STUDY OF CELLS AND TISSUE – LIVE CELLS

MICROSPECTROSCOPY AND MULTIVARIATE DATA ANALYSIS

By

MILOS MILJKOVIC

A dissertation submitted to the Graduate Faculty in Chemistry in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2006

UMI Number: 3205008

UMI Microform 3205008

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

This manuscript has been read and accepted for the Graduate Faculty in Chemistry in satisfaction of the dissertation requirements for the degree of Doctor of Philosophy.

Max Diem

_____ _____

Date Chair of Examining Committee

Gerald Koeppl

_____ _____

Date Executive Officer

Ronald Birke

Charles Michael Drain

Richard Mendelsohn

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

STUDY OF CELLS AND TISSUE – LIVE CELLS MICROSPECTROSCOPY AND MULTIVARIATE DATA ANALYSIS

By

MILOS MILJKOVIC

Adviser: Professor Max Diem

The first part of this thesis focuses on *in vivo* infrared (IR) spectroscopic methods for detection of spectral changes in live cells during the division cycle. Detecting these changes is of prime importance for the development of spectroscopic methods of diagnosis.

The second part deals with coupling infrared spectral imaging, multivariate data analysis, such as hierarchical cluster analysis (HCA), and classification by artificial neural networks (ANN). These methods promise to be a powerful tool for the detection and identification of cancer cells within lymph nodes.

HeLa cells were grown either in a specially designed liquid cell or deposited on IR microscope slides and sealed in a custom made IR cell. Perkin-Elmer Spotlight 300 and Smiths Detection IlluminatIR FT-IR microspectrometers were used for data acquisition.

Sections of lymph nodes, approximately 5 μm thick, were placed onto IR microscope slides for infrared analysis. Data acquisition was performed on an FT-IR imaging system, Perkin-Elmer Spotlight 300. IR spectra were imported to CytoSpec® software to execute data preprocessing for subsequent ANN data analysis, which was performed using NeuroDeveloper® software.

The author demonstrates that it is feasible to collect IR spectral data from live cells in an aqueous environment. This opens a wide variety of experiments on subjects such as drug uptake, drug mechanism and apoptosis.

The second part of the thesis indicates that IR spectral imaging, in conjunction with hierarchical cluster analysis and ANN data classification, offers potential for a quick, automated screening and diagnosis of cancer in lymph nodes.

Acknowledgements

During the years of my work toward the Ph. D., many people have been involved in it, directly or otherwise. The reader is advised not to give any meaning to the ordering of acknowledgements.

One could not have wished for a better research mentor than Max. His assistance and counsel were invaluable. He has the ability to transform a work space of a scientist into a place that is fun to be in. For all the reasons mentioned above I am indebted to him. I hope that all my future bosses will be as knowledgeable, friendly, easygoing and generous as he has been.

Klaus is a man whose love for organic chemistry I have never shared, but fortunately we share the same passion for beer. He helped me navigate murky and traitorous waters of the Hunter College and the Graduate Center administration. Over the years, his professional and personal advice has guided me to bringing the research to a conclusion, and for this I am grateful.

The Levis: Jelena, Pavle and their kiddo Luka hold a special place in my heart and I will fondly remember the years of friendship we have shared. Thank you *tatatatirci*!

As a Ph. D. student, the biggest parts of my days were spent in the lab. I was fortunate enough to share this time with my dear colleagues, and even more proud that I now call them friends. The long hours spent together I will

remember as a pure fun, even when we were disagreeing about matters of science, politics, soccer… The fact that I had Christian, Melissa, Susie, Sebastian and Ivana (not associated with the Diem group, but considered a member at large) by my side makes me a happy and rich man. My memories of us at the events and places we have been together I wouldn't trade for all the beer in the world.

To Peter and Jürgen I owe a big "Thank you" note for all fruitful discussions we had and for their help in making the software packages work seamlessly.

Half of this thesis would not exist without generous help from Raj. Her medical expertise was of a paramount importance and for that I am thankful.

I am grateful to the Committee Members Mike, Rich and Ron who had enough patience to sit through the annual meetings and dispense valuable advices and suggestions.

Last but not least I am grateful to my family, and humbled by their appreciation and understanding for what I have been doing in the past years. Mom and Dad, Dule, Daca, Uros and Jovan thank you for your love.

Table of Contents

List of Tables

List of Figures

1 Introduction

For *in vivo* spectroscopic methods of medical and biochemical diagnosis and the cell cycle analysis, detection of spectral changes during a cell's division cycle in live cells is of prime importance. The cell cycle is a closely controlled and programmed series of events whose execution leads to the creation of two identical daughter cells. During the division a cell goes through several drastic morphological, physiological and biochemical changes, i.e. cell's size, composition and metabolic activity is subject to change. These events are at the focus of above mentioned infrared spectroscopic investigations which have helped in gaining a more thorough understanding of a complex system such as human cell. Experimental results are in agreement with the infrared analysis of exfoliated human cells, which has shown a large spectral heterogeneity that can be partially attributed to the variation of spectra from cells in different stages of the cell cycle. The same rationale is used to explain the differences observed between normal and cancerous sections of diseased human tissues; the number of actively dividing cells may be the root cause of the registered spectral feature differences.

The goal of the second part of this thesis was to establish a basis for a new type of histopathological diagnosis to supplement procedures and methods

currently in use. Coupling of Fourier transform infrared microspectroscopy (FT-IR MSP), multivariate data analysis and artificial neural network (ANN) pattern recognition is tested and compared to the classical medicinal diagnostics. Using unsupervised methods for spectral data processing, in conjunction with supervised machine-learning algorithms, more objective diagnostic tool can be created compared to the presently used procedures which are burdened with visual pattern recognition, data storage in the human brain and inherent subjectivity. Lymph nodes were chosen as tissue sample material for a number of reasons. Lymph nodes act as a barrier against the spread of metastatic cancer cells and in the course of events can become foci of secondary tumors. Frequent excision of the tissue during operations for the purpose of staging of diseases, low number of metastatic cells which may be hard to detect and large backlog of samples delaying diagnosis also motivated the research presented in following chapters.

2 Biology Background

2.1 Eukaryotic Cell

The cell is the basic structural and functional unit of living organisms. All cells have to perform necessary functions such as respiration, nutrients consumption and metabolic byproduct excretion to sustain life [1]. Cells are divided in two types: prokaryotes and eukaryotes. Prokaryotes do not have the genetic material compartmentalized in cellular organelles (specialized subcellular unit) called the nuclei and lack other complex cell structures. Eukaryotes, i.e. fungi, plant, animal and human cells, have nuclei and other organelles whose structure and function in humans will be discussed in following section.

2.1.1 Human Cell

Human cells vary in shapes and sizes, they may range from a configuration that is about 10 to 15 μm on edge and nearly cubic in shape, to stratified (flattened) morphology up to 60 μm in diameter and 5 to 10 μm thick, depending on the organ of origin.

The water molecule is the major component of the cell, accounting for about 70 % of its weight. Inorganic ions, sugars, amino acids, nucleotides, lipids and other small molecules make up 8 % of the total mass and the remaining 22 % consists of macromolecules, proteins, nucleic acids and polysaccharides. By dry weight, cells consist of about 60 % protein and 25 % nucleic acids, with the rest composed of other constituents mentioned above. In most healthy cells the RNA/DNA ratio is about 5. This composition of cells may vary depending on the organ, the stage of the cell division cycle and other factors.

Figure 2.1 shows a diagram of the structural organization of a typical human (eukaryotic) cell.

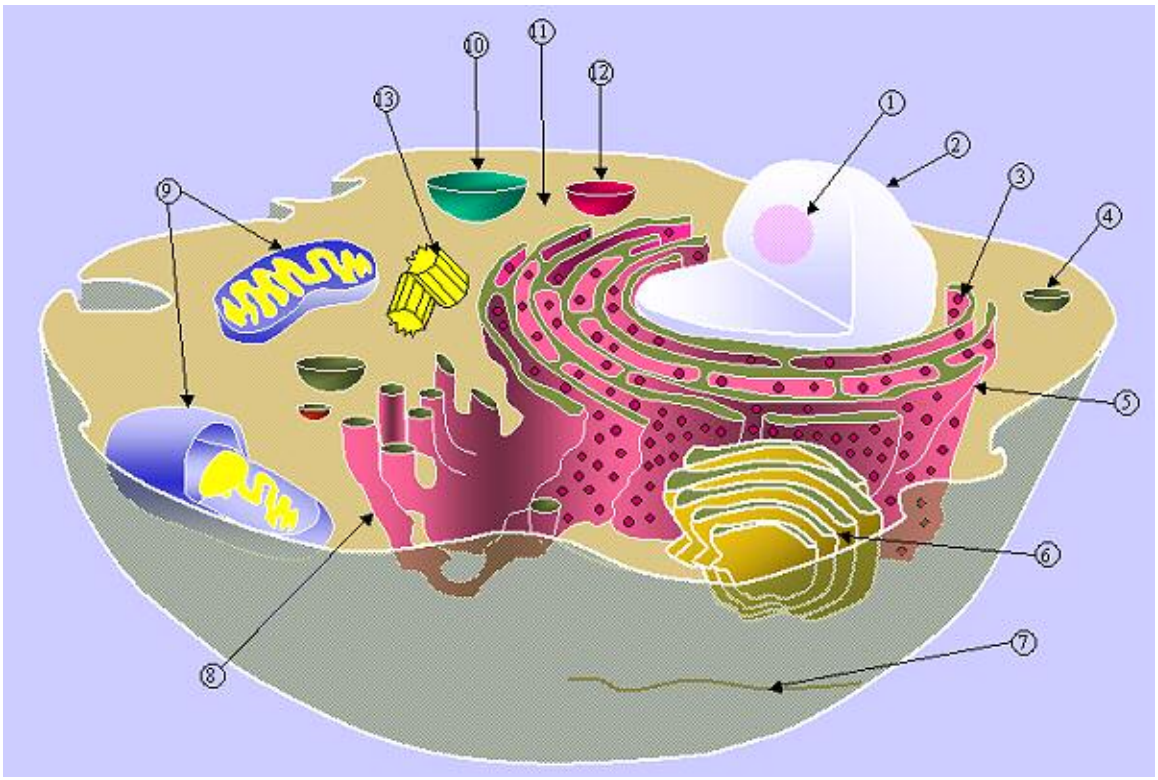

Figure 2.1 Diagram of the structural organization of a human cell: (1) Nucleolus; (2) Nucleus; (3) Bound ribosome; (4) Vesicle; (5) Rough endoplasmatic reticulum; (6) Golgi apparatus; (7) Cytoskeleton; (8) Smooth endoplasmatic reticulum; (9) Mitochondrion; (10) Vacuole; (11) Cytoplasm; (12) Lysosome; (13) Free ribosome [taken from reference 2].

Organelles can be divided into membranous and non-membranous. The membranous organelles are cell membrane, nucleus, endoplasmatic reticulum, Golgi apparatus, mitochondria, lysosomes, endosomes and peroxisomes. The non-membranous organelles are microtubules, filaments, centrioles and ribosomes.

2.1.2 Cell Membrane

The plasma membrane is more than a simple boundary structure; it takes a role in numerous cell functions. It is a double layer with the molecular level organization currently described as the modified fluid mosaic model. The membrane consists of phospholipids, cholesterol and proteins. The lipid molecules are arranged in such manner to form a bilayer in which the fatty acid chains face each other making the inner part of the membrane hydrophobic and the outer surface hydrophilic. Although the membrane serves as a barrier to the outside world to the interior of the cell, materials pass through the membrane, so the interior is not isolated. Low molecular weight substances such as water and other solvents can move through the membrane by diffusion. High molecular weight molecules – nutrients and proteins – can be brought into the cell through the membrane through active transport systems by means of specific protein channels opening to allow in- and out-take. The cell membrane also acts as an anchor point for specific receptor proteins which participate in signal transduction and intercellular communication.

2.1.3 Nucleus

The nucleus is a cell organelle which contains most of the cellular genetic material. Its primary function is the storage of information needed for cellular division; a secondary role is control of chemical reactions within the cytosol. The nucleus is enclosed in a two-layer membrane which fuses at evenly spaced points to form nuclear pores which facilitate the transport of water soluble molecules across the nuclear envelope. The outer membrane is fused to the rough endoplasmatic reticulum. The volume of the nucleus is filled with a matrix called nucleoplasm which contains one or more nucleoli, chromatin, enzymes, proteins and transcription factors. The nucleolus is made of proteins and ribosomal ribonucleic acids (rRNA); its function is to transcribe the genetic material in order to form polymeric rRNAs which, after the synthesis, exit to the cytosol through the nuclear pores. Genetic material is present in the nucleus as a protein-DNA complex called the chromatin. The DNA is made of a number of discrete units called the chromosomes. There are two types of chromatin: euchromatin and heterochromatin. Euchromatin is the least compact form of DNA, and the regions of DNA which constitute euchromatin contain genes which are frequently expressed by the cell. In heterochromatin, DNA is more tightly compacted and regions of DNA which constitute heterochromatin generally contain genes which are not expressed often by the cell. In multicellular

organisms, cells are highly specialized to perform specific functions, hence different sets of genes are required and expressed; the regions of DNA that constitute heterochromatin vary between cell types.

2.1.4 Endoplasmatic Reticulum

The endoplasmatic reticulum consists of an extensive membrane network which forms pocket-like structures. The membrane delineates a volume of space, acting as a gateway to the cytosol from the parts of it which are a continuation of the outer membrane of the nuclear envelope. Parts of the reticulum are covered with ribosomes (nodules assembling amino acids into proteins based on instructions from the nucleus) – due to the rough appearance under electron microscopy it is referred to as the rough endoplasmatic reticulum (RER); other parts are free of ribosomes and are called smooth endoplasmatic reticulum (SER). The RER synthesizes and transports proteins which make organellar membranes and excretes free proteins. The SER has functions in several metabolic processes: synthesis of lipids, carbohydrate metabolism and other enzymatic reactions.

2.1.5 Mitochondrion

The mitochondrion can be described, in simple terms, as a power plant of the cell since its primary function is to produce energy in the form of adenosine triphosphate (ATP). The number of mitochondria in the cytosol depends on the cell's level of metabolic activity, more activity means more mitochondria. The mitochondrion is an oval-shaped organelle, ranging from 1 to 4 μm in length. It is enclosed in a membrane made of the inner and outer membranes which are composed of phospholipid bilayers and proteins. The two membranes have very different properties. The outer mitochondrial membrane, which encloses the entire organelle, is composed of about 50 % phospholipids by weight, and contains a wide spectrum of proteins involved in many diverse metabolic activities. The inner mitochondrial membrane has a very high protein to phospholipid ratio, more than 3:1 by weight. Besides the primary function of converting organic materials into energy in the form of ATP, mitochondria play a prominent role in a number of important metabolic tasks, such as apoptosis, cellular proliferation, heme and steroid synthesis, heat production etc.

2.1.6 Golgi Aparatus

The Golgi apparatus consists of a membranous structure similar to the one found in the endoplasmatic reticulum. Its primary function is to process and transport proteins targeted to the plasma membrane, lysosomes or endosomes, and those that will be secreted from the cells. In this way, it functions as a central protein distribution and delivery system for the cell.

2.1.7 Lysosomes, Endosomes and Peroxisomes

Lysosomes are sack-like organelles involved in digesting matter brought into the cell from the outside (endocytosis), the autophagy pathway (digestion of the cell's internally produced substances or particles), and hydrolysis of cellular components. When lysosomes contact the structure which contains substances to be digested, it fuses with it to form a new secondary lysosome. At the heart of the digestive processes lay a number of enzymes: proteases, nucleases, glycosidases, lipase, phosphatases, etc.

The vesicles formed by pinocytosis (the cytoplasmic membrane invaginates and pinches off placing small droplets of fluid and the foreign body in a pinocytic vesicle, afterwards the liquid contents of the vesicle is slowly transferred to the cytosol) are generally referred to as endosomes.

Peroxisomes are spherical membrane limited organelles about 0.5 μm in diameter. They contain an enzyme that hydrolyzes hydrogen peroxide (H_2O_2). The primary function of peroxisomes is to protect the cell against H_2O_2; they also play a role in dergradation of purine bases, amino acids and lipids.

2.1.8 Microtubules, Filaments and Centrioles

Microtubules are hollow cylinders about 20 nm in diameter and hundreds of micrometers in length; the main constituent is a protein called tubulin. They are distributed throughout the cytoplasm in different patterns. These organelles are involved in the elongation and movement of the cell, intracellular transport and movement of chromosomes during mitosis.

Filaments range in diameter and size and are divided in two groups: microfilaments and intermediate filaments. Both classes are composed mostly of two proteins: actin and myosin. Filaments are responsible for the contractile and viscoelastic properties of the cytoplasm; they also have supporting and general structural function.

The spherical structural components of the cell center that are involved in mitosis are the centrioles; they are equally important in forming the new microtubular system in daughter cells.

2.1.9 Ribosome

A ribosome is an organelle composed of rRNA (synthesized in the nucleolus) and ribosomal proteins. Its function is to translate the messenger ribonucleic acid (mRNA) into a polypeptide chain, i.e. a protein. Ribosomes can float freely in the cytoplasm or bind to the endoplasmic reticulum.

Free ribosomes occur in all cells, and also in mitochondria in eukaryotic cells. They usually produce proteins that are used in the cytosol or in the organelle they occur in.

Bound ribosomes, in general, produce proteins that are used within the cell membrane or are expelled from the cell via exocytosis. When proteins are synthesized by a bound ribosome, they become "membrane-bound", associated with the membrane of the nucleus or the rough endoplasmic reticulum (RER). The newly produced polypeptide chains are inserted directly into the ER, from where they are transported to their destinations.

2.2 Cell Division and Cycle

Division of cells can proceed in two ways: mitosis and meiosis. These processes are different in nature and serve two very distinct purposes. Somatic

cells undergo mitotic cellular division, in this way they preserve diploid set of chromosomes (2n) for each of the daughter cells. Meiosis imposes a reduction in number of chromosomes from 2n to n producing gametes and assures interchange of genetic material. In short, mitosis preserves a continuation of hereditary material, while meiosis enables diversification of hereditary material.

The cell cycle is defined as the interval between the completion of mitosis of the parent cell and the completion of mitosis in one or both of the daughter cells. Human cells can be classified in three populations: static, stable and renewing according to the displayed mitotic activity. Static cell populations include cells that no longer (nerve and skeletal muscle cells) or rarely (cardiac muscle cells) divide. Stable cell populations are characterized by low mitotic activity but are able to divide at higher rate under certain conditions (cells such as fibroblasts, osteoblast and epithelial cells of organs). Renewing cell populations engage in regular mitotic activity (blood cells and epithelial cells of skin).

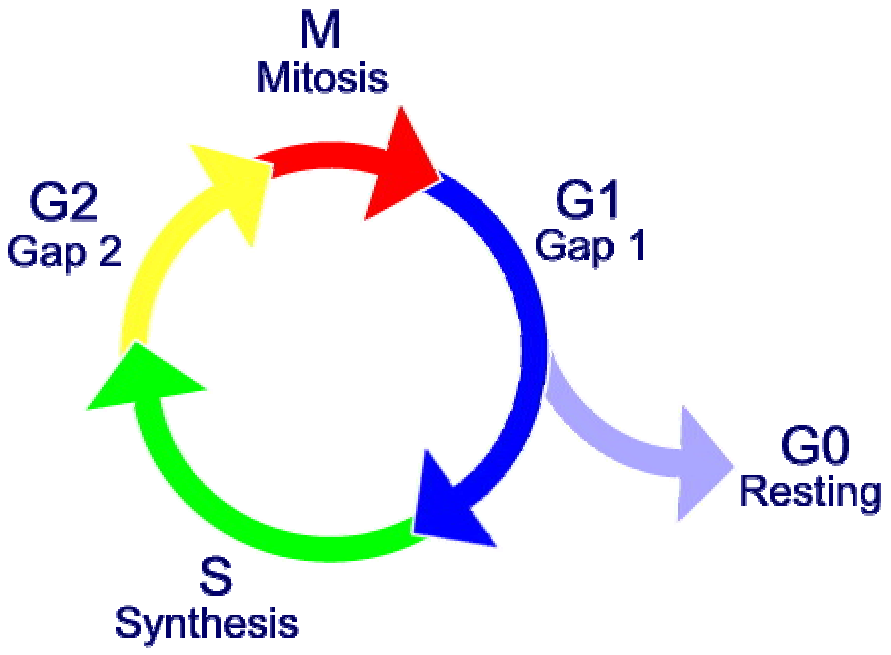

Figure 2.2 Schematic representation of the cell cycle showing the sequence and period of time it takes to complete each of the stages [taken from reference 3].

The cycle is divided into two principle stages: mitosis and interphase according to the changes in the DNA synthesis during repeating divisions. Interphase can be further subdivided into gap 1 (G1), synthesis (S) and gap 2 (G2) stages. Cells that do not undergo division (static cell population) in their lifespan are locked into gap 0 (G0) phase – a period in the cell cycle where cells exist in a quiescent (rest) state. A surveillance system, so-called "checkpoints", monitors the cell for DNA damage and failure to perform critical processes. Checkpoints can block progression through the phases of the cell cycle if certain conditions are not met. The time needed to complete a cell cycle depends on the

type of the cell and function it performs. Human cells grown in vitro usually go through one cycle in 24 hours; mitosis takes ~1 hour, gap 1 (G1) ~8 hours, synthesis (S) ~10 hours and gap 2 (G2) ~5 hours.

