Strategies for Untargeted Biomarker Discovery in Biological Fluids

Michel Boisvert

A Thesis

In the Department

of

Chemistry and Biochemistry

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (chemistry) at

Concordia University

Montreal, Quebec, Canada

January 2012

CONCORDIA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Michel Boisvert**

Entitled: **Strategies for Untargeted Biomarker Discovery in Biological Fluids**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Chemistry)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

____Chair
Dr. P. Gulick

____ External Examiner
Dr. P. Wentzell

____ External to Program
Dr. D. Walsh

____ Examiner
Dr. L. Cuccia

____ Examiner
Dr. Y. Gélinas

____Thesis Supervisor
Dr. C. Skinner

Approved by ____
Chair of Department or Graduate Program Director
Dr. H. Muchall, Graduate Program Director

April 5, 2012 ____
Dr. B. Lewis, Dean, Faculty of Arts and Science

ABSTRACT

Strategies for Untargeted Biomarker Discovery in Biological Fluids

Michel Boisvert, Ph. D.
Concordia University, 2012

The health status of an organism modulates the dynamic and complex interplay of biochemical species that make-up the body and fluids of the organism. As such, these biological fluids are routinely used for diagnostic testing, yet they are often not used to their full potential. For instance, amniotic fluid (AF), the fluid that surrounds the fetus during gestation, is collected primarily for genetic testing from women with identified risk factors. The AF proteome and/or metabolome are seldom considered and represent a largely untapped wealth of relevant clinical information. Extensive, multi-analyte data can be collected from biological samples with modern analytical instrumentation. However, sophisticated data preprocessing and analysis (*i.e.* chemometrics) are required to reveal the relationships between the biochemical signals and the health status. This thesis seeks to demonstrate that untargeted biomarker discovery strategies can be efficiently applied to the task of finding novel biomarkers and complement the traditional hypothesis driven approaches.

In the work underlying this thesis, a chemometric data analysis strategy was developed to search for biomarkers in capillary electrophoresis (CE) separations data. The absorbance data from amniotic fluid samples (n=107) collected at 15 weeks gestation, at 195 +/- 4 nm, was normalized, time aligned with Correlation Optimized Warping and reduced to a smaller number of variables by Haar transformation. The reduced data was then classified into normal or abnormal health classes by using a Bayes classifier algorithm.

The chemometric data analysis was first employed to find biomarkers of gestational diabetes mellitus (GDM) and revealed that human serum albumin (HSA) could predict the early onset of disease. The same approach was successfully used to identify cases of large-for-gestational age (LGA) with the same AF CE-UV data. It was also employed for the classification of embryos with high and low reproductive potential using *in vitro* fertilization (IVF) culture media analyzed by CE-UV.

Overall, a chemometric method was developed to perform untargeted biomarker discovery in biological samples and provide new means to detect GDM pregnancies, LGA neonates and viable embryos in IVF. The method was successful at identifying biomarkers of interest and showed high flexibility and transferability to other biological fluids.

Acknowledgements

First, and foremost, I thank my supervisor, Cameron Skinner, for the opportunity he gave me to pursue research in his laboratory. I could always count on him to provide insightful ways to overcome difficulties and finds solutions to the challenges of research. I particularly appreciate the freedom I was given with respect to my thesis work.

A large part of my thesis pertains to chemometrics and I have to thank David Burns for the time he has given me. He is an extremely busy person and yet still found the time to help me get through impasses with some of the more challenging data analysis step.

I also have to thank lab members, past and present, for being there to bounce ideas around and troubleshoot, for being more than colleagues, for being friends. So, in order of appearance, thank you Jean-Louis Cabral, Wei Lin, Maria Kaltcheva, Alexander Lawandi, Lynn Miller and John Chin, I have learned from you, I have laughed with you and I will carry fond memories of all of you and our time at Concordia.

I have met some great friends and colleagues during my time at Concordia. I can’t name them all. I thank them and apologize for not naming them.

I need to thank my family for being there, for being encouraging and supportive of the somewhat crazy undertaking that is a Ph.D.

