from their local area as possible and should encompass all organs of the body to improve the research ;use filter to remove the noise from sequences; use sequence pair distance to calculate the distance between sequence.

References

1. N.M. Luscombe , D. Greenbaum, M. Gerstein, "*What is bioinformatics? An introduction and overview" ,* Department of Molecular Biophysics and Biochemistry, Yale University New Haven, USA.
2. Alberts, Bruce ; Alexander Johnson; Julian Lewis; Martin Raff; Keith Roberts; Peter Walters "The Shape and Structure of Proteins" . Molecular Biology of the Cell; Fourth Edition. New York and London: Garland Science. ISBN 0-8153-3218-1. (2002).
3. Dietary reference intakes for energy, carbohydrate, fiber, fat, fatty acids, cholesterol, protein and amino acids.*J Am Diet Assoc.*
4. Cieplaka M. and Niewieczerza S. Hydrodynamic interactions in protein folding. Journal of Chemical Physics. 130 2009.
5. R.W. Carrell, D.A. Lomas
6. Lisewski AM, "Random Amino Acid Mutations and Protein Misfolding Lead to Shannon Limit in Sequence-Structure communication",PLoS ONE. Vol. 3 2008.
7. Pauline C. Ng, Steven Henikoff, Fred Hutchinson Cancer Research Center, Seattle,Washington 98109.
8. Ali Mansoori, PiroozMohazzabi*, Percival McCormack, et al., *World Review of Science, Technology and Sustainable Development, Vol. 4, Nos.* Pauline C. Ng, Steven Henikoff, Fred Hutchinson Cancer Research Center, Seattle,Washington 98109. *2/3, 2007.*
9. AvrilCoghlan, A Little Book of R For Bioinformatics, July 20, 2014.
10. *Bioinformatics Toolbox User's Guide ,* COPYRIGHT 2003–2006 by The MathWorks, Inc.
11. Genome.tugraz.at/../ws11_chapter03.
12. Medical –dictionary .the freedictionary.com.
13. www.angelfire.com/ill/protein/.
14. www.bioinformaticsworld.com/blast2.htm.
15. http//www.Ncbi.Nlm.Nih.Giv/Nuccore/Nc_000011.9 ?Report=Fasta &From=5246454& To=5248541&Strand=True
16. Http://www.Ncbi.Nlm.Nih.Gov/Nuccor/528476600?Report=Fasta