- G1 phase covers the period between mitosis and DNA replication. In this phase cytoplasmatic growth occurs and the cell is preparing its enzymatic machinery for the next stage.
- S phase is the period during which DNA replication takes place doubling the amount of genetic material.
- G2 phase is the period between DNA replication and mitosis. During this stage a rapid integrity check of the replicated DNA is performed.

Mitosis is a dynamic process but for purposes of description it is broken down into five subphases: prophase, prometaphase, metaphase, anaphase, and telophase.

- Prophase – chromatin in the nucleus begins to condense and becomes visible in the light microscope as chromosomes. The nucleolus disappears. Centrioles begin moving to opposite ends of the cell and fibers extend from the centromeres. Some fibers cross the cell to form the mitotic spindle.
- Prometaphase – at this stage the nuclear envelope breaks down so there is no longer a recognizable nucleus. Some mitotic spindle fibers

elongate from the centrioles and attach to kinetichores (protein based anchoring point on the chromosome). Other spindle fibers elongate and overlap each other at the cell center.

- Metaphase – spindle fibers align the chromosomes along the middle of the cell, this line is called to the metaphase plate. This organization helps to ensure that in the next phase of the chromosome separation, each new nucleus will receive one copy of each chromosome.
- Anaphase – The paired chromosomes separate and move to opposite sides of the cell. Motion results from a combination of kinetochore movement along the spindle microtubules and through the physical interaction of polar microtubules.
- Telophase – chromatids (daughter chromosomes) arrive at opposite poles of cell, and new membranes form around the daughter nuclei. The chromosomes disperse and are no longer visible under the light microscope. The spindle fibers disperse, and cytokinesis (partitioning of the cell) may also begin during this stage.

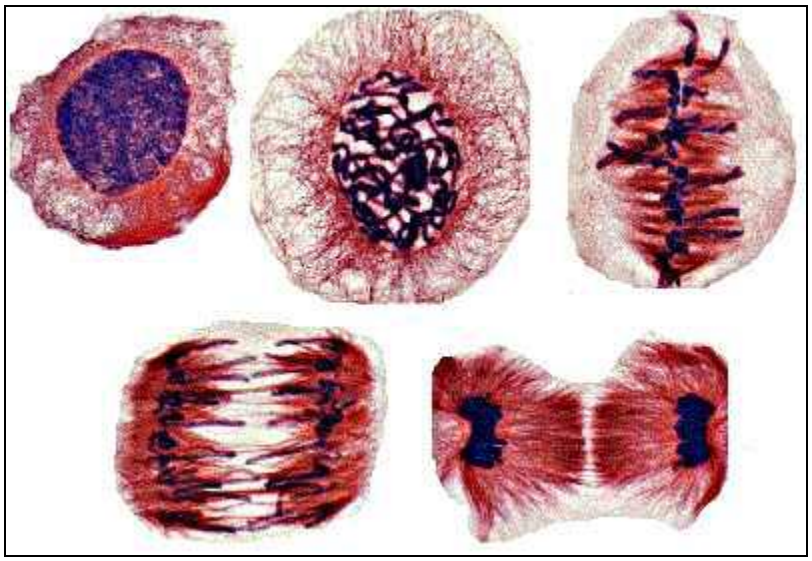

Figure 2.3 Eukaryotic cell undergoing mitosis, starting from the top left, phases are: prophase, prometaphase, metaphase, anaphase, and telophase [taken from reference 4].

2.3 Lymphatic System

The lymphatic system can be used as an anatomical term to identify the part of the circulatory system whose purpose is collection and drainage of lymph from various tissues of the body [1]. The lymphatic tissues and organs include the thymus, bone marrow, lymphatic nodules (lymphoid follicles), lymph nodes and spleen. The term lymphatic system is sometimes used to describe the immune system due to its function in the human body – surveillance and defense from exogenous bodies (free substances, viruses, microorganisms or cells) that enter the human body and endogenous constituents (cancer cells) of the body.

The major function-specialized constituents of lymphatic tissue are lymphocytes and macrophages which are contained in the framework made of reticular cells and reticular fibers.

2.3.1 Lymph Nodes

Lymph nodes are oval, encapsulated organs ranging in size from about 1 mm to 2 cm in the longest dimension. They are placed in the pathway of lymphatic vessels and serve as filters and barriers through which lymph flows on its way to the blood.

A schematic diagram of a lymph node is shown in Figure 2.4. The substance of the lymph node is divided into a cortex (domain of T-lymphocytes), including paracortical region and a medulla. The whole organ is enveloped in the capsule comprised of dense connective tissue. The cortex contains spherical aggregates of lymphocytes called lymph nodules which are the domain of B-lymphocytes; in an active node these, lighter in color, centers are called germinal centers. The medulla extends all the way to the hilus (stem of a node) and consists of regions of lymphatic tissue, appearing as irregular cords, separated by the medullary sinuses. The dense population of lymphocytes between the cortex and the medulla constitutes paracortex which contains postcapillary venules.

Trabeculae extend from the capsule into the substance of the node; adjacent to them are the lymph sinuses designated as subcapsular or cortical and trabecular respectively. Afferent lymphatic vessels empty into subcapsular sinuses which communicate with trabecular and medullary sinuses. The efferent lymphatic vessel is the point where the lymph exits a node.

Lymph nodes function as filters of the lymph. Particular matter undergoes phagocytosis by the phagocytic cells and this can be considered as the basic aspect of filtration. When antigens conveyed in the lymph percolate through the sinuses and penetrate the lymph nodules some of them are trapped on the surface of specialized cells known as follicular dendritic cells. Trapped in this way, the antigens are subjected to immunosuppression by the memory B-cells. The involvement of T-cells may become necessary to facilitate activation of the B-cells. Activated B-cells migrate to germinal centers and undergo transformation to immature immunoblasts. These cells proliferate and give rise to plasma and memory cells. The plasma cells migrate to the medullary cords where they synthesize and release specific antibodies into the lymph flowing through the sinuses. The memory cells may leave the lymph node and circulate to the various regions of the human body.

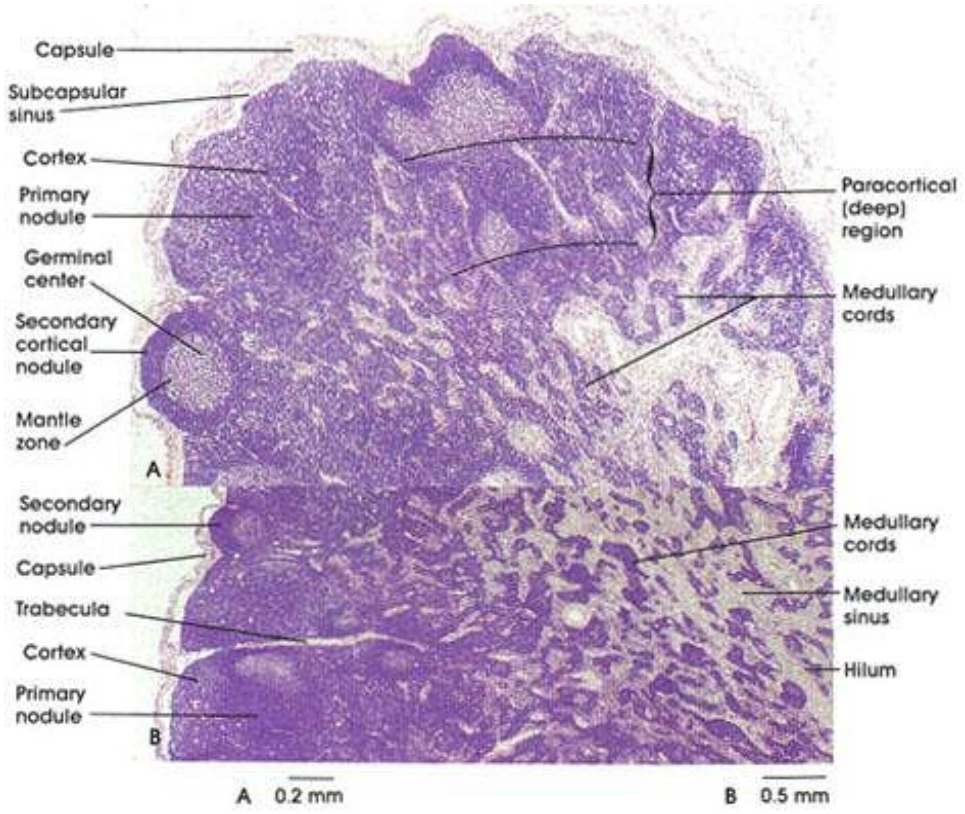

Figure 2.4 Microscopic images of a tissue section excised from the human lymph node. The slice was fixed in formalin and stained using hemotoxin and eosin histo-pathological dyes (H&E). Microscope magnification for section A was 40x and 25x for section B [taken from reference5].

2.3.2 Lymphocytes

Lymphocytes are migratory cells which circulate throughout the lymphatic vessels and bloodstream. The mobility enables them to move from one part of the lymphatic system to another and reach sites in the body where they are needed. There are two types of lymphatic cells: T-lymphocytes (T-cells) and B-lymphocytes (B-cells).

2.3.2.1 T-lymphocytes

The T-lymphocytes originate from the bone marrow and later differentiate into immunocompetent cells in the thymus; they have long life span. The T-cells are involved in cell-mediated immunity response. There are three fundamentally different types of T-cells:

- Cytotoxic lymphocytes (CTL, also know as killer T-cells) – they serve as the primary effector cells that recognize other cells having foreign antigens on the surface and kill them by lysis.
- Helper T-lymphocytes (TH cells) – a major function of TH cells is to recognize foreign antigens and by secretion of lymphokines stimulate B-cells, CTLs or macrophages.
- Suppressor T-lymphocytes (TS cells) – they act as function regulators of the immune response by suppressing the activity of T- and B-lymphocytes.

2.3.2.2 B-lymphocytes

B-lymphocytes originate from the bone marrow from where they later differentiate into immunocompetent cells, they have variable life span. Coming

in a contact with an antigen, B-lymphocytes become activated and transform into immunoblasts which start proliferating and at then differentiate into plasma cells or memory cells. The plasma cells synthesize and secrete antibody which binds to the antigen and eliminates it in a variety of ways. The memory cells do not take any role in the primary response to a specific antigen, they are programmed to be ready to respond to the same antigen should it appear again.

2.3.3 Lymphomas and Secondary Cancer of the Lymph Nodes

Lymphoma is a term used generally for malignancies of lymphocytes. Secondary cancer of the lymph nodes relates to a condition when a cancer from some other part of human body or organ has spread to the lymphatic system and infected the lymph nodes.

Lymphomas are traditionally divided in two groups: Hodgkin's and non-Hodgkin's lymphomas. The way to differentiate between these two diseases is to microscopically investigate thin sections of the lymph node excised for the biopsy, which have been stained with immuno-histological reagents.

In case of Hodgkin's lymphoma, microscopic examination of the lymph node biopsy reveals the completely or partially effacement of the lymph node architecture by scattered large malignant cells known as Reed-Sternberg cells.

Hodgkin's disease (now also known as Hodgkin's lymphoma) can be subclassified into four pathologic subtypes based upon Reed-Sternberg cell morphology: nodular sclerosing subtype, lymphocyte predominant subtype, lymphocyte depleted subtype, and mixed-cellularity subtype. After diagnosing the disease, a patient is staged – a series of tests and procedures which determine what areas of the body are affected – into on of the categories ranging from Stage I to IV [6].

- Stage I indicates that the cancer is located in a single region, usually one lymph node and the surrounding area.
- Stage II indicates that the cancer is located in two separate regions, an affected lymph node or organ within the lymphatic system and a second affected area, and that both affected areas are confined to one side of the diaphragm - that is, both are above the diaphragm, or both are below the diaphragm.
- Stage III indicates that the cancer has spread to both sides of the diaphragm, including one organ or area near the lymph nodes or the spleen.
- Stage IV indicates that the cancer has spread beyond the lymphatic system and involves one or more major organs, possibly including the bone marrow or skin.

All other lymphomas which do not fit Hodgkin's lymphoma description are grouped together and called non-Hodgkin's lymphomas. Classification of non-Hodgkin's lymphoma is primarily descriptive; in 2001 the World Health Organization suggested a classification scheme according to cell type [7], i.e. the normal cell type that most closely resembles the tumor. They are classified in three large groups: the B-cell tumors, the T-cell and natural killer cell tumors. Staging is the same as for Hodgkin's lymphoma.

Cancer has the ability to spread away from the area where it first started. The place where the cancer first started is known as the primary site. The cancer cells travel either in the bloodstream or in the lymphatic system. When cancer cells break away from the primary site and settle in the lymph nodes near the primary tumor (this is called lymph node involvement or regional disease) or elsewhere in the body, they can stay dormant for many years or can grow into a secondary tumor or metastasis. Morphologically these cells generally look similar to cells from the primary site. Classification is carried out according to the primary tumor type and aggressiveness of the cancer, and staging is the same as mentioned in the previous paragraph.

It is estimated that in the USA in 2005, 4.6 % of all cancer cases will be Hodgkin's lymphoma, and out of these 4.1 % will be non-Hodgkin's lymphoma [8]. The number of deaths caused by lymphoma is estimated to be

20,610 [8]. The National Cancer Institute's (NIC's) investment in lymphoma research will increase to $118 million in fiscal year 2005 [9].

The exact cause of lymphoma remains unknown. Some forms of lymphoma are more likely to develop in people who have taken drugs to prevent rejection of an organ following a transplant or who have reduced immunity (people with HIV or AIDS), infection with the Epstein-Barr virus increases the likelihood of the disease, and there is a rare type of lymphoma which is known to be caused by a type of bacterial infection known as *Helicobacter pylori*. It is thought that certain chemicals in insecticides and wood preserving materials may cause lymphoma. Lymphomas are not contagious and cannot be passed down through families (not inherited).

3 Methods and Materials

3.1 Infrared Spectroscopy

Spectroscopy is the study of interaction between matter and electromagnetic (EM) radiation – more precisely the way in which atoms and molecules absorb or emit electromagnetic waves.

If a beam of infrared light is passed through a medium and then decomposed into the monochromatic components, it can be observed that attenuation of the incident light occurs at particular wavelengths. The root cause of this phenomenon is selective absorption by molecules or ions. If intensity of the transmitted light is plotted against the wavelength we get an infrared spectrum.

The infrared range covers wavelengths between visible (~ 800 nm) and microwave (~ 1 mm) portion of EM spectrum. In the IR literature it is more common to use the wavenumber unit (ν in cm^{-1}) so in the following discussions the text will follow this convention. According to the type of energy transition IR radiation range can be further subdivided into near- (12500 – 4000 cm^{-1}), mid- (4000 – 650 cm^{-1}) and far-infrared (200 – 100 cm^{-1}) spectral regions.

3.1.1 Beer's Law

Whenever a beam of polychromatic radiation passes through a medium some loss of intensity occurs. A number of processes go on during the time it takes the incident radiation to traverses the medium. First, reflection takes place at the phase boundaries as a result of refractive index differences between the medium and its surroundings. Second, scattering caused by inhomogeneities or by thermal fluctuations in the bulk of the medium produces an additional small loss of power from the main beam. In general, neither of these is as significant in accounting for the intensity decrease as the fact that the medium will absorb the radiant frequencies that promote energy changes within its atoms, molecules or ions.

For each wavelength present in the beam, the absorbance depends on the nature of the medium, i.e. its composition, and the length of optical path in the medium. The dependence is expressed by Beer's law:

$$\log \frac{P}{P_0} = -\varepsilon bc \qquad \text{Equation 3.1}$$

$$P = P_0 10^{-\varepsilon bc} \qquad \text{Equation 3.2}$$

Taking the reciprocal of the ratio P / P_o removes the negative sign and gives:

$$\log \frac{P_0}{P} = A = \varepsilon bc \qquad \text{Equation 3.3}$$

where P is intensity of incident radiation, P_o is intensity of transmitted radiation, ε is molar absorptivity, b is the path length and c is concentration of the absorbing species. The ratio P / P_o is also called transmittance and it is expressed in units of %. The assumptions made in deriving the law are following:

- The incident radiation is monochromatic.
- The absorbing species (atoms, molecules or ions) act independently of one another regardless of number and kind.
- The absorption is limited to a volume of uniform cross-section.

It should be noted that for substances following Beer's law the molar absorptivity is independent of concentration and optical path length. It is a constant characterizing a substance at a particular wavelength when dissolved or dispersed in a given medium. The restriction that absorptive species do not interact with themselves or others present causes Beer's law to be applicable mainly to dilute solutions or mixtures.

3.1.2 Theoretical Basics of Vibrational Spectroscopy

Taking into consideration factors that determine the frequency of vibration of a chemical bond, i.e. the frequency of IR radiation that it absorbs, it is a common practice, for the simplicity of analogy, to represent a diatomic

molecule in a form of mechanical harmonic oscillator (two weights connected by a spring) [10]. In this model the weights with masses m_1 and m_2 represent atoms constantly vibrating between and the spring connecting them has the role of a chemical bond. Applying Hooke's law for mechanical harmonic oscillator, the expression for the natural frequency of vibration is derived:

$$\nu = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}} \qquad \text{Equation 3.4}$$

where k is the force constant of chemical bond and μ is the reduced mass:

$$\mu = \frac{m_1 m_2}{m_1 + m_2} \qquad \text{Equation 3.5}$$

By examining Equation 3.4., it can be concluded that the frequency of the absorbed IR radiation is dependant on the strength of chemical bond and masses of atoms involved in vibration. In order to absorb IR radiation, a molecular vibration must cause a change in the dipole moment of the molecule. The absorption frequency depends on the molecular vibrational frequency, the absorption intensity depends on how effectively the impinging infrared photons can interact with the molecule and this depends on the change of the dipole moment that occurs as a result of molecular vibration.

Absorption of IR radiation, like all other EM radiation absorptions, is a quantized process. Energy transitions between vibrational levels, induced by infrared absorption, cover the range of 8 – 40 kJ/mol. These vibrational energy

levels are characterized by vibrational quantum numbers $n = 0, 1, 2\ldots$ Applying quantum mechanics in conjunction with Hook's law to diatomic molecule gives the expression for calculating potentional energy of vibrational levels:

$$E_n = (n + \tfrac{1}{2})\frac{h}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \text{Equation 3.6}$$

$$E_n = (n + \tfrac{1}{2})h\nu \quad n = 0, 1, 2\ldots \qquad \text{Equation 3.7}$$

Energy needed for transition from the ground state to first excited level ($n = 0 \rightarrow n = 1$) equals the energy with natural frequency of particular chemical bond:

$$E_{0\rightarrow 1} = E_1 - E_0 = h\nu \qquad \text{Equation 3.8}$$

Selection rules, which deal with probabilities of transitions between different vibrational levels, allow only those transitions where $\Delta n = 1$, most of bands in an IR spectrum are of this kind – other types of transitions ($\Delta n > 1$) are not allowed according to the selection rules and mostly do not occur. If there are non-allowed changes of vibrational energy levels, the IR bands are weak in intensity – they are called overtones. It is also possible to have combination and difference bands show up in IR spectrum at positions that correspond to sum or difference in frequencies of fundamental bands ($\nu_1 + \nu_2$ or $\nu_1 - \nu_2$), they also are not allowed by the selection rules. When an overtone or combination band has almost the same value as the frequency of another fundamental vibrational

transition, two new relatively strong bands may appear in the IR spectrum where only one strong band for the fundamental was expected. This effect is called Fermi resonance.

The total number of all possible vibrations in a molecule consisting of *N* atoms can be calculated taking into consideration degrees of freedom in three-dimensional space, reduced by the number of translations.

3N – 6 (Nonlinear molecule)

3N – 5 (Linear molecule)

There are two major types of vibrations: stretching and bending.

The number of absorption maxima, which correspond to the vibrational transitions, in an IR spectrum is usually smaller than the number of all possible ones. This is due to following reasons:

- IR radiation is absorbed only by vibrational energy transitions that cause a change in the dipole moment – these are called IR active.
- In highly symmetrical molecules a number of vibrational transitions can absorb at the same frequency (degenerate vibrations), this in turn will give a rise in intensity to usually weak IR bands.
- Frequencies of vibrations involving heavy atoms, thus having high reduced mass, are low and can be out of spectrometer's working range.

- Some fundamental bands can not be detected due to very weak intensities.

3.1.3 Fourier Transform Infrared Spectroscopy

In dispersive infrared spectrometers every wavelength is being sampled one at the time by the detector. Fourier transform infrared (FT-IR) spectrometers overcome this and several other disadvantages of dispersive systems by incorporating following design features [10]:

- All wavelengths are present and sampled simultaneously – Fellgett or multiplex advantage.
- Optical beams are ordinarily wide, not restricted by slit – Jacquinot or throughput advantage.
- Mirror travel is monitored by laser with high precision – Connes or precision wavelength advantage.

At the heart of all FT instruments lies the interferometer, most of the commercially available instruments are based on the Michelson interferometer, in the past some were implementing a Fabry-Perot interferometer.

Schematic layout of a FT-IR spectrometer is shown in Figure 3.1. The basic parts are source, interferometer – consisting of IR beam-splitter, fixed and

moving mirror and laser – and detector. The IR beam goes from the source to the beam-splitter which splits it in two. One part is transmitted to the moving mirror; the other part is reflected to the fixed mirror. The moving mirror moves back and forth at constant velocity. This movement is timed according to the very precise laser wavelength in the system which also acts as an internal wavelength calibration. The two beams are reflected from the mirrors and recombined at the beam-splitter. The beam from the moving mirror has traveled a different distance than the beam from the fixed mirror. When the beams are combined the interference pattern is formed, since some of the wavelengths recombine constructively (mirror retardation equals integral number of wavelengths $x = m\lambda$) and some destructively (mirror retardation equals $x = (m + ½)\lambda$). This interference pattern is called interferogram.

