# BIOPHYSICAL CHARACTERIZATION OF PROTEINS IN SOLUTION AND HUMAN FLUIDS FOR CANCER DIAGNOSIS APPLICATIONS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SHERIF ABBAS MOUSA ABBAS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
BIOLOGY

MAY 2016

Approval of the thesis:

**BIOPHYSICAL CHARACTERIZATION OF PROTEINS IN SOLUTION AND HUMAN FLUIDS FOR CANCER DIAGNOSIS APPLICATIONS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY**

submitted by **SHERIF ABBAS MOUSA ABBAS** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Biological Sciences Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Orhan Adalı _____
Head of Department, **Biological Sciences**

Prof. Dr. Feride Severcan _____
Supervisor, **Biological Sciences Dept., METU**

Prof. Dr. Mete Severcan _____
Co- Supervisor, **Electric and Electronic Eng. Dept., METU**

**Examining Committee Members:**

Assoc. Prof. Dr. Çağdaş D. Son _____
Biological Sciences Dept., METU

Prof. Dr. Feride Severcan _____
Biological Sciences Dept., METU

Assoc. Prof. Dr. Hakan Altan _____
Physics Dept., METU

Prof. Dr. Deniz Köksal _____
Department of Chest Diseases, Hacettepe university, Ankara.

Asst. Prof. Dr. Filiz Korkmaz _____
Physics Division, Atılım University

**Date:** 25/05/2016

**I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.**

Name, Last name: Sherif Abbas Mousa ABBAS

Signature:

# ABSTRACT

## BIOPHYSICAL CHARACTERIZATION OF PROTEINS IN SOLUTION AND HUMAN FLUIDS FOR CANCER DIAGNOSIS APPLICATIONS USING FOURIER TRANSFORM INFRARED SPECTROSCOPY

Abbas, Sherif Abbas Mousa
Ph.D., Department of Biological Sciences
Supervisor: Prof. Dr. Feride Severcan
Co-Supervisor: Prof. Dr. Mete Severcan
May 2016, 89 pages

Proteins play very important roles in cells regulation and structure. Understanding of proteins structures greatly help in understanding of the mechanism of action of these proteins. The optical spectroscopic techniques such as Fourier transform infrared (FTIR) and circular dichroism (CD) spectroscopy can be used to study proteins in its native environment without complicated sample preparations that is required for a high resolution technique such as x-ray. In order to calculate the protein secondary structure from amide I band in FTIR spectra, different methods can be used such as curve fitting and deconvolution. However, these techniques have some disadvantages due to the noise and dependence on the operator. In this study, a protein FTIR dataset of known proteins structures was produced using FTIR transmission mode. This database was used as a training set for an artificial neural networks (ANNs). Because of the limited number of our proteins in the dataset (35 proteins), a leave-one-out approach for training and testing our neural networks was performed. To achieve generalized ANNs in a limited number proteins dataset, discrete wavelet transform (DWT) was successfully used as a data reduction technique for amide I spectra. The results of

ANNs predictions showed 96.88%, 93.92% and 95.98% success in β-sheets, α-helix and other structures respectively. In the second part of this thesis, Human Apo- and Holo-transferren structure and their thermal denaturation behavior in dilute and artificial crowded environment was studied using FTIR spectroscopy. Dextran 70 and Ficoll 70 as a "molecular crowder" did not have a major effect on the secondary structure of transferrin as deduced from the analysis of the amide I band. However, it does alter the tertiary structure since significant differences in hydrogen-deuterium exchange was seen by monitoring the intensity of the residual amide II band as a function of time. The study of transferrin thermal denaturation using 2D-IR showed two different aggregated secondary structures patterns in dilute and in an artificial crowded environment. Finally, the proteins secondary structure of human pleural fluid accumulated due to malignant pleural mesothelioma (MPM), lung cancer (LC) and benign transudate (BT) was studied using attenuated total reflectance FTIR spectroscopy. Wavelet analysis was performed to extract the amide I spectral features. The extracted features were used as an input for the previously trained artificial neural network in the first part of this thesis. The ANNs results indicated significant differences in protein content of BT, LC and MPM pleural fluid samples. The chemometric results of the plural fluid proteins secondary structure lead to an accurate, cost effective method for the diagnosis of MPM from lung cancer and benign transudate with 88% sensitivity and 100% specificity.

**Keywords:** Proteins, molecular crowding, malignant pleural mesothelioma, ATR-FTIR spectroscopy, 2D-IR correlation, ANNs.

# ÖZ

Fourier Dönüşüm Kızılötesi Spektroskopisi Kullanarak Proteinlerin Çözelti İçindeki Formları ile Kanser Teşhisi Uygulamalarında İnsan Sıvılarınının Biyofiziksel Karakterizasyonu

Abbas, Sherif Abbas Mousa
Doktora, Biyolojik Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Feride Severcan
Ortak Tez Yöneticisi: Prof. Dr. Mete Severcan

Mayıs 2016, 89 Sayfa

Proteinler hücre içinde yapısal ve düzenleyici olarak önemli rol alırlar. Protein yapılarını anlamak, onların etki mekanizmalarını anlamada oldukça yardım sağlar. Protein çalışmalarında, FTIR ve CD gibi optik spektroskopik teknikler, X-ışınları gibi yüksek çözünürlüklü tekniklerin gerek duyduğu karışık örnek hazırlama yöntemleri olmaksızın kullanılabilir. Amid I bandından protein ikincil yapılarını tahmin etmek için eğri benzeştirme ve dekonvolüsyon gibi farklı teknikler kullanılabilir. Fakat bu teknikler uzman kullanıcılara gereksinim duyar ve sonuçlar kullanıcıya bağlıdır. Bu çalışmada, FTIR geçirgenlik modu kullanılarak bilinen proteinlere ait bir FTIR protein veri seti oluşturuldu. Bu veri seti, yapay sinir ağı için eğitici set olarak kullanıldı. Proteinlerin sayısı sınırlı (35 protein) olduğu için, sinir ağlarının eğitimi ve testi için leave-one-out yaklaşımı kullanıldı. Sınırlı sayıda proteinler veri kümesi içinde genelleştirilmiş YSA elde etmek için, kesikli dalgacık dönüşümü (DWT) amid I bandı için veri indirgeme tekniği olarak başarıyla kullanılmıştır. YSA'ların tahmin sonuçları,

sırasıyla tabaka, heliks ve diğer yapılarda 96.88%, 93.92% ve 95.98% başarı göstermiştir. Bu tezin ikinci bölümünde, insan Apo ve Holo-transferrin yapısı ve termal denatürasyon davranışları seyreltik ve yapay kalabalık ortamda FTIR spektroskopisi kullanılarak incelenmiştir. Amid I bandı analizlerinden çıkarıldığı üzere moleküler kalabalıklık olarak Dekstran 70 ve Ficoll 70 transferrinin ikincil yapısı üzerinde önemli bir etkisi olmamıştır. Bununla birlikte, üçüncül yapıda değişikliğe yol açmıştır, çünkü zamanın fonksiyonu olarak artık amid II bandın şiddetinin izlenmesi ile hidrojen-döteryum dönüşümünde önemli farklılıklar görülmüştür. 2D-IR korelasyonu kullanılarak transferrinin termal denatürasyon çalışması, seyreltik ve yapay kalabalık bir ortamda iki farklı toplanmış ikincil yapı desenleri gösterdi. Son olarak, malign plevral mezotelyoma (MPM), akciğer kanseri (LC) ve benign transüda (BT) hastalıkları için insan plevral sıvısı proteinlerinin ikincil yapısı ATR-FTIR spektroskopisi kullanılarak incelenmiştir. Dalgacık analizi, amid I spektral özelliklerini çıkarmak için uygulandı. Çıkarılan özellikler önceden eğitilmiş yapay sinir ağı için bir girdi olarak kullanılmıştır (bu tezin 1. bölümü). YSA sonuçları, BT, LC ve MPM plevral sıvı örneklerinin protein içeriğinde önemli farklılıklar göstermiştir. Plevral sıvı proteinleri ikincil yapı kemometrik sonuçları, MPM'in akciğer kanseri ve benign transüdadan ayrımı için % 88 duyarlılık ve % 100 özgüllük ile doğru, uygun maliyetli bir yönteme sebep olmuştur.

**Anahtar Kelimeler:** Proteinler, moleküler kalabalıklık, malign plevral mezotelyoma, ATR-FTIR spectroskopi, 2D-IR korelasyonu, yapay sinir ağı.

*Dedicated to my beloved wife and Son,*

*Radwa OSMAN and Seifeldin ABBAS*

## ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest appreciation to my supervisor Prof. Dr. Feride SEVERCAN for her endless support, patience motivation and immense knowledge. Without her inspiring guidance and encouragement, I couldn't have accomplished this thesis successfully. I am also grateful to my co-supervisor Prof. Dr. Mete Severcan for her valuable suggestions and guidance. I would also like to express my deep thanks to my PhD thesis committee members Prof. Dr. Deniz Köksal, Assoc. Prof. Dr. Çağdaş D. Son, Asst. Prof. Dr. Filiz Korkmaz, and Assoc. Prof. Dr. Hakan Altan.

My deepest thanks to Prof. Dr. Ibrahim Hassan, Prof. Dr. Sami Hindawy, Prof. Dr. Abdelsattar Salam, Prof. Dr. Elsayed Mahamoud, Prof. Dr. Mona Salah from Biophysics group, Physics Department Ain Shams University, Cairo, Egypt. I would like to express my special thanks to Dr. Parvez Haris for his precious and endless help, concern and support during my study at DMU. My deepest thanks for Dr. Nihal Simsek, Dr. Ozlem bozkurk, Dr. Pinar Demir. I would like to extend my thanks to my friends in Lab-146, Dilek YONAR, Rafig GURBANO, Nuri ERGEN, Seher Gok and Fatma KÜÇÜK for their great help, support and considerable advices.

I would like to give my deepest thanks to my wife Radwa Osman, my son Seifeldin Abbas and my best friend Ahmed Elagamy for their endless patience, support and love during my all academic life. Words failed me to express my appreciation to my mother Salma Hamad and my father Abbas Elghazawy for their love, constant support, care and understanding throughout my life. I wouldn't be here without their encouragement.

I would like to thanks the supporters of my PhD study, Egyptian cultural affairs and missions sector (http://www.mohe-casm.edu.eg/), Residency for Turks abroad and related communities (YTB), Scientific and Technical Research Council of the Turkish Republic (TÜBİTAK).

## TABLE OF CONTENTS

# LIST OF TABLES

**TABLES**

## LIST OF FIGURES

**FIGURES**

## LIST OF ABBREVIATIONS

| | |
|---|---|
| FTIR | Fourier transform infrared |
| ATR | Attenuated Total Reflectance |
| HCA | Hierarchical cluster analysis |
| PCA | Principal component analysis |
| 2D-IR | Two dimensions infrared |
| CD | Circular dichroism |
| NMR | Nuclear magnetic resonance |
| TF | Transferen protein |
| HTF | Holo-transfeerin |
| ATF | Apo-transfferin |
| DWT | Discrete wavelet transform |

# CHAPTER 1

# INTRODUCTION

Proteins are large, complex biomolecules which play important roles in the body. For example, they are required for the structure, function, and regulation of the body's tissues and organs. Amino acids are the building unit for all types of proteins. There are 20 different types of amino acids which can be combined in one chain or more to make proteins. Amino acids are linked to each other by peptide linkages to form primary structure of the protein. This primary structure produces the secondary, tertiary and quaternary structure which will be discussed in details in this chapter. The functions of protein are highly structure dependent; this means that any small variation of the protein structure can strongly affect its function. Proteins can be classified according to their functions in the body to be: Antibody, Enzyme, Messenger, Structural component, Transport/storage. On the other hand, the structure of proteins secreted from cells and tissues can be affected in case of diseases. Because of this the study of proteins structure excreted from the cells can help in diagnosis of many diseases.

Proteins structures can be investigated using different techniques such as X-ray, NMR, Raman and FTIR. Each technique has its advantages and disadvantages, however only X-ray and NMR can predict the complete proteins 3D structures. X-ray technique requires a highly pure protein crystal which is not possible for all protein specially membrane proteins. NMR is limited to low molecular weight proteins. Because of those limitations in high resolution techniques (X-ray and NMR) the low resolution techniques such as (Raman and FTIR spectroscopy) can help in estimation of proteins structures variations in vivo.

## 1.1 Proteins structure

### 1.1.1 Proteins primary structure

Amino acids (the building unit of protein) usually contains amine (-NH2), carboxylic acid (-COOH) functional groups and a specific side chain group (Figure 1). Two or more amino acids linked by peptide bonds will form a polypeptide (Figure 2). Polypeptides or proteins consist of a backbone and side chains. The backbone contains the amide nitrogen, the alpha carbon and the carbonyl carbon that are contributed by each amino acid unit. The side chains contain the "R" groups which is differs according to the amino acid type.

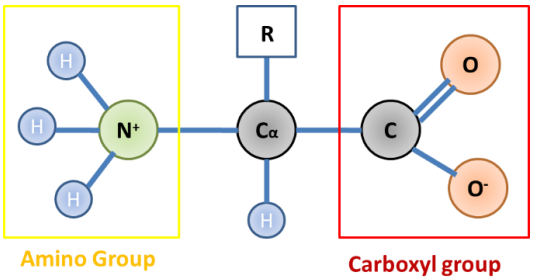


Figure 1: Amino Acid typical structure

A polypeptide is considered as a protein when it is folded into a well-defined 3-dimensional structure. The 3d structure is required for protein in order to do its functions.

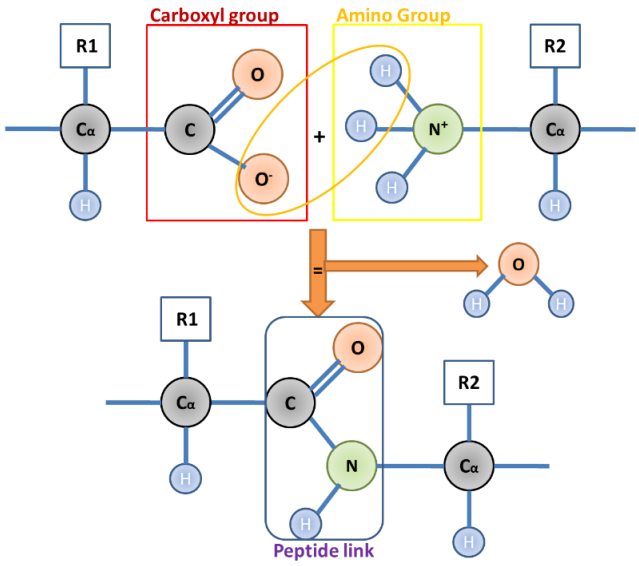


Figure 2: Formation of peptide from two amino acids.

### 1.1.2 Proteins secondary structure:

In 1950s the American chemist Linus Pauling discovered the two most important secondary structures (α-helix and β-sheet). Pauling recognized that the bond angles and planar configuration are preserved because of the folding of the peptide bond. This also keeps atoms at fixed since they repel of each other through van der Waal's interactions.

#### 1.1.2.1 α-helix

In α-helix the H-bond are regularly spaced along the polypeptide chain. For H-bond the amide hydrogen is H-bond donors and the carbonyl oxygen are the acceptors (Figure 3a).

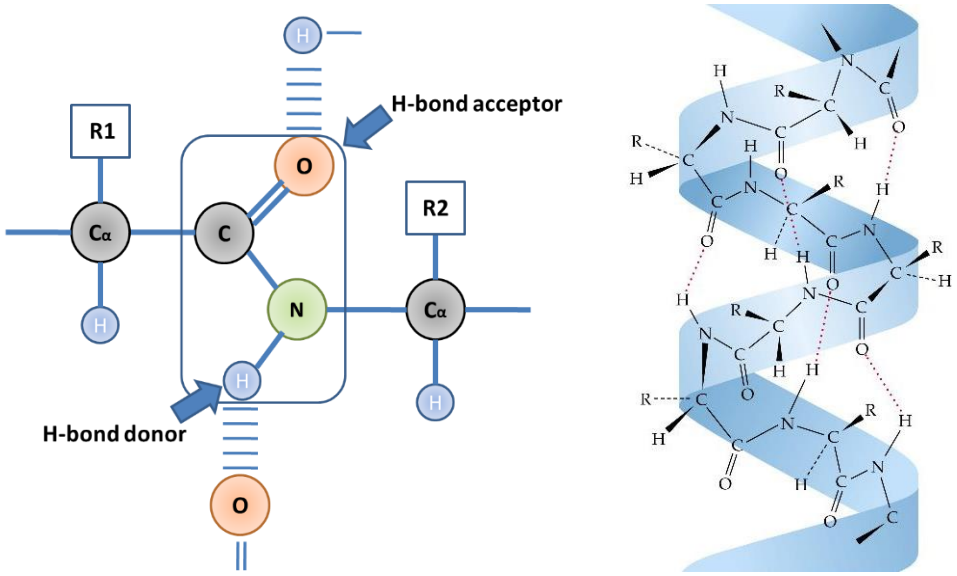


Figure 3: a) Hydrogen bonds in peptide link B) α-helix structure*

* B) Adopted from http://itech.dickinson.edu/chemistry/?p=381#more-381

As the α-helix turns, the carbonyl oxygen of the peptide bond point upwards toward the downward-facing amide protons, making the hydrogen bond. The R groups of the amino acids point outwards from the α-helix (Figure 3b).

#### 1.1.2.2 β-sheet:

In β-sheets protein secondary structure the backbone residues forms H-bonding. This structure occurs when a part of a polypeptide chain overlap on another one and form a row of hydrogen bonds. β-sheets can be either parallel, which means both chains point to the same direction. Or antiparallel, which means both chains point to opposite direction when represented by the amino- to carboxyl- terminus (Figure 4).

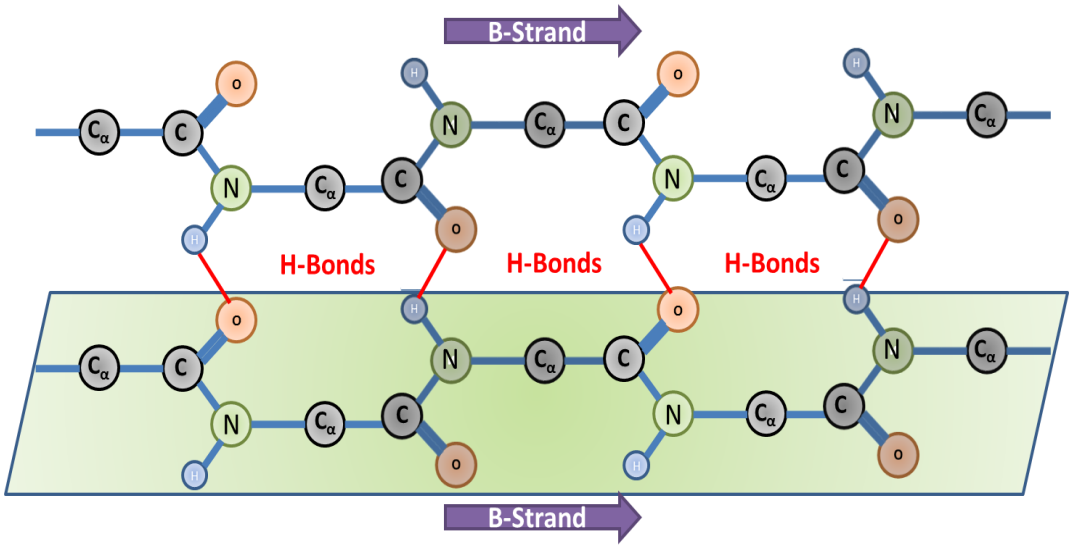


Figure 4: Parallel B-Sheet structure

### 1.1.3 Proteins tertiary and quaternary structure

The three-dimensional structure of a protein is known as Tertiary structure. Different sorts of bonds are included for keeping up the tertiary structure of proteins. As an example of these bonds there are hydrogen bonds, dipolar bonds, electrostatic bonds, and disulfide bonds. The disulfide bond can be considered as the strongest one from them. This disulfide bond consists of covalent bond between two cysteine amino acid side chains. The Hydrogen bond can be formed between any two appropriate atoms. The attraction of oppositely charged groups present in two amino acid chains can form electrostatic bonds. Also the interaction between electron clouds can form the Van der Waals bonds (Whitaker 1994, Platis et al. 2006. Damodaran et al. 2007). For the Quaternary structure; it can be consist of more than one polypeptide chain to form the protein final structure (Damodaran et al. 2007) as shown in figure 5.

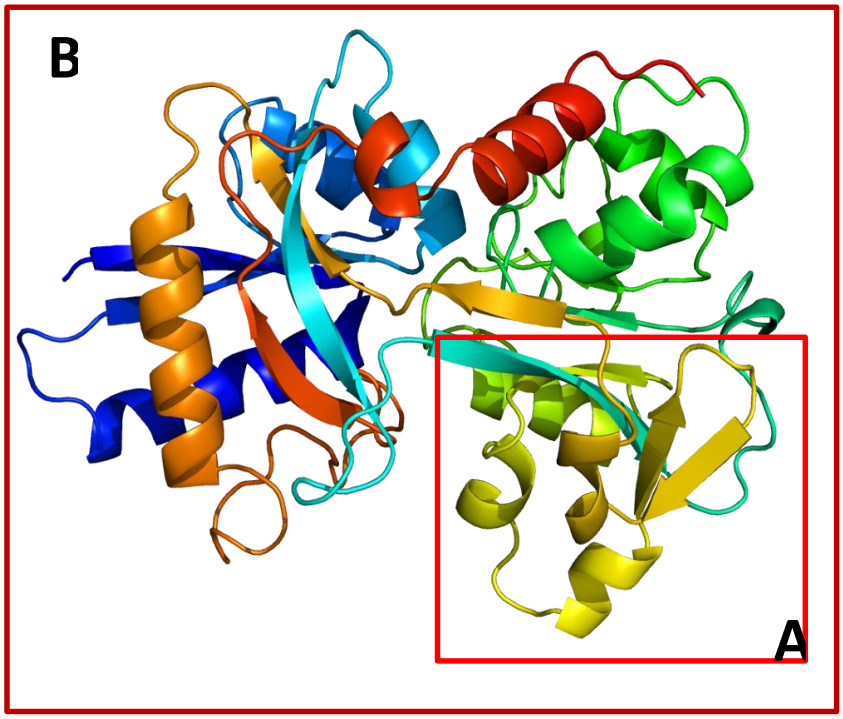


Figure 5: Structure of the transferrin protein based on PyMOL rendering of PDB 1a8e. A) Tertiary and B) Quaternary structure

## 1.2 Fourier Transform Infrared Spectroscopy (FTIR)

### 1.2.1 The electromagnetic spectrum:

Electromagnetic radiation consists of an oscillating electrical and magnetic fields perpendicular to each other and perpendicular to their traveling direction as shown in figure (Figure 6). Electromagnetic radiation can be classified according to wavelength, wavenumber, frequency or energy. (Table 1) shows the main classification of Electromagnetic radiation.

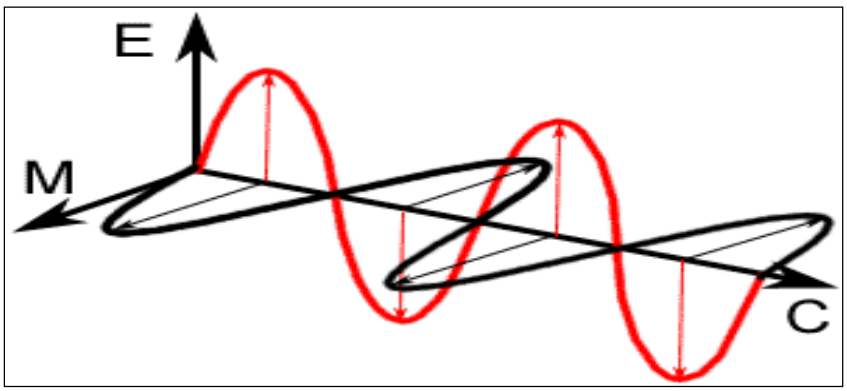


Figure 6: Electrical field (E) and magnetic field (M) (taken from Stuart, 1997 )

Table 1: Summary of the electromagnetic radiation.

| Region | Wavelength (centimeters) | Wavenumber ($cm^{-1}$) | Frequency (Hz) | Energy (eV) |
|---|---|---|---|---|
| **Radio** | > 10 | < 0. 1 | $< 3 \times 10^{9}$ | $< 10^{-5}$ |
| **Microwave** | 10 - 0.01 | 0.1 – 100 | $3x10^{9}$ - $3x10^{12}$ | $10^{-5}$ - 0.01 |
| **Infrared** | 0.01 - $7 \times 10^{-5}$ | 100 – 14285 | $3x10^{12}$- $4.3x10^{14}$ | 0.01 - 2 |
| **Visible** | $7 \times 10^{-5}$ - $4x10^{-5}$ | 14285 - 25000 | $4.3x10^{14}$- $7.5x10^{14}$ | 2 - 3 |
| **Ultraviolet** | $4 \times 10^{-5}$ - $10^{-7}$ | 25000 - $10^{7}$ | $7.5x10^{14}$- $3x10^{17}$ | 3 - $10^{3}$ |
| **X-Rays** | $10^{-7}$ - $10^{-9}$ | $10^{7}$ – $10^{9}$ | $3x10^{17}$ - $3x10^{19}$ | $10^{3}$ - $10^{5}$ |
| **Gamma Rays** | $< 10^{-9}$ | $>10^{9}$ | $> 3 \times 10^{19}$ | $> 10^{5}$ |

Figure 7 shows graphical representation of electromagnetic spectrum ranges from the shorter wavelengths to the longer wavelengths.

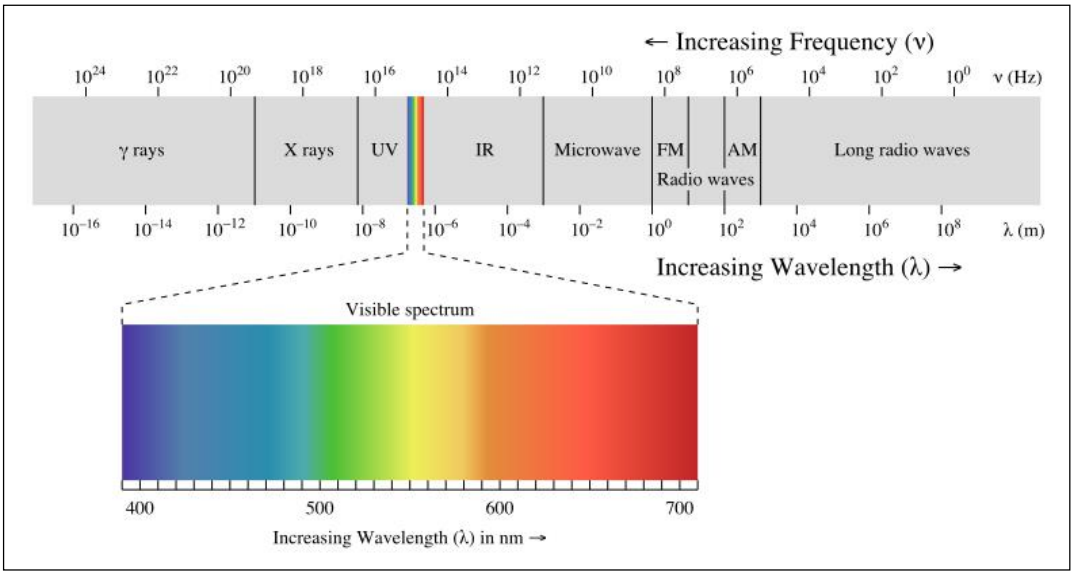


Figure 7: The electromagnetic spectrum (*created by Philip Ronan* )

### 1.2.2 Energy levels:

The molecules can be *excited* by absorbing the energy of light. According to quantum mechanics, the excited molecule absorbs a discrete amount (quanta) of energy to be excited. This amount of energy (quanta) is equal to the difference between the energy level of excited and ground state (lowest energy level).