Finally, and most importantly, I would like to thank my life partner, lover and friend Alessandra Chan. She has been very patient as we have put many projects on ice until I complete my degree. The scientific process seems to be a series of failure from which we hopefully learn something and ultimately succeed. She gave the strength and

encouragement I needed to complete my thesis. Having her by my side to get through more difficult times and enjoy the good times makes me a very lucky man.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AF: amniotic fluid
AGA: appropriate for gestational age
ART: assisted reproduction technique
BGE: Background electrolyte
BPH: benign prostatic hyperplasia
CE: capillary electrophoresis
cIEF: capillary isoelectric focusing
COW: correlation optimized warping
DET: double embryo transfer
DWT: discrete wavelet transform
ELBW: extremely low birth weight
EOF: electroosmotic flow
FCA: fetal cardial activity
GA: genetic algorithm
GDM: gestational diabetes mellitus
hCG: human chorionic gonadotropin
HETP: height equivalent of a theoretical plate
HSA: Human serum albumin
HT: Haar transform
ILS: inverse least-squares
IVF: in vitro fertilization
LBW: low birth weight
LGA: large for gestation age
LIF: laser induced fluorescence
MS: mass spectrometry
NIR: near infrared
OGTT: oral glocuse tolerance test
OS: oxidative stress
PAGE: polyacrylamide gel electrophoresis
PAPP-A: pregnancy-associated plasma protein A
PDA: photo diode array
PSA: prostate specific antigen
RSD: relative standard deviation
SDS: sodium dodecyl sulfate
SET: single embryo transfer
SGA: small for gestational age
UV: ultraviolet
VLBW: very low birth weight

Chapter 1

Introduction

The term biomarker is a contraction of the two words *biological* and *marker* and refers to any biological species (protein, peptide or small molecule) that has predictive power towards a particular biological state/outcome, such as a disease. A significant focus of clinical research, as evidenced by the large number of publications, has been the discovery and validation of biomarkers with the primary aim of facilitating disease diagnosis. The problem encountered in biomarker discovery is that any of the multitude of species present in the biological system is a potential biomarker, but only a very few are. Modern science has developed a few effective strategies that can be applied to biomarker discovery. This thesis seeks to demonstrate that untargeted biomarker discovery strategies could be applied to find novel biomarkers in amniotic fluid separated by capillary electrophoresis.

The primary motivation that fuels biomarker discovery efforts is the hope that early detection of a disease will allow treatment to begin earlier. This is believed to help mitigate the negative consequence of the disease and will ultimately lead to better health outcomes. From a societal point of view, early disease detection can also translate into a reduced burden on the health care system by lowering costs associated with illnesses. Additionally, discovering new biomarkers should reveal new and important information about a particular disease and its progression, which can then help researchers better understand it.

While each disease is a unique biological and physiological progression from health to illness, it is important to appreciate that there is a latency period when the disease is active but may not be displaying overt symptoms. This usually causes a multitude of alterations to the normal biological/physiological pathways. These alterations can have a profound impact on the genetic and biochemical profile of a patient. Unfortunately, by the time that overt symptoms present, the disease is well underway. Finding biological indicators of a disease when the disease is well advanced is useful in understanding its progression and impacts, but not necessarily helpful in predicting its onset. Finding useful biomarkers indicative of the early stages of a disease is non-trivial since patients may not be afflicted with obvious symptoms, so why would biological samples be collected from these people? Fortunately, for researchers, many people have established risk factors for the disease under investigation and are enrolled in medical studies as will be seen in this thesis for the amniotic fluid chapters.

The medical literature is replete with examples of single biomarkers used to detect a disease. Perhaps one of the most famous biomarkers is the prostate specific antigen (PSA). As originally applied by physicians, results showing abnormal levels of PSA were sufficient to confirm a diagnosis of prostate cancer. Another example of a powerful and common biomarker is glucose. Diabetes can be diagnosed, or at least confirmed, based on fasting blood glucose levels. However, using just one biomarker for the determination of a disease has the potential for misdiagnosis. The inherently high "biological" variability in the concentration of those species, from individual to individual, and over time within a single individual is difficult to overcome.