Appendix

The code:-

```
bn='GTGCTGGAATTACAGATGTGAGCCACAATGCCCGGCCTTATTTTCT
ACAACTTTGGTAACTTTAGCATATACCCCAAATCTGTAAGACATAATAT
TATAATTCAAATGCAACTCATGGCTTCTCTTTGTACTCTTTCTCTAGCTT
TTGAATTATTTATTCTAATACCAGTTTTAATTCTGACACAAAATCATGG
GAGTTCTAATCAAAATCCAACCTTTTATCATAAAAACTATGAAGAAATT
ATGAGTAGAATTTAAAAAGGAAAATAGGCCTATTAATTAGATTTGTCTT
TGTAGCATTTAACTCTATAATAAATAATATTTTATGCCTATGAGTCCCC
AACAAAGCCTCCAGCTTCTATTTAGATATAAACTGTAAAAGTCACTACT
GGATCCACAAGCAAGACTATGGTAAATAAATTTCTCCACCTAACCAGC
TTCTTTTACATGATGTTACATGTTTCTTTTGTTTTTTCATTTTGGCAAATA
TTGATTGTCATCTTCGTGTTTGTCTATGTCCTAAGTGCTGGGATACAGA
ATCTGAAAAGATGGACACAGGACCTGCCTTCAAGTTCACCCCCTTTTTT
TTTTTTTTTGAGATGCAGTTTTGCTCTTGTCACCCAGGCTGGAGTGTAAT
GGTGAGATCTCTGCTCACTGCAACCTCCACCTCCAGGGTTCAAGTGATT
CTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTACAGGTCCCAGCCACCA
CGCCTAGCTAATTTTTGTATTTTTAGTAGAGACAGCGTTTCATCATGTTG
GTCAGGCTGGTCTCGAACTCCTAACCTCAGGTAGTCGACCCACCTCGGC
CTCCCACAGTGCTGAGATTACAGGCATGAGCCACCACGCCCTGCTAGG
AGTTCACGCTTTAGTTGGGGAAAATATACAATAAGCAAGCCAGTTTTTA
AAATGAGAACTGCAATTAGAGTTAAATGCTACAAAGACAAACTCACAG
GAAGATGGGATGTAGAATGATAAGGCTCTCAGAATAGTAAGAGAAACT
ATTGCTTCTTACGATGTTTGTCTTTCTTTGTATCGGTGCTCAGCTGAGTC
TGCAGTGCTTCAGAGGCAGCTTTCATTTTATAAAAATCTATGATTTCTC
CTTCCAGTTGTTTTTTCTCTTCCTCGAGCTTCCTTATCTCCTCCTGTTGAA
TCATTTTAAGATGCTCGAACTTGTCCTGCAGCTGTGAAACCAATGTGCA
GTTGTGACACCAAAGCAGTGTGGCTGAACACCTAAAAGAATACGCTTT
TTTTCTGATTATCAAACAAACCCAAATCATCACAGTAGAGCACGATCTT
AATAACAATCTCAAAAACTCAGGAGTAAACACTCAGATATGGAATTTT
TCTTTTCTTTCTTTTTTCCTTTTATAAGATGGAGTCTCACTCTGTTGCCCA
GGCTGGAGTGCACTGGTGCGATCTCAGCTCACTGCAACCTCCATCTCCC
AGTTCAAGTGATTCTCCTGCCTCAGCCTCTTGAGTAGCTGGGACTATAG
GCATGCACCACCACTACAGCGTGTGCCACCACACCTGGCTAATTTTTGT
ATTTTTAGTAGAGATGGGGTTTTGCCATGATGGCCAGGCCGGTCTCGAA
CTCCTGACCTCAGGTGATCCTCCCGCTTTGGCCTCCCAAAGACTTTTTTT
TTTTTTTTTAATATAGAGACAAGTTCTCAGTACGTTGCCCAGGCTGGTCT
CAAACTCCTGAGCTCAAGTGATCCTCCCACCTCAGCTTCCCAAAGTGCT
```

```
GGGACTGACTGGATGCAGTGGCTCATGCTTGTAAACTCAGCACTTTGGG
AGGCCAAGGTGGGAGGATCGCTTGAGCCCAGGAGTTCAAGACCAGACT
GGGTGATATAACACAATAGTCAACTTCAACAGGAGAGAGAATCTGTAA
ACTTGAATATAGATCTTCCGAAATTATCCAGTCAGAGGACAGAGAAAA
AAGAATAAAAGAGAGAAAAGAAGGCTGGGTGTGGTGGCTCAAGCCTG
TAATCCCAACACTTTGGGAGGCCGAGGCAGGCAGATTAAGAGGTCAGG
AGTTCAAGACCAGCCTGTCCAACATGACAAAGCCCCATCTCTACTAAA
AATACAAAAATTAGCCGGGTGTGGTGGCACACACCT';
bn1=nt2aa(bn);
bn8=aacount(bn1,'chart','bar');
title('Amino acid count for blood normal before filteration');
figure;
bn2=strrep(bn1,'*','F');
ntdensity(bn2);
title('Density of nucleotides for blood normal');
figure;
bn3=aacount(bn1,'chart','bar');
title('Amino acid count for blood normal after filteration');
bn4=atomiccomp(bn2);
bn5=molweight(bn2);
bn6=isoelectric(bn2,'chart',true)
title('Isoelectric point of amino acid sequence for blood normal');
figure;
bc='TTTTTAGTAGCAATTTGTACTGATGGTATGGGGCCAAGAGATATAT
CTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACTCCTAAGCCAGTGCC
AGAAGAGCCAAGGACAGGTACGGCTGTCATCACTTAGACCTCACCCTG
TGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGG
GCAGGAGCCAGGGCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGC
TTACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGAC
ACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGT
GGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGT
TGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGG
CATGTGGAGACAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTC
TCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGGTCTACCCT
TGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATG
CTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCG
GTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTT
TGCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGA
GAACTTCAGGGTGAGTCTATGGGACGCTTGATGTTTTCTTTCCCCTTCTT
TTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTAC
```

```
AGTTTAGAATGGGAAACAGACGAATGATTGCATCAGTGTGGAAGTCTC
AGGATCGTTTTAGTTTCTTTTATTTGCTGTTCATAACAATTGTTTTCTTTT
GTTTAATTCTTGCTTTCTTTTTTTTTCTTCTCCGCAATTTTTACTATTATA
CTTAATGCCTTAACATTGTGTATAACAAAAGGAAATATCTCTGAGATAC
ATTAAGTAACTTAAAAAAAAACTTTACACAGTCTGCCTAGTACATTACT
ATTTGGAATATATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTA
TTTTCTTTTATTTTTAATTGATACATAATCATTATACATATTTATGGGTT
AAAGTGTAATGTTTTAATATGTGTACACATATTGACCAAATCAGGGTAA
TTTTGCATTTGTAATTTTAAAAAATGCTTTCTTCTTTTAATATACTTTTTT
GTTTATCTTATTTCTAATACTTTCCCTAATCTCTTTCTTTCAGGGCAATA
ATGATACAATGTATCATGCCTCTTTGCACCATTCTAAAGAATAACAGTG
ATAATTTCTGGGTTAAGGCAATAGCAATATCTCTGCATATAAATATTTC
TGCATATAAATTGTAACTGATGTAAGAGGTTTCATATTGCTAATAGCAG
CTACAATCCAGCTACCATTCTGCTTTTATTTTATGGTTGGGATAAGGCT
GGATTATTCTGAGTCCAAGCTAGGCCCTTTTGCTAATCATGTTCATACC
TCTTATCTTCCTCCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTG
GCCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATC
AGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAGTATCACT
AAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCT
AAGTCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCT
GGATTCTGCCTAATAAAAAACATTTATTTTCATTGCAATGATGTATTTA
AATTATTTCTGAATATTTTACTAAAAAGGGAATGTGGGAGGTCAGTGCA
TTTAAAACATAAAGAAATGAAGAGCTAGTTCAAACCTTGGGAAAATAC
ACTATATCTTAAACTCCATGAAAGAAGGTGAGGCTGCAAACAGCTAAT
GCACATTGGCAACAGCCCCTGATGCATATGCCTTATTCATCCCTCAGAA
AAGGATTCAAGTAGAGGCTTGATTTGGAGGTTAAA';
bc1=nt2aa(bc);
bc8=aacount(bc1,'chart','bar');
title('Amino acid count for blood cancer before filteration');
figure;
bc2=strrep(bc1,'*','F');
ntdensity(bc1);
title('Density of nucleotides for blood cancer');
figure;
bc3=aacount(bc2,'chart','bar');
title('Amino acid count for blood cancer after filteration');
bc4=atomiccomp(bc2);
bc5=molweight(bc2);
bc6=isoelectric(bc2,'chart',true)
title('Isoelectric point of amino acid sequence for blood cancer');
```

```
figure;
seqdotplot(bn2,bc2);
xlabel('normal blood');
ylabel('cancer blood ');
[score,alignment]=nwalign(bn2,bc2);
showalignment(alignment);
```