The main molecular energy levels are Electronic transition, Vibrational, Rotational and transitional (Figure 8). The Energy of a molecule ($E_{total}$) can be calculated using by:

$$E_{total} = E_{transition} + E_{rotation} + E_{vibration} + E_{electronic} + E_{electron\ spin\ orientation} + E_{nuclear\ spin\ orientation}$$

*Each E in the equation represents the appropriate energy as indicated by its subscript.*

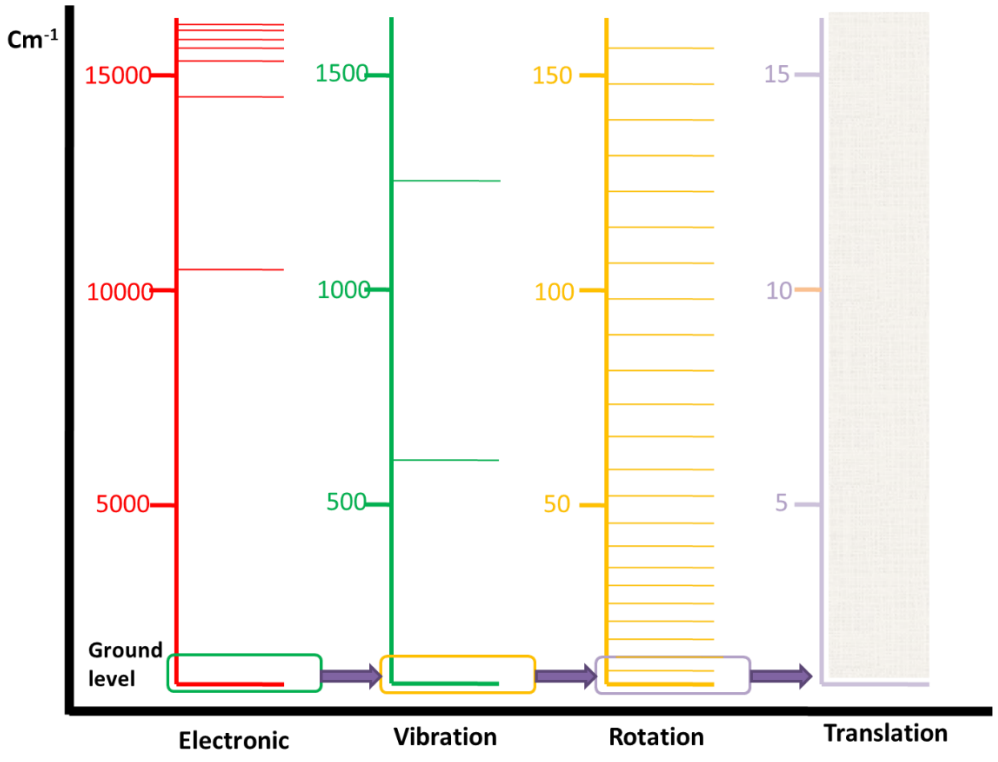


Figure 8: Molecular energy levels

The separations between respective energy levels of $E_{translation}$, $E_{electron\ spin\ orientation}$ and $E_{nuclear\ spin\ orientation}$ are very small because of this their contributions are

usually negligible. The energy levels corresponding to $E_{rotation}$, $E_{vibration}$ and $E_{electronic}$ are associated with the microwave, infrared and ultraviolet-visible region of the electromagnetic spectrum, respectively (Campbell and Dwek, 1984). Molecules absorb energy if the intermolecular distance of two or more atoms changes. Stretching and bending are the two types of oscillations correspond to the normal modes of vibration in atoms. Stretching oscillations can be symmetric or antisymmetric rhythmical movement along the bond. The bending oscillations happen when the bond angle between two atoms occurs. Also, the bending oscillations occure when a group of atoms change relative to the remainder atoms in the molecule. These bending motions can be scissoring, wagging, rocking, and twisting as shown in figure (**Figure 9**) (Marcelli et al., 2012).

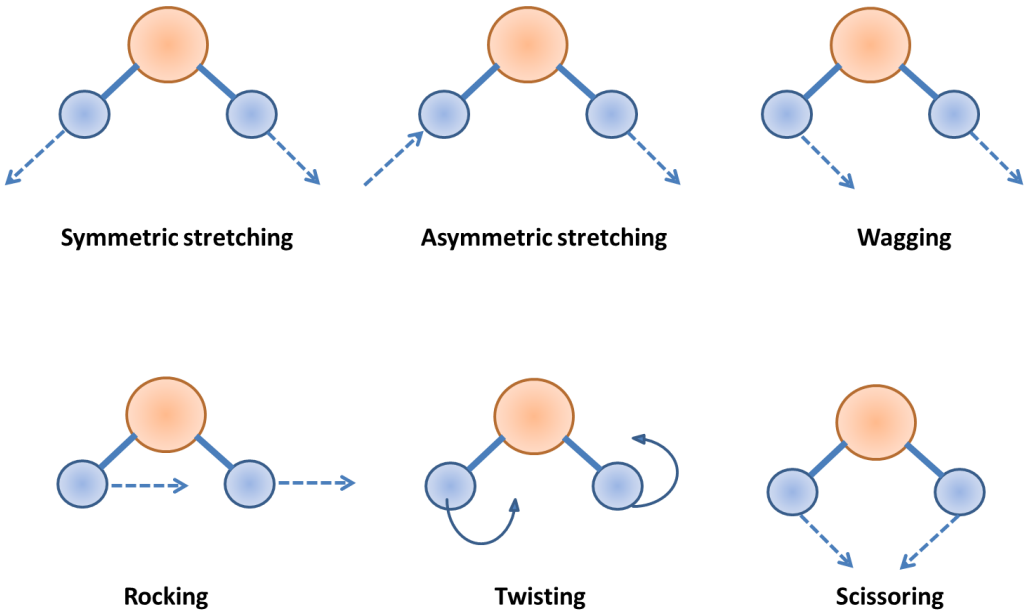


Figure 9 The vibrational modes associated to a molecular dipole moment change detectable in an IR absorption spectrum .

### 1.2.3 Infrared spectroscopy

Spectroscopy defined as the study of interaction of electromagnetic radiation with matter. In spectroscopy the sample irradiate with electromagnetic radiation to measurement the scattering, absorption, or emission in order to get some parameters such as peak height, peak wavenumber or peak area. Interpretation of these parameters can leads to useful information about the studied sample. Infrared spectroscopy (IR

spectroscopy) is concerned with the transition between vibrational energy levels. The electromagnetic spectrum of Infrared can be divided into near-, mid- and far- infrared according to their energy. The lowest energy far-infrared (400–10cm$^{-1}$) is adjacent to microwave spectrum region and may be used for rotational spectroscopy. The mid-infrared (4000–400 cm$^{-1}$) can be used to study the fundamental vibrations and associated rotational-vibrational structure. The higher-energy near-IR (14000–4000 cm$^{-1}$) can be used to study the overtone or harmonic vibrations.

#### 1.2.3.1 Michelson Interferometer and FTIR Technology

Michelson Interferometer is used in FTIR spectroscopy machines. Using of Michelson Interferometer the IR beam split into two optical beams. Then one of these two beams directed to a settled mirror and the other beam directed to the moving mirror. The moving mirror moves some millimeters to and away from the beam splitter. After that the two IR beams superimposed on each other at the beam splitter after their reflections. When the two IR beams interact with each other the interferogram is formed. In interferogram, all the frequencies are simultaneously measured at the same time. This gives the advantage of highly fast measurements of FTIR. In order to convert the interferogram to intensity-versus-frequency spectrum, a mathematical function known as Fourier transformation is used. This Fourier transformation conversion can be performed using a computer by a plot of the IR intensity versus wavenumber (cm$^{-1}$). This plot can be used for further analysis (Figure 11).

A background spectrum is collected before the collection of sample spectrum for relative scaling of absorption intensity (Figure 10). The background spectrum is compared the sample spectra to calculate the percent transmittance. After these calculations the sample spectra can be produced.

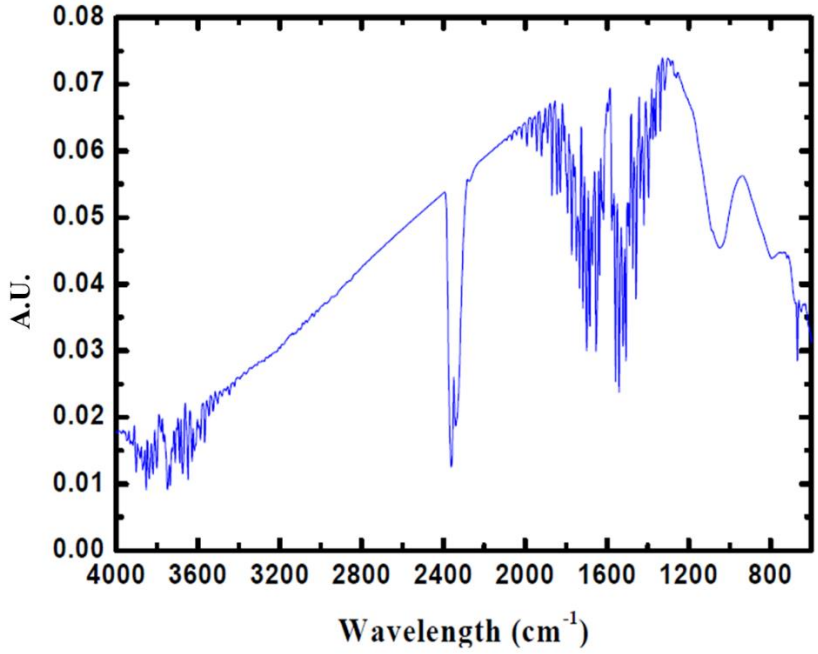


Figure 10 : A background spectrum from air.

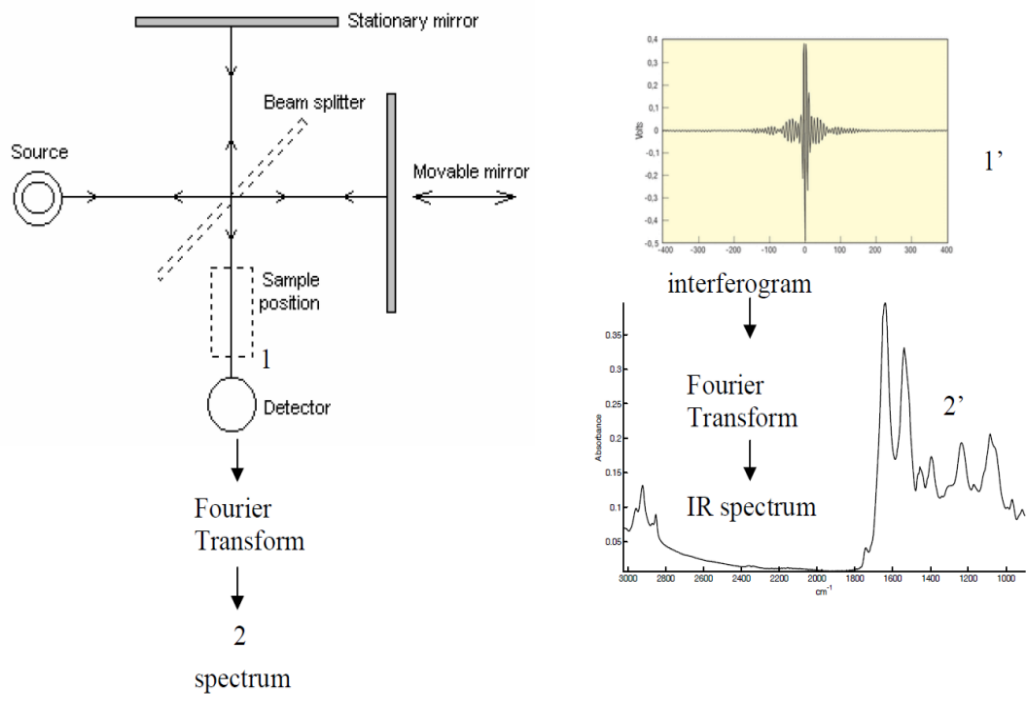


Figure 11 Basic principles of FTIR spectrometer (Adopted from Gasper, 2010).

### 1.2.3.2 ATR-FTIR Spectroscopy

In ATR-FTIR mode the refractive index of ATR-FTIR crystal must be much greater than the refractive index of sample. On the other hand, sample and crystal should have very good contact. Example of ATR crystals are diamond, germanium, silicon and zinc selenide (ZnSe) (Kazarian and Chan, 2006). When IR beam drop on the surface of ATR-FTIR crystal, part of the IR beam pass into the sample which is directed on the crystal. This wave is being attenuated when the sample absorbs the related spectral energy. Then attenuated energy is reflected back to the IR beam and leaves the crystal from opposite end and finally reaches the detector to generate the infrared spectrum (Figure 12) (Goormaghtigh et al., 1999, Gasper, 2010). Using ATR-FTIR a wide range of liquids or solids samples can be easily measured without complicated preparations.

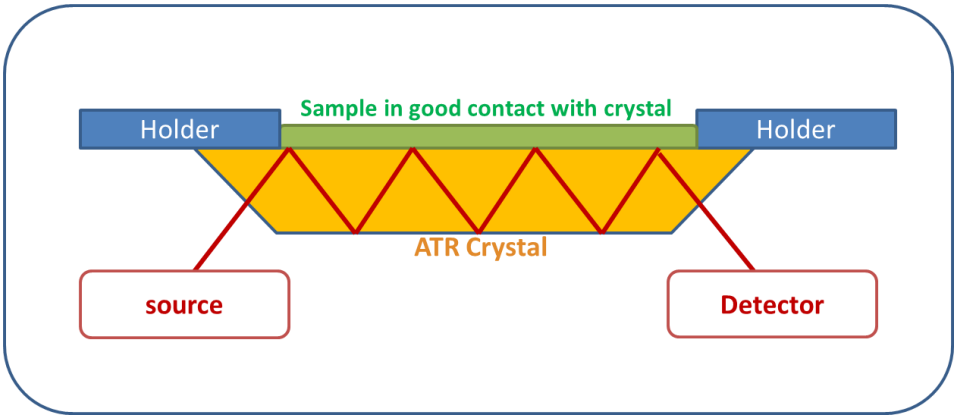


Figure 12 ATR mode in IR spectroscopy.

## 1.3 Spectra processing and classification methods

For spectra processing, OPUS spectroscopy software (Bruker Corporation) allows to do various spectral analysis steps as outlined in the following sections:

### 1.3.1 Baseline Correction

Due to instrument drift or inappropriate choice of background, the spectral baseline may be become not flat. By using the baseline correction function, we can correct sloping and curved baselines to make them flat. Figure 13 shows an example of an absorption spectrum before and after baseline correction using “Rubberband” and “concave rubberband” corrections.

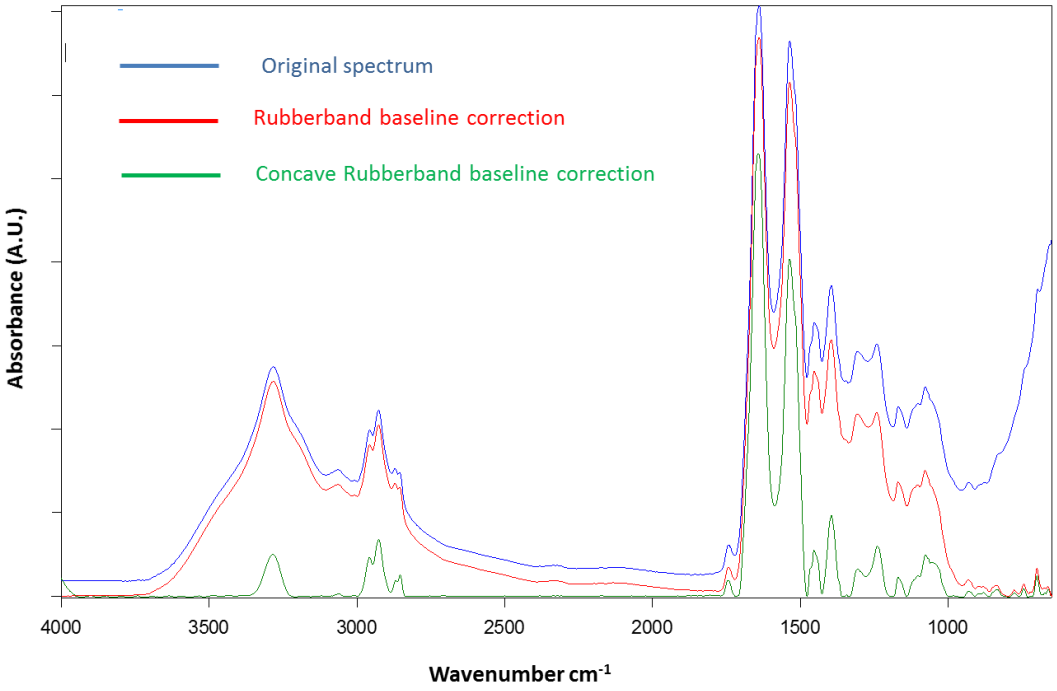


Figure 13: An absorption spectrum before and after baseline correction using Rubber band baseline correction.

### 1.3.2 Normalization

The spectra are normalized for the comparison with respect to each other. With OPUS software there are different spectral normalization methods such as:

- Min/Max normalization, the spectrum is changed so that the minimum intensity value becomes 0 and the maximum intensity value is expanded to 2 absorbance units.
- Vector normalization – The vector normalization is achieved using the following steps:

  1) Calculate the average y-value of the spectrum.

  2) Subtract the average y from the spectrum to pull the middle of the spectrum to y=0.

  3) Calculate the sum of the squares of all y-values.

  4) Dividing the spectrum by the square root of this sum.

### 1.3.3 Discrete Wavelet Transform (DWT)

Wavelets are a mathematical tool used to analyze signals like Fourier transform. It has been applied to many different problems in engineering, computers science and scientific research including image processing, heart rate and ECG (electrocardiogram) analyses. Wavelet transform is effective processes for signal analysis and feature extraction (Wan et al. 2014). In wavelet transform, the results of signal analysis are wavelet coefficients which contain valuable information about the signal. Because of this, the wavelet coefficients can be used as features for the signal. There are two types of wavelet transform one is continuous wavelet transform (CWT) and the other is discrete wavelet transform (DWT). The wavelets used in this study is DWT with wavelet known as Daubechies (db2), this wavelet was developed by Daubechies in the 1990's [26]. Details of DWT will be explained in materials and methods chapter 2.

### 1.3.4 2D-IR Correlations spectroscopy

Two dimensional correlation Infrared spectroscopy (2D-IR) is a powerful tool for spectral analysis, as it is able to reveal correlations between spectral changes and to deconvolve overlapping peaks. There are two types of (2D-IR) correlation: synchronous and asynchronous.

Synchronous spectrum reflects the simultaneous changes occurs in measured spectral series. In the synchronous spectrum the peaks can be found on the diagonal which known as autocorrelation peaks. The out of diagonal peaks are always symmetrical along the diagonal. Intensity of autocorrelation peaks is indicating of the strength of this peak. The peaks present in the out of diagonal are called cross-peaks which represent a degree of correlation between two peaks in the spectra. When this cross-peaks is positive then both the peaks, that it is formed from, are changing in the same direction i.e both are decreasing or increasing. When this cross-peaks is negative, this means the peaks, that it is formed from, are changing in the opposite way i.e one is increasing and the other is decreasing or vice versa (Figure 14).

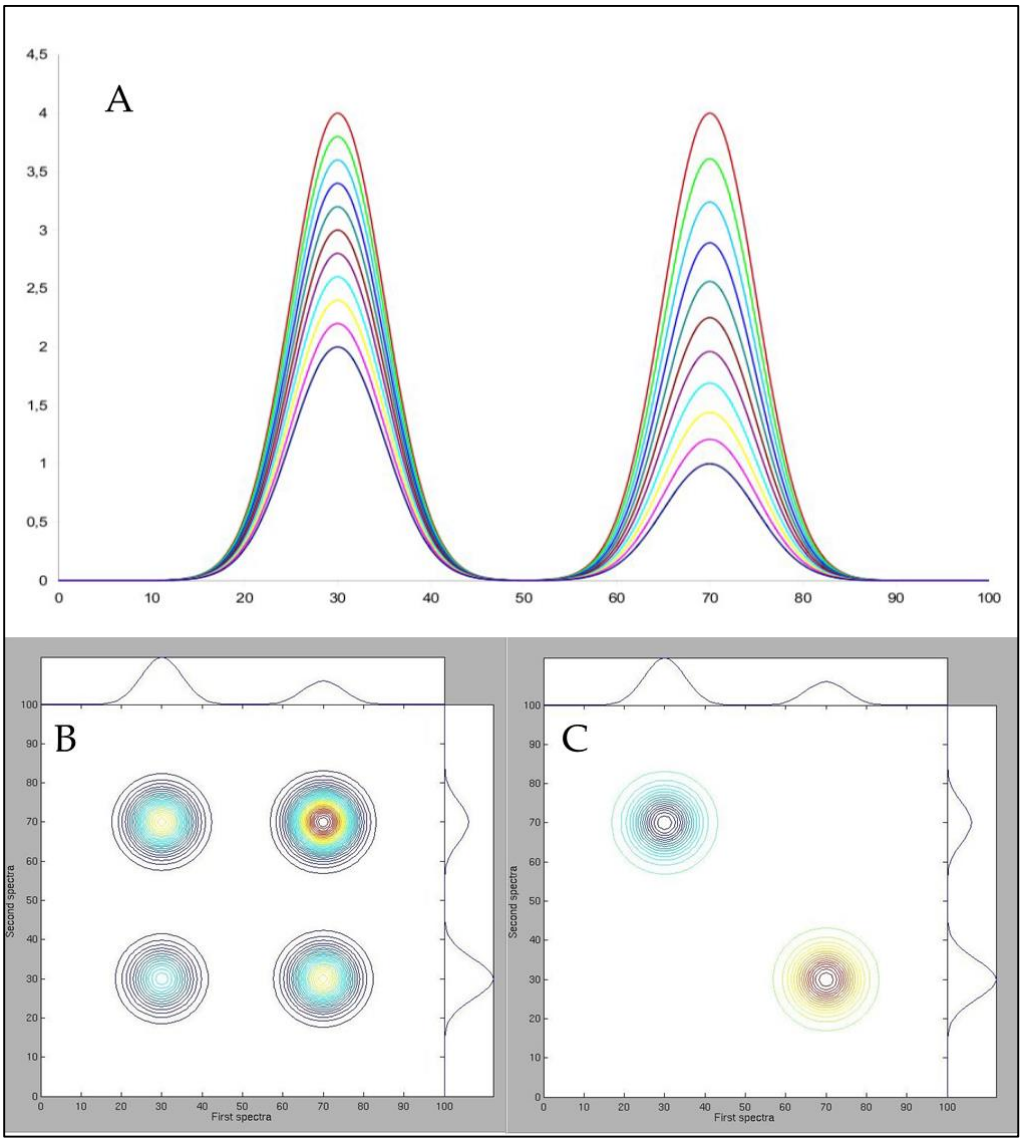


Figure 14: Example of 2D-IR correlation. A) Set of spectra representing 2 growing Gaussian peaks. B) Synchronous C) Asynchronous. 2D-IR correlation (Adopted from T. Pazderka and V. Kopeck´y Jr.).

Asynchronous spectrum is indicating the sequential changes of the measured IR spectral series. In asynchronous spectra no peaks on the diagonal are present. Also the peaks are always antisymetrical along the diagonal. When the cross-peak is positive then a band from the first spectra is growing earlier or more intensive then a band from second spectra and vice versa.

### 1.3.5 Chemometrics analysis of FTIR spectra

The application of statistical and or mathematical calculations in order to gain the information from the spectra of vibrational spectroscopy known as chemometrics analysis (Lavine, 2000). The vibrational spectra are very informative because they contain many molecular associated spectral peaks. The Multivariate data analysis can be used to find the meaningful data from the IR spectra. In general, multivariate analysis methods can be divided into two groups one named as unsupervised and the other is supervised chemometric approaches (Brereton, 2003).

#### 1.3.5.1 Unsupervised chemometric approaches

In unsupervised chemometric approaches, there is no need for priori information about studied samples. Among the unsupervised methods, principal component analysis (PCA) and hierarchical cluster analysis (HCA) are used in this study.

##### 1.3.5.1.1 Principal Component analysis (PCA)

Principal Component analysis (PCA) is a powerful technique for dimension reduction of multivariate data. PCA can be used to reduce the dimensionality of the data in order to generate a figure of data with groups clustering. Generally there are two types of PCA plots:first onw known as score and the second one known as loading plots. The degree of contributions of the spectral variations between sample groups can be shown by loading plots. Furthermore, the relationship between the samples can be deduced from the score plots. Based on the principal components (PCs), loading and score plots can be created. The advantage of PCA is that there are no need for information about the samples groups is required for the PCA calculation (Severcan and Haris, 2012).

##### 1.3.5.1.2 Cluster analysis

Cluster analysis is used to check if there is discrimination between sample groups or not. In cluster analysis spectra tend to group according to their characteristic (Mun et al. 2008). Similar spectra will classify in a same group which are shown as a dendrograms. The distances between groups calculated by Euclidean Distance value in Ward's algorithm which is a commonly used algorithm. The heterogeneity values indicate the differences between the clusters in dendogram.

#### 1.3.5.2 Supervised chemometric approaches

##### 1.3.5.2.1 Soft Independent Modeling of Class Analogy (SIMCA)

SIMCA is a statistical method for supervised classification of data developed by CAMO and included in Unscrambler multivariate data analysis commercial software. In this approach Principal Component Analysis (PCA) is run on the whole spectra dataset to identify the groups of the spectra. Then local models are estimated for each spectra group. Based on these local models the new spectra are classified to one of the established models. The advantage of SIMCA is that the unknown spectrum is assigned to the group which has high probability only. If the variance of a spectrum exceeds the upper limit for all modeled datasets, the spectra will not assign to any of the groups because, it is either an outlier or comes from a class that is not represented in the dataset. Another very important feature of SIMCA is it can work with few samples in each group which is an important consideration (Pirhadi et al. 2015).

##### 1.3.5.2.2 Artificial Neural Networks (ANNs)

Neural network or Artificial Neural Network, is a mathematical model of the biological neural networks present in living animals. The neural network consists of an interconnected group of neurons. These neurons processes the information using a certain computation calculations (Figure 15). Commonly, neural network can be considered as an adaptive system which means it will change its characteristics during the learning stage. The complex relationships between inputs and output can be predicted using ANN.