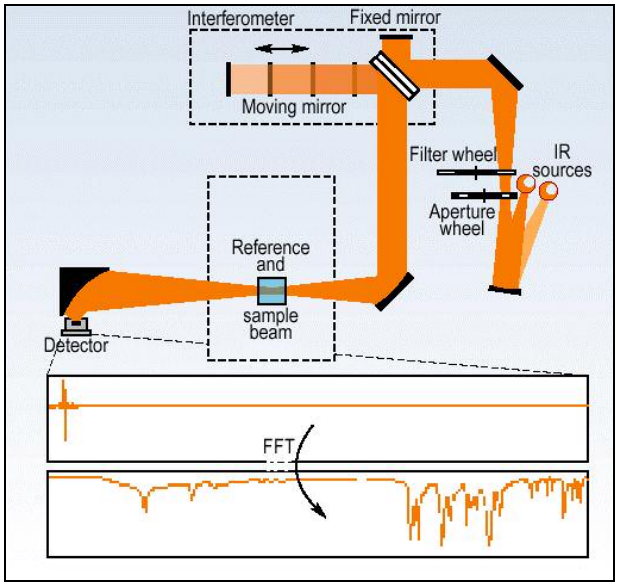

Figure 3.1 Illustration of a Fourier transform infrared spectrometer.

3.1.4 Fourier Transform

To extract chemical information from the interferogram *J(x)* (Equation 3.9) which is an intensity distribution as a function of the movable mirror retardation *x* to *S(ν)* (Equation 3.10), i.e. intensity distribution as a function of frequency, the Fourier transform (FT) of *J(x)* has to be performed [10].

$$J(x) = \int_{-\infty}^{\infty} S(\nu)\cos(2\pi\nu x)d\nu \qquad \text{Equation 3.9}$$

$$S(\nu) = \int_{-\infty}^{\infty} J(x)\cos(2\pi\nu x)dx \qquad \text{Equation 3.10}$$

The transformation of an interferogram to frequency domain is nowadays performed by the fast Fourier transform (FFT). Interferograms are not sampled continuously but at short intervals (discrete mirror retardation positions at discrete times) and principles of discrete Fourier transform (DFT) are applied. As a consequence the process of taking the Fourier transform can be written as:

$$g(kT) = \frac{1}{N}\sum_{n=0}^{N-1} G(n/NT)e^{2\pi ink/N} \qquad \text{Equation 3.11}$$

$$G(n/NT) = \sum_{k=0}^{N-1} g(kT)e^{-2\pi ink/N} \qquad \text{Equation 3.12}$$

where *N* is the total number of data points, *T* is the sampling interval and *n* and *k* are running indexes in *g* and *G* space, respectively. If we set $n / NT = m$ and $kT = p$ Equation 3.12 can be written as:

$$G(m) = \sum_{k=0}^{N-1} g(p) W^{mp} \qquad \text{Equation 3.13}$$

where

$$W^{mp} = e^{-2\pi imp} \qquad \text{Equation 3.14}$$

Equation 3.14 can be written in matrix notation:

$$\mathbf{G} = \mathbf{W}^{mp}\mathbf{g} \qquad \text{Equation 3.15}$$

For a large data sample set, a matrix manipulation would be prohibitively slow; this problem was solved by Cooley and Tukey, for detailed discussion see [10].

3.1.5 Fourier Transform Infrared Microspectroscopy

Coupling of FT-IR spectrometer with the microscope has made possible to perform IR spectroscopic experiments on the microscopic level [11]; this has created a whole new field of biospectroscopy .

A schematic representation of a FT-IR optical bench connected to a microscope is shown in Figure 3.2. Infrared radiation is channeled from the interferometer via the transfer optics and focused by the microscope onto a sample; the visible optical train is parafocal and collinear with the IR radiation.

The general design of a modern IR microscope employs an infinity-corrected optical layout which offers versatility unparalleled by finite conjugate

systems. This enables easy upgrade of base units with multiple objective mounting, Koehler illumination, polarized light filters, refractive objectives, fluorescence attachment, etc. One of the biggest advantages on infinity-corrected design is that it allows the use of flat mirror relay optics in the IR beam path. A novel approach in the IR optical train construction is use of dual confocal aperturing (alternatively single aperture in dual pass configuration) which maintains FT-IR microspectroscopy performance advantages – high IR signal throughput, stray light reduction and use of optical contrasting techniques. The resolution achievable in systems with confocal layout is limited only by the diffraction limit of infrared light and at best can reach 3 μm for the wavelength of 3000 cm^{-1} [12].

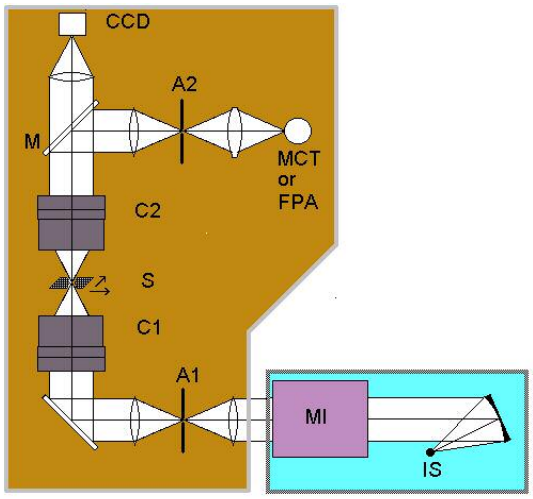

Figure 3.2 Schematic view of an FT-IR microspectrometer. FT-IR spectrometer: (IS) infrared source; (MI) Michelson interferometer. Infrared microscope: (A) aperture; (C1) and (C2) Cassegrain condenser and objective, respectively; (S) sample on computer-controlled microscope stage; (M) moveable mirror; (MCT) HgCdTe defector or (FPA) focal plain array detector; (CCD) visual camera for image capture.

The pathways of visible light and IR beams in the microscope are collinear and pass through the same optical elements; this requires use of all-reflecting objectives and condensers. They can have spherical (Cassegrain design) or elliptical mirror surfaces (Schwarzchild design). Accommodating demands for high signal throughput, good visual image quality and high enough magnification, typical numerical apertures utilized (sine of the angle of the most extreme ray entering the system) are about 0.5 [12,13].

IR microscopes can operate in two modes: transmission and reflection. Data acquisition in transmission mode can be used for IR and visible light transmissive samples. In cases of thick, opaque or non-transparent objects, FT-IR microimaging systems can be used in reflectance mode. The incident IR beam is channeled only through the upper part of the optical pathway and one half of the objective. First, IR radiation traverses the sample and then gets reflected from the suitable IR reflecting sample support, following this it traverses the sample one more time and goes to the detection system via the other half of the optical train.

If a system is equipped with a computer controlled stage, it is possible to perform mapping experiments in addition to point mode measurements. An IR map is collected by scanning the computer controlled microscope stage in a raster pattern in increments corresponding to the aperture size. At each data point a preset number of interferograms is coadded, depending on the aperture

size in order to produce the signal-to-noise ratio required for the subsequent analysis. Every data point in a map is a 4-dimensional representation of raw data in which the dimensions for each pixel point are pixel coordinates X and Y, the wavenumber axis and the intensity axis. These data sets are often called a spectral hypercube.

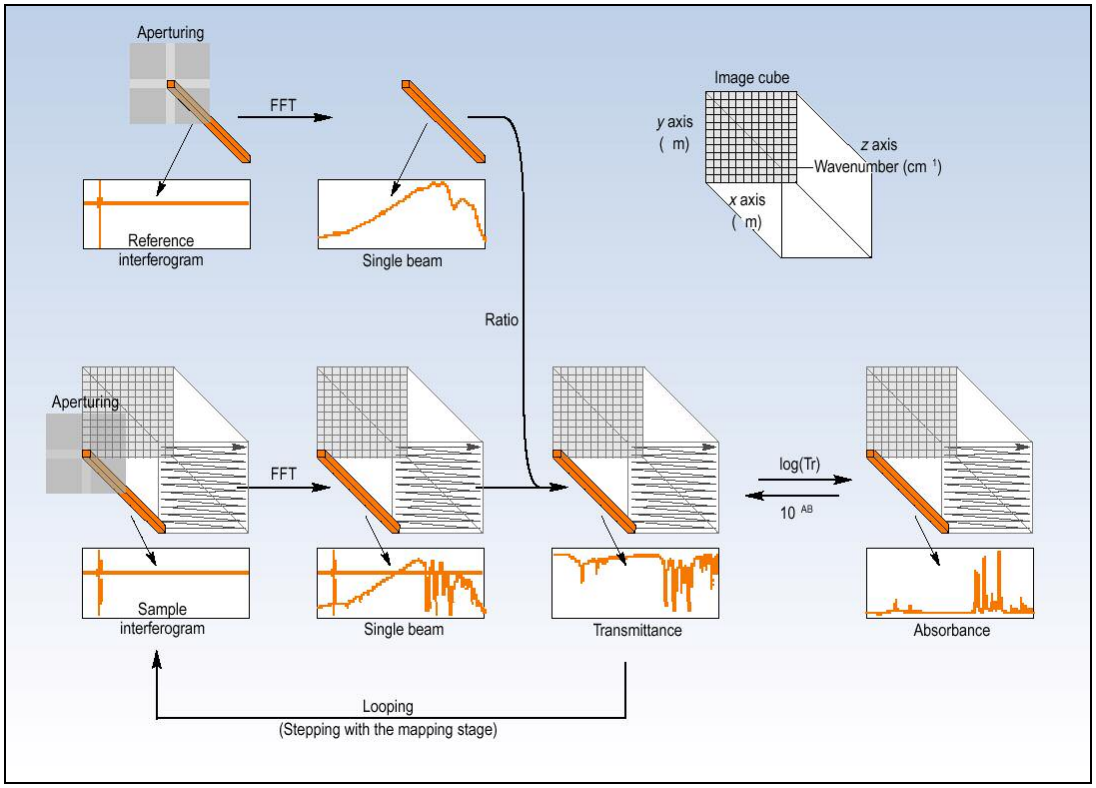

Figure 3.3 Diagram of the mapping process. Every data point in a map is a 4-dimensional representation of raw data in which the dimensions for each pixel point are pixel coordinates X and Y, the wavenumber axis and the intensity axis. These data sets are often called a spectral hypercube.

3.2 Spectral Data Preprocessing and Processing

For the analysis of mapping data sets and further input of variables into the artificial neural network data classification, the spectral hypercube is subjected to a two-step processing routine.

The data manipulation consists of preprocessing step and data correlation step known as hierarchical cluster analysis (HCA). In preprocessing step raw spectra for each pixel are automatically corrected for water vapor and CO_2 at the instrument acquisition level, quality test is performed to test for sample thickness variations and discard pixels with to low signal-to-noise ratio, normalization is applied to compensate for sample thickness variations that have passed the quality test and alternatively spectra can be derivatized (to the first or second derivative) and smoothed. Hierarchical cluster analysis is an unsupervised procedure that establishes and weighs in the degree of similarities of spectral patterns, i.e. individual spectra within a map; and groups close resembling spectra to create a 2D false color map that shows related spectra in the same color.

3.2.1 Atmospheric Correction

Two of the constituents of air are strongly IR absorbing species: water and carbon-dioxide. IR bands belonging to water and CO_2 overlap with bands of interest coming from the mapped tissue samples and can interfere in data analysis. If the concentration of water in the air (relative humidity) stays constant, there is no need for dynamic water and carbon-dioxide subtraction from spectra since it would be subtracted as the background spectrum already containing H_2O / CO_2 bands. Modern instruments implement active data processing for water and carbon-dioxide bands removal and it is done dynamically during the data acquisition process.

3.2.2 Quality Test

The absorbance, integrated over a large spectral region, can be used to assess the sample thickness. To avoid inaccuracies due to too thin samples with low absorbance and poor S/N, or due to too high absorbance with non-linear detector response, spectra may be tested using the "sample thickness" criterion. Indicating the spectral region to be used for obtaining the integral and by defining an upper and a lower threshold for the integral, a spectrum can fail the

sample thickness test if an integration value is determined to be higher or lower then the defined thresholds.

3.2.3 Normalization

To be able to compare different spectra they need to be scaled and this is achieved by normalization function. There is a number of different normalization techniques that can be used for processing of absorbance spectra.

Offset correction is a linear method of correction of the entire spectrum, or it can be applied to a predefined spectral range, in such way that within the designated number of data points the minimum occurring absorbance value is set to zero. Spectra are not scaled in this mode.

Min-max normalization achieves scaling of all spectra between zero and one; the maximal absorbance value in the selected spectral region equals one and the minimal one is set to zero. It can be used to perform simple band normalization.

Vector normalization is carried out in following way: the average value of absorbance of the spectrum is calculated first. This average value is then subtracted from the spectrum so that the middle of it is pulled down to zero. The sum of squares of all absorbance values is then calculated and the spectrum is

divided by the square root of this sum. The vector norm of the resulting spectrum is equal to one.

3.2.4 Derivatization of Spectra

An infrared spectrum can be considered as a zeroth order function. By deriving an IR spectrum to the first or second order certain advantages can be gained; small and hard to notice changes in IR band envelope shapes are more pronounced and apparent, also, the changes and shifts in baseline are corrected and offset by the derivatization procedure. If there is a low signal-to-noise ratio for a given spectrum, derivatizing it may reduce the spectral quality by filtering in the noise as more prominent feature of the 1^{st} or 2^{nd} order spectrum.

The Savitzky-Golay algorithm [14] uses a generalized moving average smoothing. Smoothing filter coefficients are derived by performing an unweighted liner least squares fit using a polynomial of a given degree that acts as a convoluting function.

3.2.5 Hierarchical Cluster Analysis

Hierarchical cluster analysis [15] is an unsupervised procedure for establishing similarities between elements of a group into a hierarchy of nested

partitions. The hierarchy can be visualized by a tree-like chart known as a dendrogram in which each cluster is nested into the next one. Hierarchical algorithms can be categorized in two types:

- Agglomerative procedure – starts with n clusters equal to the number of elements in the group and iteratively reduces the number of clusters by merging the two most similar elements or clusters respectively until only one is remaining ($n \rightarrow 1$).
- Divisive procedure – starts with one cluster and iteratively splits it to reduce heterogeneity as much as possible ($1 \rightarrow n$).

Divisive procedures have the advantage of being more robust, while agglomerative ones are more efficient in terms of computational time.

The basic steps of agglomerative algorithms are:

- Calculate the distance between all objects and construct the similarity distance matrix.
- Find two clusters i and j with the minimal distance to each other.
- Merge the clusters i and j and replace i with the new cluster, delete j and recalculate all distances.
- Repeat last two steps until the total number of cluster becomes one.

3.2.6 Correlation Measurements

All clustering algorithms use similarity measurements between two vectors, mostly by calculating linear correlation coefficients, for consequent association to a new element. There are a number of algorithms available [16]:

$$d_{jk} = \sqrt{\sum_{i=1}^{n} (x_{ji} - x_{ki})^2} \qquad \text{Equation 3.16}$$

Euclidian distance

$$\left.\begin{aligned} d_{jk} &= \sqrt{\sum_{i=1}^{n} \left(\frac{x_{ji} - x_{ki}}{std(x_i)} \right)^2} \\ std(x_i) &= \sqrt{\frac{1}{p-1} \sum_{m=1}^{p} (x_{mi} - \overline{x_i})^2} \end{aligned}\right\} \qquad \text{Equation 3.17}$$

Normalized Euclidian distance

$$r_{jk} = \frac{\sum_{i=1}^{n} (x_{ji} - \overline{x_{ji}})(x_{ki} - \overline{x_{ki}})}{\sqrt{\sum_{i=1}^{n} (x_{ji} - \overline{x_{ji}})^2 \sum_{i=1}^{n} (x_{ki} - \overline{x_{ki}})^2}} \qquad \text{Equation 3.18}$$

Pearson's correlation coefficient

$$\left.\begin{aligned} r_{jk} &= \frac{\sum_{i=1}^{n}(x_{ji}-\overline{x_{ji}})(x_{ki}-\overline{x_{ki}})}{\sqrt{\sum_{i=1}^{n}(x_{ji}-\overline{x_{ji}})^2\sum_{i=1}^{n}(x_{ki}-\overline{x_{ki}})^2}} \\ d_{jk} &= 1000(1-r_{jk}) \end{aligned}\right\}$$ Equation 3.19

D-values

3.2.7 Linkage Rules

At the beginning of a cluster analysis each object represents its own cluster and the distances between objects are defined by the chosen distance measure. Once several objects have been linked together, a linkage or amalgamation rules is need to determine if two clusters are sufficiently similar to be linked together. There are numerous linkage rules in use and here are some of the most commonly used [16]:

- Single linkage – the distance between two clusters is determined by the distance of two closest objects (nearest neighbors) in the different clusters.

  $$d_{jk} = \min(d_{ji}, d_{ki})$$ Equation 3.20

- Complete linkage – the distance between two clusters is determined by the greatest distance between any two objects in the different clusters (furthest neighbors).

$$d_{jk} = \max(d_{ji}, d_{ki}) \qquad \text{Equation 3.21}$$

- Group average – the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters.

$$d_{i,jk} = \left(\frac{n_j}{n_j + n_k}\right) d_{ji} + \left(\frac{n_k}{n_j + n_k}\right) d_{ki} \qquad \text{Equation 3.22}$$

- Centroid linkage – the centroid of a cluster is the average point in the multidimensional space, in other words, it is the center of gravity for the respective cluster. This method calculates the distance between two clusters as the difference between centroids.

$$d_{i,jk} = \left(\frac{n_j}{n_j + n_k}\right) d_{ji} + \left(\frac{n_k}{n_j + n_k}\right) d_{ki} - \left(\frac{n_j n_k}{(n_j + n_k)^2}\right) d_{jk} \qquad \text{Equation 3.23}$$

- Ward's algorithm – this method is based on minimizing the sum-of-squares of any two clusters that can be formed at each step.

$$d_{i,jk} = \frac{1}{n_i + n_j + n_k}\left[(n_i + n_j) d_{ji} + (n_i + n_k) d_{ki} - n_i d_{jk}\right] \qquad \text{Equation 3.24}$$

3.3 Artificial Neural Networks

Artificial neural networks (ANN) are nonlinear mapping systems whose structure is loosely based on principles observed in the nervous systems of human and animals [17]. In general terms, an ANN consists of a large number of simple processors linked by weighted connections – by simple analogy those are called neurons. Each unit receives input from many other nodes and generates a single scalar output that depends only on the locally available information, either stored internally or arriving via weighted connections. The output is distributed to and acts as an input to other processing nodes.

A single unit is not powerful by itself, the power of the system emerges from the combination of many units in a network. A network is specialized to implement different functions by varying the connection topology and values of connecting weights.

A network can have an arbitrary structure, but layered architectures are the most common ones. The multilayered perceptron is a structure where units are arranged in layers and connected so that units in layer L receives inputs from the preceding layer $L–1$ and sends outputs to the following layer $L+1$. External inputs are applied to the first layer and system outputs are taken at the last layer. Internal layers not observable from the inputs or outputs are called hidden layers.

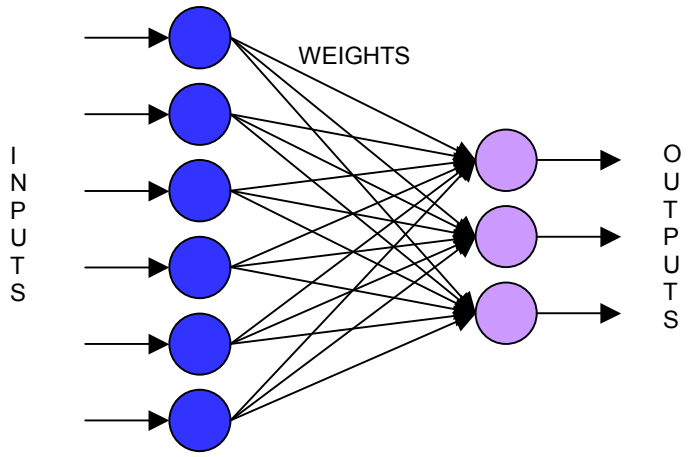

Figure 3.4 Illustration of a single layer, feed-forward artificial neural network.

The network shown in Figure 3.4 has a feed-forward structure, meaning that there are no connection loops that would allow outputs to feed back to their inputs and change the output at a later time. The network implements a static mapping that depends only on its present inputs and is independent of previous system states. Recurrent networks with feedback have a wider range of possible behaviors, but analysis and training are more difficult. The network in Figure 3.4 is fully connected; that is, every node in layer *L* receives inputs from every node in the preceding layer *L–1* and projects outputs to every node in the following layer *L+1*.

One of the reasons why artificial neural networks are so interesting is that they "learn" from examples – given examples of input patterns and desired outputs, algorithms exist that adjusts the weights so the network produces the correct output for each training input. If a system is adaptive it's also potentially

self-repairing, small errors may be tolerated because the weights can be adjusted to compensate.

In machine learning, supervised learning has come to mean the process of adjusting a system so it produces specified outputs in response to specified inputs. It is often posed as a function approximation problem. Given training data consists of pairs of input patterns **x** and corresponding desired outputs (targets) *t*, the goal is to find a function $y(\mathbf{x})$ that matches the desired response for each training input. The functional relationship is usually unknown so the idea is to start with a system flexile enough to implement many functions and adjust to fit the given data set.