The prostate specific antigen case is worth discussing further as there has been some controversy concerning its proper use and its evolution as a biomarker [1]. A serum PSA level above 3-4 ng/mL is sufficient cause for further testing such as a prostate biopsy. The controversy emanates from two studies that have not been able to establish a net benefit in screening for PSA [2, 3]. Part of the problem is that prostate cancer is not the only cause of elevated PSA in serum; other diseases such as benign prostatic hyperplasia (BPH) can cause elevated PSA. In 2009, Sarrats *et al.* showed that PSA has several subforms (we will later re-define this term as protein isoforms) and that the amount of one subform compared to the others can be indicative of PSA or BPH [4]. The key point is that it is risky to establish a diagnosis based on a single measurement as there can be varied cause for one abnormal measurement.

Continuing with the second example, diabetes is essentially a disease where the body is unable to properly manage glucose and therefore high blood glucose is a sign of diabetes. No matter what, high blood glucose is a sign that the patient's metabolism is not functioning properly, yet it gives little indication as to why. What type of diabetes is present? More tests are required to establish a clear diagnosis. Furthermore, by the time the blood glucose is at the level where diabetes is diagnosed, many biochemical pathways are already affected and damage to organs already underway. When glucose levels are used in an attempt to predict the early signs of diabetes (prediabetes) the results are not very reliable and may fail to detect diabetes up to 50% of the time [5]. This begs the question: are there better, more reliable biomarkers of diabetes that can provide early detection of the disease?

Diseases, whether they be triggered by pathogens, genetic anomalies, age or poor health, can initiate and modulate a complex series of biological and biochemical processes within the organism. This leads to the realisation that more than one biochemical species will be modulated by the disease. Thus, a multimarker approach is warranted to improve the efficacy of disease prediction and identification. The rationale for a multimarker approach for diagnosis is that it can improve the chances of early disease detection by combining risk factors and multiple biomarkers. This approach has been applied to various diseases, or abnormal health statuses, yielding minimal improvements for the prediction of cardiovascular events [6], but some notable improvement for prediction of gestational diabetes mellitus [7] or large-for-gestational age neonates [8]. Ultimately, the viability of this approach remains to be proven. To do so, novel biomarkers need to be found and evaluated, not only for their diagnostic value, but also to evaluate if their use actually improves the clinical outcome [9].

The traditional approach to biomarker discovery is to use a hypothesis driven approach where sophisticated knowledge of the biological system is used to determine if a chosen (bio)chemical species should be tested to assess its predictive value for a specific condition, *e.g.* is blood glucose predictive of diabetes? This approach comes at the risk of missing previously unknown biochemical relationships and limits finding markers to those few species investigated or which science has some level of knowledge. However, the level of sophistication attained by the modern biological human model [10] is a testament to the effectiveness of hypothesis driven approaches. Yet hypothesis driven approach shows some limitations when the number of genetic (genes $2x10^4$; mRNAs $>10^6$), proteomic (proteins $>10^6$; modified proteins $>10^7$) and metabolic ($3x10^3$)

components [11] in the human body are considered. Of course, many possibilities can be eliminated based on the current state of biomedical knowledge, yet there remains a considerable number of unknown species and species with unknown functions [12, 13]. Those unknown species simply cannot be investigated efficiently using hypothesis driven approaches. The proposed solution to this limitation of the hypothesis driven approach is to employ untargeted biomarker discovery strategies to search through the complexity of biological samples for useful disease biomarkers [14]. For this approach to be successful experiments are designed in such a way as to maximize the number of species detected quantitatively to generate large datasets that can be used to compare healthy and disease states for distinctive profiles.

The challenge in untargeted biomarker discovery lies in managing and analyzing the large datasets produced with this form of experimentation. Chemometrics is a field of chemistry concerned with relating measurements made on a chemical system, or process, to the state of the system by employing statistical or mathematical methods. These methods can be employed to improve the quality of the data collected (experimental design [15], optimization of experimental parameters [16], signal processing [17]) and the extraction of relevant information from the collected data [18, 19] (pattern recognition, modeling and multivariate calibration). As modern analytical instrumentation becomes more and more sophisticated, the data generated by them becomes information rich and complex. It also becomes less and less practical, and even feasible, to analyze data “manually”. This is particularly true for biological sample analysis where genetic, proteomic and metabolomic species are present in great number and diversity. It also demands instruments that have large linear dynamic ranges and low limits of detection that can measure a large portion of the species present in a sample. Obviously, chemometrics becomes an indispensable part of the data analysis to extract the relevant information from these datasets and is going to be more and more prevalent in bioanalytical chemistry in general [20].