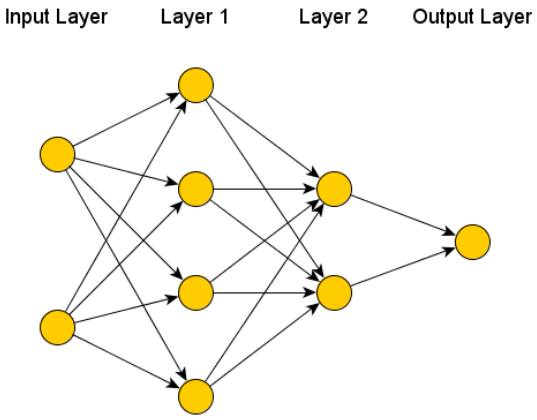


Figure 15: Artificial Neural Network

In an artificial neural network, simple artificial nodes, called neurons are connected together to form a network. These artificial neural network mimics a biological neural network (Mallick 2015). There is no certain definition of what is an artificial neural network is. In generally, the ANN can involve a network of neurons that show complex behavior by the connections between the neurons and neurons parameters. The algorithms of ANN designed to alter the strength of the connections in the network in order to produce a certain signal flow. “Similar to biological neural networks, the Neural networks are also in that functions are performed collectively and in parallel by the units, rather than there being a clear delineation of subtasks to which various units are assigned. The term "neural network" usually refers to models employed in statistics, cognitive psychology and artificial intelligence” (Filip Zavoral, Jakub Yaghob, Pit Pichappan 2010). Neural network models which emulate the central nervous system are part of theoretical neuroscience and computational neuroscience. In modern software implementations of artificial neural networks, the approach inspired by biology has been largely abandoned for a more practical approach based on statistics and signal processing. “In some of these systems, neural networks or parts of neural networks (such as artificial neurons) are used as components in larger systems that combine both adaptive and non-adaptive element” (Hassoun 1995a). While the more general approach of such adaptive systems is more suitable for real-world problem solving, it has far less to do with the traditional artificial intelligence connectionist models. What they do have in common, however, is the principle of non-linear, distributed, parallel and local processing and adaptation. “Historically, the use of neural networks models marked a paradigm shift in the late eighties from high-level (symbolic) artificial intelligence, characterized by expert systems with knowledge embodied in if-then rules, to low-level (sub-symbolic) machine learning, characterized by knowledge embodied in the parameters of a dynamical system” (Hassoun 1995b).

## 1.4 Estimation of proteins secondary structure in dilute solution using infrared spectroscopy

Protein secondary structures are built up from the combinations of some secondary structural parts name as α-helices and β-sheets. These two structure and others form the core region and can be connected by loop at the surface. x-ray crystallography and NMR can be used to obtain the secondary and tertiary structure of

proteins. However, NMR can be used only for small proteins i.e. up to 15 kilodalton proteins (Berg et al. 2002). Also for x-ray crystallographic approach there are some problems raised due to sample preparations. For example, the analysis of x-ray data can only reflect the static structure of the proteins. Because of this, the structure of a proteins in the crystal form may not reflect their actual structure in solution. Furthermore, some proteins such as membrane proteins are difficult to be crystallized which means there is no possibility to predict the structure using x-ray crystallography. On the other hand, x-ray crystallography need highly sophisticated expensive machines which may not be easily available. Because of these disadvantages of high resolution techniques, the development of low resolution techniques for proteins structure estimation is necessary. Fourier transform infrared spectroscopy is one of those techniques which is widely used in protein secondary structure estimation because of its sensitivity and rapidity. Different conformational types such as α-helix, sheet, turns, etc. result in different absorption bands (Figure 16).

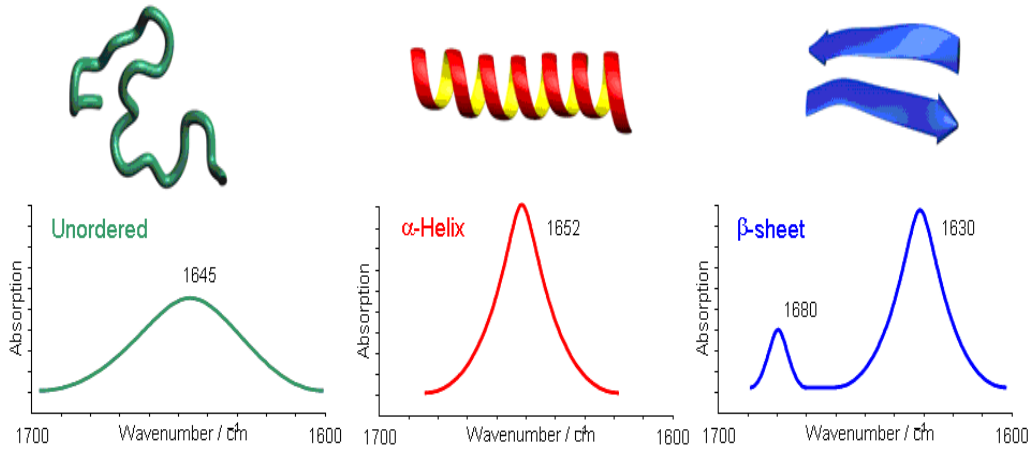


Figure 16 Secondary structure motifs in the amide I region of IR spectrum (Figure is adopted from Barth and Zscherp, 2002).

These bands are overlapping under the broad amide I band located in between 1700-1600 $cm^{-1}$. Unfortunately, it is not easy for accurate quantification of protein secondary structure from FTIR spectra. Various techniques such as curve fitting, second derivative and factor analysis have been used to predict secondary structure of

proteins from their infrared spectra however each of these techniques has its limitation beside that they all need an expert user.

## 1.5 Biophysical studies on Human Transferrin protein in artificial crowded environment

### 1.5.1 Human Transferrin

Human serum transferrin (TF) (80 kDa) is an iron transport glycoprotein that involved in the regulation and balance of iron content in blood plasma and cells. X-ray crystal structure of transferrin shows that it contains two iron-binding sites (Cheng et al. 2004). One is located in the N-terminal lobe, the other in the C-terminal one. The secondary structure of iron bonded transferrin (Holo-Transferin or HTF) using its x-ray crystal structure and pdbsum (3qyt) informatics tool shows 17.8% β-sheets, 33.1 % alpha α- helix and 49% others structures. This secondary structure makes the protein very flexible. In the iron free form of transferrin (Apo-Transferrin or ATF) the two domains open up into a 'V' shaped conformation ready to trap the iron inside. Upon the iron uptake, large conformational changes in the protein occur (Kilár and Simon 1985). Little is known about the effect of molecular crowding on the structure, dynamics and aggregation of Transferrin, because of that, this study will utilize the FTIR spectroscopy in order to investigate the effect of Dextran 70 and Ficoll 70 as macromolecular crowders on Holo and Apo- Transferrin.

### 1.5.2 Macromolecular crowding

Traditionally, studies of protein stability, in vitro, have been done using dilute buffer with low concentrations of macromolecules. This environment is totally different than native environment of proteins which known as “macromolecular crowded environment” such as Ficoll and dextran (Figure 17) . Crowding environment provides a non-specific force and affect the total excluded volume according to excluded volume theory. According to excluded volume theory; an increase in the melting point and a change in thermo-dynamic of proteins on crowded environment is expected. Also, measurements of the proteins properties that are made in dilute solutions (in vitro) may be different by many orders of magnitude from the true values seen in living cells crowded environment (in vivo). Because of this, addressing the effect of crowding on protein stability experimentally is of great current interest.

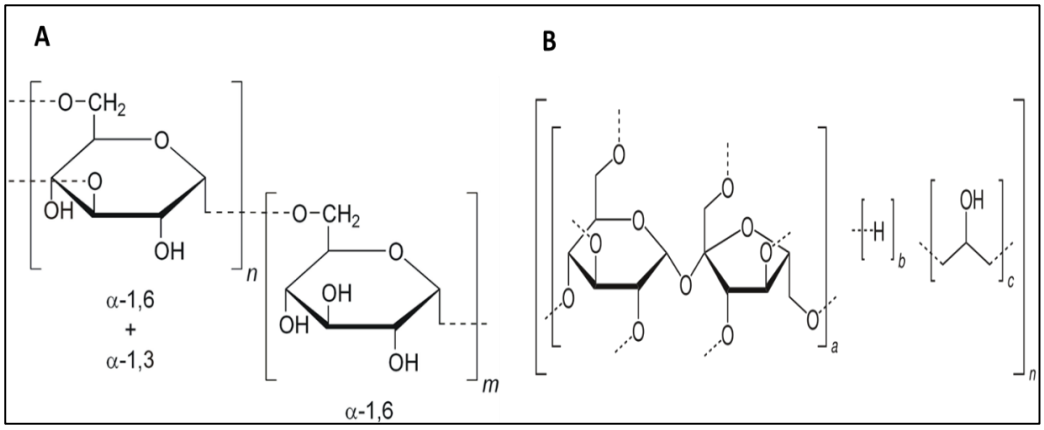


Figure 17: Chemical structure of A) Dextran 70 B) Ficoll 70

Adopted from: https://commons.wikimedia.org/w/index.php?curid=23268355 and https://commons.wikimedia.org/w/index.php?curid=15527143

## 1.6 Studies of proteins secondary structure in biological fluids using Infrared spectroscopy.

FTIR spectroscopy and imaging is useful tool for the identification and characterization of the molecular components of biological processes in cells (Krafft and Sergo 2006). Figure 18 shows a typical mid-FTIR spectra and their bands assignment which can provide information about the proteins, lipids, carbohydrates of biological fluids (Krafft et al. 2006; Krafft et al. 2008). As these molecular features change during carcinogenesis the spectra can be monitored sensitively as phenotypic markers for cancer diagnosis. Since the information is obtained label-free and non-destructively, the methods can also be applied under in vivo conditions for screening (Baker et al. 2014a). Information related to protein composition and secondary structure can be obtained by performing qualitative and quantitative analysis of amide bands. The amide I band (1700-1600 $cm^{-1}$) is the most sensitive and accurate peak for secondary structure determination.

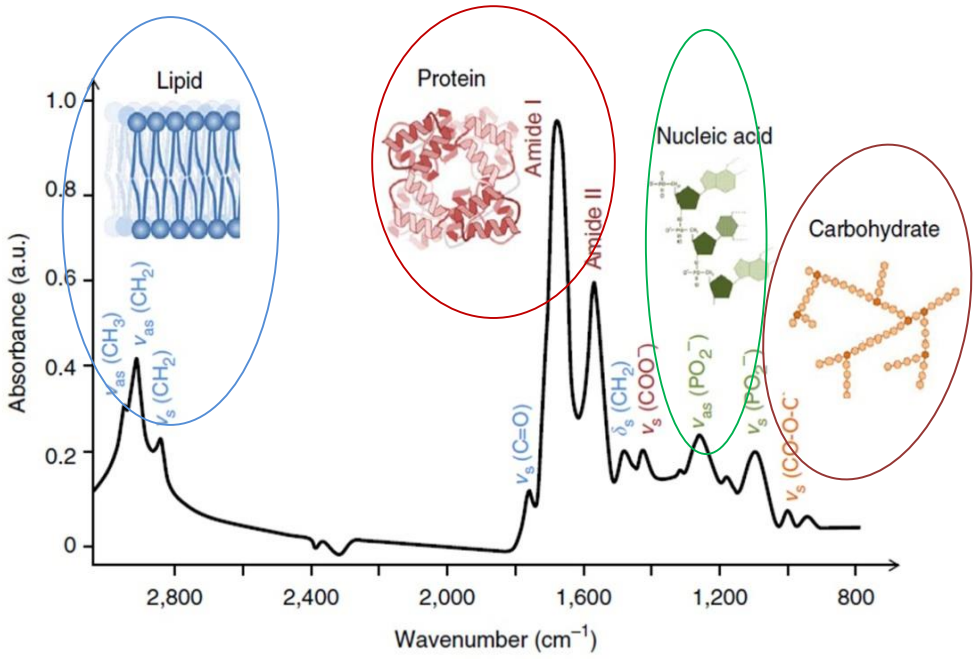


Figure 18 Typical biological spectrum showing biomolecular band assignments in the 3,000–800 cm $^{-1}$ region.

where ν = stretching vibrations, δ = bending vibrations, s = symmetric vibrations and as = asymmetric vibrations. The spectrum is a transmission-type micro-spectrum from a human breast carcinoma. (adopted from Baker et. al.,2014).

### 1.6.1 Malignant Pleural Mesothelioma

Mesothelioma or malignant pleural mesothelioma (MPM) is a rare form of lung cancer that originated from cells in the mesothelium. The mesothelium is the protective layer covers the lung and some others organs. The most common site for mesothelioma is the lung pleura which is the inner layer of the chest wall and lungs. The inhalation of asbestos dust and fibers can develop mesothelioma. Also there are no clear link between mesothelioma and tobacco smoking, however smoking may increases the risk of the other asbestos-induced cancers (McCarthy et al. 2012). Shortness of breath due to pleural effusion or chest wall pain can be considered as a signs or symptoms of mesothelioma. X-ray and CT scan can be used for the diagnosis of MPM. But the diagnosis must be confirmed pathologically, either with serious effusion cytology or

with a biopsy and microscopic examination (Sutedja 2003). Early and accurate diagnosis of MPM can greatly help to increase of patients survival rate.

## 1.7 Aim of the Study

Proteins structure and dynamics are very important in understanding of many human diseases. The study of proteins secondary structure using Fourier transformed infrared (FTIR) in dilute, artificial crowded environment and in a human pleural fluid can greatly improve the understanding of many proteins related diseases. In order to accurately predict the protein secondary structure from FTIR spectra, a novel method for extraction of protein FTIR features using wavelet transform analysis and artificial neural network (ANN) has been introduced.

For the training of ANN, we produced a proteins database from known structures of proteins in dilute 7.4 phosphate buffer solution condition. Then this database was used as a training set for an artificial neural networks (ANNs). Because of the limited number of our proteins dataset (35 proteins), we developed a leave-one-out approach Matlab algorithm for the training and testing our neural networks.

In order to understand the effect of crowded environment on the proteins structure, we investigated the secondary structure, thermal denaturation, aggregation and hydrogen-deuterium (H/2H) exchange of Apo and Holo transferrin in the presence and absence of the molecular crowding agents.

Malignant Pleural Mesothelioma is hard to diagnose and aggressive cancer type. Due to the inability of early and accurate diagnose of MPM, it has a high mortality rate both in Turkey and throughout the world. Diagnosis of this disease usually done by cytological methods from pleural fluids and by histochemical and immunohistochemical methods from biopsy samples. However, the sensitivity and specificity of these methods are not very high. Early diagnosis of the disease together with the application of appropriate and effective treatment strategies is crucial for the decrease in the mortality rate of the disease and the increase in the survival of patients. Therefore, there is a need for non-invasive methods having a high sensitivity and specificity, which can be used for the screening and diagnosis of MPM disease. To

identify whether the human pleural fluid is accumulated due to MPM or LC or BT diseases, the details of pleural fluid's proteins content and their secondary structure has been studied using Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR) spectroscopy.

# CHAPTER 2

# MATERIALS AND METHODS

## 2.1 Estimation of Proteins secondary structure using wavelet based ANN

### 2.1.1 Preparation of proteins solutions

FTIR spectra of 35 water-soluble proteins were collected in phosphate buffer solution. All protein samples were obtained from (Sigma-Aldrich, UK) and were used without further purification. The secondary structure contents of these proteins are known from X-ray crystallography (Table 2). For infrared measurements, samples were prepared by dissolving certain amount of protein in phosphate buffer to obtain a final concentration of 20 mg/ml protein in pH 7.4.

### 2.1.2 FTIR experimental setup and scanning parameters

Infrared spectra were recorded by using a Vector 22 Bruker spectrometer equipped with DTGS detector with (128 scans at 4 $cm^{-1}$ resolution). 4 µl of each aqueous sample was placed in between the two special $CaF_2$ windows with 6 um pathlength. The temperature of the protein was maintained at 25 $^{o}C$ using of a circulating water. In order to eliminate the effect of water vapor, the FTIR spectrometer has been purged with dry air.

Table 2: The 35 proteins dataset with their secondary structure using X-ray and bioinformatics tool (Joosten et al. 2011).

| Proteins | Pdb code | X-ray based results % | | |
|---|---|---|---|---|
| | | β-sheets | α-helix | Others |
| (MOUSE) - Acetylcholinesterase | 1C2O | 15.2 | 35.6 | 49.2 |
| (YEAST) - Alcohol dehydrogenase 1 | 2HCY | 29.1 | 25.9 | 45 |
| (HORSE) - Alcohol dehydrogenase | 4DXH | 24.6 | 27 | 48.4 |
| (BOVIN) - Chymotrypsinogen A | 1YPH | 34.4 | 2.3 | 63.4 |
| (RAT) - Chymotrypsinogen B | 1KDQ | 33.1 | 0.0 | 66.9 |
| (HUMAN) - Alpha-lactalbumin | 1A4V | 6.5 | 43.9 | 49.6 |
| (HORSE) - Ferritin light chain | 4DE6 | 0 | 74.4 | 25.6 |
| (HUMAN) - Fibroblast growth factor 2 | 1BFG | 38.9 | 9.5 | 51.6 |
| (BOVIN) - Serum albumin | 4F5S | 0 | 73.6 | 26.4 |
| (HUMAN) - C-reactive protein | 1B09 | 39.8 | 8.8 | 51.5 |
| (BOVIN) - Carbonic anhydrase 2 | 1V9E | 29 | 14.3 | 56.8 |
| (CANEN) - Concanavalin-A | 3CNA | 40.5 | 0 | 59.5 |
| (BOVIN) - Cytochrome c | 2B4Z | 0 | 39.4 | 60.6 |
| (BOVIN) - Cytochrome c oxidase | 1V54 | 0.4 | 70.4 | 29.2 |
| (THUAA) - Cytochrome c | 3CYT | 0 | 41.7 | 58.3 |
| (HORSE) - Cytochrome c | 1HRC | 0 | 41.3 | 58.7 |
| (PIG) - Chymotrypsin-like elastase | 1QNJ | 34.2 | 10 | 55.8 |
| (YEAST) - Enolase 1 | 3ENL | 16.5 | 42.5 | 41.1 |
| (HUMAN) - Coagulation factor XIII | 1F13 | 39.8 | 14.4 | 45.8 |
| (HUMAN) - Fibrinogen | 3GHG | 1.7 | 70.6 | 27.6 |
| (LEUME) - Glucose-6-phosphate 1-dehydrogenase | 1DPG | 34.2 | 46.4 | 42.3 |
| (HORSE) - Hemoglobin | 1NS6 | 0 | 76.6 | 23.4 |
| (HUMAN) - Hemoglobin | 1HHO | 0 | 70.9 | 29.1 |
| (HUMAN) - Serotransferrin | 3QYT | 17.8 | 33.1 | 49 |
| (RABIT) - L-lactate dehydrogenase | 3H3F | 21.5 | 43.2 | 35.3 |
| (CHICK) - Lysozyme C | 2LYZ | 6.2 | 41.1 | 52.7 |
| (HORSE) - Myoglobin | 2V1H | 0 | 74.5 | 25.5 |
| (CHICK) - Ovalbumin | 1OVA | 32.1 | 32.1 | 35.8 |
| (PAEPO) - Bacillolysin | 4GER | 17.4 | 43.4 | 39.1 |
| (BOVIN) - Ribonuclease pancreatic | 4AO1 | 34.7 | 20.1 | 45.2 |
| (BACLI) - Subtilisin Carlsberg | 3UNX | 17.9 | 29.6 | 52.6 |
| (RABIT) - Triosephosphate isomerase | 1R2R | 15.4 | 44.3 | 40.2 |
| (BOVIN) - Cationic trypsin | 1TGS | 37.8 | 8.4 | 53.8 |
| (SOYBN) - Trypsin inhibitor A | 1BA7 | 36.4 | 1.8 | 61.8 |
| (BOVIN) - Cationic trypsin | 1S0R | 32.3 | 8.5 | 59.2 |

#### 2.1.2.1 ANN Training parameters

The neural network initiated using random weights and the following parameters has been used:

Performance goal= 0

Performance function= sum squared error (sse)

Maximum number of epochs to train= 1000

Learning rate= 0.01

For implementation of these parameters in Matlab please refer to appendix.

## 2.2 Biophysical characterization of proteins in artificial crowded environment

Apo-transferrin, Holo-transferrin $D_2O$ and Dextran 70 was purchased from sigma aldrich and used without further purification.

### 2.2.1 Sample preparation

Stock Dextran solution prepared by dissolving the amount of dextran powder required to obtain 200 mg/ml concentration at pH/D 7.4 buffer saline solution.

The samples for infrared measurements were prepared by dissolving the required amount of protein to obtain a final concentration of 50 mg/ml protein in pH/D 7.4 buffer saline solutions and in dextran stock solution.

### 2.2.2 Infrared spectroscopy

In H2O: 4 µl of each sample was placed in between the two special $CaF_2$ windows with 6 um pathlength. All spectra (128 scans for each sample) were recorded with a Vector 22 Bruker spectrometer equipped with DTGS detector. Although the system continuously purged by dry air to reduce the water vapor noise, the vapor spectra (before and after each experiment) has been recorded to be subtracted from the sample spectrum. For thermal denaturation study, each sample heated from 25C to 85 C for Apo-Transferrin (ATF) and from 25C to 95C for Holo-Transfferin (HTF) in 5 C steps. At each temperature certain time allowed for temperature stabilization before recording the spectra. The OPUS 6.5 software and custom written Matlab scripts was used for the one- and two-dimensional analysis of the 1700-1600 cm-1 spectral region. The amide I region, which consists of overlapping bands, was resolved by using Fourier self-deconvolution with a band width with factor 2 and a resolution

enhancement of 2. The deconvoluted spectra were normalized with respect to 1515 $cm^{-1}$ Tyrosine peak. The second derivative of each spectra has been calculated using the same software. In order to calculate both synchrone and asynchrone 2D plots, temperature has been used as perturbation to induce time-dependent spectral fluctuations for deconvoluted.

In $D_2O$ for H/D exchange: 20 µl of each sample was placed in between the two flat $CaF_2$ windows with 50 µm spacer. The first spectrum recorded after 3 minutes from onset time of dissolving protein in $D_2O$ buffer or $D_2O$ dextran buffer solution. Then spectra recorded in two minutes intervals for an hour and each spectrum takes about one minute to average 64 scans. The spectra were normalized with respect to Amid I using OPUS 6.5. The intensity of Amid II' at 1546 cm-1 calculated for each spectra using custom Matlab script. According to literature (Dong et al. 1996) using the following formula: $F= (A_2-A_{2\infty})/A_1\omega$. where $A_1$ and $A_2$ are the absorbance maxima of the amide I and amide II bands, respectively. $A_{2\infty}$ is the amide II absorbance maximum of fully denaturated protein, and ω is the ratio of $A_{20}/A_{I0}$ , with $A_{20}$ and $A_{10}$ being the respective absorbance maxima for the amide II and amide I bands of the proteins in $H_2O$.

### 2.2.3 2D-IR correlation:

There are two types of 2D IR correlation plots. The synchronous 2D IR represents in-phase variation between the spectral components to an applied perturbation (Temperature in our case) (Noda 1990). The asynchronous 2D IR represent out-phase variation between the spectral components to an applied perturbation. The asynchronous 2D IR plot together with the synchronous plot provides details about the sequence of events following an applied perturbation (Paquet et al. 2001). The 2D IR plots are either symmetric (synchronous) or antisymmetric (asynchronous) with respect to the diagonal only peaks above the diagonal are discussed as it is also including information about the other part. In this thesis 2D-IR peaks are identified as Y vs X $cm^{-1}$, where Y and X represent the wavenumber respectively. The sign of a cross peak, either to be positive or negative, determines the sequential relationship between two peaks according to the rules proposed by Noda (15).

## 2.3 Proteins in biological fluids

### 2.3.1 Subject recruitment/ Study Subjects

The current study protocol was approved by Hacettepe University Ethics Committee (HK 12/131-36). Before collection of pleural fluid samples, a written informed consent was taken from all patients following the ethical norms of the institute. The samples were collected from patients with Malignant Pleural Mesothelioma (MPM, n=24), lung cancer (LC, n= 20), and benign transudate (BT n=26). BT was considered as control group since these pleural fluids were due to benign diseases such as congestive heart failure. MPM and LC diagnosis were confirmed by standard Hematoxylin and eosin stain (HE) and immune-histochemical staining of biopsy specimens from the tumor sites. The diagnosis of BT was confirmed according to Light's criteria via the analysis of protein and lactate dehydrogenase (LDH) levels in both pleural fluid and serum.

### 2.3.2 ATR-FTIR Spectroscopy

#### 2.3.2.1 Sample preparation and spectral acquisition for FTIR spectroscopy

Before FTIR measurements, frozen samples were thawed at the room temperature**.** Infrared spectra of the samples were collected using the one bounce ATR mode on a PerkinElmer Spectrum 100 FTIR spectrometer (PerkinElmer Inc., Norwalk, CT, USA) equipped with a universal ATR accessory. Briefly, 1 $\mu$l of pleural fluid was placed on the top of diamond/ZnSe crystal of ATR spectroscopy and dried with mild nitrogen gas for 3 min to remove the excess unbound water. 128 scans were collected for each spectral measurement in the spectral range between 4000 and 650 $cm^{-1}$ with the resolution of 4 $cm^{-1}$. Since water molecules in the air affect the IR spectrum (Mitchell et al. 2014), the spectrum of the empty diamond/ZnSe crystal was recorded as background and subtracted automatically by using appropriate software (Spectrum 100 software, Perkin Elmer). Recording and analysis of the spectral data were performed using the Spectrum One software from Perkin Elmer. Randomly chosen three portion of the sample were scanned and their spectral average was used in further analysis.

#### 2.3.2.2 Spectral Pre-processing:

Raw IR spectra from all pleural fluid samples were concave rubber band baseline corrected with 64 baseline and 50 iteration points. Then spectra were vector

normalized in order remove the effect of overall scaling or electronic gain effects may happened during samples measurements (Martens et al. 2003; Baker et al. 2014b). These pre-processed spectra were used for further chemometric analysis.

All data manipulations were carried out by OPUS 6.5 software (OPUS, Bruker Optics, Ettlingen, Germany

### 2.3.3 Chemometric Analysis

To identify the spectral differences among BT, LC and MPM groups and to classify them, unsupervised and then supervised chemometric approaches were performed using Unscrambler X (Camo Software, Inc.) program.

### 2.3.4 Unsupervised Chemometric Analysis

To identify spectral differences and relationships among BT, LC and MPM groups, unsupervised chemometric approaches such as hierarchical cluster analysis (HCA) and principal component analysis (PCA) were performed using Unscrambler X (Camo Software, Inc.) for HCA analysis of the groups (BT, LC, MPM). HCA enables to assess the similarity between samples by measuring the distances between the points in the measurement space (Owens et al. 2014). Similar samples lie close to one another, whereas dissimilar samples are distant from each other (Muehlethaler et al. 2014). PCA is commonly used as an unsupervised technique for data compression and visualization. The relationships between samples are defined by using their principal components (PCs). (PCs) are simply linear combinations of the variables that explain the greatest variance, the next greatest variance, etc. Thus by using PCA, the n-dimensional data set can be plotted in a smaller number of dimensions. This allows the observation of clusters, which can define the structure of the data set (Owens et al. 2014). Based on HCA results, to measure the performance of discrimination method, sensitivity and specificity were calculated as described in Table 3 (Gok et al. 2016).