Training refers to the adaptation process by which the system learns the relationship between the inputs and targets. This is usually a repetitive process guided by an optimization algorithm. The process is supervised in the sense that an external teacher must specify the correct output for each and every input pattern. During training, each input pattern is presented and propagated through the network to produce an output. Unless the network is perfectly trained, there will be the differences between the actual and desired output. The significance of these deviations is measured by an objective function whose output rates the quality of the networks response – a term cost or error function is also used. The overall goal is to find a system that minimizes the total error for

the given training set. When defined in this way, training becomes a statistical optimization problem and there are a number of factors to be considered:

- Variable selection and representation.
- Selection and preparation of training data.
- Network structure selection.
- Choice of error function.
- Choice of optimization method.
- Prior knowledge and heuristics.
- Generalization.

Single-layer networks (Figure 3.4) have just one layer of active units. Inputs connect directly to the outputs through single layer of weights. Each unit produces its output by forming a weighted linear combination of its inputs which is then passed through a linear or nonlinear function:

$$u_i = \sum_j w_{ij} x_j \qquad \text{Equation 3.25}$$

$$y = f(u) \qquad \text{Equation 3.26}$$

This can be expressed more compactly in vector notation:

$$y(\mathbf{x}) = f(\mathbf{w}^T \mathbf{x}) \qquad \text{Equation 3.27}$$

Where **x** and **w** are column vectors with elements x_j and w_j.

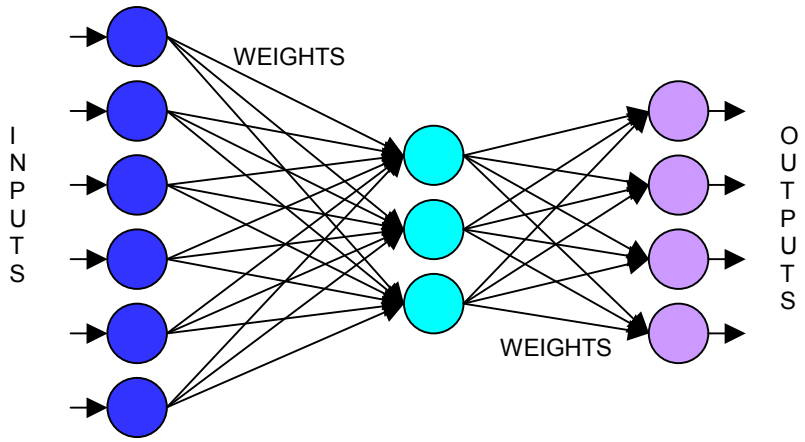

Figure 3.5 Illustration of a multilayered, feed-forward artificial neural network.

The standard multilayer perceptron (MLP) is a cascade of single layer perceptrons (Figure 3.5). There is a layer of input nodes, a layer of output nodes and one or more intermediate layers. The interior layers are sometimes called hidden layers because they are not directly observable from the system inputs and outputs. Nodes in each layer are fully connected to the nodes of the preceding and following layers, there are no connections between units in the same layer, neither connections from one layer back to a previous one, nor shortcut connections that skip over intermediate layers.

Back-propagation is, by far, the most commonly used method for training multilayer feed-forward networks. The term back-propagation refers to two different things. First, it describes a method to calculate the derivatives of the network training error with the respect to the weights by an application of the derivate chain-rule. Second, it describes a training algorithm which is very

similar to the gradient descent optimization, for using those derivatives to adjust the weights to minimize the error.

As a training algorithm, based in the principles described in the previous paragraph, it is a supervised technique in the sense that for every input pattern there is an externally specified "correct" output which acts as a target for network to imitate. Any difference between the network output and the target is treated as an error to be minimized. It is an off-line algorithm, where training and normal operation occurs at different times.

To train a network, it is necessary to have a set of input patterns and corresponding desired outputs, plus an error function that measures the differences between the network outputs and desired values. The basic steps are:

- Present a training pattern and propagate it through the network to obtain the outputs.
- Compare the outputs with the desired values and calculate the error.
- Calculate the derivatives $\delta E / \delta w_{ij}$ of the error with the respect to the weights.
- Adjust the weights to minimize the error.
- Repeat until the error is acceptably small or time is exhausted.

Back-propagation can be applied to any feed-forward network with differentiable activation functions.

Lets assume the nodes are indexed so that $i > j$ implies that node i follows node j in terms of dependency. The state of node i may depend, perhaps indirectly, on the state of node j, but node $j < i$ does not depend on node i. The advantage of this format is that it works in any feed-forward network, including those with irregular structure and shortcut (layer skipping) connections. In simulations it facilitates the dealing with each layer separately, keeping track of layer indexes.

Since the dependencies are transmitted by the connection weights, connections are allowed from nodes with low indexes to nodes with higher indexes, but not vie versa. If we denote a weight from node i to node j by w_{ij} then any forward link where $j < i$ is allowed but backward links are prohibited, $w_{ji} \equiv 0$.

3.3.1 Forward Propagation

In the forward pass, the network calculates an output based on its current inputs. Each node i computes a weighted sum a_i of its inputs and passes this through a nonlinearity to obtain the node output y_i.

$$a_j = \sum_{j<i} w_{ij} y_j \qquad \text{Equation 3.28}$$

$$y_i = f(a_i) \qquad \text{Equation 3.29}$$

Normally *f* is a bounded monotonic function such as *tanh* or sigmoid. Arbitrary differentiable functions can be used, but sigmoid-like "squashing" functions are standard.

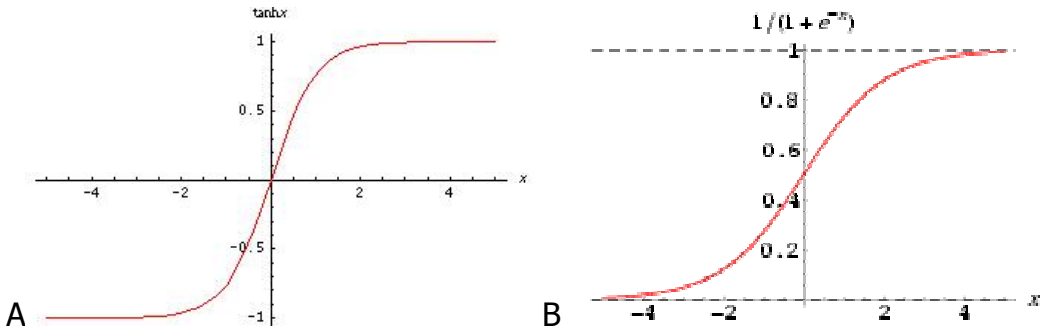

Figure 3.6 "Squashing" functions (A) hyperbolic *tan* and (B) sigmoid used for calculating an output (weighted sum of inputs) at a given node.

Every node is evaluated, starting with the first hidden node and continuing to the last output node. In layered networks, the first hidden layer is updated based on the external inputs, the second hidden layer is updated based on the outputs of the first hidden layer and so on to the output layer which is updated based on the outputs of the last hidden layer. It is obvious that that node *i* does not depend on any nodes $k > i$, all inputs to node *i* will be computed when it's evaluated and at the end of a sweep the system outputs will be stored in the output nodes.

3.3.2 Error Calculation

Unless the network is perfectly trained, the network outputs will differ from the desired outputs. The significance of these differences is measured by an error (interchangeable used term is "cost") function *E*. For explanation purposes we will use SSE (sum-of-squares error) function:

$$E = \frac{1}{2}\sum_{p}\sum_{i}(d_{pi} - y_{pi})^2 \qquad \text{Equation 3.30}$$

where p indexes the patterns in the training set, i indexes the output nodes, and d_{pi} and y_{pi} are, respectively, the desired target and actual network output for the ith output node of the pth pattern. One of the reasons SSE is convenient is that the errors of different patterns are independent – the overall error is just the sum of the individual squared errors.

$$E = \sum_{p} E_p \qquad \text{Equation 3.31}$$

$$E_p = \frac{1}{2}\sum_{i}(d_{pi} - y_{pi})^2 \qquad \text{Equation 3.32}$$

3.3.3 Back-Propagation

Term back-propagation is used interchangeably to refer to the derivative calculation or weight update algorithm.

3.3.3.1 The Derivative Calculation

After obtaining the outputs and calculating errors, the next step is to calculate the derivative of the error with the respect to the weights. Since $E_{SSE} = \Sigma_p E_p$ is just the sum of the individual patter errors, the total derivative is the sum of the per-pattern derivatives.

$$\frac{\partial E}{\partial w_{ij}} = \sum_p \frac{\partial E_p}{\partial w_{ij}} \qquad \text{Equation 3.33}$$

Back-propagation (the derivative calculation) is efficient due to operation of decomposition and ordering of the steps. The derivative can be expressed as:

$$\frac{\partial E_p}{\partial w_{ij}} = \sum_k \frac{\partial E_p}{\partial a_k} \frac{\partial a_k}{\partial w_{ij}} \qquad \text{Equation 3.34}$$

where the index k runs over all output nodes and a_j is a weighted-sum input for node j calculated in Equation 3.25. It is convenient to first calculate a value δ_i for each node i:

$$\delta_i = \frac{\partial E_p}{\partial a_i} \qquad \text{Equation 3.35}$$

$$\delta_i = \frac{\partial E_p}{\partial y_i} \frac{\partial y_i}{\partial a_i} \qquad \text{Equation 3.36}$$

which measures the contribution of a_i to the error of the current pattern.

For output nodes, $\delta E_p / \delta a_k$ is calculated directly:

$$\delta_k = -(d_{pk} - y_{pk}) f_k^{'} \qquad \text{Equation 3.37}$$

The first term is calculated from Equation 3.32

$$\frac{\partial E_p}{\partial y_k} = -(d_{pk} - y_{pk}) \qquad \text{Equation 3.38}$$

The second term:

$$\frac{\partial y_k}{\partial a_k} = f'(a_k) \qquad \text{Equation 3.39}$$

is just the slope $f'_k \equiv f'(a_k)$ of the node nonlinearity at its current activation value. The sigmoid function is convenient to use because f' is a simple function of the node output $f'(a_k) = y(1 - y)$, where $y = f(a_k)$. The *tanh* function $f'(a_k) = 1 - y^2$ is also suitable.

For hidden nodes, δ_i is calculated indirectly. Hidden nodes can influence the error only through their effect on the nodes k to which they send output connections, so:

$$\delta_i = \frac{\partial E_p}{\partial a_i} = \sum_k \frac{\partial E_p}{\partial a_k} \frac{\partial a_k}{\partial a_i} \qquad \text{Equation 3.40}$$

but the first factor is just the δ_k of the node k, so:

$$\delta_i = \sum_k \delta_k \frac{\partial a_k}{\partial a_i} \qquad \text{Equation 3.41}$$

The second factor is calculated by noting that if node i connects directly to node k then $\delta a_k / \delta a_i = f'_i w_{ki}$, otherwise it is zero. The final equation for hidden nodes is:

$$\delta_i = f_i' \sum_k w_{ki} \delta_k \qquad \text{Equation 3.42}$$

Henceforth, δ_i is a weighted sum of the δ_k values of nodes k to which it has connection w_{ki} (see Figure 3.5).

Because δ_k must be calculated before δ_i, $i < k$, the calculation starts at the output nodes and works backward toward the inputs, therefore the name "back-propagation".

To summarize

$$\delta_i = -(d_{pi} - y_{pi}) f_i' \quad \text{For output nodes} \qquad \text{Equation 3.43}$$

$$\delta_i = f_i' \sum_k w_{ki} \delta_k \quad \text{For hidden nodes} \qquad \text{Equation 3.44}$$

For output nodes, δ_i depends only on the error $d_i - y_i$ and the local slope f'_i of the node activation function. For hidden nodes, δ_i is a weighted sum of the δs of all the nodes it connects to, times its own slope f'_i. Due to the indexing scheme of the nodes, all delta values can be updated in a single sweep through the nodes in reverse order. It is not necessary to calculate delta values for the input layer, so the process stops with the first hidden layer.

Once the node deltas are calculated, it is trivial to find the partial derivatives $\delta E_p / \delta w$ with the respect to the weights. The second term in Equation 3.34 is $\delta a_k / \delta w_{ij}$. Knowing that a_k is a simple linear sum, this is zero if $k \neq i$; otherwise:

$$\frac{\partial a_i}{\partial w_{ij}} = y_j \qquad \text{Equation 3.45}$$

where y_j is the output activation of the node j. From Equation 3.34, the derivative of pattern error E_p with the respect to weight w_{ij} is then:

$$\frac{\partial E_p}{\partial w_{ij}} = \delta_i y_j \qquad \text{Equation 3.46}$$

3.3.3.2 The Weight Update Algorithm

Once the derivatives are calculated, the next step is to update the weights in such way to decrease the error.

Back-propagation (the optimization method) is equivalent to gradient descent. In order to minimize E, the weights are adjusted in the opposite direction, the weight update formula is:

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \qquad \text{Equation 3.47}$$

where the learning rate $\eta > 0$ is a small positive constant; it is also called the step size parameter. If the derivative is positive (increase in w causes increase in E) then the weight change is negative and vice versa. Very small values of η infer long learning times, so larger rates are usually used, typically in the range $0.05 < \eta < 0.75$ (note this is just a rule of thumb). Networks are ordinarily initialized with small random weights from a range $[-a, +a]$ where $0.1 < a < 2$

typically. Random values are used to disrupt the symmetry, while small values are necessary to avoid a quick saturation of the sigmoid nonlinearities.

3.3.4 Batch Learning

There are two basic weight-update variations, batch mode and on-line. In batch mode, every pattern p is evaluated to calculate the derivative terms $\delta E_p / \delta w$; these are summed to obtain the total derivative:

$$\frac{\partial E}{\partial w} = \sum_k \frac{\partial E_p}{\partial w} \qquad \text{Equation 3.48}$$

and only then are the weights updated, as a consequence of the derivative sum rule. The individual $\delta E_p / \delta w$ terms are calculated by application of the method of section "Back-Propagation: The derivative Calculation" to each pattern p.

The basic steps for every pattern p in the training set are:

- Apply pattern p and forward propagate to calculate network outputs.
- Calculate the pattern error E_p and back-propagate to calculate the single pattern derivatives $\delta E_p / \delta w$.
- Add up all the single-patterns terms to get the total derivative.
- Update the weights:

  $w(t + 1) = w(t) - \eta \cdot (\delta E / \delta w)$

- Repeat.

Each of the passes through the training set is called an epoch.

The gradient is calculated exactly and weight changes are proportional to the gradient, so batch learning approximates gradient descent when the step size η is small. Generally, each weight update reduces the error by only a small amount and many epochs are needed to minimize the error.

3.3.5 On-line Learning

An alternative to batch-mode is on-line learning. In this mode, the weights are updated after each pattern presentation. Usually a pattern p is chosen at random and presented to the network. The output is compared with the target for that pattern and the errors are back-propagated to calculate the single–pattern derivative $\delta E_p / \delta w$. The weights are then updated immediately using the gradient of the single pattern error. The patterns are presented randomly, constantly changing order to avoid cyclic effects.

The steps are:

- Pick a pattern p at random from the training set.
- Apply pattern p and forward propagate to calculate network outputs.
- Calculate the pattern error E_p and back-propagate to calculate the single-pattern derivatives.

- Update the weights immediately using the single-pattern derivative:

  $w(t + 1) = w(t) - \eta \cdot (\delta E / \delta w)$

- Repeat.

An advantage of this method is that there is no need to store and sum the individual $\delta E_p / \delta w$ contributions. Each pattern derivative is calculated, used immediately and then discarded. Another possible advantage is that many more updates occur in a given amount of time. If the training set has M patterns, on-line learning would make M weight changes in the time that batch-mode learning makes only one.

A possible disadvantage, at least from an analysis point of view, is that this is no longer a simple approximation to gradient descent. On-line learning differs from a pure gradient descent in that the sum (Equation 3.48) is never evaluated exactly, because the weights change after each pattern so the individual terms are evaluated at different points – the difference is minimal when η is very small; the weight will not change much between the steps, so the effect after all patterns have been evaluated is approximately the same as if all terms had been evaluated at a single point and summed to perform a single weight change which has the same overall result.

Stochastic search properties arise from the randomness of pattern propagation through a network. When the pure gradient descent arrives at a

local minimum, it is simply stuck there. In on-line learning, the weight state tends to jitter around its equilibrium value. Instead of sitting at the minimum, it visits many nearby points and if chance allows it may bounce out of a poor minimum and find a better solution. On the negative side, if the weight vector never settles to a stable value, it is possible after having found a good minimum to wonder off.

It is common, therefore, to adjust the learning rate as training progresses. The simplest way is to start with an intermediate value, let the system train to reach convergence and then gradually reduce the learning rate to zero to allow the system to settle to the minimum. The learning rate can also be adjusted dynamically depending on the conditions met during the training process. It is desirable to maintain a balance between stochastic search and efficient progress down the error gradient.

3.3.6 Learning Rate and Momentum Interaction

A common modification of the basic update rule is the addition of a momentum term. The idea is to stabilize the weight trajectory by making the weight change a combination of the gradient descent term in Equation 3.47 plus a fraction of the previous weight change. The modified weight change formula is:

$$\Delta \mathbf{w}(t) = -\eta \frac{\partial E}{\partial \mathbf{w}}(t) + \alpha \Delta \mathbf{w}(t-1) \qquad \text{Equation 3.49}$$

The weight change is a combination of a step down the negative gradient, plus a fraction $0 \leq \alpha < 1$ of the previous weight change. The learning rate η and momentum α directly affect the training process and it is imperative to choose good values. A number of factors affect the training, but typical range for η is $0.05 \leq \eta \leq 0.75$ and $0 \leq \alpha \leq 0.9$; these are closer to being the guidelines rather than rules.

This introduces a small amount of inertia – the direction of change of the weight vector will be the same unless opposed by the gradient term. Momentum trends to reduce the oscillations in the weight trajectory and accelerate learning in regions where $\delta E / \delta w$ is small.

3.3.6.1 Learning Rate

Factors influencing the choice of learning rate are:

- Distribution of training data.
- Momentum – it amplifies the effective learning rate $\eta_{eff} = \eta / (1 - \alpha)$.
- Weight magnitudes – small initialization values have to be offset by larger learning rates to speed up the training process, as weights

increase during the training reduction of the learning rate is necessary to prevent instability and saturation.

- Shape of the error surface $E(\boldsymbol{w})$ – the gradient $\delta E / \delta w$ may change drastically as $\boldsymbol{w}$ changes. Large learning rates are used to cross flat areas, but small values are needed to avoid entrapment in local minima.
- Stochastic nature of on-line training acts as noise and it is necessary to gradually reduce η to zero to achieve convergence of weights to stable values.

A prior choice of a good learning rate is not possible since many of the factors mentioned above are correlated and this difficulty is the reason of the popularity of adaptive methods. The speed advantage of a good adaptive algorithm stems from faster convergence than a simple back-propagation with a poorly chosen fixed learning rate. The speed advantage of the adaptive algorithm can be insignificant compared to the training with an optimal constant η; on the other hand they do not demand a search for the best learning rate. It is possible to further extend this model by dynamic adaptation of η to changing conditions in the network and local association of different rates to different parts of the network.

The optimal global learning rate for gradient descent on a linear problem is $\eta = 1 / \lambda_{max}$ where λ_{max} is the largest positive eigenvalue of the Hessian matrix $\boldsymbol{H}$ of the second derivatives of the error with the respect to the weights. Unless $\boldsymbol{H}$ is analyzed, λ_{max} is unknown and an estimate has to be used. If $\boldsymbol{H}$ is nonnegative-definite than all eigenvalues are nonnegative $\lambda_i \geq 0$ and λ_{max} can be bounded by:

$$\lambda_{\max} \leq \sum_i \lambda_i = trace(\mathbf{H}) \text{ where } (\mathbf{H}) = \sum_i h_{ii} \qquad \text{Equation 3.50}$$

Estimates of the h_{ii} can be obtained from the diagonal Hessian approximation. This overestimates λ_{max} and the resulting η may be smaller than the necessary, but it is a reasonable starting point for an adaptive method

The problem is that the $\boldsymbol{H}$ is not nonnegative-definite in general. However, there are standard methods (outer-product, diagonal and finite-difference approximations) to make the approximation nonnegative.

3.3.6.2 Momentum

Introducing momentum term in the back-propagation is the equivalent of applying a smoothing factor to the gradient descent. The rational behind this is to stabilize the weight trajectory by combining the gradient descent term in Equation 2.47 with a fraction of the previous weight change. The weight update rule is:

$$\Delta \mathbf{w}(t) = -\eta \frac{\partial E}{\partial \mathbf{w}}(t) + \alpha \Delta \mathbf{w}(t-1) \qquad \text{Equation 3.51}$$

For $\alpha \approx 1$, opposing weight changes components will lead to canceling each other, while complementary will sum effectively smoothing the weight trajectory. If successive gradients are pointing in the same direction, they will reinforce each other giving rise to acceleration in learning. The smoothing effect can be expressed by expansion of Equation 3.51:

$$\Delta \mathbf{w}(t) = -\eta \frac{\partial E}{\partial \mathbf{w}}(t) + \alpha \Delta \mathbf{w}(t-1)$$

$$\Delta \mathbf{w} = -\eta \frac{\partial E}{\partial \mathbf{w}}(t) + \alpha \left(-\eta \frac{\partial E}{\partial \mathbf{w}}(t-1) + \alpha \left(-\eta \frac{\partial E}{\partial \mathbf{w}}(t-2) + \ldots \right) \right)$$

$$\Delta \mathbf{w}(t) = -\eta \sum_{k=0}^{\infty} \alpha^k \frac{\partial E}{\partial \mathbf{w}}(t-k) \qquad \text{Equation 3.52}$$

With the momentum term introduced, the weight update is an exponential average of all the previous gradient terms and since $\alpha < 1$, the contribution from earlier derivative terms decays with each epoch, so the sum is dominated by most recent terms.