The remainder of the Introduction is divided in three sections: capillary electrophoresis, chemometrics and a description of amniotic fluid. The very basic concepts of capillary electrophoresis (CE) will first be described with an emphasis on one of the most vexing difficulties encountered in CE which is migration time variations. Then, the chemometrics section will provide a conceptual explanation of the chemometrics data treatment methods employed. This also includes data preprocessing techniques such as correlation optimized warping (COW) that overcomes the aforementioned migration time variations. The Haar wavelet transform (HT), the genetic algorithm (GA) optimization routine, and the Bayesian classification processes will then be introduced within the context of identifying normal and abnormal states (samples). The last section will provide an overview of amniotic fluid (AF), the biological sample of primary interest in this study, with particular focus on the current state of knowledge of AF biomarkers. The Introduction was designed to provide sufficient background knowledge to understand the core thesis work that is presented in the next four chapters as readers might not have expert knowledge of all three subjects.

1.1 Analytical/Instrumental strategies for biological samples

Biological samples contain complex mixtures of analytes: thousands of proteins, nutrients, salts, hormones, *etc*. Instruments capable of measuring a significant cross-section of such complex samples and outputting data representative of the sample complexity are highly desired for biomarker discovery. Ideally, the instrument should also be simple, easy to use, cost efficient and relatively unbiased. Capillary electrophoresis (CE) is particularly appealing because it meets these criteria.

The following sections will review the many different phenomena that give rise to a separation in CE and also the processes that interfere and degrade the separation. However, to ease the reader into the complexity of a CE analysis a very simplistic overview is presented in Figure 1 where an ideal separation is subjected to increasing realistic processes. If we imagine the CE separation of a simple two component mixture where all band broadening and non-ideal processes were turned off, the resulting CE separation would look like the top trace in Figure 1. The peaks are fully resolved and have a peak profile identical to the injection plug. The unfortunate reality is that many processes contribute to band broadening, non-ideal peak shape and variations in the separation profile as suggested by the text in Figure 1 and explained in greater detail in the following text. All of these undesirable processes complicate the analysis and pose problems for the development of chemometric approaches to analysing CE data sets.