Table 3. Definitions for sensitivity and specificity

| Cluster analysis results based on ATR-FTIR data | | | |
|---|---|---|---|
| Positive* | | Negative * | |
| LC/MPM | A | B | Sensitivity=A/(A+B) |
| Control (BT) | C | D | Specificity=D/(C+D) |

* Positive and negative values are determined as follows:

A: the number of LC or MPM patients identified in LC/MPM groups (true positive).

B: the number of LC or MPM patients identified in control group (false negative)

C: the number of BT patients identified in LC/MPM groups (false negative).

D: the number of BT patients identified in BT group (true negative).

### 2.3.5 Supervised Chemometric Analysis

To classify the studied groups, Soft Independent Modeling of Class Analogy (SIMCA), a supervised classification technique, was performed by UnscramblerX (CAMO Software, Inc.). Firstly, the PCA models for each group were created and then 3 samples from each group were randomly selected and tested to validate the developed methods. In SIMCA, each class of data set is modeled by principal component analysis (PCA) and then new samples are tested to each class set whether they are similar or dissimilar (Mueller et al. 2013).

# CHAPTER 3

# RESULTS AND DISCUSSION

This thesis is composed of three parts in order to study proteins structure and dynamics in dilute, artificial crowded environments and in real biological liquids. First part presents the method of proteins secondary structure estimation in dilute solution using artificial neural networks. The second part presents the study of Transferrin protein structure and dynamics in crowded environment. The second part also include the study of Transferrin thermal denaturation in dilute and crowded environment using 2D-IR correlation. The third part is the study of protein secondary structure in real human biological liquids with application to Malignant Pleural Mesothelioma (MPM).

## 3.1 Estimation of Proteins secondary structure using wavelet based ANN

### 3.1.1 Features extraction using DWT

In this study DWT has been used as data reduction for Amide I. The results of DWT have been used as an input vector for ANN. The amide I spectra are first range scaled to the interval between 0 and 1 then rubberband baseline corrected using OPUS program. By using Matlab algorithm "decROW" (appendix) the amide I of each protein decomposed to 7 Level decomposition using Near symmetric wavelet (db2, db3, db10 and haar). In order to find the best level of wavelet decomposition that represents the amide I with the lowest number of wavelet coefficient, all amide I bands of the proteins dataset have been decomposed to 7 levels.

Figure 19 shows the seven level of two different amide I signals. From the figure we can conclude that levels 7,6 and 5 are not acceptable because they loss most of the amide I features. Levels 2,3 and 4 show a better representation of amide I signals. However, level 2 consists of 35 coefficients which is not suitable with our limited

number of proteins for ANN training. After several trials and signals examination, we found that the decomposing of amide I to the 3rd level gives the best representation of amide I with lowest number of coefficients.

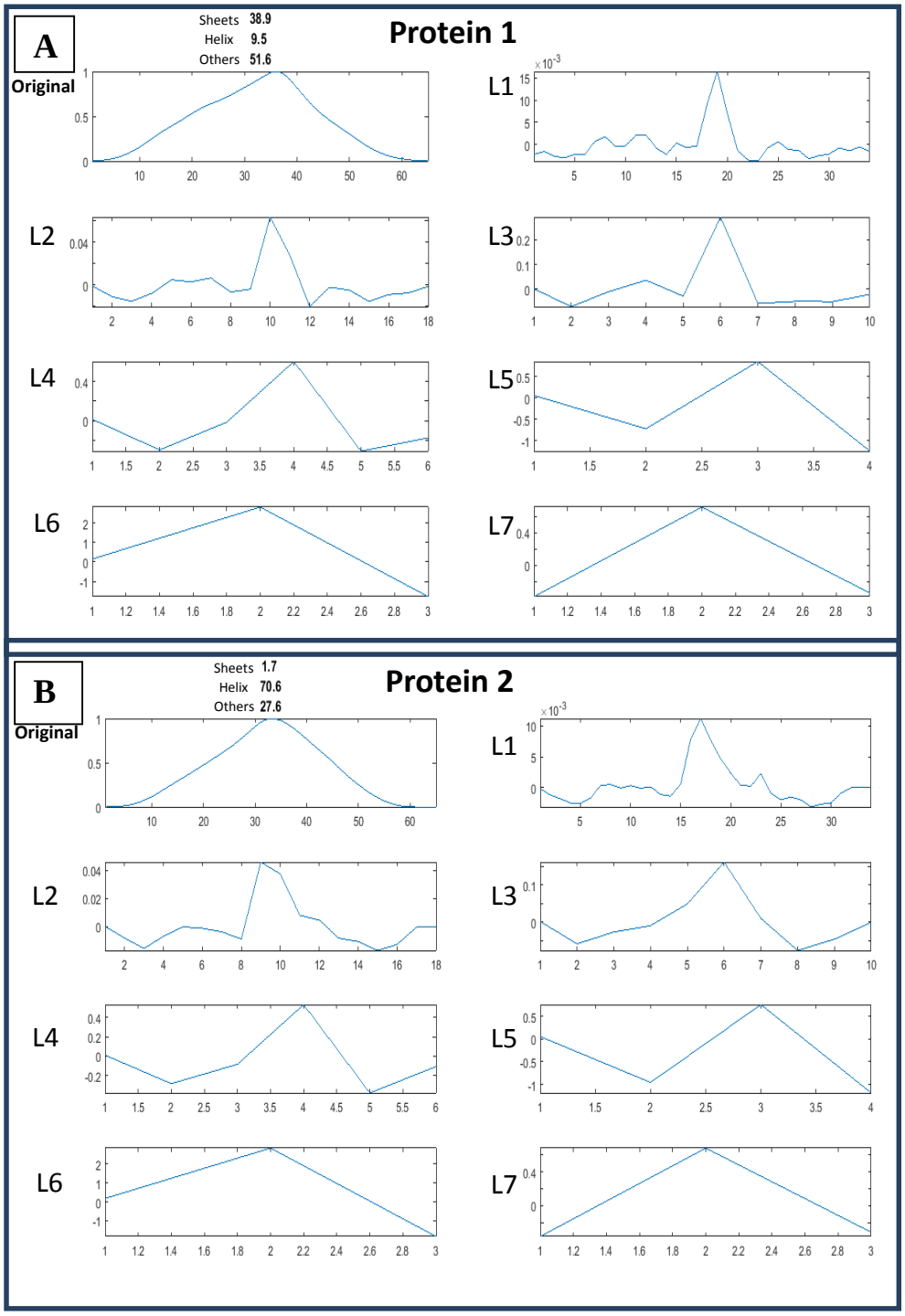


Figure 19: Seven level (L1-L7) of wavelet decomposition for two different secondary structures proteins. A) Mainly sheet protein (protein 1). B) Mainly α-helix (Protein 2)

The wavelets used in this project are the Daubechies D2 developed by Ingrid Daubechies in the 1990's (Daubechies 1992). The major difference between the Daubechies and the Haar wavelets is that the Daubechies wavelets do not have jump discontinuities and as such represent signals in frequency or scale space with better localization (Quellec et al. 2008). In order to decide which wavelets type can be used for amide I,

Figure 20 shows the amide I and its wavelet coefficients for three different types of proteins mainly β-sheets, α- helix and random structure.

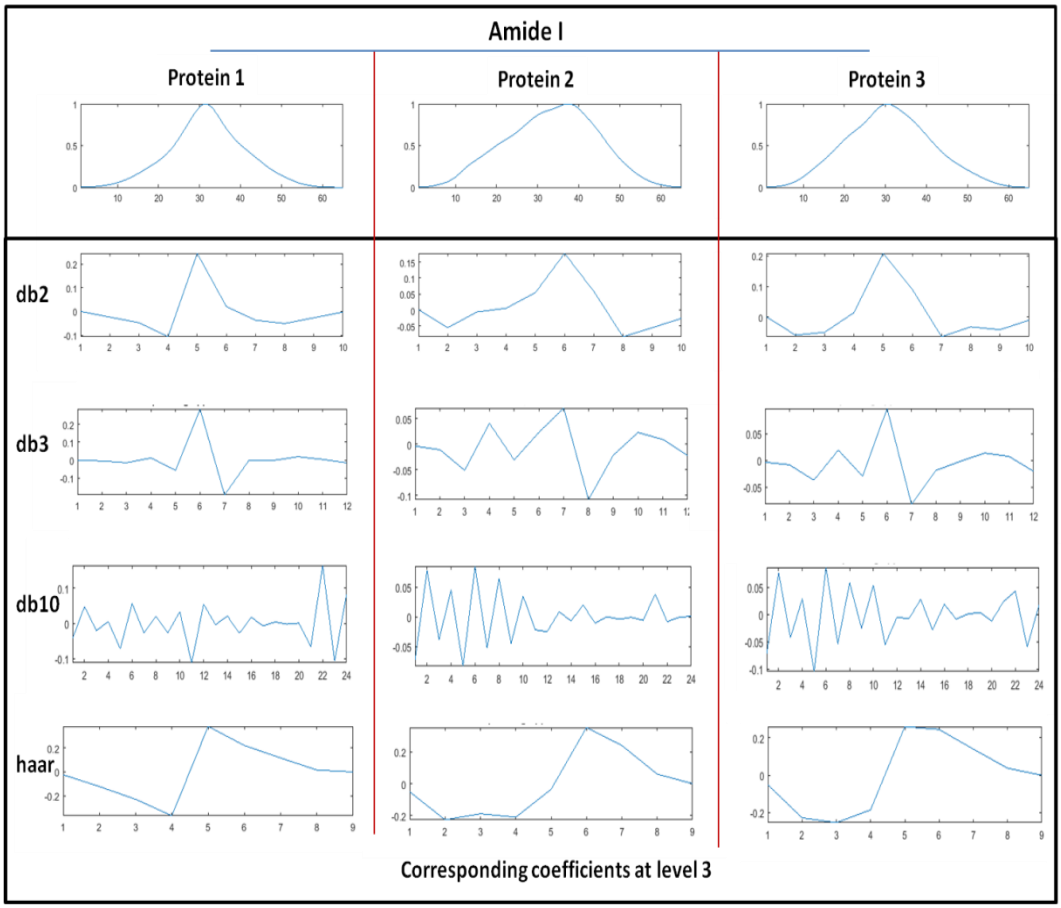


Figure 20: Wavelets coefficients at level 3 using db2, db3, db10 and haar for three different proteins

The analysis of these amide I wavelets at level 3 for the three proteins show that db2 coefficient gives the best representation reflecting the amide I characteristics. Each db2 coefficient vector at level 3 consist of 10 features which has been used as input for the neural network.

Wavelets coefficients and plots has been produced using Matlab wavelet toolbox and custom written Matlab script (appendix). Amide I signal can be decomposed to different level using DWT. The first level contains the highest number of wavelets coefficients that represent the amide I signal. Using these coefficients, amide I signal can be accurately reconstructed from its wavelets coefficient (Alzubi et al. 2011). The second level of wavelets coefficients contains a less number of coefficient which represent a less features of amide I by neglecting some minor features. By the same logic, as the level of decomposition increase the number of coefficients will decrease and the reconstructed signal will contain less features about Amide I.

Although the first level has the highest accuracy for amide I signal reconstruction, previous works (Hering et al. 2004b; Hering et al. 2004a) show that not all amide I wave numbers are useful for proteins secondary structure prediction. Furthermore, the other wave numbers in amide I can act as a noise affecting the prediction rate of the protein secondary structure. On the other hand, because of the limited number of proteins in our ANN training set, the less number of wavelets coefficients can lead to a better training and prediction rate for the ANN. In this study artificial Neural network has been trained using Matlab Neural Network Toolbox. For the training of ANN the resilient backpropagation has been used. Finally, the ANN consist of 10 inputs, one hidden layer, and an output layer with three neurons. We found that three neurons in the hidden layer gives the best resulted with smaller prediction errors. The predictions obtained for the protein dataset are presented in Table 4.

### 3.1.2 ANN Leave-one-out training approach

A feed forward neural network has been constructed using "feedforwardnet" matlab function (appendix1). A good trained neural network should make good predictions when data from outside the training set is used as an input. This is known as neural network generalization which is the most important issue in developing a neural network (Ahmed 2005). If neural network has very few hidden units, it can fail to predict the complicated data set leading to under fitting. In contrast, if the neural network has too much hidden units it can be overtraining specially for limited size training dataset. The overtraining neural network tend to memorize the training set rather than learning itself (Toney and Vesselle 2014). Both of under and over training

cause that the neural network will fail to predict the tested samples. In this study, only FTIR spectra of 35 proteins are available for the proteins dataset. Because of this, a neural network with only one unit in the hidden layer has been used initially. The results of one hidden unit neural network showed under fitting prediction. The two hidden units' neural network showed a better prediction. However, with three hidden units it showed much more better prediction. A neural network with high (more than 3) hidden units was not used because of our limited dataset and to avoid the overtraining problem. Finally the ANN consisted of one input layer with 10 neurons, one hidden layer with 3 neurons and one output layer with 3 neurons too (Figure 21).

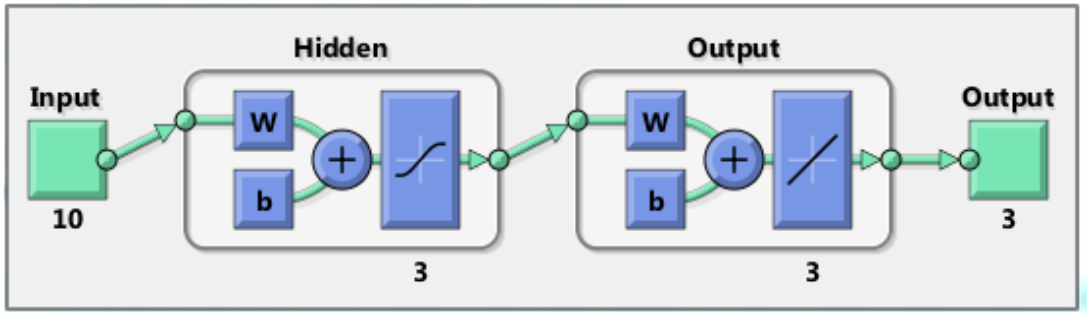


Figure 21: Structure of the feedforward ANN

The leave-one-out approach is usually used to train a neural network with small training dataset. In this method one vector (spectra) was removed from the dataset and considered as (blind test vector) for testing the neural network after training. From the remaining dataset (34 vector) one vector was used as a test and another one was used for validation then the neural network was trained by the 32 vectors. After training, the neural network was tested using the test vector and the root mean square error (rms) has been calculated from the outputs and targets values. The process of selection of one vector (as a test) was repeated for all vectors and the overall rms has been calculated. If the overall rms of the leave-one-out neural network blow a certain threshold value (calculated experimentally) the blind test vector was used to test the performance of the developed neural network. Figure 22 shows a schematic diagram for the leave-one-out learning process.

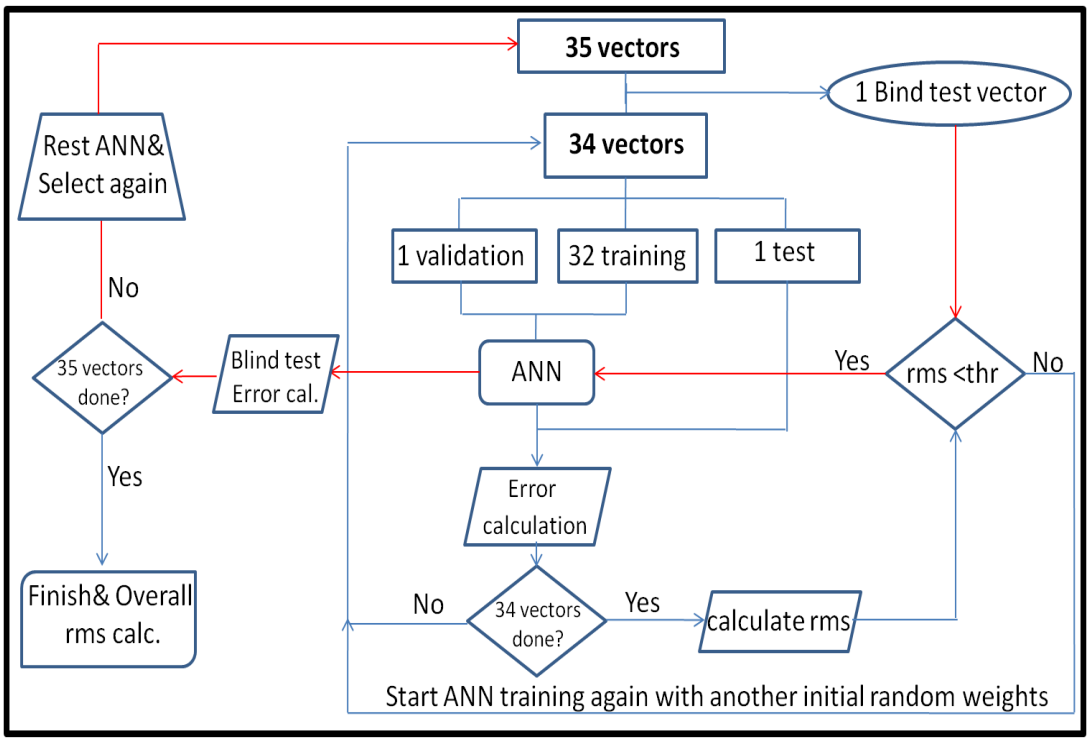


Figure 22: Schematic diagram for leave one out method used for proteins secondary structure prediction

Then, another blind test vector was removed from the dataset and the previous whole leave one out process has been repeated again. In order to overcome the problem of error terms of networks that have fallen to different local minima, a jury of 10 networks has been used for each blind test. Then the mean of the rms of blind test was calculated and considered as indicator for the neural network prediction success.

Table 4: Comparison between the X-ray based and ANN based proteins secondary structures.

| Pdb code | X-ray based results | | | ANN based results | | | Errors | | |
|---|---|---|---|---|---|---|---|---|---|
| | β-sheets | α-helix | Others | β-sheets | α-helix | Others | β-sheets | α-helix | Others |
| 1C2O | 15.2 | 35.6 | 49.2 | 17.76 | 32.37 | 54.04 | 2.56 | -3.23 | 4.84 |
| 2HCY | 29.1 | 25.9 | 45 | 29.05 | 29.23 | 44.27 | -0.05 | 3.33 | -0.73 |
| 4DXH | 24.6 | 27 | 48.4 | 29.05 | 19.45 | 49.17 | 4.45 | -7.55 | 0.77 |
| 1YPH | 34.4 | 2.3 | 63.4 | 36.31 | 2.44 | 64.70 | 1.91 | 0.14 | 1.30 |
| 1KDQ | 33.2 | 9.6 | 57.3 | 34.43 | 12.96 | 52.69 | 1.23 | 3.36 | -4.61 |
| 1A4V | 6.5 | 43.9 | 49.6 | 6.92 | 50.64 | 41.85 | 0.42 | 6.74 | -7.75 |
| 4DE6 | 0 | 74.4 | 25.6 | 2.88 | 73.22 | 20.62 | 2.88 | -1.18 | -4.98 |
| 1BFG | 38.9 | 9.5 | 51.6 | 39.87 | 5.59 | 52.05 | 0.97 | -3.91 | 0.45 |
| 4F5S | 0 | 73.6 | 26.4 | 0.26 | 78.27 | 22.61 | 0.26 | 4.67 | -3.79 |
| 1B09 | 44.7 | 6 | 49.3 | 47.04 | 5.90 | 47.10 | 2.34 | -0.10 | -2.20 |
| 1V9E | 29 | 14.3 | 56.8 | 32.91 | 8.09 | 58.95 | 3.91 | -6.21 | 2.15 |
| 3CNA | 40.5 | 0 | 59.5 | 35.94 | 6.07 | 56.37 | -4.56 | 6.07 | -3.13 |
| 2B4Z | 0 | 39.4 | 60.6 | 0.00 | 44.02 | 56.02 | 0.00 | 4.62 | -4.58 |
| 1V54 | 0.4 | 70.4 | 29.2 | 0.00 | 90.55 | 22.37 | -0.40 | 20.15 | -6.83 |
| 3CYT | 0 | 41.7 | 58.3 | 0.00 | 62.90 | 49.92 | 0.00 | 21.20 | -8.38 |
| 1HRC | 0 | 41.3 | 58.7 | 2.80 | 35.17 | 59.54 | 2.80 | -6.13 | 0.84 |
| 1QNJ | 34.2 | 10 | 55.8 | 40.85 | 0.18 | 59.63 | 6.65 | -9.82 | 3.83 |
| 3ENL | 16.5 | 42.5 | 41.1 | 18.76 | 30.62 | 44.61 | 2.26 | -11.88 | 3.51 |
| 1F13 | 39.8 | 14.4 | 45.8 | 35.70 | 7.58 | 56.06 | -4.10 | -6.82 | 10.26 |
| 3GHG | 1.7 | 70.6 | 27.6 | 0.00 | 68.27 | 35.54 | -1.70 | -2.33 | 7.94 |
| 1DPG | 34.2 | 46.4 | 42.3 | 33.59 | 44.86 | 43.28 | -0.61 | -1.54 | 0.98 |
| 1NS6 | 0 | 76.6 | 23.4 | 0.00 | 62.89 | 36.32 | 0.00 | -13.71 | 12.92 |
| 1HHO | 0 | 70.9 | 29.1 | 0.00 | 75.25 | 25.62 | 0.00 | 4.35 | -3.48 |
| 3QYT | 17.8 | 33.1 | 49 | 18.04 | 29.84 | 51.77 | 0.24 | -3.26 | 2.77 |
| 3H3F | 21.5 | 43.2 | 35.3 | 25.24 | 40.86 | 32.93 | 3.74 | -2.34 | -2.37 |
| 2LYZ | 6.2 | 41.1 | 52.7 | 1.08 | 53.41 | 44.89 | -5.12 | 12.31 | -7.81 |
| 2V1H | 0 | 74.5 | 25.5 | 10.83 | 63.40 | 25.49 | 10.83 | -11.10 | -0.01 |
| 1OVA | 32.1 | 32.1 | 35.8 | 29.64 | 40.45 | 31.10 | -2.46 | 8.35 | -4.70 |
| 4GER | 17.4 | 43.4 | 39.1 | 17.01 | 41.53 | 39.31 | -0.39 | -1.87 | 0.21 |
| 4AO1 | 34.7 | 20.1 | 45.2 | 30.64 | 20.63 | 47.17 | -4.06 | 0.53 | 1.97 |
| 3UNX | 17.9 | 29.6 | 52.6 | 18.19 | 27.80 | 56.38 | 0.29 | -1.80 | 3.78 |
| 1R2R | 15.4 | 44.3 | 40.2 | 20.61 | 30.07 | 47.16 | 5.21 | -14.23 | 6.96 |
| 1TGS | 37.8 | 8.4 | 53.8 | 47.34 | 0.00 | 62.90 | 9.54 | -8.40 | 9.10 |
| 1BA7 | 36.4 | 1.8 | 61.8 | 31.94 | 6.79 | 63.61 | -4.46 | 4.99 | 1.81 |
| 1S0R | 32.3 | 8.5 | 59.2 | 29.75 | 9.54 | 60.80 | 0.00 | 1.04 | 1.60 |

As seen from the table, the estimation accuracies we obtained are very good, and can be compared to those reported by previous works (Severcan et al. 2001; Severcan et al. 2004a; Khanmohammadi et al. 2009) for some proteins. The overall average rms

errors for α- helix, β-sheet and turns in our case were 3.12, 6.08, and 4.02 respectively. The artificial Neural networks (ANNs), as a method to predict protein secondary structure, has been developed in a previous work by our group that is able to provide predictions of secondary structure of proteins in solution better than previously used methods (Severcan et al. 2001; Hering et al. 2002a; Hering et al. 2004c; Severcan et al. 2004a; Hering and Haris 2009).The main difficulty in all of these approaches was the limitation of the available spectral data (18 proteins) for training of the NNs. In (Severcan et al. 2001), Bayesian regularization was used in order to train the ANN. Also, leave one out approach was used to show the applicability of the method. The networks have been tested and standard error of prediction (SEP) has been reported as 4.19% for α-helix, 3.49% for β-sheet, and 3.15% for turns have been achieved. Enhanced neural network by (Hering et al. 2002b; Hering et al. 2002a) revealed that by providing part of the amide I region in combination with appropriate pre-processing of the spectral data can produced a good prediction results. Their results showed a standard error of prediction with 6.16% for β-sheet, 4.47% for α-helix and 4.61% for turns. (Hering et al. 2003) showed that proteins can be accurately classified into two main classes "all alpha proteins" and "all beta proteins" merely based on the amide I band maximum position of their FTIR spectra using of specialized neural networks architecture combining an adaptive neuro fuzzy inference system. The standard errors of prediction (SEPs) in % structure were improved by 4.05% for α-helix structure, by 5.91% for sheet structure, by 2.68% for turn structure, and by 2.15% for bend structure. (Severcan et al. 2004b) produced an artificially proteins training dataset using linear interpolation in order to improve the generalization ability of the neural networks. The networks have been tested and standard error of prediction (SEP) of 4.19% for a α-helix, 3.49% for b sheet, and 3.15% for turns have been achieved. (Hering et al. 2004a) used a reference set composed of FTIR spectra recorded at different laboratories to investigate possible effects on prediction accuracy by neural network analysis. The SEP results show small difference between the datasets recorded at different laboratories suggests that FTIR may be safely combined into one reference set. (Severcan et al. 2001; Hering et al. 2002a; Hering et al. 2004c; Severcan et al. 2004a; Hering and Haris 2009) studies has been done on 18 proteins dataset. In order to increase the accuracy of ANN prediction, we increased the number of proteins in our

database to be 35 proteins. Up to best of our knowledge this is the first study uses that wavelets coefficient as a feature for amide I FTIR signal.

Table 5: Comparison between x-ray, NMR and ANN based protein secondary structure for some proteins in protein dataset.

| **X-ray based results** | | | | **NMR based results** | | | | **ANN based results** | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Pdb code | β-sheets | α-helix | Others | Pdb code | β-sheets | α-helix | Others | β-sheets | α-helix | Others |
| 1R2R | 15.4 | 44.3 | 40.2 | 1ypi | 16.2 | 43.0 | 40.9 | 20.61 | 30.07 | 47.16 |
| 2V1H | 0 | 74.5 | 25.5 | 1myf | 0 | 73.2 | 26.8 | 10.83 | 63.40 | 25.49 |
| 1HRC | 0 | 41.3 | 58.7 | 1akk | 0 | 37.5.3 | 62.5.7 | 2.80 | 35.17 | 59.54 |
| 1NS6 | 0 | 76.6 | 23.4 | 2h35 | 0 | 62.4 | 37.6 | 0.00 | 62.89 | 36.32 |
| 4AO1 | 34.7 | 20.1 | 45.2 | 1FS3 | 34.7 | 20.1 | 45.2 | 30.64 | 20.63 | 47.17 |

Some of proteins in our dataset have been studied using NMR, because of this we compared the secondary structure of these proteins based on x-ray, based on NMR and based on our ANN. The results indicated that for four proteins in our data set (indicated in Table 5), the secondary structure prediction based on NMR are almost similar to their proteins secondary structure prediction based on x-ray. However, for Hemoglobin (pdb:1NS6), which is large molecular weight protein, The NMR showed a closer result to our ANN based proteins secondary structure analysis. This could indicate that the large proteins have different secondary structure in solution than the crystal form. This show the importance of the proteins prediction in solution rather than in crystal form in order to reflect the real proteins structure in native form. Finally, our results show that ANN base proteins secondary structure method is very promising and can be used for proteins secondary structures estimation from FTIR spectra.