Momentum has the effect of amplifying the learning rate from η to the effective value of $\eta_{eff} = \eta / (1 - \alpha)$. Lets consider the case where the derivative is a constant $\delta E / \delta w(t) = J$, a reasonable approximation when η is very small so w

does not change much with each epoch and applicable to flat areas of $E(\mathbf{w})$ where the gradient is small. Then:

$$\Delta\mathbf{w}(t) = -\eta\sum_{k=0}^{\infty}\alpha^k\,\frac{\partial E}{\partial\mathbf{w}}(t-k)$$

$$\Delta\mathbf{w}(t) = -\eta J\sum_{k=0}^{\infty}\alpha^k \quad if \quad \sum_{k=0}^{\infty}\alpha^k = \frac{1}{(1-\alpha)} \quad \text{for } |\alpha| < 1$$

$$\Delta\mathbf{w}(t) = -\frac{\eta J}{1-\alpha} \qquad \text{Equation 3.53}$$

To sum up, momentum has the following effects:

- It smoothes weight changes – momentum tends to suppress cross-stitching across the error surface since the consecutive opposing weight changes tend to cancel out.
- If a long sequence of weight changes is unidirectional, momentum amplifies the effective learning rate leading to faster convergence.
- Momentum can help the system traverse local minima by giving the state vector enough inertia to coast over small ridges on the error surface.

3.3.7 Weight Decay

Another common modification is the addition of a weight decay term. Weight decay is applied to adapt the complexity of the network to the difficulty

of the problem – if the network is too complex, then it should be possible to delete many weights without significantly increasing the error. The way to do this is to introduce a tendency for weights to drift to zero by iteratively reducing their magnitudes. The update rule with weight decay is:

$$\Delta \mathbf{w}(t) = -\eta \frac{\partial E}{\partial \mathbf{w}}(t) - \rho \mathbf{w}(t) \qquad \text{Equation 3.54}$$

where $0 \leq \rho << 1$ is the weight decay parameter. In case that $\delta E / \delta w_i$ equals to zero for weight w_i then w_i will decay to zero exponentially, if $\delta E / \delta w_i$ is nonzero, weight bears significance and the two terms in Equation 3.54 will balance out at some point, preventing the weight from decay to zero.

3.3.8 Error Function

The network output is a function of its weights, therefore the error is a function of $\boldsymbol{w}$. In general, $E(\boldsymbol{w})$ is a multidimensional and impossible to visualize. If it was possible to plot it as a function of $\boldsymbol{w}$, E might look like a landscape with hills and valleys. Many advanced optimization methods assume the error surface is locally quadratic. The quadratic assumption is not reasonable for large networks, but usually near a minimum this assumption is a good approximation.

Regression problems involve modeling the conditional distribution of the output variables, based on the input variables. The use of "sum-of-squares"

function (Equation 3.32) is beneficial due to its properties. Classification problems are faced with modeling the posteriori probabilities of class membership, based on the input variables. Although the "sum-of-squares" error function can be used for classification there are other more appropriate functions such as cross-entropy error function:

$$E = \sum_{p} \sum_{i} t_{pi} \ln y_{pi} + (1 - t_{pi}) \ln(a - y_{pi}) \qquad \text{Equation 3.55}$$

Here the y_{pi} is the estimated probability that the pattern p belongs to a certain class i and $t_{pi} \in \{0, 1\}$ is the target.

The sum-of squares error function assumes a model where errors have a Gaussian distribution. The cross-entropy corresponds to a classification model with binomial or polynomial distribution of errors. These functions are well understood and their advantages include easy differentiability and independence. It is possible to use SSE function for classification problems despite it assumes Gaussian distribution due to the fact that the determination of the output weights in a network represents a linear optimization problem.

3.3.8.1 "Sum-of-Squares" Error Function

To derive the final equation for the sum-of-squares error function let's assume that the target variable t_k is given by some deterministic function of the input vector $\boldsymbol{x}$ with added Gaussian error ε [18]:

$$t_k = h_k(\mathbf{x}) + \varepsilon_k \qquad \text{Equation 3.56}$$

Assuming that error ε_k have a normal distribution with zero mean and standard deviation σ which does not depend on $\boldsymbol{x}$ or k, the distribution is given by equation:

$$p(\varepsilon_k) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\varepsilon_k^2}{2\sigma^2}\right) \qquad \text{Equation 3.57}$$

Trying to model the functions $h_k(\boldsymbol{x})$ by a neural network with outputs $y_k(\boldsymbol{x};\boldsymbol{w})$ where $\boldsymbol{w}$ is the set of weight parameters governing the neural network mapping, the probability distribution of target variables is:

$$p(t_k \mid \mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\left(y_k(\mathbf{x};\mathbf{w}) - t_k\right)^2}{2\sigma^2}\right) \qquad \text{Equation 3.58}$$

This leads to the following expression for the error function:

$$E = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \sum_{k=1}^{c} \left(y_k(\mathbf{x}^n;\mathbf{w}) - t_k^n\right)^2 + Nc \ln\sigma + \frac{Nc}{2} \ln(2\pi) \qquad \text{Equation 3.59}$$

Noting that the second and third term in the equation are independent on $\boldsymbol{w}$ and so is $1 / \sigma^2$, we derive the familiar expression:

$$E = \frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{c}\left(y_k(\mathbf{x}^n;\mathbf{w}) - t_k^n\right)^2 \qquad \text{Equation 3.60}$$

which is the same as Equation 3.32 only written in different notation.

3.3.9 Weight Initialization Techniques

The learning process can be speeded up by choosing good initial weight values. There are two groups of methods. One consists of consists of methods for choosing parameters controlling the distribution of random initial weights. The goal of these methods is to avoid sigmoid saturation that incurs slow training – they do not use domain-specific information. The second group consists of techniques for initializing the system from an approximate solution found by another modeling system, such as rule based systems, decision trees or nearest neighbor classifiers. The goal here is to reduce the training times and probability of convergence to poor minima by starting the system that is close to a good solution – these methods are generally faster and it is possible to embed domain-dependent information.

3.3.9.1 Random Initialization

The initialization procedure involves setting weights to reasonable small random values, where the randomness is used to break the symmetry and small values to avoid node saturation.

If all nodes had identical weight vectors (symmetry), then all nodes in one layer would respond identically as if it were comprised of only one node. During the back-propagation nodes would receive the same error information and never have a chance to diverge.

Relatively small weights are needed to avoid immediate saturation – large weights could amplify a moderate input to produce an extremely large weighted sum at the inputs of the next layer. This would impinge onto the flat region of nodes' nonlinearities and slow down the learning due to small derivatives. One should avoid a start with too small weights, since the learning speed would be limited by the small δ values back-propagated through the weights.

Methods and mathematical models used in deriving and explaining random initialization techniques are beyond the scope of this thesis, it is appropriate to direct the reader to suitable books and papers dealing with these matters:

- Calculation of the initial weight range – discussed [17].
- Initialization to maximize back-propagation deltas – discussed in [19].

- Constrained random initialization – discussed in [20] and [21].

3.3.9.2 Nonrandom Initialization

An alternative way to initialize the weights is to start the system with approximate values provided by another type of classifier or fitting system. The underlying idea is that the system can be easily fine trained if the set of variables it started with was reasonable good.

The advantage is that the initialization time can be negligible compared to training times of randomly initialized networks. The actual training time may be much shorter if the initial solution is good and problems of convergence to local minima may be avoided because the system is already an approximate solution when it starts. Another reason for using nonrandom initializations is to embed domain-dependent information in a network. It is also possible, in case of sparse data, to construct a network that generalizes well due to externally applied bias supplied by the initial solution.

Methods and mathematical models used in deriving and explaining nonrandom initialization techniques are beyond the scope of this thesis; then it is appropriate to direct the reader to suitable books and papers dealing with these matters:

- Principal components initialization – discussed in [22].
- Discriminant analysis initialization – discussed in [17].
- Nearest neighbor classifier initialization – discussed in [23] and [24].
- Initialization from a decision tree classifier – discussed in [25].
- Initialization from symbolic rule systems – discussed in [26] and [27].
- Initialization from fuzzy rule systems – discussed in [28].

3.3.10 Application of Adaptive Techniques in Back-propagation

The most common complaint about back-propagation is that it can be very slow. There are many other factors influencing learning time: network structure, the input-output representation, the choice of error function and so on. At the beginning of the training, these choices will have already been made and only through application of variations of the aforementioned algorithm can one accelerate the learning process. Many methods speed up the training by using adaptive learning rate schemes. The often recommended learning rate of $\eta = 0.1$ is an arbitrarily chosen value and may be completely inappropriate for a given problem. The magnitude of the gradient depends how the targets are scaled and this is related to η. If "sum-of-squares" is used, the size of the error and thus the learning rate may depend on the size of the training set. Redundancies such as

duplication of training patterns and correlation between different elements of the same pattern may lead to the amplification of the effective learning rate.

Faced with the difficulty of choosing a good learning rate a priori, it is sensible to start with a conservative value and adjust it accordingly to the system behavior. Some methods apply a single global rate while others modularly assign different learning rates. The unifying idea for all of the methods mentioned above is to increase the step size when the error is decreasing consistently and decrease it when significant error decrease occurs.

3.3.10.1 Rprop – Resilient Propagation

What differentiates Rprop from most other heuristic back-propagation variations is that the learning rate adjustments and weight changes depend only on the sign of the gradients terms, not their magnitudes. If the error surface is a complicated nonlinear type, the gradient magnitude is unpredictable a priori and there is no reason why the step size should be proportional to the magnitude per se. It can be argued that the step size should be inversely proportional in order to quickly traverse areas where the gradient is small and to take small steps where the gradient is large.

Rprop is a batch update algorithm. Each weight w_{ij} has its own step size or update value Δ_{ij} which varies with time according to:

$$\Delta_{ij}(t) = \begin{cases} \eta^{+}\Delta_{ij}(t-1) & if \quad \frac{\partial E}{\partial w_{ij}}(t-1)\frac{\partial E}{\partial w_{ij}}(t) > 0 \\ \eta^{-}\Delta_{ij}(t-1) & if \quad \frac{\partial E}{\partial w_{ij}}(t-1)\frac{\partial E}{\partial w_{ij}}(t) < 0 \\ \Delta_{ij}(t-1) & otherwise \end{cases} \qquad \text{Equation 3.61}$$

where $0 < \eta^{-} < 1 < \eta^{+}$. A change is sign of the partial derivative corresponding to weight w_{ij} indicates that the last update was too big and the system has jumped over a minimum, so the update value Δ_{ij} is decreased by a factor η^{-}. Consecutive derivates with the same sign indicate that the system is moving in one direction, so the update value is increased slightly to accelerate convergence in shallow regions.

The weights are changed according to:

$$\Delta w_{ij}(t) = \begin{cases} -\Delta_{ij}(t) & if \quad \frac{\partial E}{\partial w_{ij}}(t) > 0 \\ +\Delta_{ij}(t) & if \quad \frac{\partial E}{\partial w_{ij}}(t) < 0 \\ 0 & otherwise \end{cases} \qquad \text{Equation 3.62}$$

Examining the expression above, it is obvious that the change depends only on the sign of the partial derivative and is independent of the magnitude. If

derivative is positive, the weight is decreased by $\Delta_{ij}(t)$, if the derivative is negative, the weight is increased by $\Delta_{ij}(t)$.

All update values are initialized to a constant $\Delta_{ij} = \Delta_0$, which determines the size of the first weight change – the reasonable value is $\Delta_0 = 0.1$.

The range of update values is limited to $\Delta_{min} = 10^{-6}$ and $\Delta_{max} = 50$ to avoid floating point underflow/overflow problems. Limiting Δ_{max} to smaller values may give smoother decreases in the error at the cost of slow convergence.

The value $\eta^- = 0.5$ was chosen based on following rational; when the system overshoots a minimum, it will be half way between the current and the previous weights, on average, so the step size should be reduced to a half of its previous value.

The value $\eta^+ = 1.2$ is a compromise. It has to be large enough to allow fast growth in the flat regions of the error surface, but not so large that the system has to immediately reduce the update value in the next step.

The wide spread of Rprop for a range of problems can be explained by two factors. First, the error tends to decrease exponentially towards the minimum and with Rprop the step size does not depend on the magnitude of the gradient, so learning does not slow to a crawl in the final stages. Second, in layered networks, back-propagation causes attenuation of derivatives as they propagate from the output layer towards the inputs. Each layer inserts a

nonlinearity derivative factor that is less than *0.25* for sigmoids and less than *1* for *tanh* nodes. This produces small $|\delta E / \delta w|$ for weights far from the outputs and learning is slow. Application of deep networks with many layers has been made possible, in part, due to aforementioned properties of Rprop algorithm.

3.3.10.2 Other Adaptive Learning Rate Methods

There are many other adaptive learning rate algorithms, it is appropriate to direct the reader to suitable papers describing these methods in details:

- Vogl's method – discussed in [29].
- Delta-bar-delta – discussed in [30].
- Silva and Almeida – discussed in [31].
- SuperSAB – discussed in [32]
- Quickprop – discussed in [33].

3.3.11 Generalization

In view of artificial neural networks and supervised learning, generalization can be described as an interpolation or approximation problem. Having no restrictions, we can store the training patterns in a look-up table. The problem is that although the look-up table "learns" the training data perfectly, it

can not cope with novel patterns. A desirable system would have the capability to generalize from the training examples to the underlying target function so it produces correct or at least reasonable outputs in response to new patterns that have not been seen before. A system that learns the training data and does well on new data is considered to generalize well.

3.3.11.1 Network Complexity versus Target Complexity

To generalize, a system has to be sufficiently powerful to approximate the target. If it is too simple to fit even the training data then generalization to new data is likely to be poor. If the network is powerful enough, then the good generalization is possible if not limited by other factors. Large network may perform better since it is more powerful and better able to the true target. An overly complex system may be able to approximate the data in many different ways that produce the similar errors and is unlikely to choose the one that will generalize best unless other constrains are imposed.

If the target function is completely unknown, there is no way to determine a priori if the network is complex enough and additional data is need to avoid the possibility of under- or over-fitting.

3.3.11.2 Number of Examples versus Target Complexity

In cases where the training and test sampling densities match, generalization may suffer if the training set does not capture all significant features of the underlying target. In general, the number of examples required depends on the constraining knowledge. More complex targets will generally require more samples and in many cases the number of samples required may be unreasonable large, depending on the constraints. Finally, the constraining knowledge must be available in a form useful to the training procedure.

3.3.11.3 Number of Examples versus Network Complexity

Large and complex systems can compute more functions but need more examples to determine which function choose to get to the target (so called curse of dimensionality). If the number of examples is small compared to the number of degrees of freedom of the network, then the network may be able to implement many functions consistent with the examples. Some of these may generalize better than the others, but the learning algorithm guided only by errors on the samples has no way of telling and choosing the one that generalizes best. Increasing the number of samples decreases the number of hypothesis functions consistent with the samples, but it will never reduce it to just one

function unless the set of hypothesis functions is finite or there are constraints imposed on the system.

The general rule of thumb is that the number of samples should be several times larger than the number of weights in the network. Of course, this simple rule may not hold for more complex targets and networks.

3.4 Spectroscopy of Biological Molecules

As it was stated before, the water molecule is the major component of the cell, accounting for about 70 % of its weight. Inorganic ions, sugars, amino acids, nucleotides, lipids and other small molecules make up 8 % of the total mass and the remaining 22 % consists of macromolecules, proteins, nucleic acids and polysaccharides. By dry weight, cells consist of about 60 % protein and 25 % nucleic acids, with the rest composed of other constituents mentioned above. In most healthy cells the RNA/DNA ratio is about 5. This composition of cells may vary depending on the organ, the stage of the cell division cycle and other factors.

The average protein spectra found inside cells are dominated by the amide I band at ~1650 cm^{-1} that is primarily associated with the stretching motion of the C=O group. This peak is sensitive to the environment of the

peptide linkage and also depends on the protein's overall secondary structure. The amide II vibration, mainly a coupled C–N stretching and C–NH deformation coordinate, occurs at 1530 cm^{-1}. Weaker protein vibrations include the amide III peak (coupled C–H/N–H deformation) at 1245 cm^{-1} and a number of side-chain vibrations in 1310-1450 cm^{-1} range. The IR spectra of RNA and DNA are shown in Figure 3.7. These spectra also depend on the state of hydration and the nucleic acid secondary structure. They exhibit absorption peaks between 1580 and 1700 cm^{-1} due to the aromatic base breathing and C=O stretching vibrations. The ionized PO_2^- and ribose groups exhibit a triad of peaks that occur in DNA at 1071, 1084 and 1095 cm^{-1} with nearly equal intensities. DNA peaks are also observed at 965 and 1245 cm^{-1} (the phosphodiester vibration). In RNA the peak at 1085 cm^{-1} is stronger than the two other peaks in this triad and the intensity ratio of this triad to the phosphodiester peak at 1245 cm^{-1} is about 1:0.7, whereas it's about the same in DNA [34]. An overview of the characteristic IR bands of some major cellular constituents is given in Table 3.1 [35, 36].

Spectra of biomolecules are affected not only by their primary structure, i.e. chemical composition, but also by the way of the spatial arrangement of long polymeric chains – known as the secondary structure. Proteins can assume α-helical or β-sheet arrangements, or can have a mixture of these domains within one molecule; consequently this will give distinctive spectral features of

generally similar looking IR spectral envelope. The polymeric nature of the nucleic acids and polysaccharides is the root cause of their polymorphism; detailed discussion of this phenomenon can be found in the biochemical literature and it will not be presented here since it is out of the scope of this thesis.

3300	N–H Stretching	1379	CH_3 Symmetric bend
2929	CH_2 Antisymmetric stretch	1340 – 1240	Amide III
2850	CH_2 Symmetric stretch	1237	O–P=O Antisymmetric stretch (PO_2^-)
1737	C=O Stretch	1173	–CO–O–C Antisymmetric stretch
1690 – 1620	Amide I	1150	C–O Stretch
1570 – 1530	Amide II	1083	O–P=O Symmetric stretch (PO_2^-)
1468	CH_2 Symmetric bend	1063	–CO–O–C Symmetric stretch
1455	$CH_{2/3}$ Deformation modes	1050	C–O Stretch
1397	COO^- Symmetric stretch	968	C–O Phosphodiester residue

Table 3.1 Frequencies and assignments of absorption peaks found in IR spectra of cells and cell fractions.

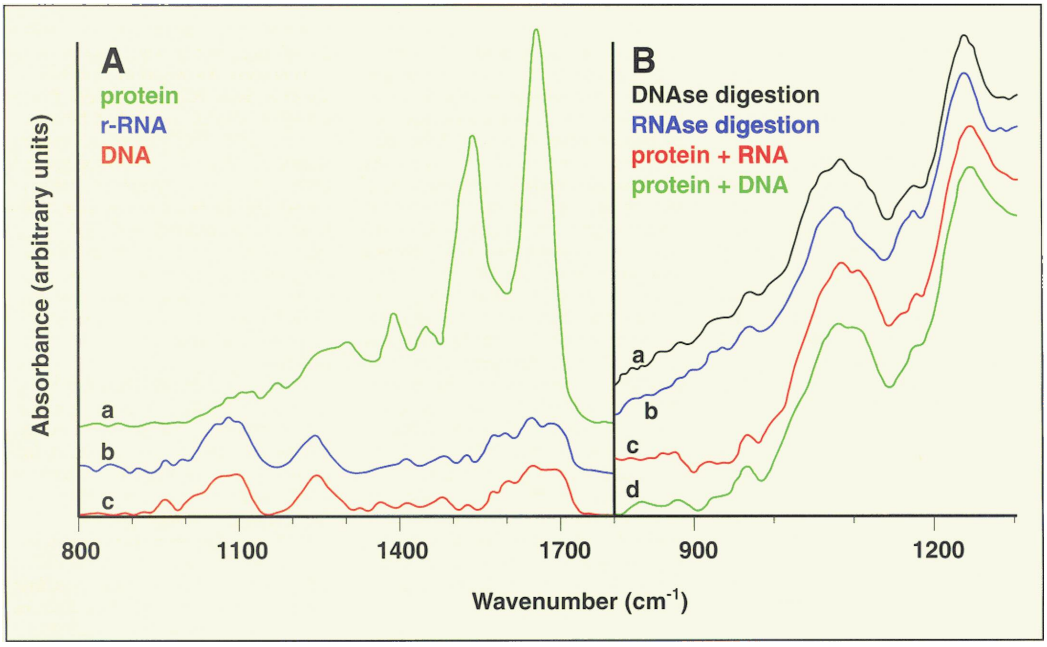

Figure 3.7 (A) IR absorption spectra in the mid-IR spectral range of (a) protein (human serum albumin); (b) DNA and (c) RNA. (B) IR absorption spectra of (a) cellular RNA after digestion of DNA; (b) nuclear DNA after digestion of RNA; (c) protein and DNA, by coaddition of appropriate spectra and (d) protein and RNA, by coaddition of appropriate spectra.

4 Results and Discussion

4.1 Live Cells

In 1999, the first FT-IR results were published that indicated that spectral variations accompany the stages of the cell cycle [37]. The spectral analysis of exfoliated cells (diagnosis of cervical disease) revealed large spectral heterogeneity, which could, at least partially, explained by the variations of cellular spectra during the division cycle. Secondly, the differences observed between normal and cancerous regions of biopsy tissue sections can be understood in terms of the spectral variations observed during the cell's division cycle. Varying abundance of dividing cells in cancerous samples may be the root cause. This can also affect the average spectra collected as a "pixel" of tissue biopsy.