### 3.2 Biophysical characterization of proteins in artificial crowded environment

The previous FTIR studies of Transferrin in dilute solution indicate that upon iron binding the conformational changes occur at tertiary structure rather than the secondary structure (Hadden et al. 1994b). Thus, this makes the study of protein dynamics in solution very important to understand the properties of this protein. Inside the cell, the environment is not like a diluted solution but there is crowded environment of macromolecules such as proteins, carbohydrates and lipids (Fulton 1982; Minton

1983; Zimmerman and Trach 1991). This crowded environment could physically affect the properties and dynamics of transferrin (Harada et al. 2012). In the current study dextran and ficoll as a model crowding agents has been used because they are uncharged, inert polymers and FTIR transparent in amid I region (Samiotakis et al. 2009). Investigate the conformation of proteins can be deduced from FTIR spectra because it is a very well technique for the analysis of the amide peaks thus the denaturation and aggregation can be monitoring (Haris and Severcan 1999; Severcan et al. 2001). The amide I band has been used extensively used to determine the secondary structure of proteins (Dong et al. 1990; Dong et al. 1995; Haris et al. 2004). Dextran has been used to produce artificial crowded environment like crowded environment inside the cells to study the transferrin. In this study thermal denaturation, aggregation and Hydrogen deuterium exchange of transferrin in the presence and absence of dextran have been studied using deconvoluted, second derivative and two-dimensional infrared correlation spectroscopy (2D-IR). In order to generate 2D-IR analysis, thermal perturbation has been used as time dependent variations in infrared spectra. Our results present a direct evidence of the molecular crowding agents on the aggregation and dynamics of Apo and Holo transferrin.

### 3.2.1 1D-IR spectroscopy

#### 3.2.1.1 Effect of molecular crowding on HTF and ATF secondary structures:

Figure 23A shows Fourier deconvolved infrared spectra of amide I and amid II regions for HTF, ATF, HTF in Dextran (200 mg/ml and 400mg/ml), HTF in Ficoll (200 mg/ml and 400mg/ml), ATF in Dextran (200 mg/ml and 400mg/ml) and ATF in Ficoll (200 mg/ml and 400mg/ml) all of which are in phosphate buffered (pH 7.4) saline solution at 25°C. The second derivative spectra (Figure 23B) of all spectra in figure 1A unraveled three bands; a dominant band at 1654 $cm^{-1}$ and two smaller bands at 1682 $cm^{-1}$ and 1635 $cm^{-1}$. Figure 23A and B also show that the Dextran and Ficoll with concentrations (200mg/ml and 400mg/ml) don't have an effect on the HTF and ATF secondary structure at room temperature. The band at 1656 $cm^{-1}$ is characteristic of the protein amide groups in alpha-helices or irregular structure while the bands at 1685 cm-1 and 1634 cm-1 are generally associated with turns and β-sheets, respectively

(Byler and Susi 1986). The quantitative analysis of HTF and ATF using a method described at (Yan et al. 2006) has been calculated using a custom written Matlab script. The secondary structure results showed 35% α-Helix and 22 % ß-sheets. These results were in good agreement with those obtained by X-ray crystal structure of HTF (pdb code: 3qyt) which shows 33.1% α-Helix, 17.8% ß-sheets calculated using pdbsum bioinformatics tool (Chang et al. 2015). Furthermore, our Transferrin secondary structure calculations were in good agreement with a previous FTIR study on Transferrin (Hadden et al. 1992).

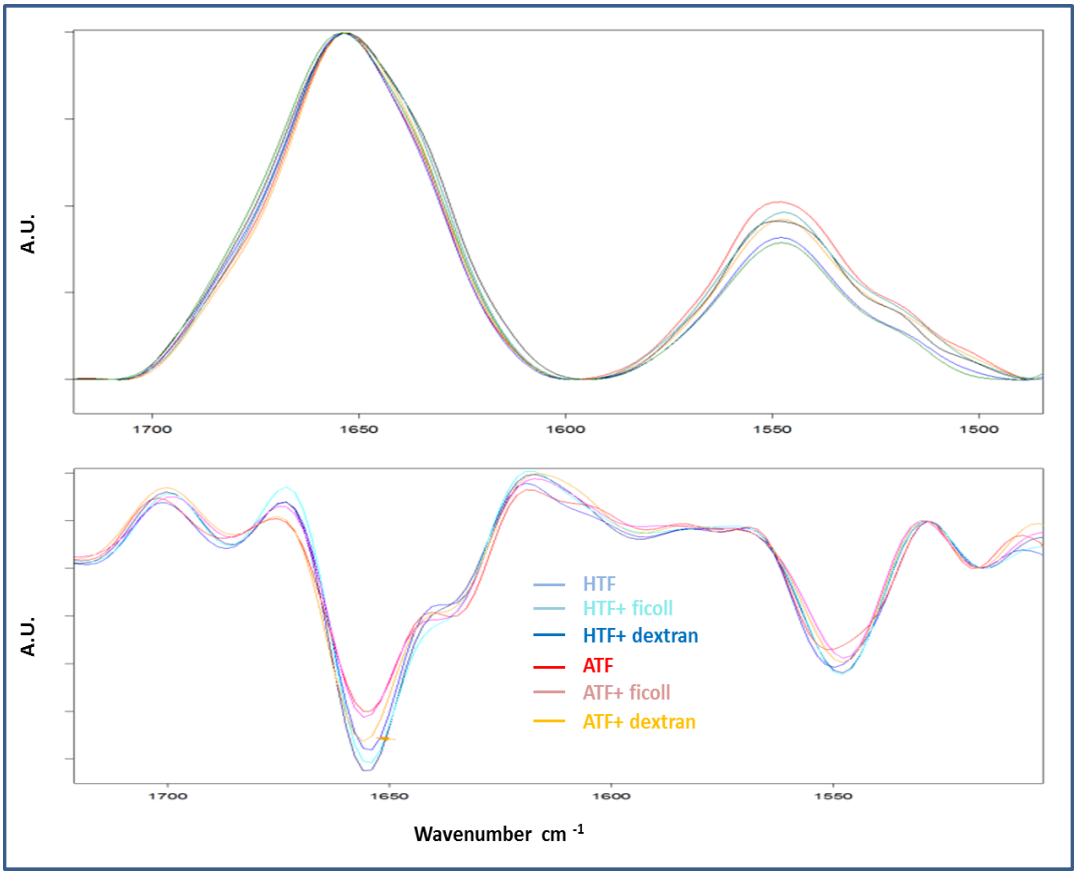


Figure 23: : Original (A) and second derivative (B) FTIR spectra for Transferrin in the amide I and II region at 30°C.

According to the analysis of ATF and HTF FTIR spectra at room temperature, the secondary structure analysis of transferrin did not show significant changes up on binding of iron. This result was in agreement with a previous small-angle neutron scattering study (Martel et al. 1980) which indicated that iron binding alters the relative position of the two lobes of the transferrin molecule rather than the structure of the lobes. The study also suggested that iron binding may result in some kind of twisting of the two lobes relative to each other.

### 3.2.1.2 Effect of molecular crowding on H/D exchange:

Figure 24A, shows the variation of the amide II peak absorption of HTF in the absence and presence of dextran at 25 ºC. These spectra were normalized with respect to the amid I band. The results obtained indicated a large decrease in H/D exchange rate in the presence of dextran.

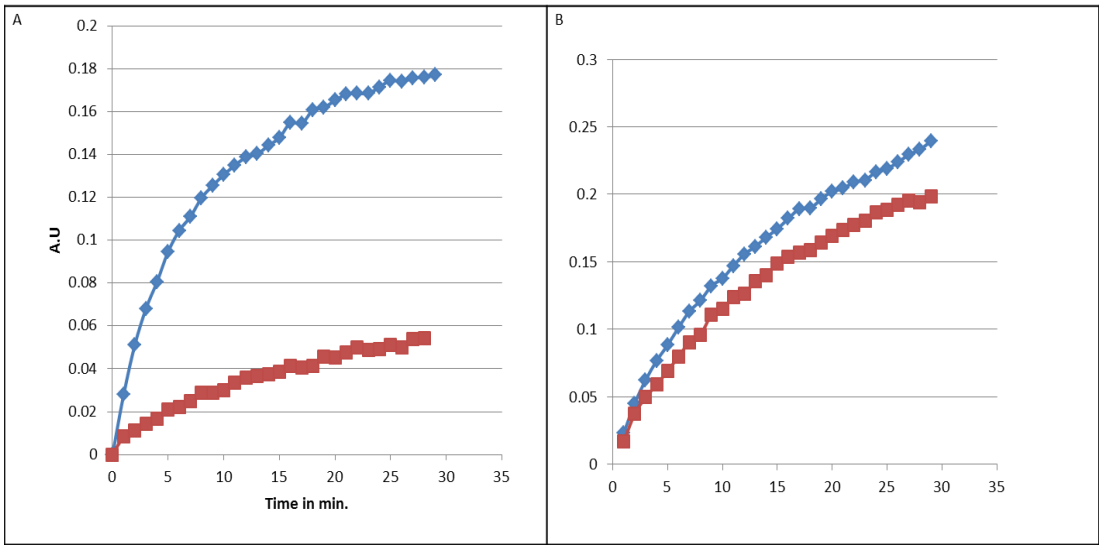


Figure 24: Time-dependent $^{1}H$-$^{2}H$ exchange of Amide II intensity for Transferrin in absence (blue) and presence (red) of dextran at 25 C. A) HTF only B) ATF only.

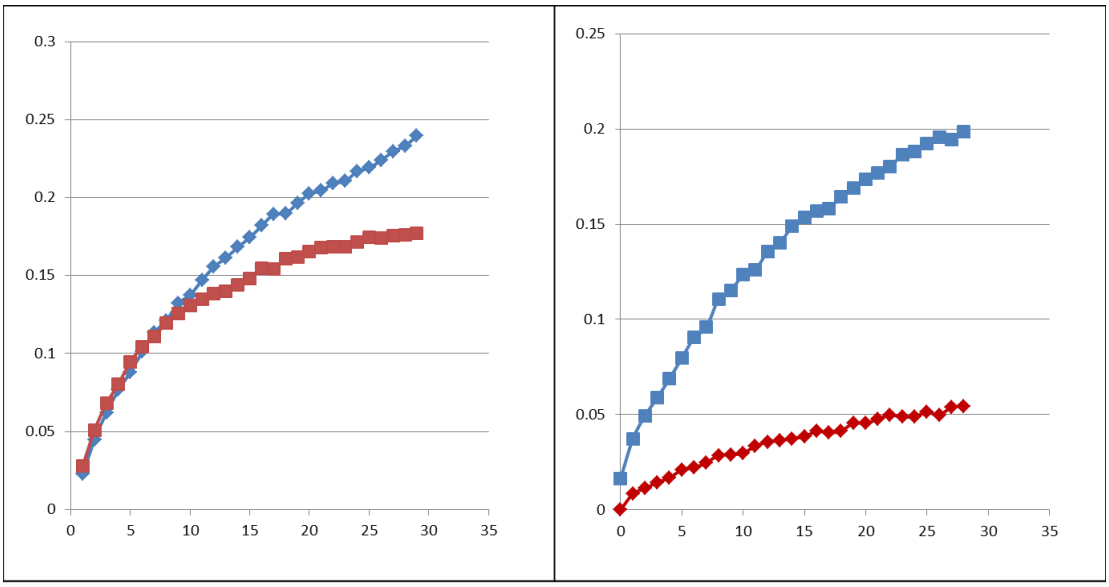


Figure 25 : Time-dependent $^{1}H$-$^{2}H$ exchange of Amide II intensity for ATF (blue) and HTF (red) at 25 C. a) diluted solution only b) in dextran solution.

The fraction of unexchanged amide protons (F) has been calculated as described in method section. After 1 hour of dissolving HTF in $H_2O$ buffer at 25 C, about 63% of amide protons in HTF were exchanged in contrast to only 21% exchange in the presence of dextran. This suggests over 3-fold reduction in hydrogen-deuterium exchange of the amide protons in HTF the presence of dextran. This could be due to an increase in the compactness of the HTF structure induced by the molecular crowding effect of the dextran molecules. The more rigid/compact structure prevents the replacement of H with $^{2}H$ within the peptide groups due either to an increase in H-bonding within the peptide groups and/or due to the inability of the D2O to penetrate the interior core of the HTF molecule.

Figure 24B shows the variation of the amide II absorption at 25 C for ATF in the absence and presence of dextran in D2O buffer (spectra normalized to amide I). The fraction of unexchanged amide protons has been calculated by the same method described in the methods section. After 1 hour of dissolving ATF in D2O buffer at 25 C, about 52% of the amide protons undergo exchange. In contrast, only 43% of the amide protons underwent exchange in the presence of dextran. The magnitude of difference in exchange was far less compared to what was seen for HTF. This suggested that the impact of molecular crowding on ATF was far less compared to

what was seen for HTF. The hydrogen deuterium exchange (H/D) results of ATF and HTF in $D_2O$ Figure 25A shows similar exchange rate at the beginning time but a decrease was seen subsequently for HTF relative to ATF. The explanation for the rapid exchange at the beginning can be attributed to the exchange of amide protons located on the surface and solvent accessible region of both ATF and HTF. The slower exchange rate corresponded to the exchange of buried solvent in-accessible residues and/or residues that were strongly hydrogen-bonded. But in case of HTF the closed structure lowered the rate of exchange. Although part of this results was similar to the previous study (Hadden et al. 1994a) to the best of our knowledge first time H/D exchange rate was used to confirm the open and closed structure of transferrin. Figure 25B showed that the H/D exchange rate of ATF and HTF in dextran solution significant which indicates the physical effect of dextran on the solvent accessibility of the transferrin molecule. This (Figure 25B) confirmed that in crowded environment, the physical properties of confined water differed considerably from those corresponding to bulk water and effect on transferrin dynamics and stability (Despa et al. 2004). We observed that the amount of H/D exchange rate decrease in HTF was higher than ATF in presence of dextran. This could be attributed to the open structure form of ATF that allowed more exchange from inside the transferrin rather than the closed structure of HTF.

#### 3.2.1.3 Effect of molecular crowding on HTF and ATF thermal denaturation:

Figure 26A shows the amide I band of HTF at 25C and 90C in dilute, dextran crowded and ficoll crowded solutions. Figure 4B show the second derivative of Figure 26A spectra. The main absorption peak of HTF at 25C is 1656 $cm^{-1}$ strongly shifted to 1647 $cm^{-1}$ at 90C. However, in the presence of dextran the main absorption peak of HTF shifted to 1653 $cm^{-1}$ at 90C.

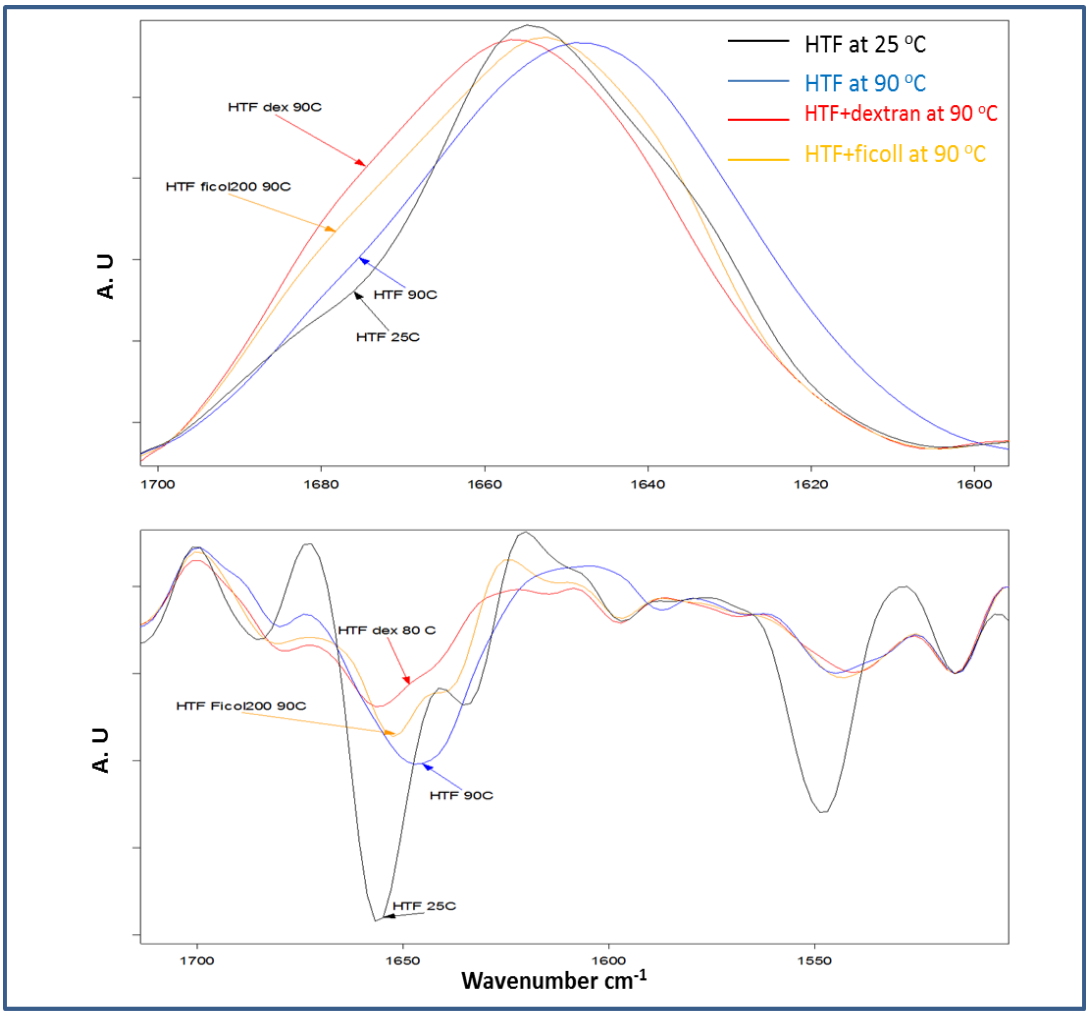


Figure 26: Amide I FTIR absorption spectra (A) and their second derivative (B) of HTF at room temperature (black) and (HTF (blue), HTF+dextran (red) and HTF+ficoll (orange) ) at 90 ºC.

In the presence of Ficoll the main absorption peak shifted only to 1655 $cm^{-1}$. The less shift (1-3 $cm^{-1}$) of the 1656 $cm^{-1}$ peak in the presence of Dextran and Ficoll at 90 C respectively may indicate the protective effect on the HTF thermal stability due to macromolecules crowded environment (Kuznetsova et al. 2014). Commonly, the protein aggregation is shown in the IR spectroscopy by the formation of two peaks at ;1618 and ;1681 $cm^{-1}$. These two peaks was previously assigned to hydrogen bonded extended intermolecular β-sheet structures. This extended intermolecular β-sheet structures was formed upon aggregation of the thermally denatured proteins (Yan et al. 2003a). However, the denaturation of HTF can be observed in the infrared spectra

by a broadening of Amid I, decrease in 1656 $cm^{-1}$ band intensity and formation of 1645 $cm^{-1}$ and 1676 $cm^{-1}$ peaks (Figure 26 A, B). This uncommon behavior of HTF has been reported before in a previous FTIR study (Hadden et al. 1994b). In the presence of Dextran and Ficoll, the HTF denaturation was observed by a broadening of amid I, a decrease in 1656 $cm^{-1}$ band and a formation of 1616±2 $cm^{-1}$, 1638±2 $cm^{-1}$ and 1676±2 $cm^{-1}$ peaks rather than 1645 $cm^{-1}$ peak. Similar results have been obtained by using Ficoll rather than Dextran at the same conditions.

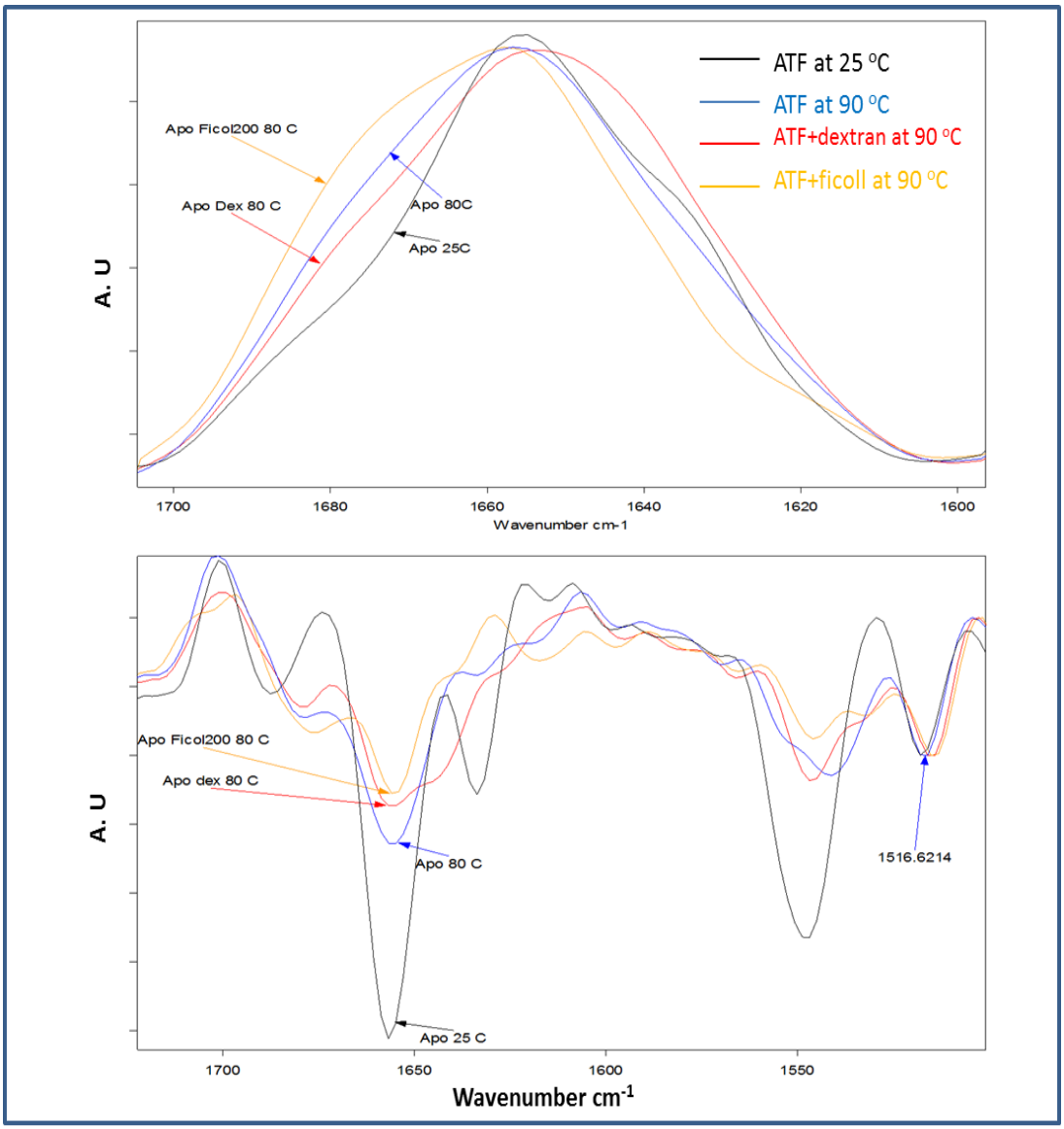


Figure 27: Amide I FTIR absorption spectra (A) and their second derivative (B) of ATF at room temperature(black) and (ATF(blue), ATF+dextran (red) and ATF+ficoll (orange) at 90 °C.

This can clearly prove that the protective effect on HTF thermal stability was due to the physical effect of the macromolecules rather than a special case due to chemical interaction between HTF and Dextran or Ficoll which has been suggested by several previous studies (Chebotareva et al. 2004; Chebotareva 2007). Similarly; ATF denaturation was reflected in the infrared spectra and its second derivative mainly by a broadening of amid I and formation of 1680 cm-1 and 1616 peaks (Figure 27A, B). In the presence of Dextran and Ficoll, the ATF denaturation was reflected by a broadening of amid I, a decrease in 1656 $cm^{-1}$ and 1634 $cm^{-1}$ bands intensity and a formation of 1680 $cm^{-1}$, 1643±2 $cm^{-1}$ and 1623±3 $cm^{-1}$ peaks. These results again indicated the effect of macromolecules on the formation of two different secondary structure of ATF in diluted and crowded environments.

### 3.2.2 2D-IR Correlation:

In order to gain more insight about the variation of HTF and ATF secondary structure with temperature in the presence and absence of both dextran and Ficoll, we performed 2D-IR correlation analysis. 2D-IR method allows to study the response of a protein on different physical-chemical impulses with use of conventional spectrometers (Yan et al. 2003b). The main 2D-IR concept is that; the peaks that undergo changes in intensity result in correlation peaks in the 2D plots, whereas the peaks that remain constant result in no or very small correlation peaks (Noda et al. 1988).

In the synchronous 2D-IR plot of HTF deconvoluted spectra (Figure 28A), four autocorrelation peaks (along the diagonal) centered at 1624, 1642, 1657, 1674 $cm^{-1}$ were observed. , indicating that the relative intensities of these bands changed with temperature. The negative cross-correlation peaks were observed at 1642 vs 1657 $cm^{-1}$ and at 1624 vs 1657 $cm^{-1}$, revealing that the unfolding of alpha-helices (1657 $cm^{-1}$) is accompanied by formation of two peaks at (1642 $cm^{-1}$ and 1624 $cm^{-1}$) which indicated the HTF protein aggregation. The positive cross-correlation peak at (1642 $cm^{-1}$ vs 1624 $cm^{-1}$) confirms that these two peaks were formed together upon HTF denaturation.

The asynchronous 2D IR plot was antisymmetric with respect to the diagonal line. Asynchronous plot had no autopeaks at the diagonal position and consisted only of cross peaks located at off-diagonal positions. In 2D-IR figures, each peak can be

denoted by its y-axis value versus its x-values such as (1656 vs 1630 $cm^{-1}$). Figure 28B shows HTF asynchronous 2D IR plot. The figure shows four positive peaks at (1674 vs 1624, 1674 vs 1640, 1685 vs 1674 and 1640 vs 1630 $cm^{-1}$). There was also one strong negative peak located at (1674 vs 1659 $cm^{-1}$) and another two weak peaks at (1657 vs 1630 and 1630 vs 1620 $cm^{-1}$). The position and sign of the cross peaks of an asynchronous 2D correlation spectrum reveal useful information about the relative temporal relationship or order of the actual sequence of reorientation. The positive peaks at (1674 vs 1624 $cm^{-1}$ and 1674 vs 1640 $cm^{-1}$) clearly indicate that the weak structures at 1674 cm-1 is highly affected by temperature and lead the formation of the aggregate structure at 1624 and 1674 $cm^{-1}$. The negative peak at (1674 vs 1659 $cm^{-1}$) shows that the rate of change in the peak intensity at 1674 is slower than the change of 1659 $cm^{-1}$.