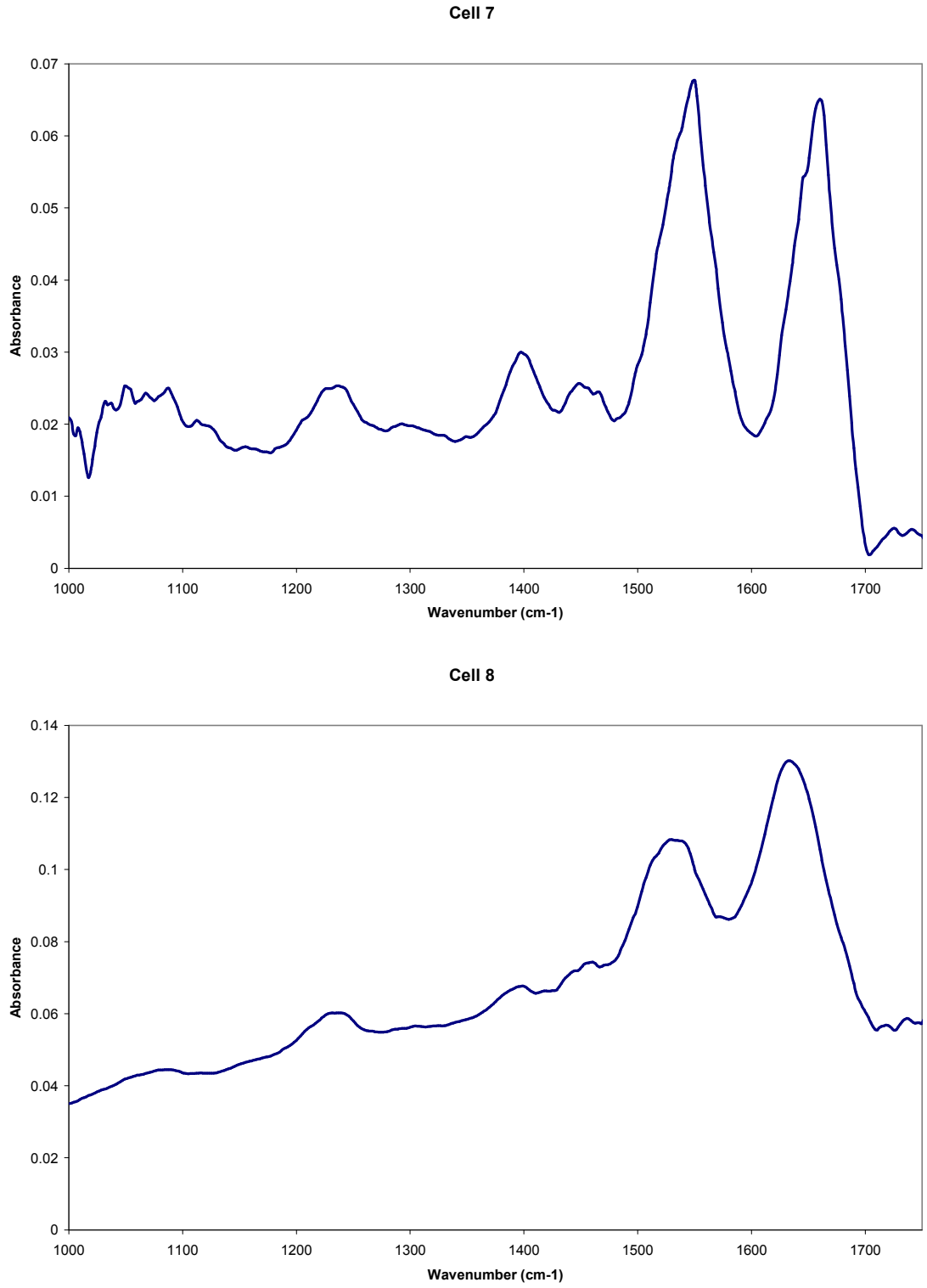

Figure 4.1 IR spectra of HeLa cells 7 and 8 suspended in buffered saline solution.

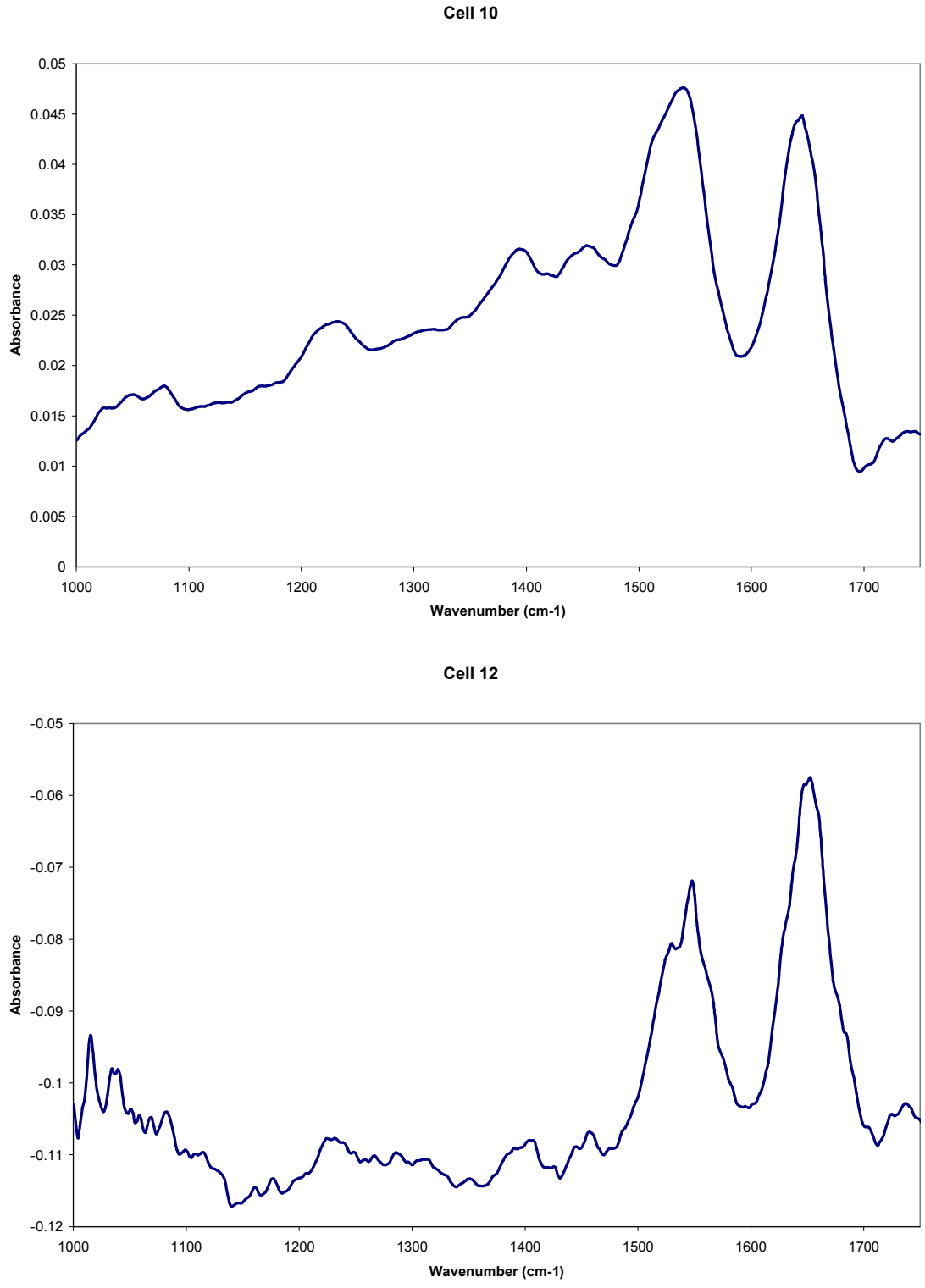

Figure 4.2 IR spectra of HeLa cells 10 and 12 suspended in buffered saline solution.

4.1.1 Single Cell Reflection-absorption Spectral Measurements

Reflection-absorption spectra of cells suspended in buffered saline solution (BSS) are shown in Figure 4.1 & 4.2 and Figure 4.3 shows visual images of the same cells taken through the objective used for IR data acquisition. It's believed that these spectra are ones of live cells, although they may have been compressed in the sample preparation process. The unusual amide I/II band profile is an artifact due to overcompensation for the water background.

Amide I/II envelope and band shape in spectra collected from cells in suspension is due to the fact that the cell contains less water than the surrounding medium or buffer. This can be avoided by scaling the amplitude of the buffer spectrum when the sample absorbance spectrum is computed or by adding a scaled buffer spectrum to the raw absorption spectrum of the sample. Since the exact difference in water content of the cell and the medium is unknown, procedures are based on visual fitting – the spectra are corrected until they display a "normal" amide I/II envelope. This is shown in Figure 4.4A, trace 1 shows the raw absorption spectrum of a cell, using the unscaled spectrum of the growth medium (trace 3) for a background. The 3400 – 3700 cm^{-1} region in the resulting raw absorption spectrum is overcompensated for the growth medium spectrum. By scaling the growth medium spectrum, a resulting

spectrum (trace 2) exhibits positive amide I/II ratio with nearly "normal" band profile, as shown in Figure 4.4B.

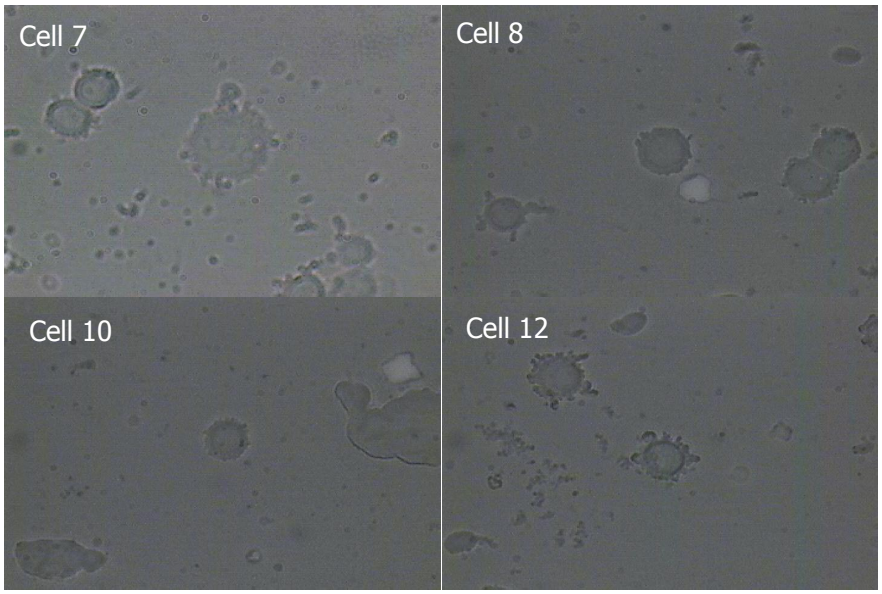

Figure 4.3 Visual images of HeLa cells suspended in buffered saline solution. Cell 7 is approximately 25 μm in diameter; cells 8, 10 and 12 are approximately 12 μm in diameter.

The above mentioned procedure is a subjective one since the exact ratio of water to cellular components is not a priori known. To overcome this obstacle, two approaches based on statistical methods were undertaken; multivariate curve resolution (MCR) and principle component resolution (PCR). Both methods produced inconclusive results, primarily due to lack of quantitatively significant IR bands in the spectrum of the cells' surrounding medium. The author believes that the only way to approach this problem is to report spectra of

live cells either as raw spectra or after applying Savitzky-Golay smoothing function.

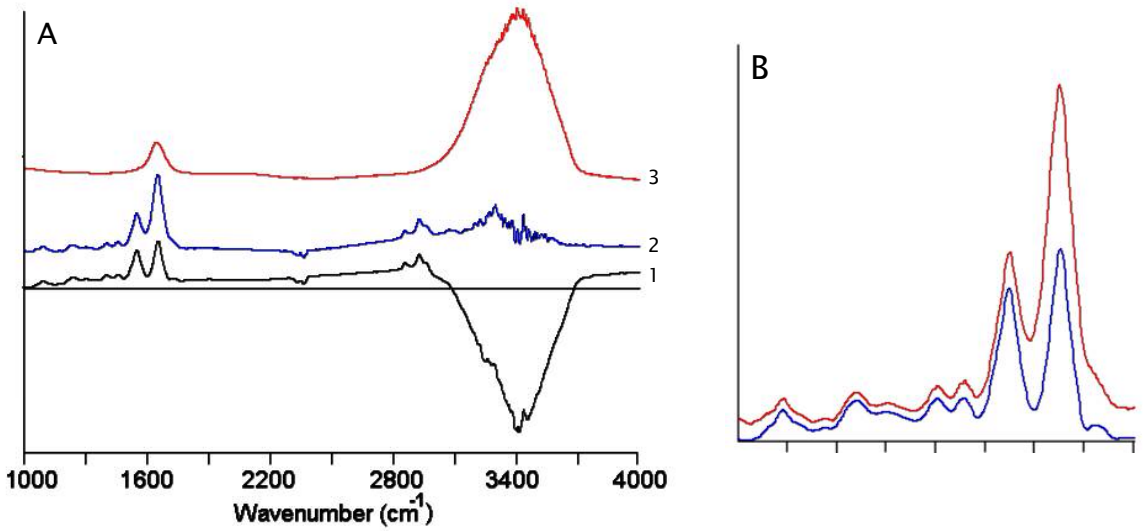

Figure 4.4 (A) Trace 1 shows the raw absorption spectrum of a cell, using the unscaled spectrum of the growth medium (trace 3) for a background. By scaling the growth medium spectrum, a resulting spectrum (trace 2) exhibits positive amide I/II ratio with nearly "normal" band profile. (B) The 1000-1800 cm^{-1} region, in blue is the raw spectrum and in red the spectrum corrected for water content.

4.1.2 Mapping Cells – Transmission Spectral Measurements

To observe cells in their native environment, spectral maps of individual cell were collected in transmission mode [38]. Cells were chosen on the basis of their size, with diameter greater than 40 μm. The spectral images are based on raw IR intensities at the indicated wavenumber. Spectral maps based on the 1650 (protein) and 1085 cm^{-1} (phosphate) spectral peaks are shown in Figure 4.5. The protein map shows the highest amide I intensity in the nucleus, a behavior

previously observed for maps of single dried cells [39], where the sample thickness is highest in the nucleus. Here, however, the path length is constant, indicating that the high protein signal is due to the large protein concentration in the nucleus, rather than a path length effect.

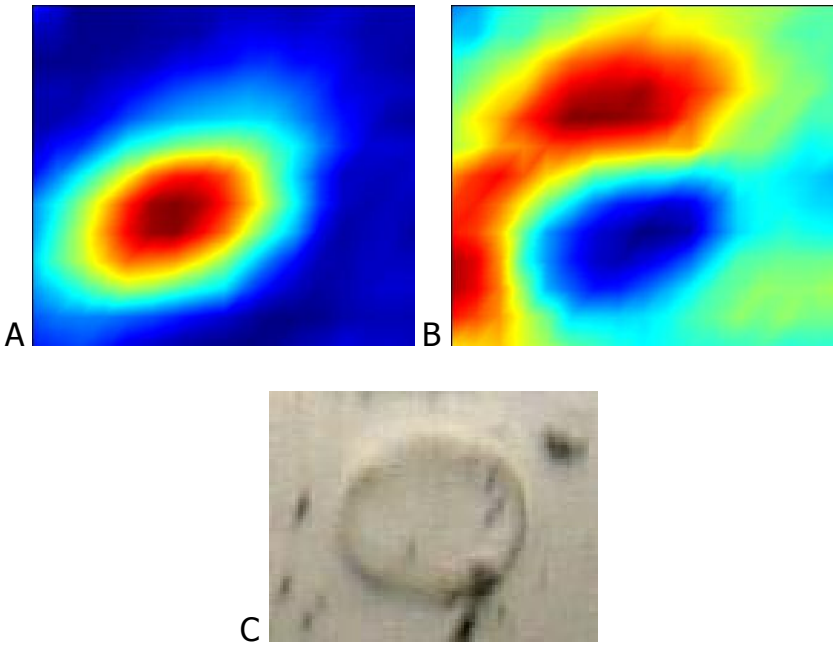

Figure 4.5 Spectral maps of a HeLa cell in growth medium based on the (A) 1650 (protein) and (B) 1085 cm^{-1} (phosphate) peaks. Blue hues correspond to the lowest and red hues to the highest intensity of the selected features. (C) Visual image of a HeLa cell in growth medium. Size of the cell is approximately 50 μm.

Figure 4.6 shows representative raw spectra from the map in Figure 4.5; spectra are extracted as a line scan along a horizontal line intersecting the cell through the middle. The fact that these spectra exhibit less distortion in the amide I/II envelope is due to shorter effective path, compared to the reflection-

absorption experiments. A sharp band in the symmetric phosphate stretching region (1085 cm^{-1}) is present in the spectra of the cell in growth medium with relatively large intensities throughout the nucleus and cytoplasm. It is not due to the growth medium itself, which has no spectral features in this region. This peak was also observed in dried cells that were grown, rather than spin-deposited, on substrates [40]. The identity of this compound is currently unknown, but its presence in both cytoplasm and nucleus suggests RNA, phosphates such as ATP, phosphorylated proteins, or others that are particularly prevalent in metabolically active cells.

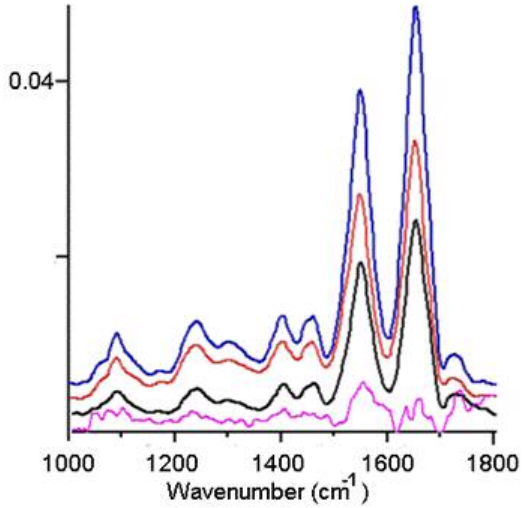

Figure 4.6 IR spectra from a line scan along a horizontal line intersecting the cell (Figure 4.5) through the middle. Spectra, from the bottom to the top, are arranged in a sequence starting at the periphery of the cell and ending in the cell's nucleus.

4.1.3 Cell Cycle – Transmission Spectral Measurements

For the in vivo spectroscopic methods of diagnosis the detection of spectral changes during a cell's division cycle in live cells is of prime importance. The cell cycle is defined as the interval between the completion of mitosis of the parent cell, and the completion of mitosis in one or both of the daughter cells. It is divided into four main phases: G1, S, G2 and M, where each phase has a particular function.

To fully understand these data, we have to relate the spectral changes to the ”extreme” biochemical and morphological changes occurring during the cell proliferation and these processes are realized in a 3D space in contrast to the microscopically observed 2D realm.

The mass of a cell, dry weight, doubles going from the "gap one" (G1) phase to mitosis (Figure 4.7). The final result of the cellular division is two daughter cells and it is reasonable to assume that the volume of an actively dividing cell doubles. The increase in volume is causing a morphological change of a cell. This is facilitated by the cytoskeletal transformation of a cell in the M phase of the cell cycle, ultimately causing the cell to become more spherical, a schematic representation is shown in Figure 4.8.

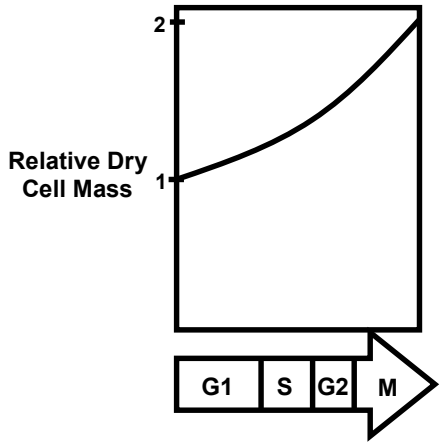

Figure 4.7 A plot of the increase in cell's mass during the cell cycle.

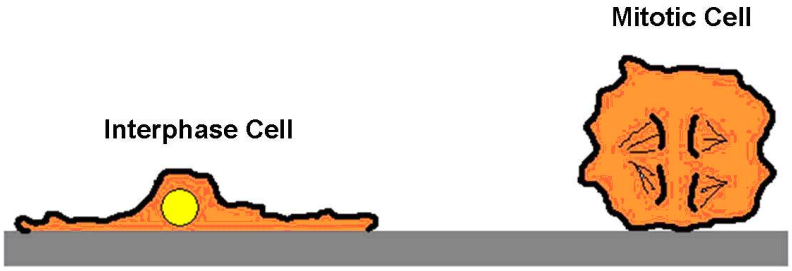

Figure 4.8 A cytoskeletal change occurs in a cell in M phase, forcing it to become more spherical.

IR spectra of actively dividing cells are shown in Figure 4.9, with the corresponding visual images in Figure 4.10. The unusual amide I/II envelope is due to the fact that the cell contains less water than the surrounding medium as explained in the previous section. The low frequency spectral region containing contributions from DNA/RNA shows out-of-ordinary absorption intensities and band shapes which are caused, in authors opinion, by very high optical density of DNA/RNA resulting in a nonlinear absorption of the IR light. During the "synthesis" phase (S) of the cell cycle, the cell's entire genome is duplicated and it is in a less condensed state than it is in the nucleus in order to facilitate

unobstructed DNA replication. At this point the IR "active" concentration of DNA/RNA is extremely high and, ultimately, causes a nonlinear absorption of IR light. Spectra supporting this supposition are the ones obtained from cells A and B shown in Figure 4.9.