In the presence of Dextran (Figure 28C) the two autocorrelation peaks at 1624 and 1657cm-1 were disappeared indicating that the presence of dextran prevents the unfolding of α-helix structure and formation of 1624 cm-1 aggregate structure. However in the presence of Ficoll (Figure 28E) the weak 1657 $cm^{-1}$ peak indicated a weak effect of the temperature on the α-helix structure. The shift of 1624 cm-1 autocorrelation peaks to 1630 $cm^{-1}$ was observed in the presence of either Dextran or Ficoll indicating that the presence of the macromolecules has an effect on the final aggregated structure. The asynchronous 2D-IR plot of HTF in the presence of Dextran (Figure 28D) showed a high similarity to the asynchronous 2D-IR plot of HTF only. However the asynchronous 2D-IR plot of HTF in the presence of Ficoll (Figure 28F) showed some differences such as disappear once of the negative peak at 1674 vs 1659 cm-1 and appearance of strong positive peak at 1640 vs 1624 cm-1. From this we can conclude that although both Dextran and Ficoll show a protective effect for the alpha α-helix structure, the Ficoll does affect the rate of the peaks intensity changes.

In the synchronous 2D-IR plot of ATF deconvluted spectra (Figure 29A), two autocorrelation peaks (along the diagonal) centered at 1634 and 1678 were observed, indicating that the relative intensities of these bands change with increasing temperature. The absence of the 1656 autocorrelation peak indicated that the attachment of Iron to Transferrin (ATF) make it stronger than HTF and prevent the complete unfolding of the alpha α-helix structure.

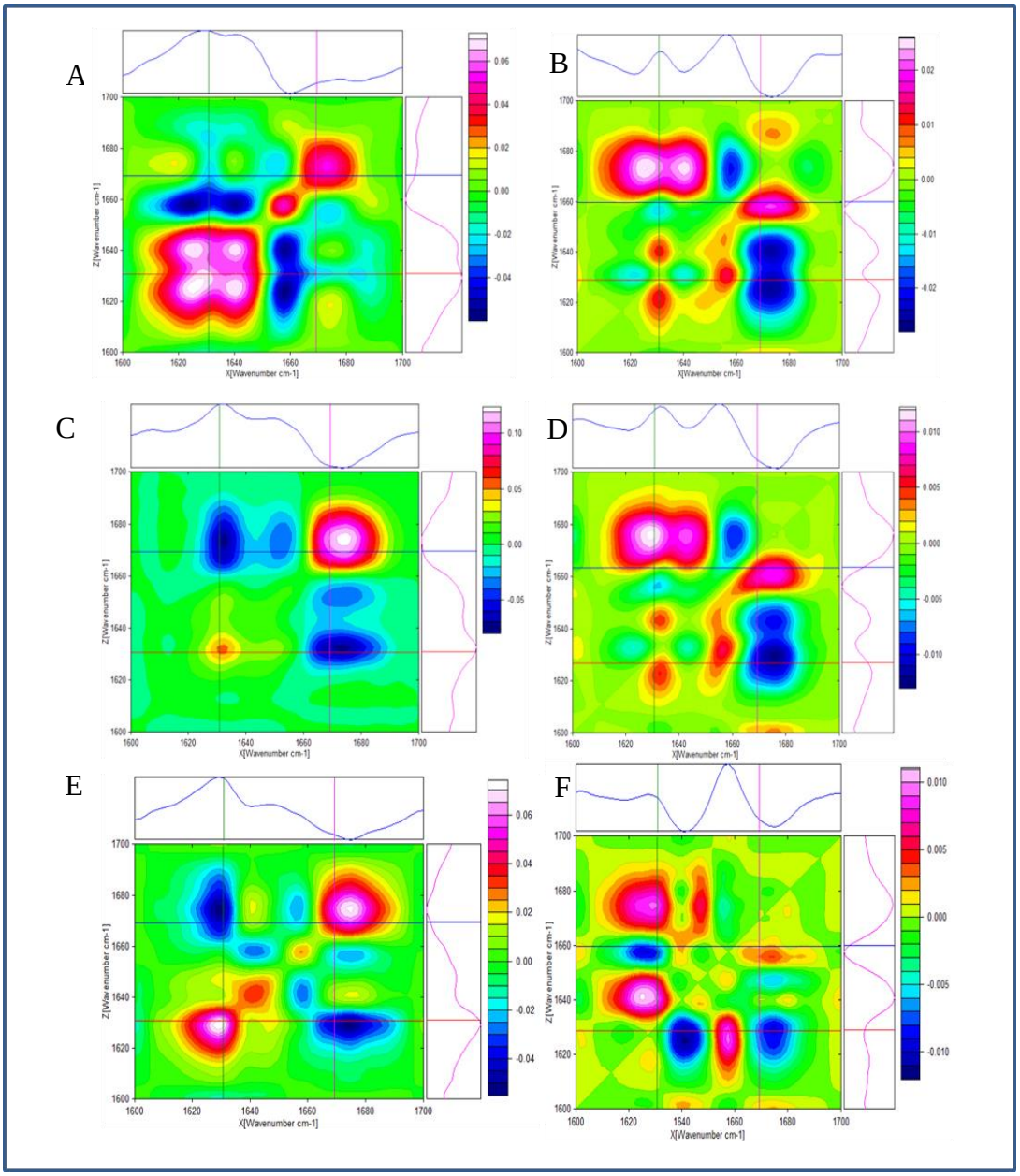


Figure 28: 2D IR (A, C, E) synchronous and (B, D, F) asynchronous plots of the amide I FTIR spectra of HTF (A,B), HTF+Dextran (C,D) and HTF+Ficoll (E,F).

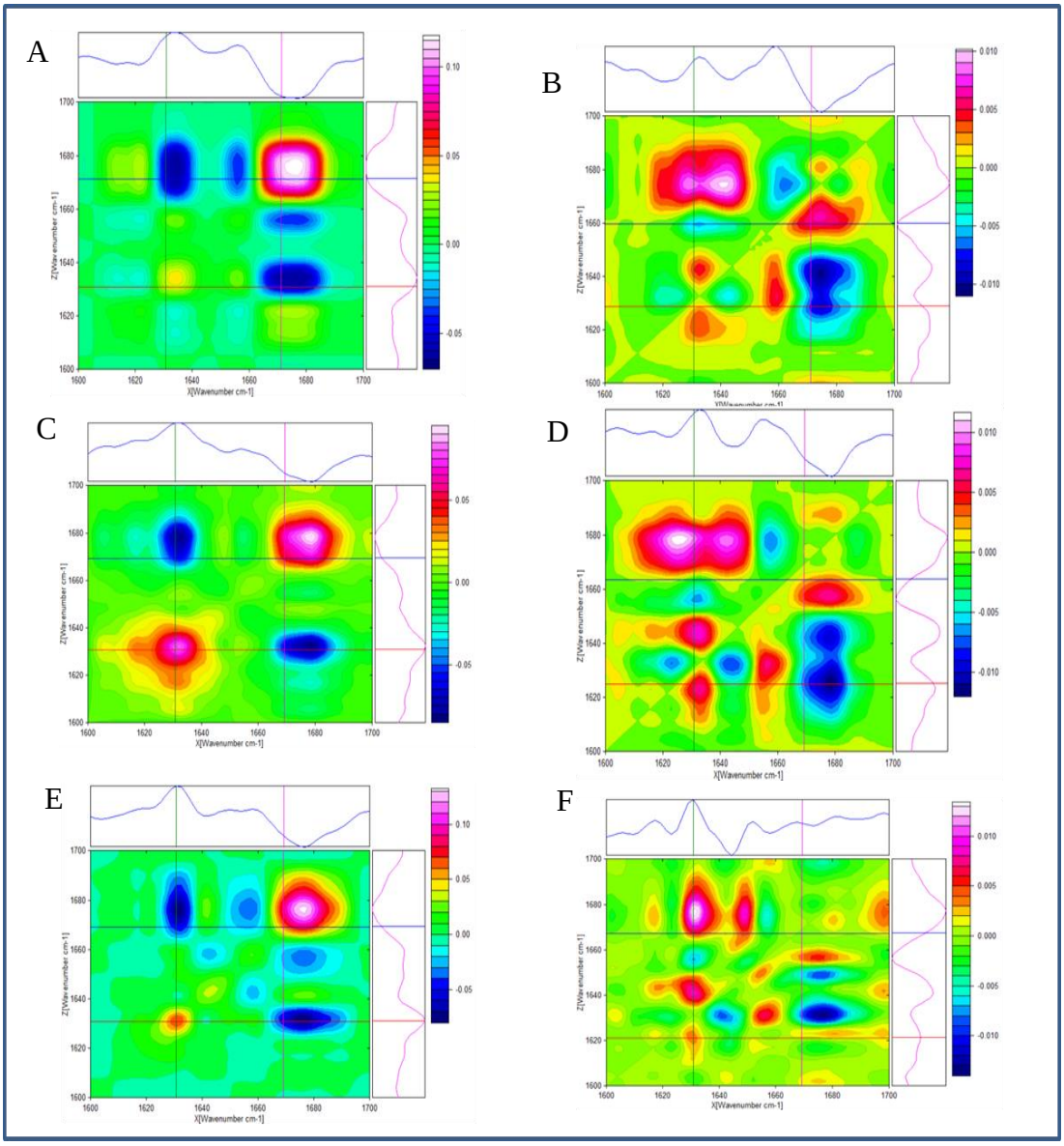


Figure 29: 2D IR (A, C, E) synchronous and (B, D, F) asynchronous plots of the amide I FTIR spectra of : ATF (A,B), ATF+Dextran (C,D) and ATF+Ficoll (E,F).

In ATF two negative cross-correlation peaks were observed at 1674 vs 1656 cm-1 and at 1674 vs 1634 cm-1, revealing that the partial unfolding of alpha-helices (1656 cm-1) were accompanied by protein aggregation (1674 and 1634 cm-1). In ATF asynchronous 2D IR plot the negative elongated peak at (1674 vs 1645 cm-1) indicated that the variation of weak structure at 1674 lead the formation of aggregated structure at 1645 and suggested that additional peaks were buried underneath. Figure 29B shows

ATF asynchronous 2D IR plot. The figure showed four positive peaks (similar to HTF) at (1674 vs 1628, 1674 vs 1640, 1682 vs 1674 and 1642 vs 1634 $cm^{-1}$). There was also one strong negative peak located at (1674 vs 1659 cm-1) and another two weak peaks at (1660 vs 1630 and 1630 vs 1622 cm-1). In the presence of Dextran (Figure 29C), the negative peak at 1657 vs 1684 cm-1 was disappeared indicating that the presence of dextran there was no more correlation between the unfolding of α-helix structure and formation of 1684 cm-1 aggregate structure. However, the autocorrelation peak at 1630 became very strong indicating a formation of large aggregated structure at 1634. In the presence of Ficoll (Figure 29C) the ATF asynchronous 2D IR plot showed one more negative peak at (1643 vs 1657) indicating that the presence of Ficoll induced the unfolding of α-helix to an aggregated structure at 1643 cm-1. The asynchronous 2D-IR plot of ATF in the presence of Dextran (Figure 29D) showed a high similarity to the asynchronous 2D-IR plot of HTF only. However, the asynchronous 2D-IR plot of HTF in the presence of Ficoll (Figure 29F) showed some differences. Similar to the situation of HTF we can conclude that although both Dextran and Ficoll showed a protective effect for the α-helix structure, the Ficoll did affect the rate of the peaks intensity changes.

Finally, the effect of macromolecular crowding on protein thermal stability can be explained based on two types of interactions which are volume exclusion and soft interactions (non-specific chemical interactions) [24]. Volume exclusion decrease the space available to the protein under study thereby it can increase protein stability as shown from our results. On the other hand, soft interactions can be destabilizing stabilizing [24]. This can explain the few differences between the effect of Dextran and the effect of Ficoll on the aggregated structure of the HTF and ATF.

## 3.3 Proteins secondary structure analysis of pleural fluids and its application in the diagnosis of MPM.

The asbestos-induced lung cancer or Malignant pleural mesothelioma (MPM) arises due to occupational and/or environmental exposures to asbestos for a long time. The main symptom of MPM is the accumulation of pleural fluid around the lungs. However, the accumulation of pleural fluid could be due to other kinds of diseases such as inflammation. Therefore, pleural fluid may contain some plasma proteins and also it can contain some proteins derived from the tissues and cells in the lung. The investigation of whether these proteins can be used as biomarkers for diagnosis of lung cancer and other diseases was shown by (Tyan et al. 2005).

In order to identify whether the pleural fluid is due to MPM or other benign diseases, the details of pleural fluid's proteins content and their secondary structure has been studied using Attenuated Total Reflection Fourier Transform Infrared (ATR-FTIR). In this study, IR spectroscopy is used to develop a non-invasive, operator independent diagnostic method for MPM from pleural fluid samples with the assist of Wavelet based ANN and Chemometrics analysis techniques. Pleural fluids spectra were collected in the wavenumber range from 4000 to 650 $cm^{-1}$.

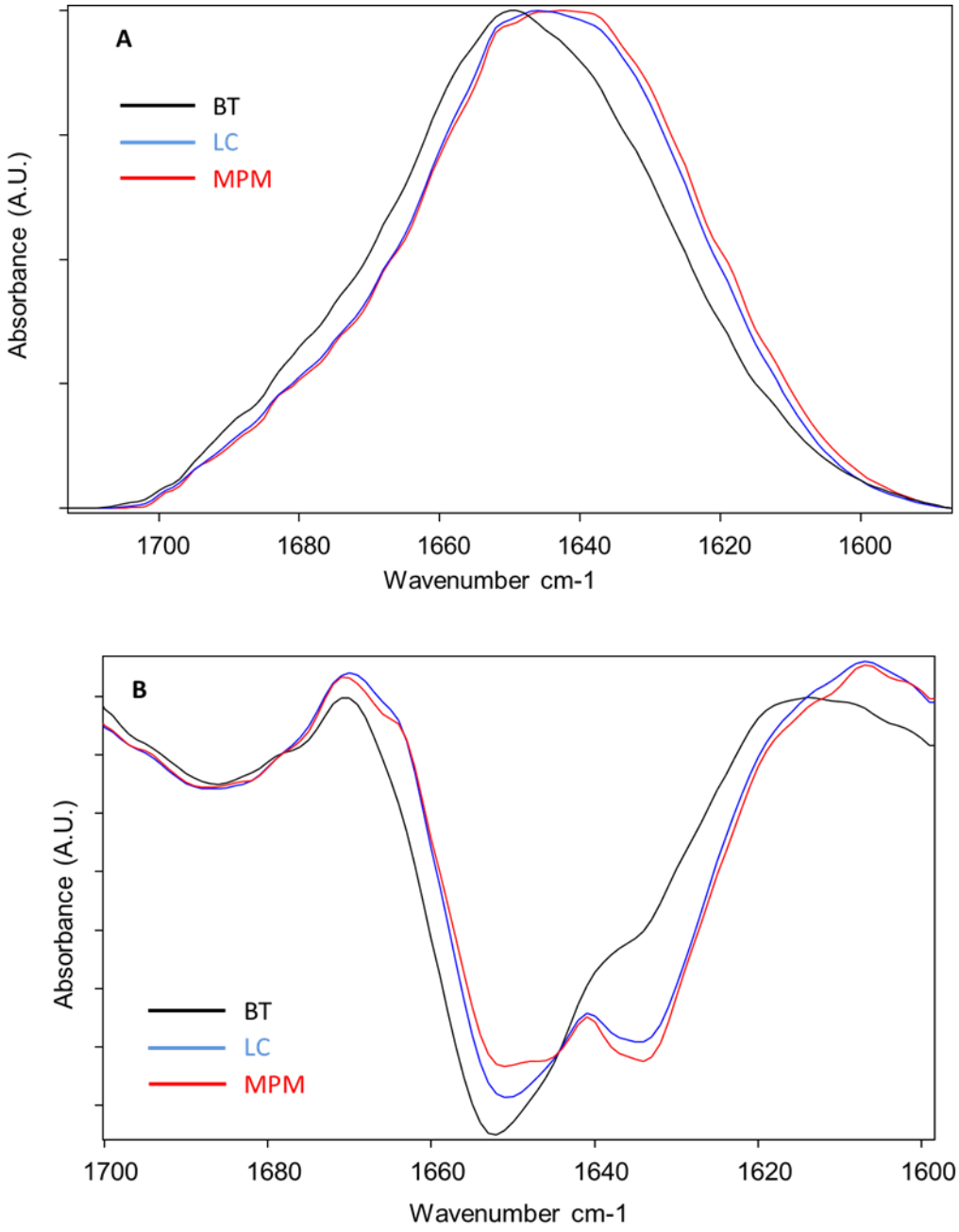


Figure 30 A and B show representative absorbance and their second derivative spectra for BT, LC and MPM, respectively of pleural fluids in amide I band (1700-1600cm$^{-1}$).

As it can be seen from (Figure 30), the amide I of pleural fluid spectra contain different characteristics peaks overlapped on each other. As can be deduced form Figure 30B, there are obvious alterations between the pleural fluid groups. For example, in BT group there is an increase in the intensities of peak located at 1656 $cm^{-1}$, which indicate the α- helix content, relative to the other groups. In absorption spectroscopy, according to the Beer-Lambert law, the increase in the intensity of the spectral band indicates increased concentration of this secondary structure content (Severcan and Haris 2012). Therefore, increased intensity in 1656 $cm^{-1}$ bands implies increment in α-helix content of BT pleural fluid samples. Similarly, in MPM group the increase in 1634 $cm^{-1}$ band indicated an increase in β-sheets content of this group. As supporting to our findings, the increase in the soluble mesothelin related protein content in pleural fluid of mesothelioma patient has been previously shown in other studies (Creaney et al.).

In order to calculate the proteins secondary structure of all pleural fluids samples, we used the ANN approach which we developed based on our proteins database (part 1 of this thesis). The secondary structure results of all samples are summarized in (Table 6).

Table 6: Proteins secondary structure analysis results of pleural fluids using ANN

| Sample No. | BT | | | LC | | | MPM | | |
|---|---|---|---|---|---|---|---|---|---|
| | Sheets | Helix | Others | Sheets | Helix | Others | Sheets | Helix | Others |
| 1 | 12.48 | 39.78 | 49.92 | 31.2 | 26 | 42.8 | 44.1 | 5.2 | 47.33 |
| 2 | 13.26 | 38.11 | 62.41 | 31.62 | 21.84 | 46.54 | 44.44 | 6.18 | 46.57 |
| 3 | 5.15 | 42.23 | 39.21 | 34.68 | 34.96 | 30.36 | 46.2 | 11.44 | 44.2 |
| 4 | 15.3 | 40.8 | 53.75 | 57.54 | 26.25 | 16.21 | 43.26 | 2.15 | 44.94 |
| 5 | 15.3 | 42.64 | 45.36 | 32.55 | 26.25 | 41.2 | 24.44 | 10.2 | 50.02 |
| 6 | 12.36 | 38.48 | 26.73 | 30.16 | 25.5 | 44.34 | 43.26 | 27.07 | 23.44 |
| 7 | 16.64 | 40.95 | 42.32 | 49.87 | 24.48 | 25.65 | 48.3 | 14.42 | 44.91 |
| 8 | 23.39 | 60.92 | 66.38 | 28.56 | 24.72 | 46.72 | 44.88 | 9.18 | 46.06 |
| 9 | 7.85 | 46.35 | 37.19 | 29.29 | 25.75 | 44.96 | 27.94 | 19.55 | 50.69 |
| 10 | 14.7 | 43.86 | 45.46 | 40.16 | 24.72 | 35.12 | 34.88 | 6.24 | 53.86 |
| 11 | 12.12 | 43.68 | 50.02 | 26.78 | 23.69 | 49.53 | 43.05 | 5.1 | 78.87 |
| 12 | 34.28 | 45.32 | 66.86 | 33.99 | 26.25 | 39.76 | 18.3 | 5.2 | 47.3 |
| 13 | 14.28 | 46.35 | 61.08 | 42.32 | 25.25 | 32.43 | 47.47 | 4.04 | 48.24 |
| 14 | 17.34 | 43.43 | 51.9 | 26.4 | 21.84 | 51.76 | 42.42 | 5.05 | 44.71 |
| 15 | 6.48 | 29.78 | 68.05 | 30.9 | 44.24 | 42.85 | 44.1 | 13.65 | 44.92 |
| 16 | 23.52 | 40.56 | 48.33 | 33.66 | 24.15 | 21.11 | 59.46 | 4.2 | 46.51 |
| 17 | 25.44 | 38.85 | 52.38 | 6.52 | 23.69 | 69.79 | 46.35 | 5.2 | 53.65 |
| 18 | 14.7 | 50.95 | 27.82 | 13.33 | 21.63 | 65.04 | 56.2 | 28.08 | 55.56 |
| 19 | 19.64 | 13.86 | 33.26 | 31.62 | 12.05 | 56.33 | 41.82 | 7.28 | 50.44 |
| 20 | 10.1 | 43.43 | 49.49 | 29.12 | 24.72 | 46.16 | 45.15 | 4.16 | 56.02 |
| 21 | 30.1 | 45.32 | 40.66 | | | | 29.35 | 3.14 | 49.36 |
| 22 | 16.16 | 43.43 | 47.78 | | | | 47.84 | 10.4 | 47.18 |
| 23 | 10.5 | 20.56 | 42.73 | | | | 40.46 | 12.12 | 48.7 |
| 24 | 15.75 | 43.43 | 52.57 | | | | 41.74 | 18.08 | 49.88 |
| 25 | 10.5 | 41.82 | 42.8 | | | | | | |
| 26 | 14.84 | 43.42 | 52.47 | | | | | | |

The evaluation of the usefulness of pleural total protein measurements for differentiating between exudates and transudates has been shown by (Patel and Choudhury). They showed that the total proteins can be used to differentiate between BE and BT and the results were comparable to Light's criteria methods. In this study and as shown from

Figure 31A, the β-sheets content of LC and MPM was significantly higher relative to control group BT. Figure B shows a significant decrease the α- helix structure in LC and MPM relative to the control group BT. These results of α- helix and β-sheets structure content of the 3 groups were in agreement with the secondary structure prediction using the second derivative analysis method.

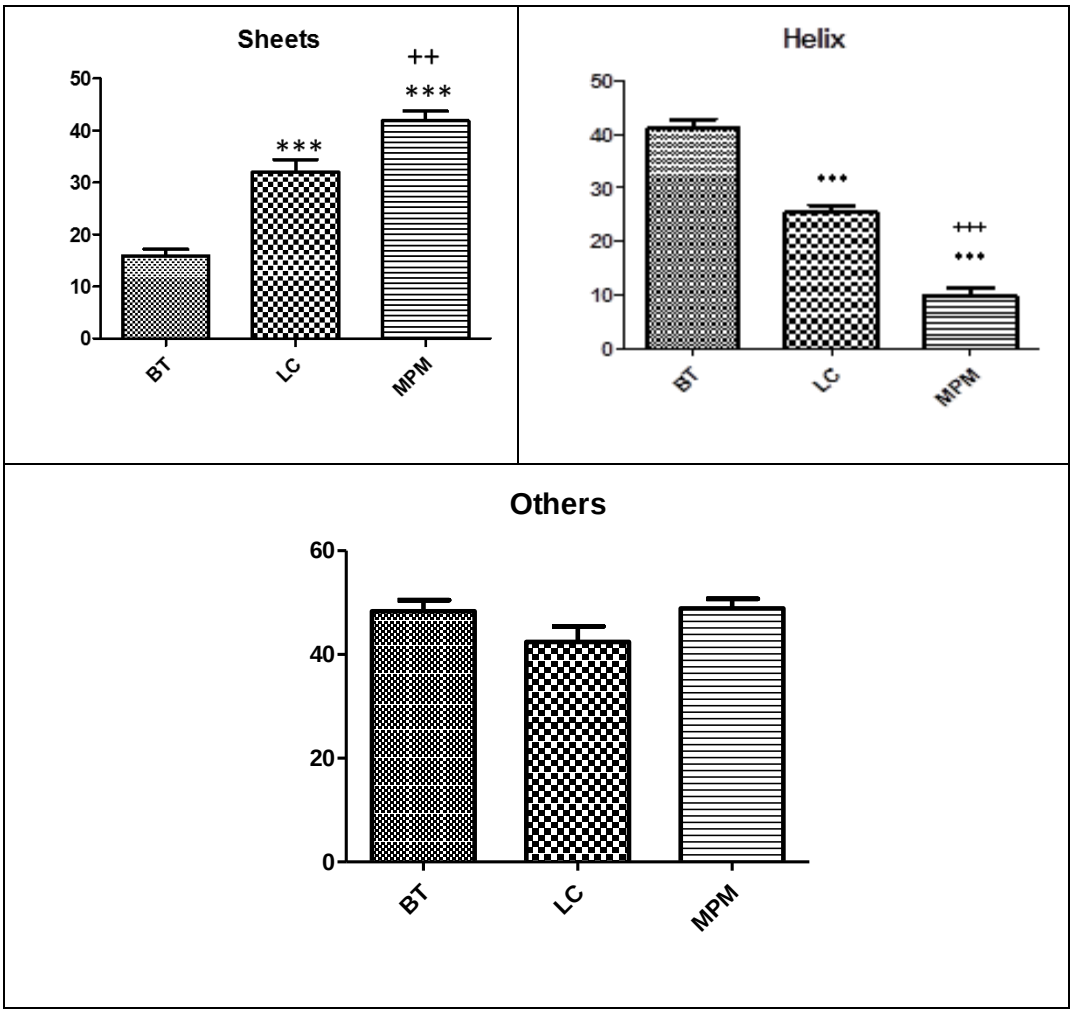


Figure 31: Means with Standered Errors of Means (SEMs) for each group. showing t-test statistical analysis between the studied groups. A) B-sheets, B) β- helix C) others proteins secondary structure of pleural fluids.

The change in proteins secondary structure may be due to the expression of new proteins in the pleural fluids such as methoselin in case of MPM group.

Figure 31 shows a detailed descriptive analysis of each secondary structure components. In a previous study (Hsieh et al. 2006), the proteomic profiles of 14 malignant and 13 transudate pleural effusions were studied using two-dimensional gel electrophoresis. ELISA and Western immunoassay studies showed that pigment epithelium-derived factor levels (mainly α- helix protein) were significantly increased in BT than in MPM. This could explain our finding that the α- helix structure in MPM was much more less than in BT group as shown in

Figure 31. The C-reactive (mainly β-sheets protein) levels in pleural fluid were significantly lower in the BT group (Yilmaz Turay et al. 2000). Again this confirms our finding that the β-sheets content of MPM is higher in comparison to BT group. (Paramothayan and Barron 2002) showed that measurement of fluid LDH values and the calculation of fluid to serum total protein ratios will aid in differentiating exudates from transudates. In benign exudate the level of LDH (mainly α- helix protein) was higher than BT group (Paramothayan and Barron 2002). This again confirm our finding that the α- helix content of benign exudate is higher than the α- helix content of BT group.

To extract meaningful data from complex data such as that of protein secondary structure of pleural fluids, advanced chemometric analysis approaches including unsupervised and supervised methods are required. Therefore in the present study , in order to accurately characterize and identify the differences in proteins secondary structure pleural fluid, firstly unsupervised chemometric analysis such as HCA and PCA were employed to the constituents of the BT, LC and MPM groups. For unsupervised approach no priori information about studied samples is required so samples are clustered into a number of classes based on their similarity degree. Therefore, different classes of samples can be identified easily. Figure 32 demonstrates HCA results of BT, LC and MPM groups achieved for the amide IR spectral region (1700-1600 $cm^{-1}$).