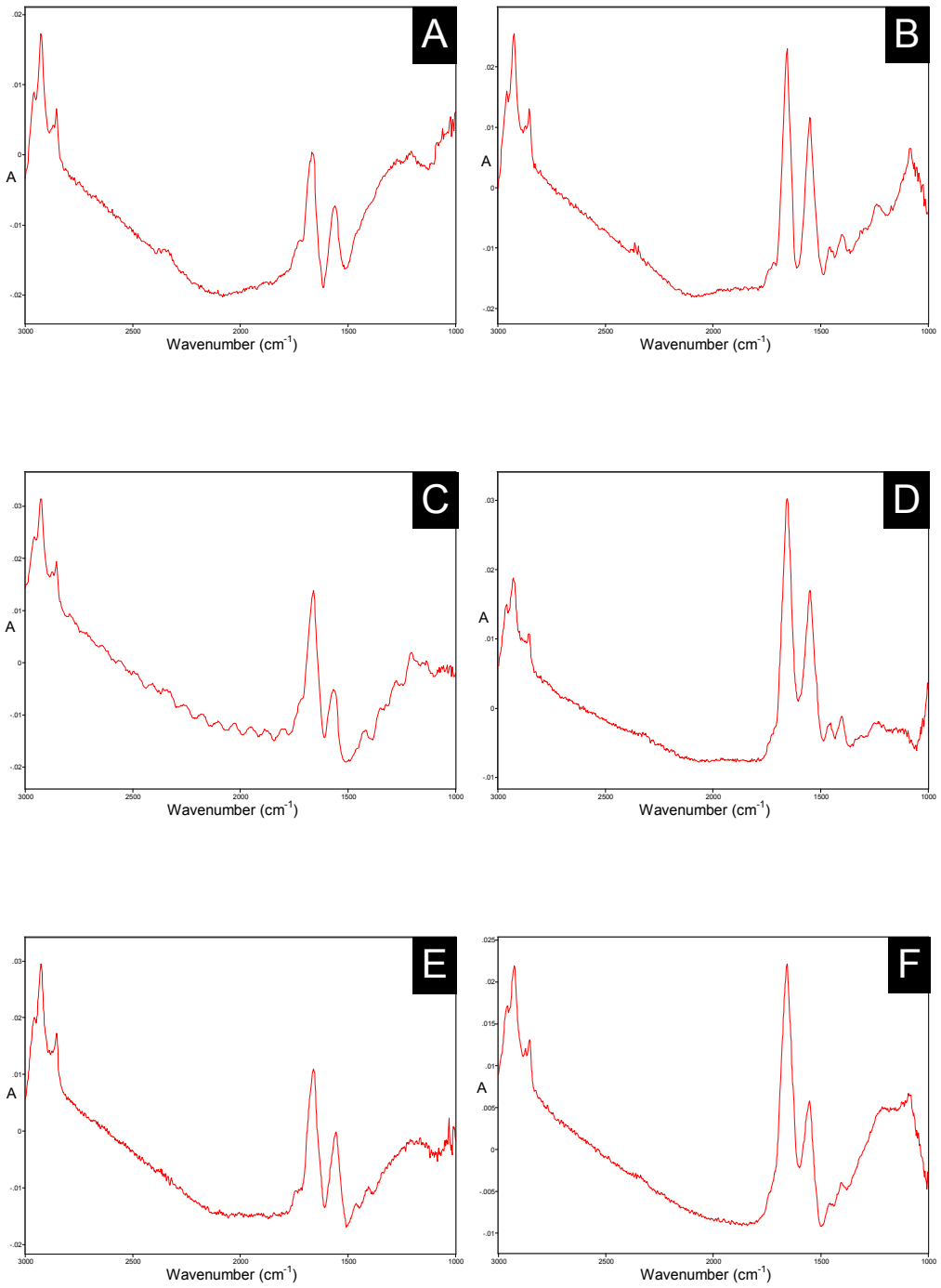

Figure 4.9 Raw IR spectra of the corresponding cells visual images are shown in Figure 4.9. (A) & (B) Cells in anaphase of the mitosis. (C), (F) & (E) Daughter cells formed by the cellular division. (D) Cell in the interphase, non-dividing cell.

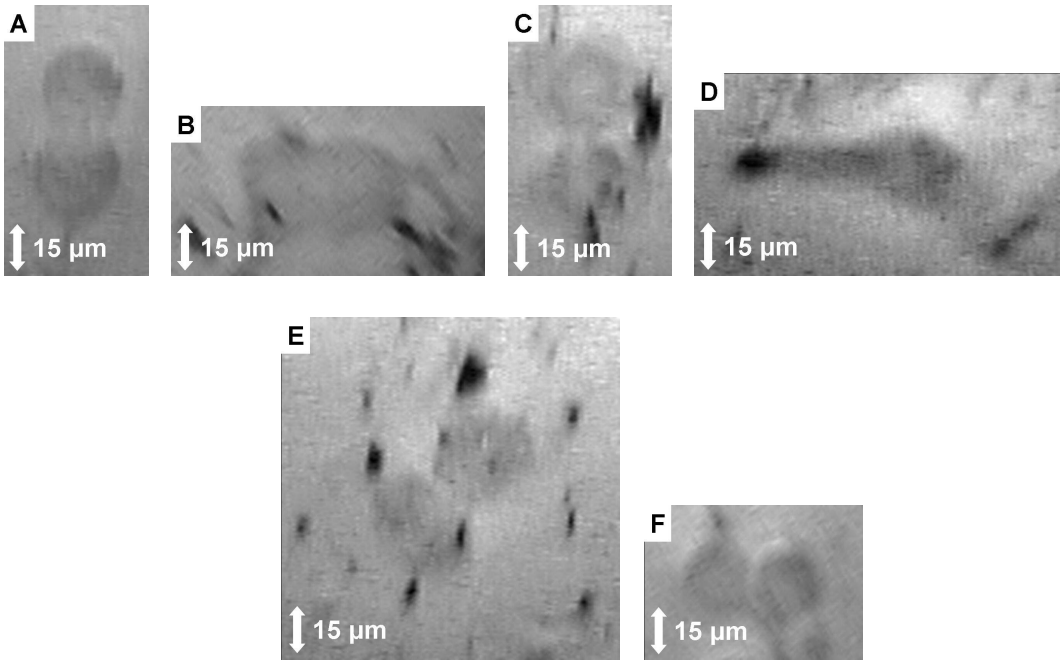

Figure 4.10 Visual images of HeLa cells undergoing various phases of the cell cycle. (A) & (B) Cells in anaphase of mitosis. (C), (F) & (E) Daughter cells formed by the cellular division. (D) Cell in interphase, non-dividing cell.

4.2 ANN Imaging

The discriminative power of FT-IR micro-spectroscopy has facilitated its widespread use in analysis of diseased and normal tissue. A number of comprehensive books and articles give an insight into the field [41-43].

In the past, several groups of authors have utilized infrared mapping or imaging of different tissues to identify, classify and grade regions and type of cancer in the respective samples [44-46].

Fourier transform infrared micro-spectroscopy, multivariate data analysis, such as hierarchical cluster analysis, and artificial neural network pattern recognition were combined in order to test and compare with classical medicinal diagnostics. Using unsupervised methods for spectral data processing, in conjunction with supervised machine-learning algorithms, more objective diagnostic tools can be created compared to conventional procedures which are burdened with visual pattern recognition, data storage in the human brain and inherent subjectivity [47, 48].

Generally, IR spectra of a tissue contain the information about total chemical composition and structure. The objectives of this thesis were not to search for distinct cancer markers to diagnose cancerous areas of a biopsy, but to utilize the mathematical procedures which search for and analyze subtle differences in the observed infrared spectral region, ultimately enabling differentiation of normal from diseased tissue. It is necessary to have a comprehensive understanding of spectral features of the cell types, maturation and cell cycle stages in order to interpret these spectral differences. A previous study demonstrated that patient-to-patient variations of spectral patterns are smaller than those between different tissue types and stages of a disease [49].

The outline of the methodology of the diagnostic ANN algorithm (ANNA) development and design is shown in Figure 4.11.

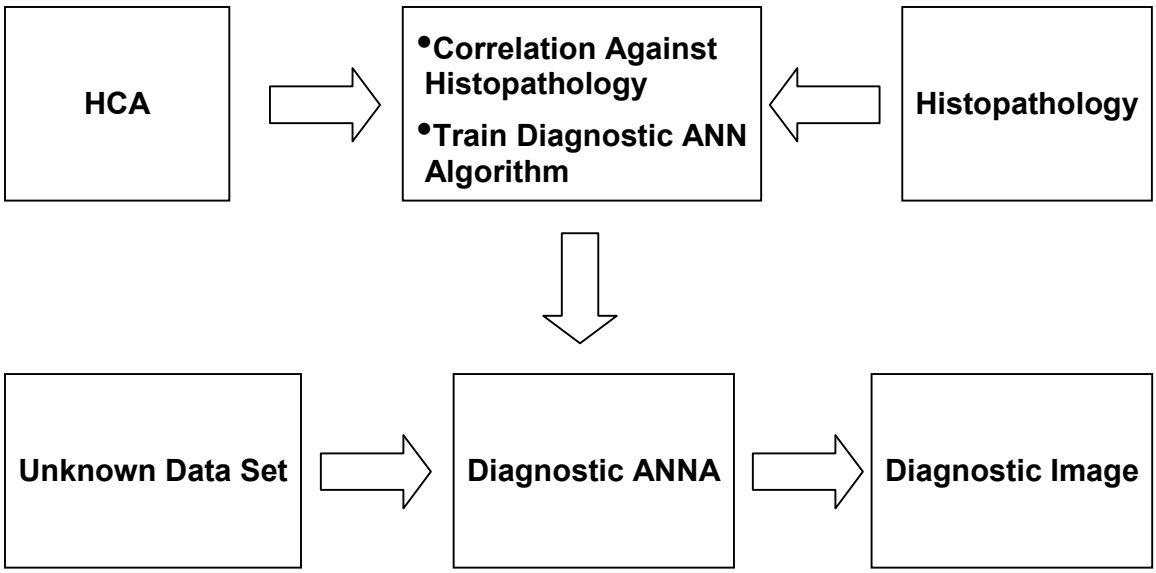

Figure 4.11 Outline of the methodology of the diagnostic algorithm development and design.

The first step in the algorithm designing was correlating histopathological diagnosis results with the mapping of tissue sections based on HCA of IR data acquired from the same samples prior to staining. Following this, determining ANN topology and training networks to the expectable level of correct class assignment had to be carried out. After extensive testing, diagnostic artificial neural network algorithm was promoted to the level of "Diagnostic ANNA". Performance was then examined on unknown samples and produced diagnostic images were compared to the H&E stained tissue sections with the known diagnoses.

4.2.1 Description of the Biopsies

A histopathological micrograph (Fig. 4.12A) of an H&E stained tissue section shows the morphology of a lymph node effaced with the metastatic breast cancer. Features such as macrophages (1), cancer (2), lymphocytes (3) and capsule and fat (4) are easily discerned by the observation of different colors and color intensities of the H&E stain. HCA was able to reproduce these patterns as it is presented in Fig. 4.12B, it is apparent that HCA preserves the tissue architectural and pathological differentiation.

A micrograph of unstained tissue slice adjacent to the one in Fig. 4.12A is shown in Fig. 4.13A. It exhibits the following features: macrophages (1), cancer (2), cortex (3) and capsule and fat (4). Mapping based on hierarchical cluster analysis results mimics the morphology as explained in the figure caption.

In Fig. 4.14A a micrograph of an H&E stained lymph node displays its various morphological features: fatty tissue (1), capsule (2), medullary cord (3), paracortex (4) and secondary lymphoid follicle (5). After applying the HCA algorithm to the IR spectra acquired from the same tissue section, prior to staining, the pseudo-color map (Fig. 4.14B) not only reproduced above mentioned architecture, but it was able to further differentiate the cortex, paracortex and germinal centers, which contained activated B-lymphocytes.

The ability of hierarchical cluster analysis to match the resolution power of visual methods, i.e. H&E staining pattern analysis, is proven again when comparing a micrograph of a section of a lymph node with colon adeno-carcinoma (Fig. 4.15). In Figure 4.15A a photomicrograph and HCA pseudo-color map (Fig. 4.15B) of a tissue section of a lymph node with colon cancer are shown, presented in brown and orange are capsule and fat, in yellow paracortex, in green cortex, in aqua glandular cancer and in blue stromal cancer.

4.2.2 Linking H&E Based Histopathology, HCA and ANN Imaging

Hierarchical cluster analysis served as a link between the diagnostic methods based on H&E stained images and artificial neural network analysis. The number of clusters for each sample was chosen in such a way that the pseudo-color maps established the maximum possible resemblance to the histopathological counterparts. Clustering was stopped at that level and IR spectra were exported to the data base to serve as members of various classes. Table 4.1 summarizes the number of spectra within each class for a lymph node containing breast cancer metastasis. Inspection of this table reveals an under-representation of spectra for some of the classes. The contents of the data base were determined by the tissue sample composition and since not all classes are evenly distributed or present in the examined lymph nodes, this led to an uneven

sample distribution. The exact number of spectra that served as the training and validating set for ANN analysis was determined on per spectral image basis and it ranged from 10 to 30% of the total number of the elements in a given class. These spectra were randomly chosen from the respected class pools. By applying a random selection procedure, the ANN classifier was able to achieve a good generalization power of the learning algorithm.

Tissue Section Type	**Number of Spectra**
Fat and Capsule	1,765
Cancer	6,255
Macrophages	3,315
Lymphocytes	4,710

Table 4.1 Database, produced from IR spectra obtained from a lymph node with breast cancer, showing an uneven distribution of spectral information due to the architectural and morphological features of the tissue.

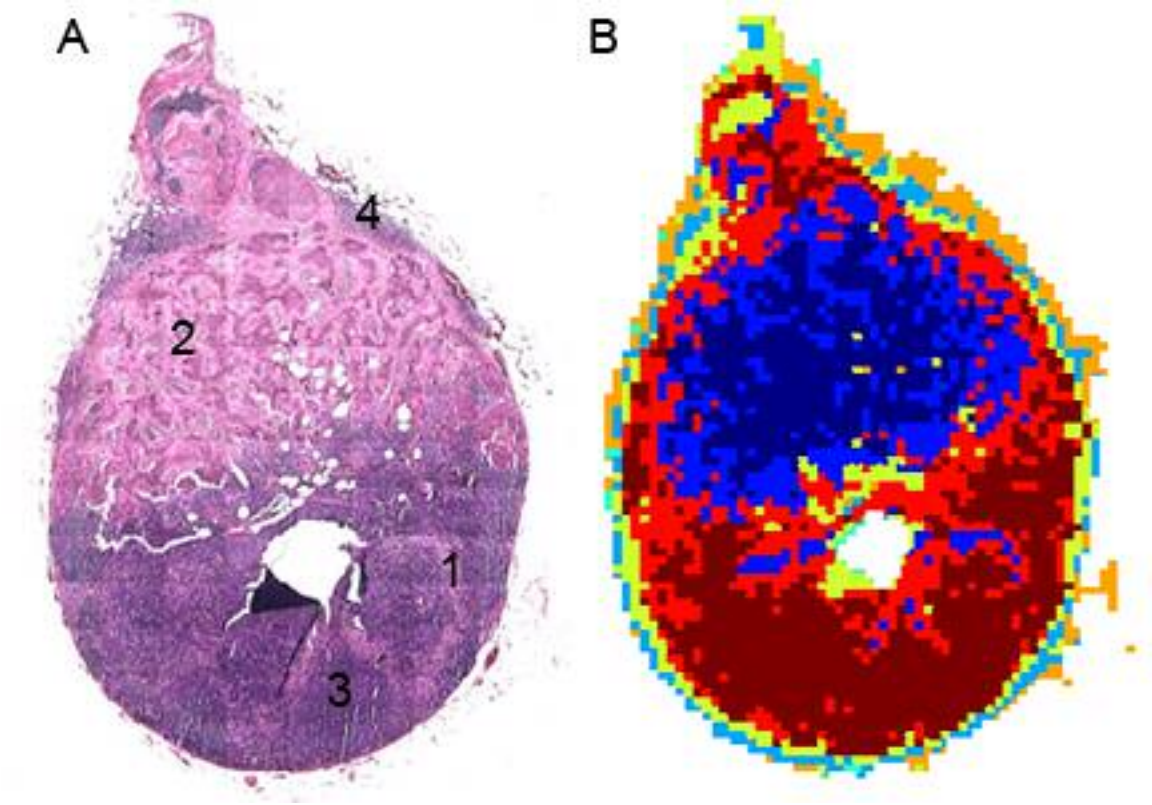

Figure 4.12 (A) Photomicrograph of an H&E stained tissue section of a lymph node effaced with the metastatic breast cancer. Marked morphological features are: macrophages-1, cancer-2, lymphocytes-3 and fat/capsule-4. (B) HCA pseudo-color map of the same lymph node shown in A obtained after IR data collection and prior to staining. Presented in red, macrophages; in dark blue/blue, cancer cells; in brown, lymphocytes; and in aqua, green and orange, fat and capsule.

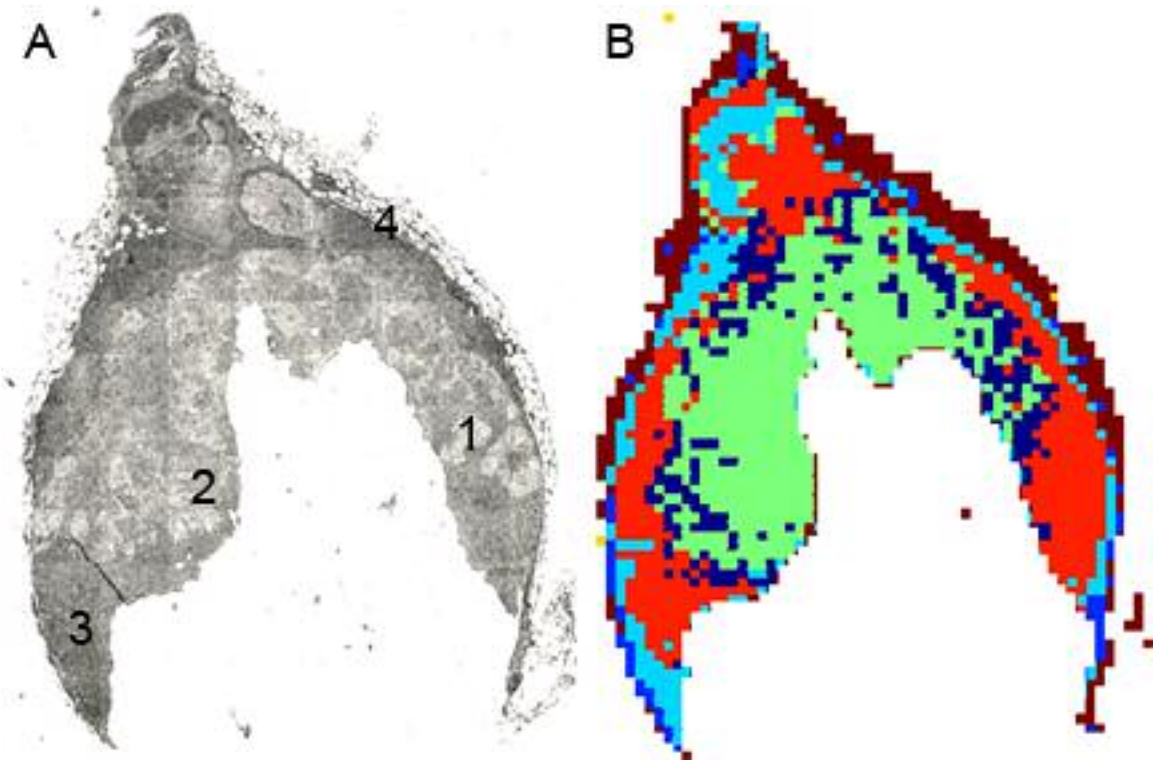

Figure 4.13 (A) Micrograph of unstained tissue slice adjacent to the one in Fig. 4.12A, exhibiting the following features: macrophages-1, cancer-2, cortex-3, and capsule and fat-4. (B) HCA pseudo-color map of the same lymph node shown in A obtained after IR data collection. Presented in brown, aqua and blue, capsule and fat; in red, cortex; in dark blue, macrophages; and in green cancer.