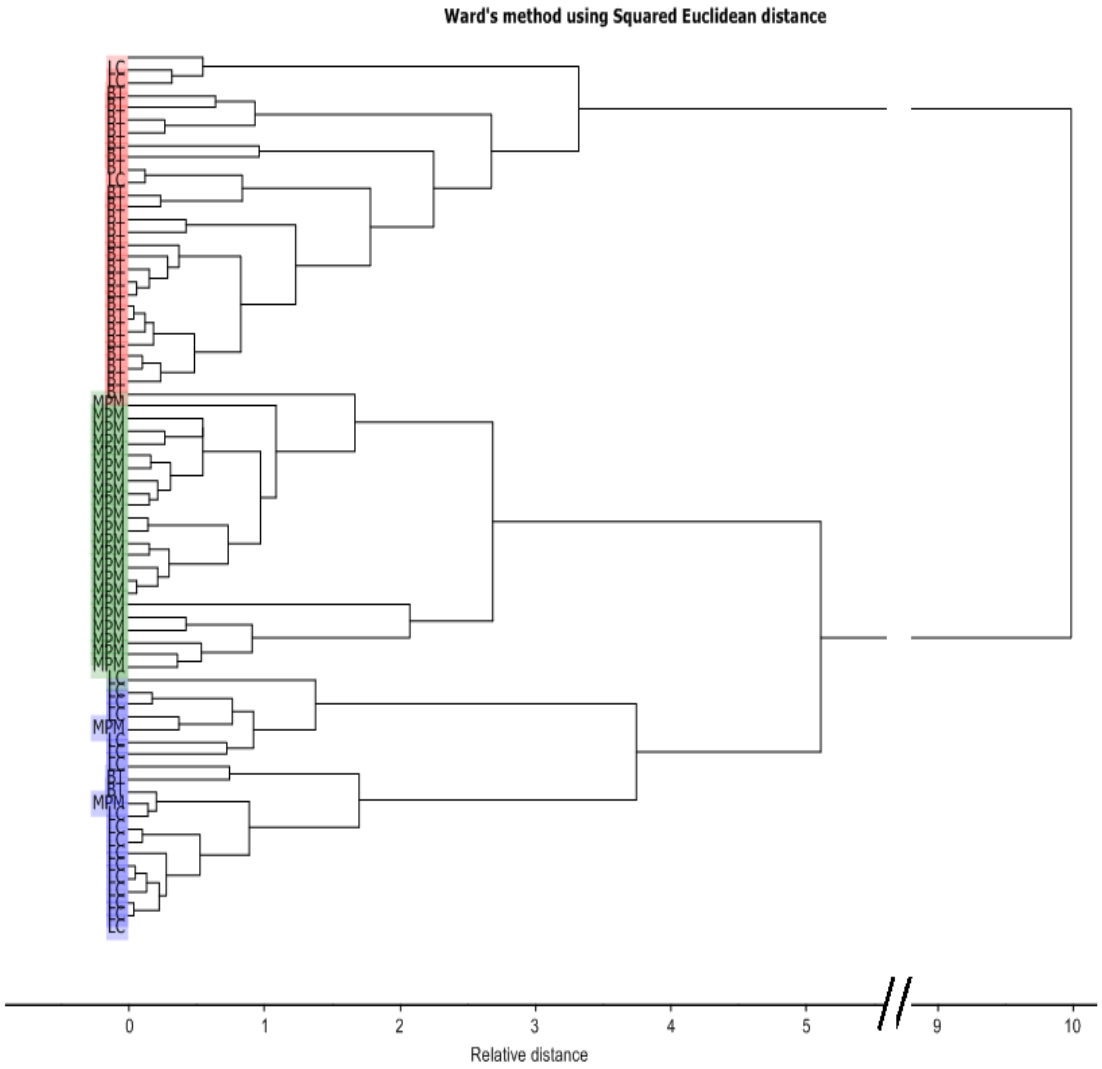


Figure 32: Hierarchical Cluster analysis of the three BT, LC and MPM groups in the amide I spectral region.

As can be seen from Figure 32, BT group was successfully clustered from the other groups with a higher heterogeneity value. To validate our HCA classification method, we calculated the sensitivity and specificity values for MPM and LC groups as described in materials and methods chapter. These terms are generally used to indicate the diagnostic performance of the models in clinical diagnostics. For MPM group, 88% sensitivity and 100% specificity were obtained. For LC, 85% sensitivity and 88.5% specificity were acquired. These high values implied successful discrimination capacity of HCA method in the MPM and LC groups from BT ones. In addition to

HCA to differentiate the groups, another unsupervised chemometric method namely PCA was applied to the IR spectra of amide I region to differentiate the groups.

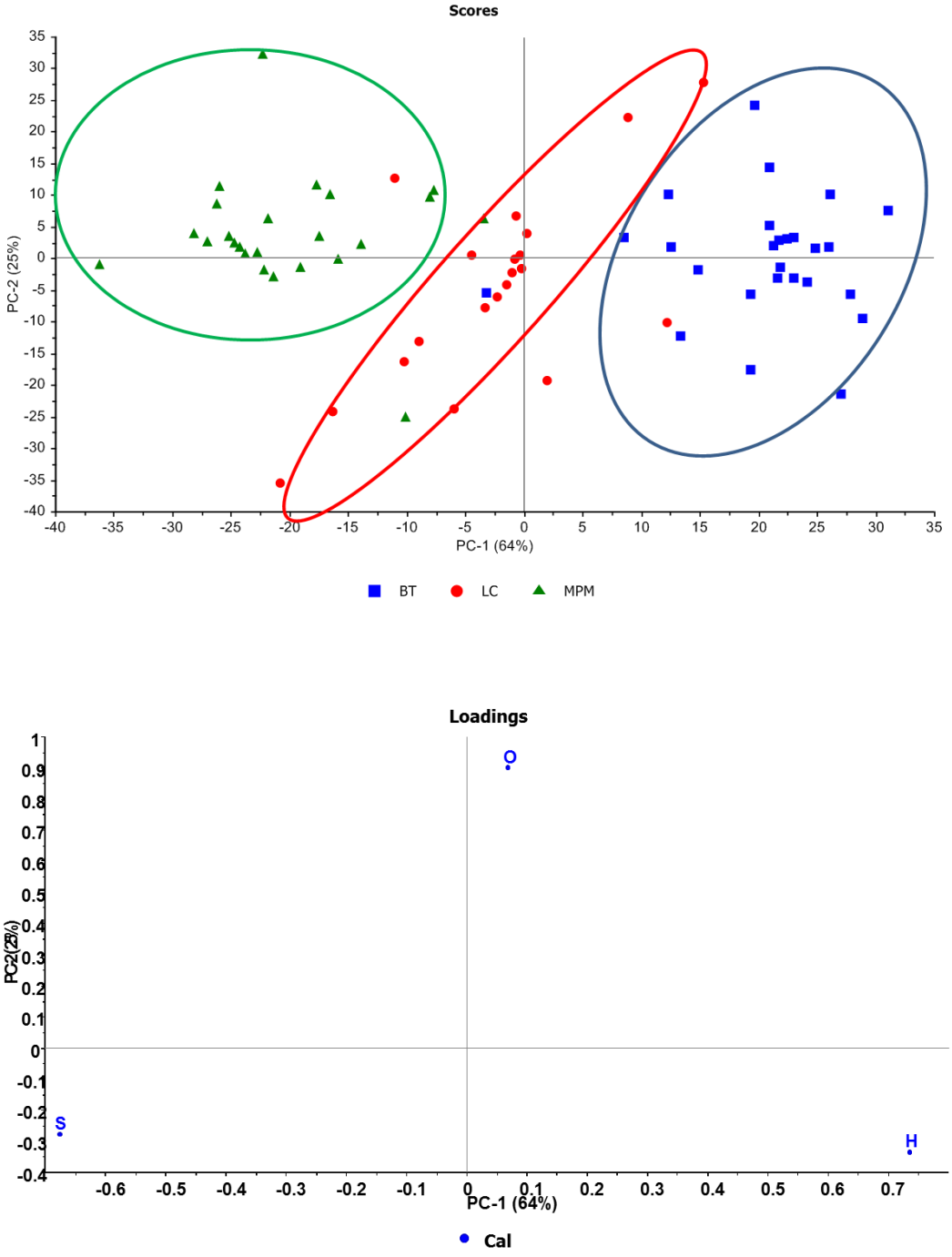


Figure 33: PCA A) Scatter plots B) Loading plot for BT, LC and MPM in amide I spectral regions.

Figure 33A represents two dimensional score plot of the first two principal components. These components represent the variation among the samples. Analysis

of the scores plot for the whole region showed 64% of the variation which was accounted by the first principal component (PC) and 25% by the second PC. As can be deduced from the Figure 33A, MPM and LC groups were distinguished from BT one and from each other.

In PCA method, the loading plot (Figure 33B) enables the analysis of IR spectra and to identify of the variant that contribute most to the variation described in the PC terms. It is clearly seen from the loading plot that there was a higher variation of Eigen vector values for both α-helix and β-sheets content in PC1 indicating that they have the major contribution for PC1. Figure 34 shows the amount of contribution for α-helix (H), β-sheets (S) and other structure (O) in each PC. For PC-1 it is clear that H and S have a high contribution in this PC but O almost has no contribution. However, in PC-2 the three structure (H, S and O) contribute equally for this PC.
In our study the number of samples are relative small, because of this the validation was used in order to obtain a PCA model can be used for the new samples.

For all PCA calculations in this study a full cross validation was used. Figure 35 shows the PCA result calculated from the calibration set (blue) and the scores corresponding to leave-one-out cross validation (red) for each sample. The PCA in Figure 35 show that the calibration and cross validation results for the studied spectra are close to each other which means the developed PCA model was reliable and can be used for new samples.

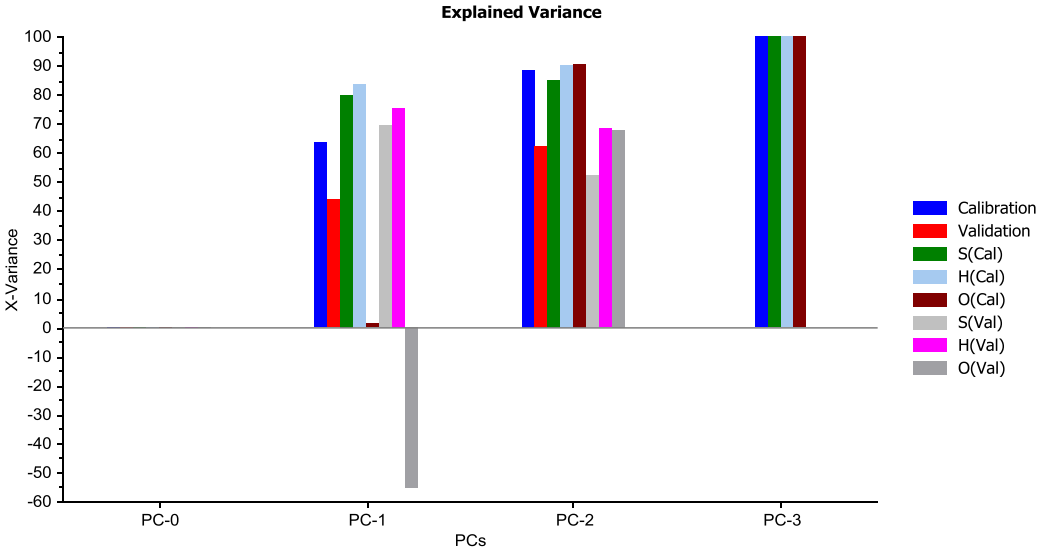


Figure 34: Amount of contribution for α- helix (H), β-sheets (S) and other structure (O) in each PC.

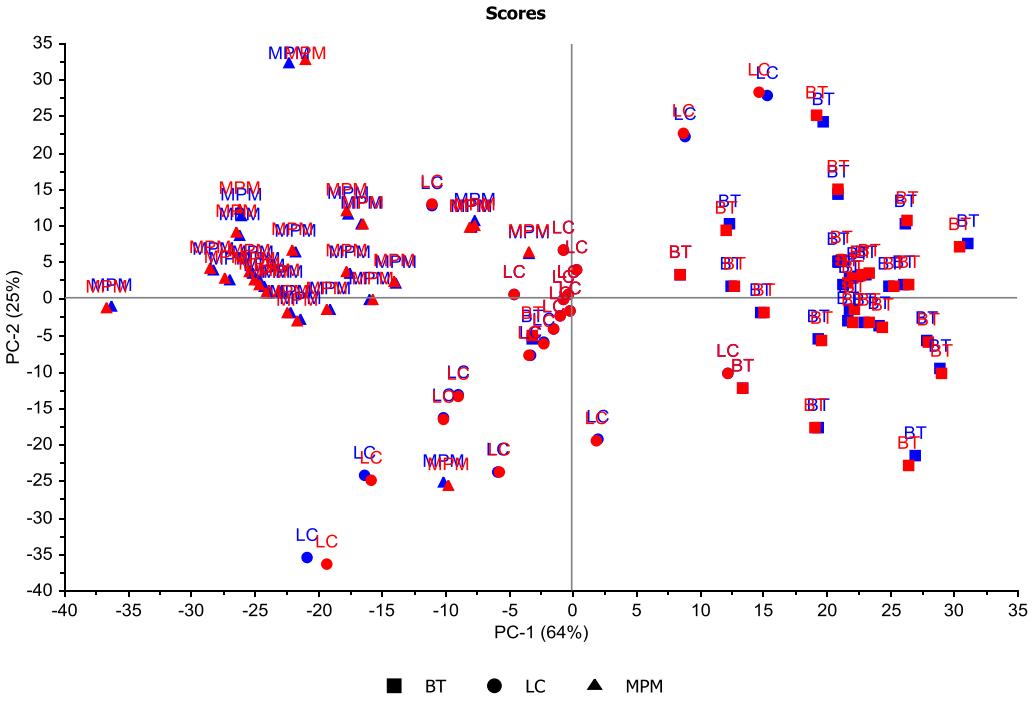


Figure 35: Leave-one-out cross validation analysis of the pleural fluids proteins secondary structure.

Following successful discrimination results obtained from HCA and PCA, to develop a more precise classification, we applied supervised chemometric analysis approach to the pleural fluid spectra of the studied samples. This approach requires initial knowledge about class of samples such as disease vs healthy and enables more accurate classification with class borders. Among supervised methods, SIMCA is commonly used for classification of spectral data since it enables good classification samples even with low sample size and high variability within-class (Bylesjö et al. 2006). Therefore, in our study, we performed SIMCA supervised chemometric analysis technique for the amide I bands to identify the class membership of unknown samples (test samples). To perform SIMCA analysis, three PCA models were developed from the spectra of the samples that make up the training data (BT:22, LC:17, MPM:21). Figure 36 demonstrates the distance of MPM and LC models to BT models. The model distance plot enables to determine how different each model from each other with respect to PC space and thus reveals the success of the classification method. In the creation of distance, total residual standard deviations are used as a measure. A model distance greater than three, indicates that models are significantly different from each other (Mouwen et al. 2005; Lu et al. 2011).

As can be seen from Figure 36, the distance of both MPM and LC models to BT model was 38 and 32 respectively, implying very robust and clear differentiation of both groups from BT. The discrimination of MPM from BT and also LC group with 10% significance level can also be seen in Cooman's plot in
Figure 37. This plot is used to show the discrimination between two classes and to test the validation and accuracy of the diagnostic models. In order to test our models, 3 samples from each group were randomly selected to form an overall 9 samples in the test group. These test group spectra were not included in SIMCA training set. The process of randomly selected samples was repeated three times and each time the percentage of correct spectra classification has been calculated. The results revealed 100%, 89% and 100% for the $1^{st}$, $2^{nd}$ and $3^{rd}$ trials, respectively with overall 96.3% correct spectra classification.

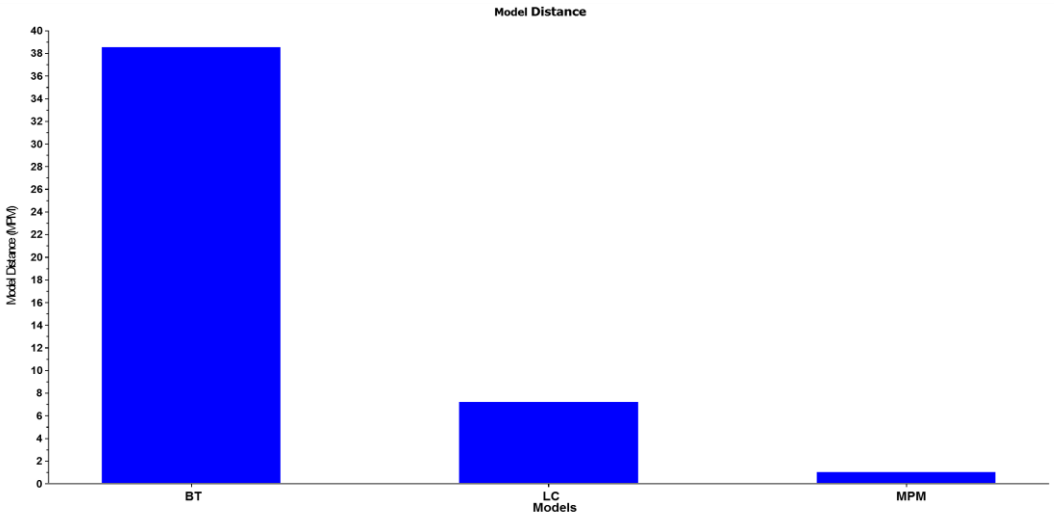


Figure 36: Distance in PCA space of BT and LC calibration models from MPM model.

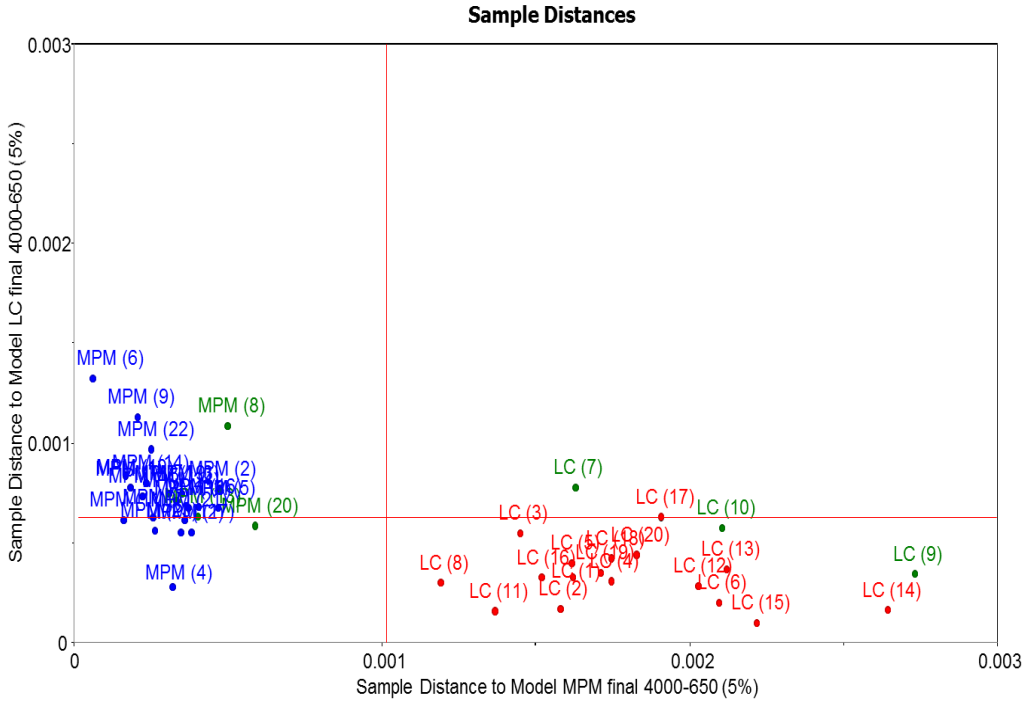


Figure 37: A) SIMCA Cooman's plot of MPM (green) LC (red) and BT (blue) for pleural fluids proteins secondary structure.

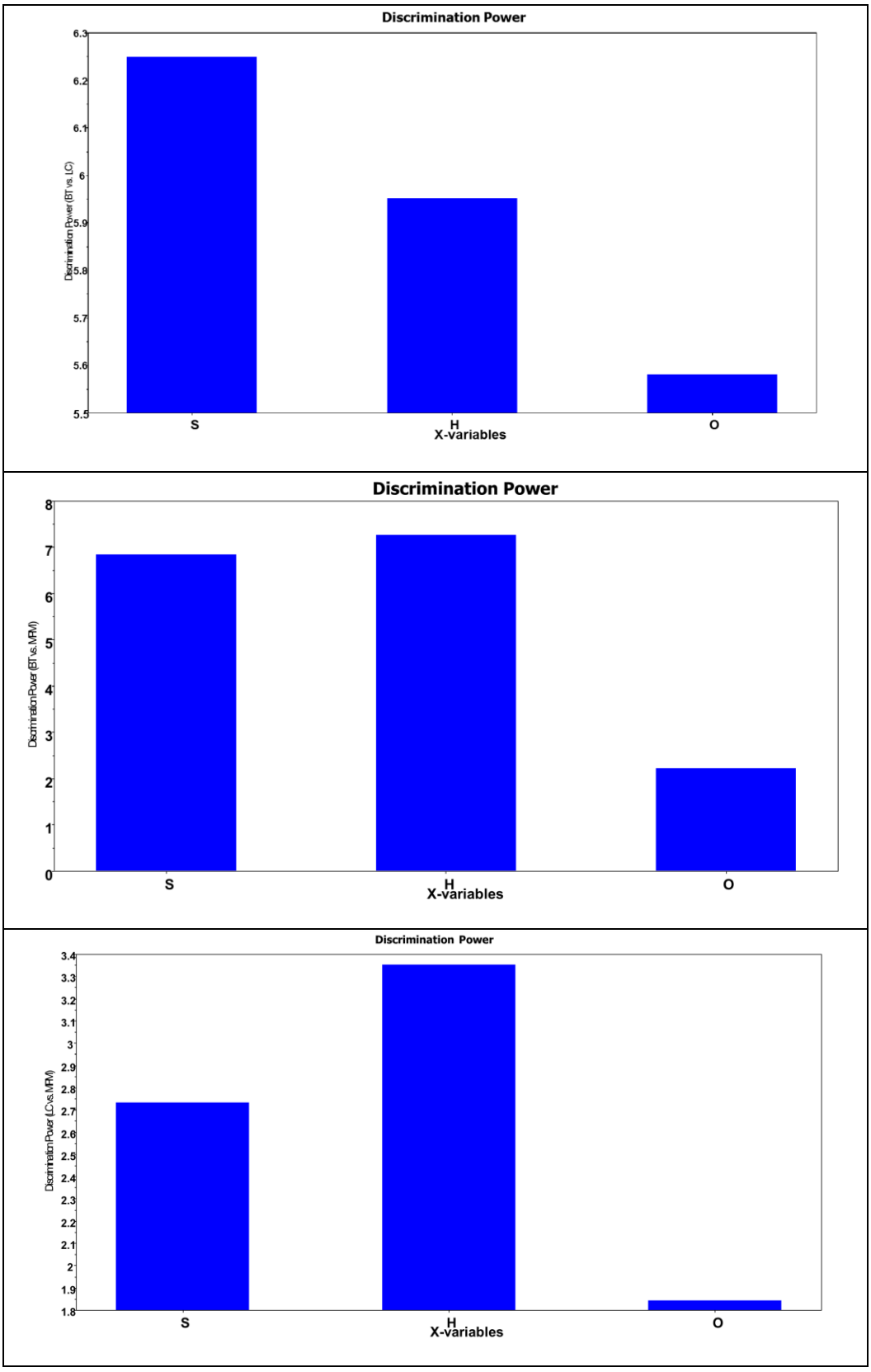


Figure 38: Discrimination power of the proteins secondary structures. A) BT from LC, B) BT from MPM and C) LC from MPM.

Figure 38A shows the discrimination power of the three proteins secondary structures to differentiate BT and LC pleural fluids. As it is clear from the figure, the β-sheets structure has more contribution to the differentiation of BT and LC. Similarly
Figure 38B shows that the both β-sheets and α-helix structures have high contribution in the differentiation of BT from MPM.
Figure 38C shows that the differentiation of MPM from LC were mainly dependent on the α-helix continent of each group. In previous study (Segal et al. 2013) they showed that the cytological diagnosis of MPM can be done with positive predictive value of 99% from pleural fluid however the absolute sensitivity was only 68%. In another study (Hegmans et al. 2009), the diagnosis methods for MPM, based on soluble mesothelin-related proteins [SMRP] revealed only 76% and 69%, sensitivity and specificity, respectively. In comparison to those previous studies, our study gave satisfactory results with 93% sensitivity of MPM diagnosis and overall 96.3% correct classification using SIMCA for the three groups. The results of this study indicated that IR spectroscopy can be considered as a promising tool for screening and diagnosis of MPM cancer disease.

# CHAPTER 4

# CONCLUSION

In this study we focused on the applications of FTIR spectroscopy for proteins structure and dynamics prediction in dilute, artificial crowded and native environments. Proteins play very important roles in cells regulation and structure. Understanding of a protein structures greatly help in understanding the mechanism of action of this protein. The alteration in a certain protein structure can be linked to some diseases; thus the study of proteins structure can help for the diagnosis of many diseases. Up to know, the most accurate techniques for proteins structure determination are x-ray crystallography and NMR. However, those techniques have some limitations in the study of certain proteins. For example, membrane proteins are very difficult to be crystalized into single crystal which is required for x-ray crystallography. In addition, static nature of a crystal that is used in x-ray technique cannot monitor the dynamic structure of a protein in native aqueous environment. The study of protein structure by NMR spectroscopy is limited to small size proteins because of NMR signals complexity. On the other hand, the optical spectroscopy such as FTIR and CD spectroscopy can deal with almost all kind of proteins without complicated sample preparation and, in some cases, in the protein native environment. In FTIR, the amide I peak plays the most important role for the proteins secondary structure predictions. In order to calculate the protein secondary structure from amide I, different methods can be used such as curve fitting, deconvolution, and signal intensity amide I second derivative spectra. One of the promising methods for the estimation of proteins secondary structures is the ANNs. Because of this, we developed a protein FTIR dataset of known proteins structures in dilute buffer solution. This database was used as a training set for ANNs. The 35 proteins have been scanned

using FTIR transmission mode in 7.4 phosphate buffer aqueous solution. For the first time, wavelet analysis has been used as a data reduction of amide I in order to reduce the number of inputs neurons for the ANNs. Because of the limited number of proteins in dataset, we used leave-one-out approach for training and testing our neural networks. To achieve generalized ANNs dataset with a limited number protein, for the first time up best to our knowledge, discrete wavelet transform (DWT) was successfully used as data reduction technique for amide I spectra. Unlike transformations techniques such as Fourier transform, the DWT preserve the local information of the amide I signal (). This feature of wavelet analysis is very important for the analysis of amide I, because each part of amide I reflects a certain protein secondary structure. The results of ANNs predictions showed 96.88%, 93.92% and 95.98% success in β-sheets, α-helix and other structures respectively. After successfully prediction of the proteins secondary structure using ANNs in dilute solution, we attempted to the answer of the following question: Does the crowded environment affect the proteins structure and its dynamics?