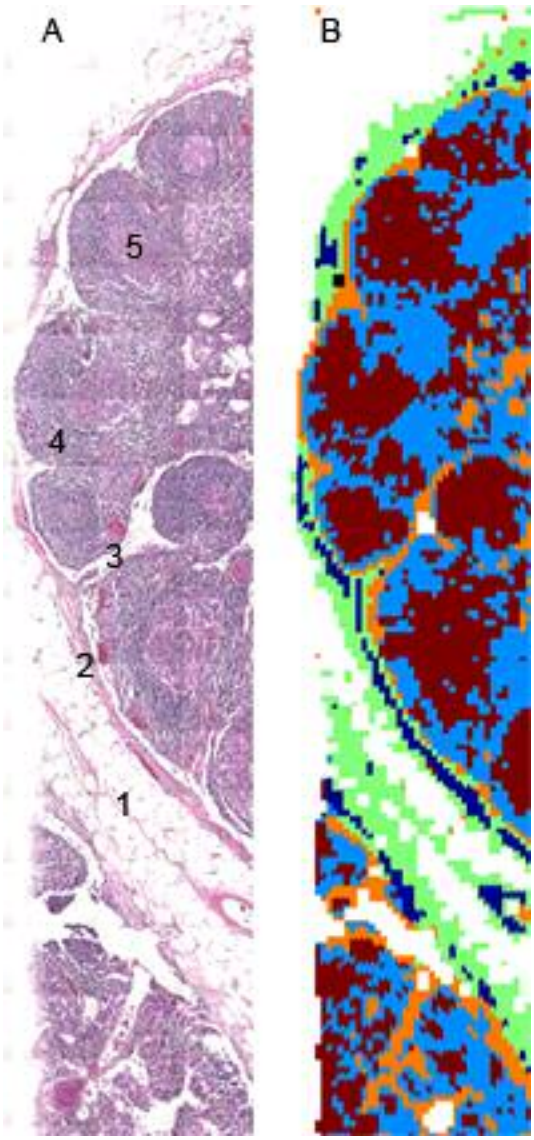

Figure 4.14 (A) Micrograph of an H&E stained lymph node displays its various morphological features: fatty tissue-1, capsule-2, medullary cord-3, paracortex-4 and secondary lymphoid follicle-5. (B) HCA pseudo-color map of the same lymph node shown in A obtained after IR data collection. Presented in green, connective tissue; in dark blue, fibrous tissue; in orange, medullary cords; in brown, B-lymphocytes; and in blue, T-lymphocytes.

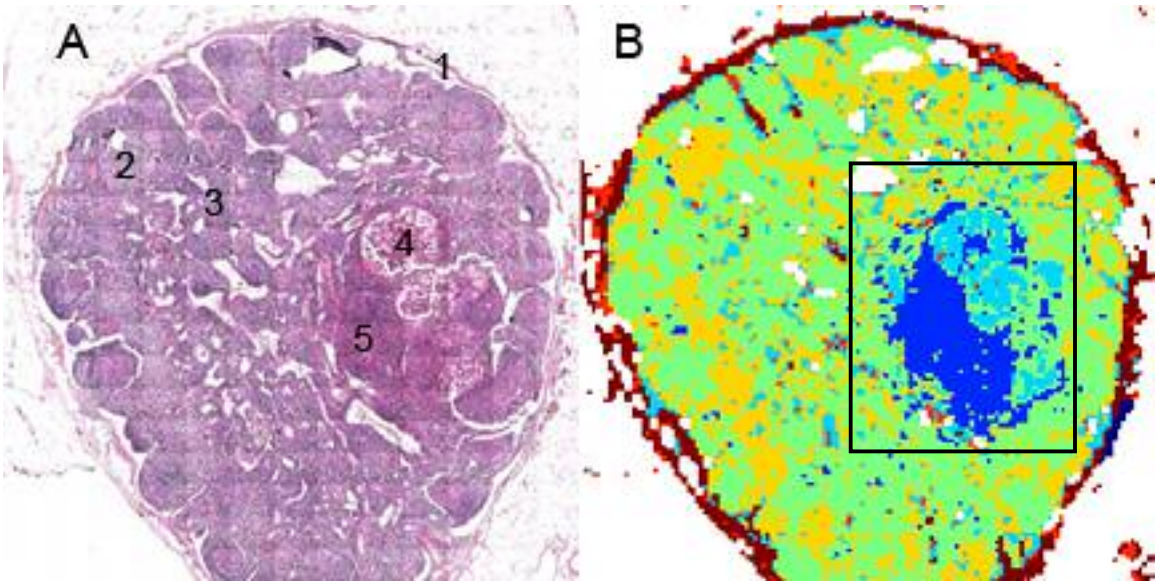

Figure 4.15 (A) Micrograph of an H&E stained tissue section of a lymph node effaced with colon cancer; capsule and fat-1, paracortex-2, cortex-3, glandular cancer-4, and stromal cancer-5. (B) HCA pseudo-color map of the same lymph node shown in A obtained after IR data collection. Presented in brown and orange, capsule and fat; in yellow, paracortex; in green cortex; in aqua, glandular cancer; and in blue stromal cancer. Area marked with the black rectangle is imaged in Figure 4.19.

Maps produced by ANN imaging showed very good to excellent correspondence to HCA based images, which have already been related to H&E stained images.

Figure 4.16 shows the results of ANN imaging of a lymph node effaced with the metastatic breast cancer. Detailed description of the color-code key is given in the figure caption. The ANN algorithm achieved sensitivity of 96 % and specificity of 98 %. All of the five classes were different enough to enable a single level ANN classification.

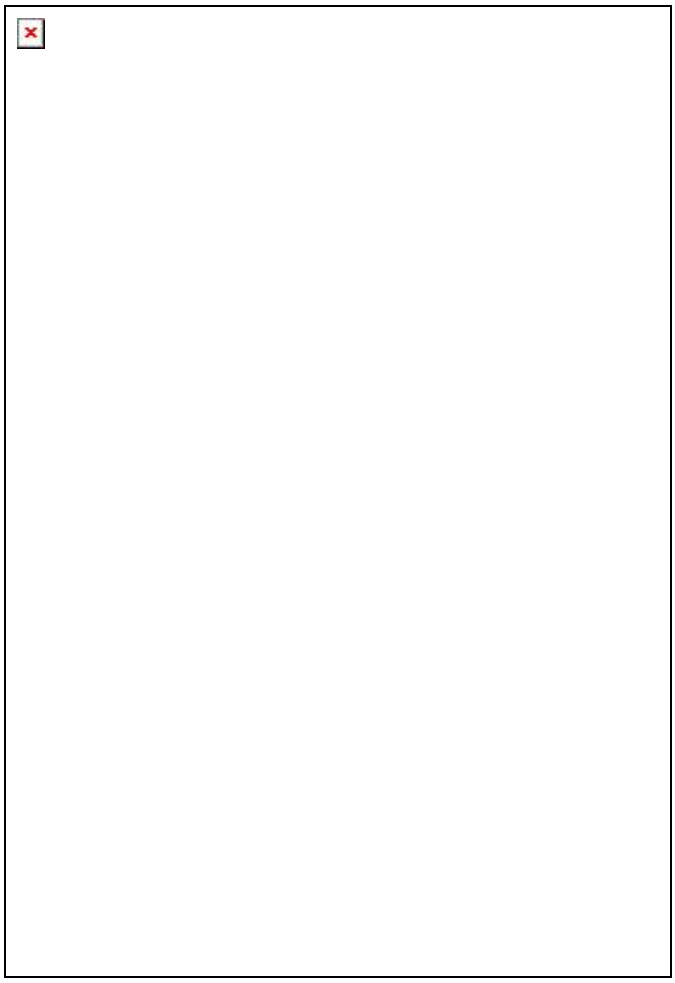

Figure 4.16 Imaging based on ANN analysis of IR data. A lymph node effaced with the metastatic breast cancer. Presented in green, macrophages; in blue/gray, cancer cells; in aqua, lymphocytes; and in red, fat and capsule.

ANN image of the tissue slice from Figure 4.13A is shown in Figure 4.17, which was the adjacent to the one described in the previous paragraph. The same ANN algorithm as the one used for imaging in Fig. 4.16 achieved sensitivity (percentage of classified spectra compared to HCA database) of 95 % and specificity (percentage of correctly classified spectra within classes) of 98 %. This

example showed that there is very small inter-patient variation when using the same algorithm.

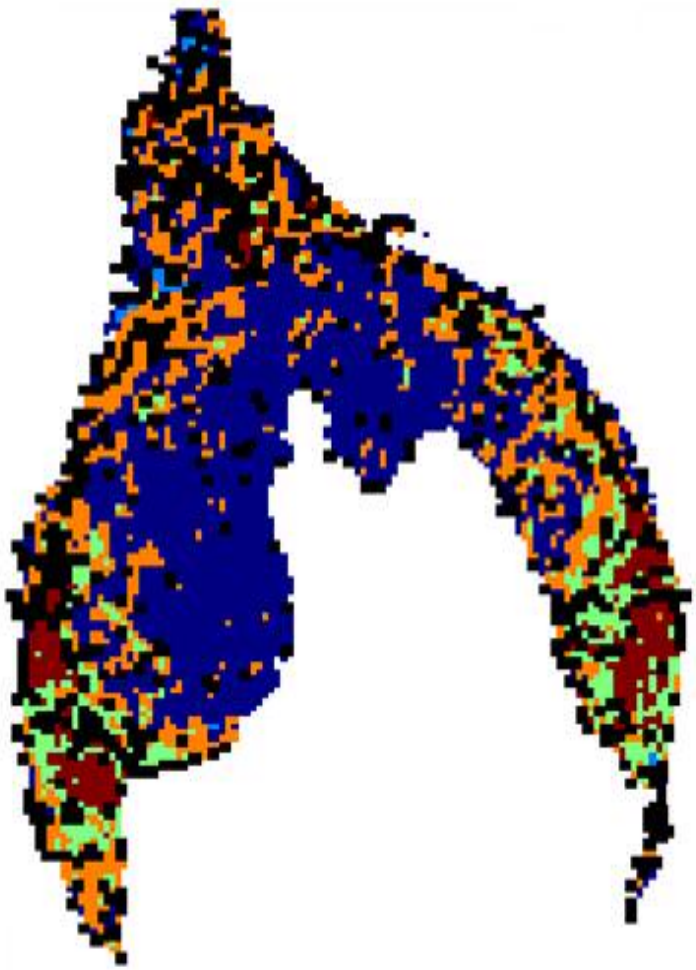

Figure 4.17 AAN pseudo-color map of the lymph node shown in Fig. 4.13A. Presented in brown, lymphocytes; in dark blue and orange, cancer; in green, macrophages; and in light blue, capsule.

Figure 4.18 shows the results of ANN imaging of a healthy lymph node lymph node, displaying its various morphological features. A detailed description of the color-code key is given in the figure caption. A sensitivity of

94 % and specificity of 97 % was achieved by the applied ANN algorithm. All of the classes were different enough to enable a single level ANN classification.

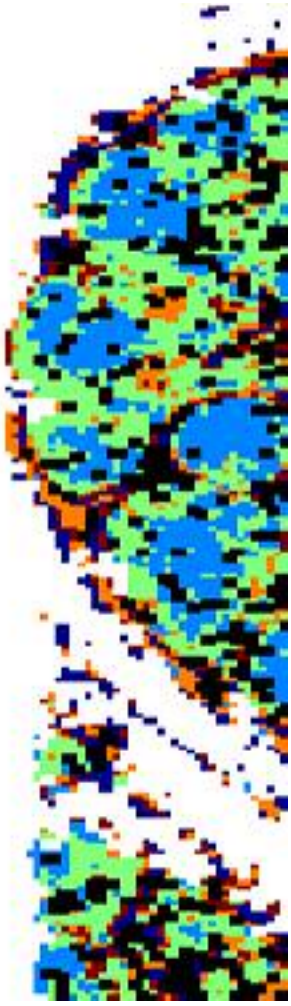

Figure 4.18 Imaging based on ANN analysis of IR data. The same lymph node shown in Fig. 4.13A, showing various morphological features, presented in light blue, B-lymphocytes; in green, T-lymphocytes; in orange, connective tissue; in dark blue, fibrous tissue; and in brown sinusoids.

ANN image of the tissue section of a lymph node effaced with colon cancer (Fig. 4.15) is shown in Fig. 4.19. Detailed description of the color-code key

is given in the figure caption. ANN algorithm achieved sensitivity of 94 % and specificity of 98 %. In this case it was necessary to use a three-level hierarchical neural network to correctly classify the spectra. A single layer artificial neural network could have reached convergence of the training and validating data set errors, ultimately classifying the spectra, but the convergence was extremely slow.

Figure 4.19 ANN pseudo-color map of a part of a tissue section of the lymph node with colon cancer (area marked by the black rectangle in Fig. 4.15A), where in green, cortex; in red, stromal cancer; in olive, glandular cancer; and in salmon paracortex.

4.2.3 Performance of ANN Algorithms

When an ANN fails to correctly classify a spectrum, the output of such a pattern is represented by a black pixel in the ANN pseudo-color map. The ANN software used for the neural network analysis, (NeuroDeveloper 2.5b) applied the following criteria for class assignment: WTA which stands for "winner takes all", "40-20-40" function and user determined maximum allowed extrapolation in winner class.

WTA classification depended on the highest output activation. A minimum value for the winner was set to 0.7 and a minimum distance to the second highest activation was set to 0.3.

The "40-20-40" function applied a different winner decision making algorithm; the activation of one neuron had to exceed 0.6 and all other activations of further classes had to be below 0.4 or the spectrum remained unclassified.

The algorithm applied by NeuroDeveloper for the extrapolation in winner class used a distance value derived from the training and validation data sets. The maximum distance of a pattern to its corresponding class is calculated and set to 100 %. During the classification of a new pattern, the distance is calculated and set into relation to the distances in the training data. In case the calculated extrapolation value of the spectrum exceeds 100%, an extrapolation occurs. The

closer the value is to 100%, the stricter the threshold. The default value to determine patterns as unclassified was set to 200%.

There are a number of reasons for failures of the ANN algorithms to correctly classify all the spectra. IR data acquisition was performed in reflection mode and this technique can introduce certain types of spectral distortions. MirrIR slides, in conjunction with certain types of tissues, can cause the appearance of a dispersion or reflectance artifact in IR spectra [50]. Artifacts caused an unusual ratio of amide I/II bands, a distorted, sloping baseline and a shift to lower wavenumbers of the bands. Secondly, Rprop algorithm is designed to deal, primarily, with data having the Gaussian distribution; nonetheless it is very robust in classification problems but not necessarily the most appropriate one [51, 52]. Thirdly, when the training and validation data sets do not compromise all classes or the entire range of a feature needed for a given classification problem, the potential for misclassification increases. With the advance of spectroscopic and statistical software and the interface between them, the above mentioned shortcomings could be overcome.

5 Bibliography

[1] Ross, M. H.; Romrell, L. J. *Histology*, 2nd ed.; Williams & Wilkins: Baltimore; 1989; pp.332-339.

[2] http://upload.wikimedia.org/wikipedia/en/6/63/Biological_cell.png

[3] http://www.med.unibs.it/~marchesi/cellcycle.gif

[4] http://tidepool.st.usm.edu/pix/mitosis.jpg

[5] http://www.vh.org/adult/provider/anatomy/MicroscopicAnatomy/ /Images/Plate166.jpg

[6] Carbone, P. P.; Kaplan, H. S.; Musshoff, K.; Smithers, D. W.; Tubiana, M. *Cancer Res.* **1971**, *31*, 1860-1861.

[7] Percy, C.; Fritz, A.; Jack, A.; Shanmugarathan, S.; Sobin, L.; Parkin, D.M.; Whelan, S. *International Classification of Diseases for Oncology (ICD-O)*, 3rd ed.; World Health Organization: Geneva; 2001.

[8] Surveillance, Epidemiology and End Results Program and the National Center for Health Statistics; http://seer.cancer.gov

[9] National Cancer Institute Financial Management Branch; http://www3.cancer.gov/admin/fmb/

[10] Diem, M. *Modern Vibrational Spectroscopy*; John Wiley & Sons: New York; 1993.

[11] Messerschmidt, R. G. *Practical Guide to Infrared Microspectrscopy*; Humecki, H. J., Ed.; Marcel Dekker: New York; 1995.

[12] Schiering, D. W.; Tague, T. J., Jr.; Reffner, J. A.; Vogel, S. H. *Anal Bioanal Chem* **2000**, *28*, 46-52.

[13] Reffner, J. A.; Martoglio, P. A. *Rev. Sci. Instrum.* **1995**, *66*, 1298-1302.

[14] Savitzky, A.; Golay, M. J. E. *Anal. Chem*. **1964**, *36*, 1627-1639.

[15] Wood, B. R.; Chiriboga, L.; Yee, H.; Quinn, M. A.; McNaughton, D.; Diem, M. *Gynecol. Oncol.* **2004**, *93*, 59-68.

[16] Lasch, P. *A Matlab based application for IR imaging*, see http://www.cytospec.com

[17] Reed, R. D.; Marks, R. J.; II *Neural Smithing*; MIT Press: Cambridge, MA; 1998.

[18] Bishop, C. M. *Neural Networks for Pattern Recognition*; Oxford University Press: New York; 1999.

[19] Rojas, R. *Proceedings of the International Conference on Artificial Neural Networks (ICANN '94)*; Springer-Verlag: London, 1994; pp 577-580.

[20] Wessels, L. F. A. *IEEE Transactions on Neural Networks* **1992**, *3*, 899-905.

[21] Nguyen, D. H.; Widrow, B. *Proceedings of the International Joint Conference on Neural Networks (San Diego)*; IEEE Press: New York, 1990; vol. 3, pp 211-226.

[22] Georgiu, G. M.; Koutsougeras, C. *Proceedings of SPIE Conference on Adaptive and Learning Systems (Orlando)*; SPIE: Bellingham, 1992.

[23] Smyth, S. G. *IEEE Transactions on Neural Networks* **1992**, *3*, 329-333.

[24] Reed, R. D.; Marks, R. J.; II *IEEE Symposium on Neuroinformatics and Neurocomputing (Rostov on Don)*; IEEE Press: New York, 1992; pp 120-127.

[25] Sethi, I. K. *Proceedings of the IEEE*; **1990**, *78*, 1605-1613.

[26] Shavlik, J. W. *Artificial Intelligence and Neural Networks: Steps Toward Principled Integration*; Honavar, V.; Uhr, L., Ed.; Academic Press: New York; 1994; pp 561-580.

[27] Sietsma, J; Dow, R. J. F. *Neural Networks*; **1991**, *4*, 67-79.

[28] Okada, H. *Proceedings of the International Joint Conference on Neural Networks (Baltimore)*; IEEE Press: New York, 1992; vol. 1, pp 239-244.

[29] Vogl, T. P.; Manigs, J. K.; Rigler, A. K.; Zink, W. T.; Alkon, D. L. *Biological Cybernetics*; **1988**, *59*, 257-263.

[30] Jacobs, R. A. *Neural Networks*; **1988**, *1*, 295-307.

[31] Silva, F. M.; Almeida, L. B. *Neural Networks, Proceedings EURASIP Workshop*; Springer-Verlag: New York; vol. 412, pp 110-119.

[32] Tollanaere, T. *Neural Networks*; **1990**, *3*, 561-573.

[33] Fahlman, S. E. *Proceedings of the 1998 Connectionist Models Summer School*; Morgan Kaufmann: San Mateo, 1988; pp 38-51.

[34] Chiriboga, L.; Yee, H.; Diem, M. *Appl. Spectrosc.* **2000**, *54*, 480-485.

[35] Naumann, D.; Schultz, C. P.; Helm, D. *Infrared Spectroscopy of Biomolecules*; Mantsch, H. H.; Chapman, D., Ed.; Willey and Sons: New York, 1996.

[36] Jackson, M.; Ramjiawan, B.; Heweko, M.; Mantsch, H. H. *Cell. Mol. Biol.*; **1998**, *44*, 89-98.

[37] Boydston-White, S.; Gopen, T.; Houser, S.; Bargonetti, J.; Diem, M. *Biospectrscopy*; **1999**, *5*, 219-227.

[38] Miljkovic, M.; Romeo, M.; Diem, M. *Biopolymers*; **2004**, *74*, 172-175.

[39] Lasch, P.; Pacifico, A.; Diem, M. *Biopolymers*; **2002**, *67*, 335-338.

[40] Romeo, M.; Matthaüs, C.; Miljkovic, M.; Diem, M. *Biopolymers*; **2002**, *67*, 168-172.

[41] Wetzel, D. L.; Levin, S. M. *Biological Application of Infrared Micro-spectroscopy*; Gremlich, H. U.; Yan, B., Ed.; Marcel Dekker: New York, 2001; pp. 1-14.

[42] Shaw, R. A.; Mansfield, J. R.; Rempel, S. P.; Low-Ying, S.; Kupiriyanov, V. V. *J. Mol. Struct.* **2000**, *500*, 129-138.

[43] Diem, M.; Boydston-White, S.; Chiriboga, L. *Appl. Spectrosc.* **1999**, *53*, 148A-161A.

[44] Fabian, H.; Lasch, P.; Boese, M.; Haensch, W. *Biopolymers* **2002**, *67*, 354-357.

[45] Lasch, P.; Schmitt, J.; Naumann, D. In *Biomedical Spectroscopy: Vibrational Spectrosopcy and other Novel Techniques*; Mahadevan-Jansen, A.; Puppels, G. J., Ed.; Proceedings of SPIE, 2000, vol. 3918; pp. 45-56.

[46] Romeo, M.; Burdon, F.; Quinn, B.; Wood, B.; McNaughton, D. *Cell. Mol. Biol.* **1998**, *44*, 179-187.

[47] Stoler, M. H.; Schiffman, M. *JAMA*, **2001**, *285*, 1500-1505.

[48] Klinkhamer, P. J.; Vooijs, G. P.; de Haan, A. F. *Acta Cytol.* **1988**, *32*, 794-800.

[49] Lasch, P.; Naumann, D. *Cell. Mol. Biol.* **1998**, *44*, 189-202.

[50] Romeo M. J., Diem, M. *Vibrational Spectrosc.* **2005**, *38*, 115-119.

[51] Riedmiller M. *Computer Standards & Interfaces* **1994**, *16*, 265-278.

[52] Riedmiller M., Braun H. *IEEE International Conference on Neural Networks (San Francisco)*; IEEE: New York, 1993, vol. 1; pp. 586-591.