The second part of this thesis is to answer the previous question, in order to highlight this phenomena, the proteins structure prediction and dynamics should be studied in an environment that mimics the native protein environment rather than the dilute solution. Proteins, in native, present in a crowded environment. Because of this, proteins structures studies should take the effect of macromolecules crowding in consideration. Human Apo- and Holo-transferren structures and their thermal denaturation behavior have been studied in dilute and artificial crowded environment using FTIR spectroscopy. Dextran 70 and Ficoll 70 as a "molecular crowder" did not have a major effect on the secondary structure of transferrin as deduced from the analysis of the amide I band. However, it does alter the tertiary structure as causing significant differences in hydrogen-deuterium exchange which was seen by monitoring the intensity of the residual amide II band as a function of time. The hydrogen-deuterium exchange is reduced in the presence of dextran which suggests that molecular crowding produces a more compact and rigid protein structure. Furthermore, the study of transferren thermal denaturation using 2D-IR spectroscopy showed two different aggregated secondary structures patterns in dilute and in artificial crowded environment. We can conclude from our study that molecular crowding does

indeed have an effect on the tertiary structure and dynamics of a protein and this may have important implications for its functional activity.

Finally, as an application for proteins secondary structure, we studied the proteins secondary structure of human pleural fluid accumulated due to malignant pleural mesothelioma (MPM), lung cancer (LC) and benign transudate (BT). In order to identify whether the pleural fluid is due to MPM, LC or BT, the details of pleural fluid's protein and their secondary structure contents has been studied using ATR-FTIR spectroscopy. For the analysis of proteins secondary structure from FTIR spectra, commonly, the amide I region (1700-1600 cm-1) is utilized. Wavelet analysis has been used to extract the amide I spectral features. The extracted features were used as an input for a previously trained artificial neural network using protein database to estimate the proteins secondary structures (part 1 of this thesis). Spectral analysis indicated significant differences in protein content of BT, LC and MPM pleural fluid samples. Furthermore, an increase of beta- sheet structure in the MPM pleural fluid has been observed which could be attributed to the presence of Mesothelin and other beta sheet structure proteins. The chemometric results of the plural fluid proteins secondary structure lead to an accurate, cost effective method for the diagnosis of MPM from lung cancer and benign transudate with 88% sensitivity and 100% specificity. This enabled an accurate and specific differentiation of MPM pleural fluid from the others two groups.

In summary, this study highlights the advantages of FTIR as a spectroscopic technique to study proteins structure. We improved the prediction accuracy using wavelet based ANNs. However, we are looking forward to establish an online proteins database that can be easily used by scientist to estimate their proteins secondary structure using our trained ANNs.

## REFERENCES


Ahmed FE (2005) Artificial neural networks for diagnosis and survival prediction in colon cancer. Mol Cancer 4:29. doi: 10.1186/1476-4598-4-29

Alzubi S, Islam N, Abbod M (2011) Multiresolution analysis using wavelet, ridgelet, and curvelet transforms for medical image segmentation. Int J Biomed Imaging 2011:136034. doi: 10.1155/2011/136034

Baker MJ, Trevisan J, Bassan P, et al (2014a) Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc 9:1771–91. doi: 10.1038/nprot.2014.110

Baker MJ, Trevisan J, Bassan P, et al (2014b) Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc 9:1771–91. doi: 10.1038/nprot.2014.110

Berg JM, Tymoczko JL, Stryer L (2002) Three-Dimensional Protein Structure Can Be Determined by NMR Spectroscopy and X-Ray Crystallography.

Byler DM, Susi H (1986) Examination of the secondary structure of proteins by deconvolved FTIR spectra. Biopolymers 25:469–87. doi: 10.1002/bip.360250307

Bylesjö M, Rantalainen M, Cloarec O, et al (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom 20:341–351. doi: 10.1002/cem.1006

Chang CCH, Tey BT, Song J, Ramanan RN (2015) Towards more accurate prediction of protein folding rates: a review of the existing Web-based bioinformatics approaches. Brief Bioinform 16:314–24. doi: 10.1093/bib/bbu007

Chebotareva NA (2007) Effect of molecular crowding on the enzymes of glycogenolysis. Biochem Biokhimiĭa 72:1478–90.

Chebotareva NA, Kurganov BI, Livanova NB (2004) Biochemical effects of

molecular crowding. Biochem Biokhimiia 69:1239–51.

Cheng Y, Zak O, Aisen P, et al (2004) Structure of the Human Transferrin Receptor-Transferrin Complex Albert Einstein College of Medicine. Cell 116:565–576. doi: S0092867404001308 [pii]

Creaney J, Dick IM, Robinson BWS Discovery of new biomarkers for malignant mesothelioma. Curr Pulmonol reports 4:15–21. doi: 10.1007/s13665-015-0106-8

Daubechies I (1992) Ten lectures on wavelets.

Despa F, Fernández A, Berry RS (2004) Dielectric modulation of biological water. Phys Rev Lett 93:228104.

Dong A, Huang P, Caughey WS (1990) Protein secondary structures in water from second-derivative amide i infrared spectra. Biochemistry 29:3303–3308.

Dong A, Hyslop RM, Pringle DL (1996) Differences in conformational dynamics of ribonucleases A and S as observed by infrared spectroscopy and hydrogen-deuterium exchange. Arch Biochem Biophys 333:275–81. doi: 10.1006/abbi.1996.0391

Dong A, Prestrelski SJ, Allison SD, Carpenter JF (1995) Infrared spectroscopic studies of lyophilization- and temperature-induced protein aggregation. J Pharm Sci 84:415–424. doi: 10.1002/jps.2600840407

Filip Zavoral, Jakub Yaghob, Pit Pichappan EE-Q (2010) Networked Digital Technologies, Part II: Second International Conference, NDT 2010, Prague, Czech Republic, July 7-9, 2010 Proceedings. Springer Science & Business Media

Fulton AB (1982) How crowded is the cytoplasm? Cell 30:345–347. doi: 10.1016/0092-8674(82)90231-8

Gok S, Aydin OZ, Sural YS, et al (2016) Bladder cancer diagnosis from bladder wash by Fourier transform infrared spectroscopy as a novel test for tumor recurrence. J Biophotonics. doi: 10.1002/jbio.201500322

Hadden JM, Bloemendal M, Haris PI, et al (1994a) Fourier transform infrared

spectroscopy and differential scanning calorimetry of transferrins: human serum transferrin, rabbit serum transferrin and human lactoferrin. Biochim Biophys Acta 1205:59–67.

Hadden JM, Bloemendal M, Haris PI, et al (1994b) Fourier transform infrared spectroscopy and differential scanning calorimetry of transferrins: human serum transferrin, rabbit serum transferrin and human lactoferrin. Biochim Biophys Acta (BBA)/Protein Struct Mol 1205:59–67.

Hadden JM, Haris PI, Srai KS, Chapman D (1992) Conformational studies on human transferrin. Biochem Soc Trans 20:200S.

Harada R, Sugita Y, Feig M (2012) Protein crowding affects hydration structure and dynamics. J Am Chem Soc 134:4842–9. doi: 10.1021/ja211115q

Haris PI, Molle G, Duclohier H (2004) Conformational Changes in Alamethicin Associated with Substitution of Its α-Methylalanines with Leucines: A FTIR Spectroscopic Analysis and Correlation with Channel Kinetics. Biophys J 86:248–253.

Haris PI, Severcan F (1999) FTIR spectroscopic characterization of protein structure in aqueous and non-aqueous media. J Mol Catal B Enzym 7:207–221. doi: 10.1016/S1381-1177(99)00030-2

Hassoun MH (1995a) Fundamentals of Artificial Neural Networks.

Hassoun MH (1995b) Fundamentals of Artificial Neural Networks. MIT Press

Hegmans JPJJ, Veltman JD, Fung ET, et al (2009) Protein profiling of pleural effusions to identify malignant pleural mesothelioma using SELDI-TOF MS. Technol Cancer Res Treat 8:323–32.

Hering JA, Haris PI (2009) FTIR spectroscopy for analysis of protein secondary structure. Adv Biomed Spectrosc 2:129–167. doi: 10.3233/978-1-60750-045-2-129

Hering JA, Innocent PR, Haris PI (2004a) Towards developing a protein infrared

spectra databank (PISD) for proteomics research. Proteomics 4:2310–9. doi: 10.1002/pmic.200300808

Hering JA, Innocent PR, Haris PI (2003) Neuro-fuzzy structural classification of proteins for improved protein secondary structure prediction. Proteomics 3:1464–75. doi: 10.1002/pmic.200300457

Hering JA, Innocent PR, Haris PI (2004b) Empirical knowledge and genetic algorithms for selection of amide I frequencies in protein secondary structure prediction. 345–350.

Hering JA, Innocent PR, Haris PI (2002a) Automatic amide I frequency selection for rapid quantification of protein secondary structure from. Proteomics 16:839–849. doi: 10.1002/1615-9861(200207)2:7<839::AID-PROT839>3.0.CO;2-L

Hering JA, Innocent PR, Haris PI (2004c) Beyond average protein secondary structure content prediction using FTIR spectroscopy. Appl Bioinformatics 3:9–20.

Hering JA, Innocent PR, Haris PI (Parvez I. (2002b) An alternative method for rapid quantification of protein secondary structure from FTIR spectra using neural networks.

Hsieh W-Y, Chen M-W, Ho H-T, et al (2006) Identification of differentially expressed proteins in human malignant pleural effusions. Eur Respir J 28:1178–85. doi: 10.1183/09031936.06.00135405

Iloro I, Chehín R, Goñi FM, et al (2004) Methionine adenosyltransferase alpha-helix structure unfolds at lower temperatures than beta-sheet: a 2D-IR study. Biophys J 86:3951–8. doi: 10.1529/biophysj.103.028373

Joosten RP, te Beek TAH, Krieger E, et al (2011) A series of PDB related databases for everyday needs. Nucleic Acids Res 39:D411–9. doi: 10.1093/nar/gkq1105

Khanmohammadi M, Garmarudi AB, Ghasemi K, et al (2009) Artificial neural network for quantitative determination of total protein in yogurt by infrared spectrometry. Microchem J 91:47–52. doi: 10.1016/j.microc.2008.07.003

Kilár F, Simon I (1985) The effect of iron binding on the conformation of transferrin. A small angle x-ray scattering study. Biophys J 48:799–802. doi: 10.1016/S0006-3495(85)83838-8

Krafft C, Codrich D, Pelizzo G, Sergo V (2008) Raman and FTIR microscopic imaging of colon tissue: a comparative study. J Biophotonics 1:154–69. doi: 10.1002/jbio.200710005

Krafft C, Sergo V (2006) Biomedical applications of Raman and infrared spectroscopy to diagnose tissues. Spectroscopy 20:195–218.

Krafft C, Shapoval L, Sobottka SB, et al (2006) Identification of primary tumors of brain metastases by SIMCA classification of IR spectroscopic images. Biochim Biophys Acta 1758:883–91. doi: 10.1016/j.bbamem.2006.05.001

Kuznetsova IM, Turoverov KK, Uversky VN (2014) What macromolecular crowding can do to a protein. Int J Mol Sci 15:23090–140. doi: 10.3390/ijms151223090

Lu X, Rasco BA, Kang DH, et al (2011) Infrared and Raman spectroscopic studies of the antimicrobial effects of garlic concentrates and diallyl constituents on foodborne pathogens. Anal Chem 83:4137–46. doi: 10.1021/ac2001498

Mallick PK (2015) Research Advances in the Integration of Big Data and Smart Computing. IGI Global

Martel P, Kim SM, Powell BM (1980) Physical characteristics of human transferrin from small angle neutron scattering. Biophys J 31:371–80. doi: 10.1016/S0006-3495(80)85065-X

Martens H, Nielsen JP, Engelsen SB (2003) Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures. Anal Chem 75:394–404. doi: 10.1021/ac020194w

McCarthy WJ, Meza R, Jeon J, Moolgavkar SH (2012) Chapter 6: Lung cancer in never smokers: epidemiology and risk prediction models. Risk Anal 32 Suppl 1:S69–84. doi: 10.1111/j.1539-6924.2012.01768.x

Minton AP (1983) The effect of volume occupancy upon the thermodynamic activity of proteins: some biochemical consequences. Mol Cell Biochem 55:119–140. doi: 10.1007/BF00673707

Mitchell AL, Gajjar KB, Theophilou G, et al (2014) Vibrational spectroscopy of biofluids for disease screening or diagnosis: Translation from the laboratory to a clinical setting. J Biophotonics 7:153–165. doi: 10.1002/jbio.201400018

Mouwen DJM, Weijtens MJBM, Capita R, et al (2005) Discrimination of enterobacterial repetitive intergenic consensus PCR types of Campylobacter coli and Campylobacter jejuni by Fourier transform infrared spectroscopy. Appl Environ Microbiol 71:4318–24. doi: 10.1128/AEM.71.8.4318-4324.2005

Muehlethaler C, Massonnet G, Esseiva P (2014) Discrimination and classification of FTIR spectra of red, blue and green spray paints using a multivariate statistical approach. Forensic Sci Int 244:170–178. doi: 10.1016/j.forsciint.2014.08.038

Mueller D, Ferrão MF, Marder L, et al (2013) Fourier transform infrared spectroscopy (FTIR) and multivariate analysis for identification of different vegetable oils used in biodiesel production. Sensors (Basel) 13:4258–4271. doi: 10.3390/s130404258

Mun EY, von Eye A, Bates ME, Vaschillo EG (2008) Finding groups using model-based cluster analysis: heterogeneous emotional self-regulatory processes and heavy alcohol use risk. Dev Psychol 44:481–95. doi: 10.1037/0012-1649.44.2.481

Noda I (1990) Two-Dimensional Infrared (2D IR) Spectroscopy: Theory and Applications. Appl Spectrosc 44:550–561.

Noda I, Dowrey AE, Marcott C (1988) Two-dimensional infrared (2D IR) spectroscopy. A new tool for interpreting infrared spectra. Mikrochim Acta 94:101–103. doi: 10.1007/BF01205847

Owens GL, Gajjar K, Trevisan J, et al (2014) Vibrational biospectroscopy coupled with multivariate analysis extracts potentially diagnostic features in blood

plasma/serum of ovarian cancer patients. J Biophotonics 7:200–209. doi: 10.1002/jbio.201300157

Paquet MJ, Laviolette M, Pézolet M, Auger M (2001) Two-dimensional infrared correlation spectroscopy study of the aggregation of cytochrome c in the presence of dimyristoylphosphatidylglycerol. Biophys J 81:305–12. doi: 10.1016/S0006-3495(01)75700-1

Paramothayan NS, Barron J (2002) New criteria for the differentiation between transudates and exudates. J Clin Pathol 55:69–71.

Patel AK, Choudhury S Combined pleural fluid cholesterol and total protein in differentiation of exudates and transudates. Indian J Chest Dis Allied Sci 55:21–3.

Pirhadi S, Shiri F, Ghasemi JB (2015) Multivariate statistical analysis methods in QSAR. RSC Adv 5:104635–104665. doi: 10.1039/C5RA10729F

Quellec G, Lamard M, Josselin PM, et al (2008) Optimal wavelet transform for the detection of microaneurysms in retina photographs. IEEE Trans Med Imaging 27:1230–41. doi: 10.1109/TMI.2008.920619

Samiotakis A, Wittung-Stafshede P, Cheung MS (2009) Folding, stability and shape of proteins in crowded environments: experimental and computational approaches. Int J Mol Sci 10:572–88. doi: 10.3390/ijms10020572

Segal A, Sterrett GF, Frost FA, et al (2013) A diagnosis of malignant pleural mesothelioma can be made by effusion cytology: results of a 20 year audit. Pathology 45:44–8. doi: 10.1097/PAT.0b013e32835bc848

Severcan F, Haris PI (2012) Vibrational Spectroscopy in Diagnosis and Screening. IOS Press

Severcan M, Haris PI, Severcan F (2004a) Using artificially generated spectral data to improve protein secondary structure prediction from Fourier transform infrared spectra of proteins. Anal Biochem 332:238–44. doi: 10.1016/j.ab.2004.06.030

Severcan M, Haris PI, Severcan F (2004b) Using artificially generated spectral data to improve protein secondary structure prediction from Fourier transform infrared spectra of proteins. Anal Biochem 332:238–44. doi: 10.1016/j.ab.2004.06.030

Severcan M, Severcan F, Haris PI (2001) Estimation of protein secondary structure from FTIR spectra using neural networks. J Mol Struct 565-566:383–387. doi: 10.1016/S0022-2860(01)00505-1

Sutedja G (2003) New techniques for early detection of lung cancer. Eur Respir J Suppl 39:57s–66s.

Toney LK, Vesselle HJ (2014) Neural networks for nodal staging of non-small cell lung cancer with FDG PET and CT: importance of combining uptake values and sizes of nodes and primary tumor. Radiology 270:91–8. doi: 10.1148/radiol.13122427

Tyan Y-C, Wu H-Y, Su W-C, et al (2005) Proteomic analysis of human pleural effusion. Proteomics 5:1062–74. doi: 10.1002/pmic.200401041

Wan C, Cao W, Cheng C (2014) Research of Recognition Method of Discrete Wavelet Feature Extraction and PNN Classification of Rats FT-IR Pancreatic Cancer Data.

Yan Y-B, Wang Q, He H-W, et al (2003a) Two-dimensional infrared correlation spectroscopy study of sequential events in the heat-induced unfolding and aggregation process of myoglobin. Biophys J 85:1959–1967. doi: 10.1016/S0006-3495(03)74623-2

Yan Y-B, Wang Q, He H-W, et al (2003b) Two-dimensional infrared correlation spectroscopy study of sequential events in the heat-induced unfolding and aggregation process of myoglobin. Biophys J 85:1959–67. doi: 10.1016/S0006-3495(03)74623-2

Yan Y-B, Zhang J, He H-W, Zhou H-M (2006) Oligomerization and aggregation of bovine pancreatic ribonuclease A: characteristic events observed by FTIR spectroscopy. Biophys J 90:2525–33. doi: 10.1529/biophysj.105.071530

Yilmaz Turay U, Yildirim Z, Türköz Y, et al (2000) Use of pleural fluid C-reactive

protein in diagnosis of pleural effusions. Respir Med 94:432–5.

Zimmerman SB, Trach SO (1991) Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. J Mol Biol 222:599–620.

## APPENDIX

Matlab code for artificial neural networks with leave-one-out approch

```
clear; clc;
for jury =1:10              % No of jury results to be avage
   load('Final35proteinsnormalized.mat');
   inputs = Final35proteinsnormalized;
   load('fnal35proteinstargetall.mat');
   targets2=fnal35proteinstarget;%(1,:); % In sheet-helix-others order

   HL= 3;   %% no. of hidden layers  example: [10 15 20] for three layes

   %% Wavelet analysis %%

   dirDec = 'c';         % Direction of decomposition
   level  = 7;           % Level of decomposition   7
   wname  = 'db2';       % Near symmetric wavelet the best is db2
   decROW = mdwtdec(dirDec,inputs,level,wname);

   wvlt3  = decROW.cd{1, 3}; % the best is 3 then 2

  %% Plot for Wevwelt coeff. %%
   for n=20:27
      figure(n),
      subplot(2,1,1);
      plot(inputs(:,n));
      title('Original Data');
      title(fnal35proteinstarget(:,n));
      axis tight

      subplot(2,1,2);
      plot(wvlt3(:,n));
      title('Corresponding approximations at level 3');
      axis tight
   end
   %% NN %%
   inputs2=wvlt3;%(2:9,:); % first and last coeff. not used
   finished=1;
   for protn=1:35                    % how many blind protein to test
     inputs2 = circshift(inputs2,1,2);     % shift inputs, one columb every run
     targets2 = circshift(targets2,1,2);   % shift outputs, one columb every run
     inputsx= inputs2(:,1:34);             % Remove the last protein for the blind test
     targetsx=targets2(:,1:34);            % Remove the last targt for the blind test
     net1 = feedforwardnet (HL,'trainrp');
     net1 = configure(net1,inputsx,targetsx); % strating nn
```

```
    currentRMSE=100;                % just to generat currentRMSE variable
    nn=0;                           % count the no. of trials that ANN doesn't show
RMSE less than thr.
    while  currentRMSE> 8           % The threshold value for accepted RMES
      nn=nn+1;

      for tp=1:34                   % Repeat training to find the best trained nn
        inputsx = circshift(inputsx,1,2);   % shift inputs one columb every run
        targetsx = circshift(targetsx,1,2); % shift inputs one columb every run

        net1.divideFcn = 'divideind';
        net1.divideParam.trainInd=1:32;
        net1.divideParam.valInd=33;         % one protein for validation
       %net1.divideParam.testInd=34;

        net1.trainParam.goal=0;
        net1.performFcn = 'sse';
        net1.trainParam.epochs=1000;
        net1.trainParam.max_fail=6;
        net1.trainParam.lr=0.01;
        net1.trainParam.showWindow = false;

        [net1,tr] = train(net1,inputsx,targetsx);

        outputsx(:,tp) = net1(inputsx(:,34));       % array of test the lefted protein
        errors(:,tp)=gsubtract(targetsx(:,34),outputsx(:,tp)); % error

      end % End of Repeat training to find the best trained nn

      errsqr=errors.^2;
      errsqrmean=mean(mean(mean(errsqr)));
      RMSE(nn)=errsqrmean.^0.5;             %Root mean squar error
      currentRMSE=RMSE(nn);
      %%% Only for folowing durin run %%%
      currentRMSEdisp(1,1)=jury;            % just for following the run (jury)
      currentRMSEdisp(1,2)=finished;        % just for following the run (protiens
done)
      currentRMSEdisp(1,3)=RMSE(nn)         % just for following the run
(current RMSE)
      %%%%%%%%%%%%%%%%%%%%%%

    end  % End of while (if RMSE still higher than the thr value)

    %%%% If the RMSE of nn is less than thr, blind protein will be tested%

    test(:,protn) = net1(inputs2(:,35));       % The blind protein tested with the nn
after training
```

```
        tok(:,protn)=targets2(:,35);              % tok is the target for blind protein
        er(:,protn)= targets2(:,35)-test(:,protn); % er is the error for the blind protein
        finished=finished+1;                      % to count no. of blind proetins done

    end  % End of testing all blind proteins

    methoderror=er.^2;
    methoderrormean=mean(methoderror')';
    methodRMSE(:,jury)=methoderrormean.^0.5;       % RMSE of the method from
one jury

end % End of jury check

jurymethodRMSE=mean(methodRMSE')'  % Avarge of error from all jurys
```

# CURRICULUM VITAE

**Name:** Sherif Abbas Mousa Abbas

**Job:** Research assistant at Physics Department, Ain Shams University, Cairo- Egypt.

## Education and Qualifications

- **2002 B.Sc.** in Biophysics **(Excellent with honor degree**) Physics Department Biophysics group - Faculty of Science – Ain Shams University.
- **2007 M.Sc** in Biophysics entitled "Computer Based System for Biophysical Classification of White Blood Cells Images" Physics Department Biophysics group - Faculty of Science – Ain Shams University
- **2010 METU Graduate Course Performance Award:** The most successful student in the PhD program of the Department of Biology with cGPA 3.83/4.00
- **2013 ERASMUS internship** grant for three months at De Montfort University - United Kingdom.
- **2014 ERASMUS student mobility** grant for one semester at De Montfort University - United Kingdom.

## Publications:

1. **Sherif ABBAS**, Nihal SIMSEK OZEK, Salih EMRI, Feride SEVERCAN. "Diagnosis of Malignant Pleural Mesothelioma from pleural fluid by FTIR Spectroscopy as a novel approach" **"manuscript is in preparation"**
2. **Sherif ABBAS**, Feride Severcan, Parvez I. Haris. "2D-IR correlation study of proteins thermal denaturation in artificial crowded environment" **"manuscript is in preparation"**

**3.** S Garip, SH Bayari, M Severcan, **S. ABBAS,** IK Lednev, F Severcan "Structural effects of simvastatin on liver rate tissue: Fourier transform infrared and Raman microspectroscopic studies" **Journal of biomedical optics 21 (2), 025008-025008 (2016).**

4. **Sherif ABBAS,** Salih EMRI, Feride Severcan "SIMCA analysis applications in biomedical science and Forensic Sciences" **2nd international congress of forensic toxicology (Abstract accepted 2016).**
5. **Sherif ABBAS,** Feride Severcan, Parvez I Haris "Effect of Molecular Crowding on the Structure and Dynamics of Human Apo and Holo Transferrin using 2D-IR Correlation Spectroscopy" **Biophysical Journal 108 (2), 16a. (2015)**
6. **S. ABBAS** "Microscopic images dataset for automation of RBCs counting" **Data in brief 5, 35-40 (2015).**
7. **S ABBAS**, D Yonar, N Şimşek-Özek, S Emri, F Severcan "A novel method for better diagnosis of asbestos-induced lung cancer (mesothelioma) from human body fluids" **Turkish Journal of Occupational/Environmental Medicine and Safety 1 (1 (2)) (2015).**
8. R Gurbanov, **S. ABBAS,** M Bilgin, F Severcan "The Effect of Selenium Treatment On-Diabetic-Induced Structural Variations in the Molecules of Rat Kidney Plasma Membrane" **Biophysical Journal 108 (2), 626a (2015).**
9. **S. ABBAS**, Nihal S. Ozek, Deniz Koksal, Mete Severcan, Salih A. Emri, Feride Severcan "Infrared Spectroscopy as a Novel Approach in Differential Diagnosis of Malignant Pleural Mesothelioma from Lung Cancer Using Pleural Fluid" **Journal of Thoracic Oncology, Volume 10, Number 9, Supplement 2, September 2015.**
10. **Sherif ABBAS**, Parvez Haris, Mete Severcan, Salih EMRI, Feride Severcan "Proteins secondary structures changes in pleural fluids due to asbestos-induced lung cancer" **1st International Congress and Workshop of Forensic Toxicology (2014).**
11. **Sherif ABBAS,** Parvez Haris, Feride Severcan, Mete Severcan." Estimation of protein secondary structure from FTIR spectra using wavelet analysis and neural networks**" Modeling of Biomolecular Systems Interactions,**

**Dynamics, and Allostery: Bridging Experiments and Computations September 10-14, 2014**. **Istanbul, Turkey.**

12. **Sherif Abbas,** Nihal Simsek, Salih Emri Feride Severan "Evaluation of FTIR Spectroscopy as a diagnostic tool for Malignant Pleural Mesothelioma from pleural fluid" **2nd International Translational Nanomedicine Conference 2014 Boston, USA.**
13. **Sherif Abbas**, Mete Severcan and Feride Severcan "Semi-automated features extraction of Fourier Transformed Infrared spectra for biomedical artificial intelligence and statistical analysis" **ECSBM 2013, Oxford UK.**
14. **Sherif ABBAS**, Feride Severcan, Parvez I. Haris "Effect of molecular crowding on the structure and dynamics of human Holo-transferrin investigated using Fourier transform infrared spectroscopy" **DrugDesign Oxford UK, (2013).**
15. Rafig Gurbanov, **Sherif ABBAS**, Mete Severcan, Mehmet Bġlgġn, Feride Severcan "Investigation of the effects of selenium on diabetic kidney cell membrane" **XIII. National conference of Spectroscopy (2013).**
16. Hindawi S. K., A. M. Abo-Zaid, I. H. Ibrahim, **Sherif ABBAS** (2007) "Automated Segmentation of White Blood Cell Images", **Egyptian J. Biophysics, Vol. 13, No. 1 (2007)**