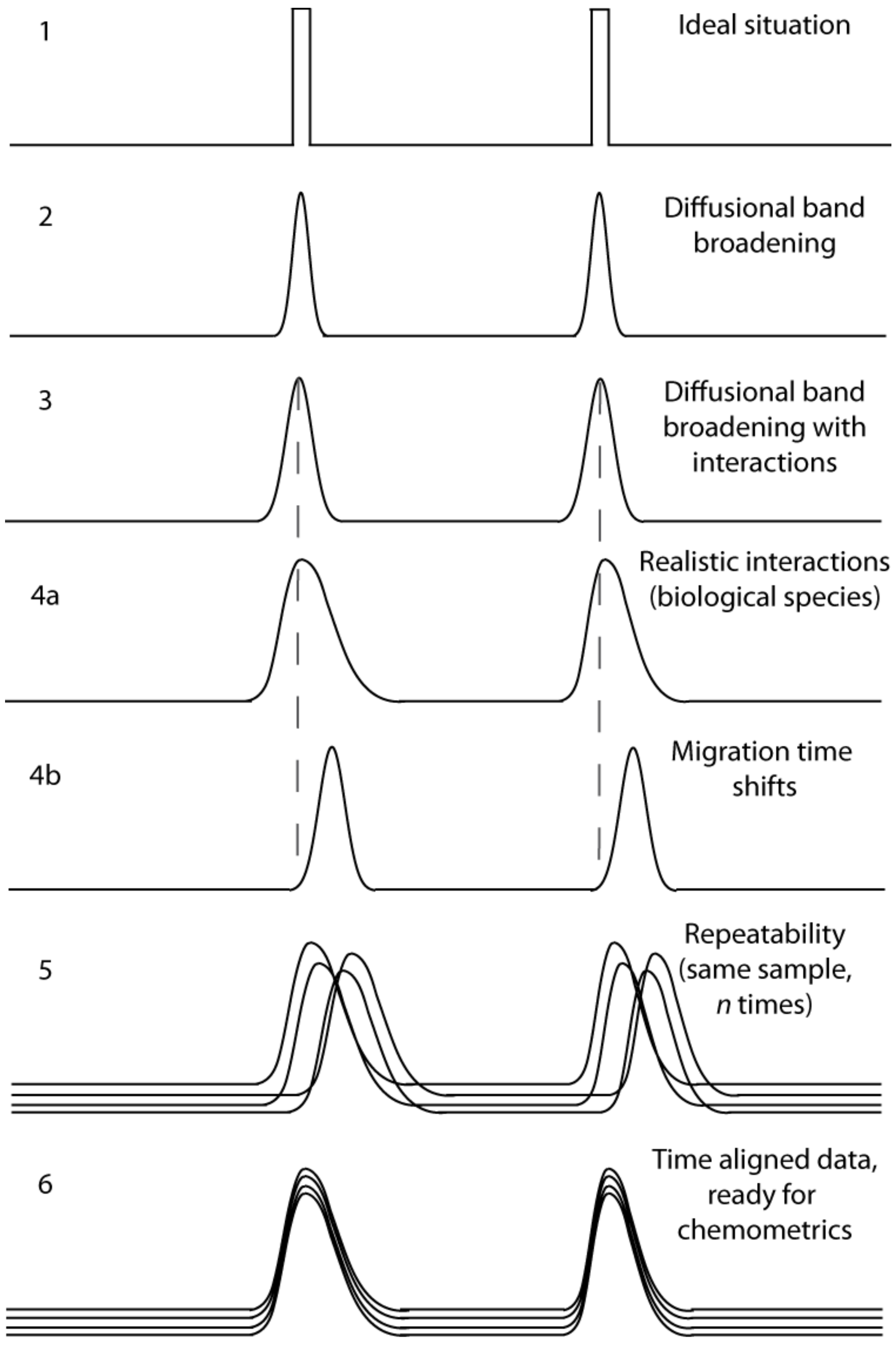

1
Ideal situation
2
Diffusional band broadening
3
Diffusional band broadening with interactions
4a
Realistic interactions (biological species)
4b
Migration time shifts
5
Repeatability (same sample, *n* times)
6
Time aligned data, ready for chemometrics

Figure 1: An "ideal" separation transitioning to a more realistic separation. An ideal separation is depicted in trace 1, followed in trace 2 by a separation where only diffusion contributes to band broadening. In trace 3, diffusion and molecular interactions explain increased band broadening. Peak shape distortions are added in trace 4a due to non-equilibrium processes. Migration time shift from one sample to the next is depicted in trace 4b. In trace 5, the typical repeatability of the same sample run n times shows that there are run-to-run variations. In trace 6, the electrophoretic data has been time aligned to largely remove migration time shifts and allow for subsequent chemometric data processing.

1.1.1 Capillary Electrophoresis

Capillary electrophoresis is an established technique for the analysis of biological fluids [21-23]. The following sub-sections are intended to provide basic concepts in CE and explain what gives rise to peak width and migration time variations. Minimizing the sources of variation is critical to the success of the following data analysis steps described in section 1.2.

1.1.1.1 The Capillary Electrophoresis Instrument

A CE instrument requires surprisingly few components to operate. It begins with a power supply capable of generating a potential difference ranging from 0 to 30 kilovolts. This difference in potential is applied across an electrolyte-filled quartz capillary with each end in electrolyte filled reservoirs. The capillary has an inner diameter that is usually less than 200 μm to minimize Joule heating and is usually coated with polyimide, but Teflon® or polymethacrylate coatings are available for specialized applications. The primary function of the coating is to impart flexibility and strength to an otherwise fragile quartz capillary. The electrical circuit is closed by platinum electrodes immersed in the two liquid reservoirs containing the background electrolyte (BGE), *i.e.* buffer and additives. The content of the buffer reservoirs is introduced into the capillary to control the pH and to generate the electroosmotic flow (EOF). The importance of the BGE and the cause of EOF are explained later in the text. Temperature control of capillary is critical to minimize the effects of Joule heating on the separation. This can be achieved by either ambient or forced air, or preferably liquid cooling. Finally some form of detection is required. Typical detectors include, but are not limited to, UV absorbance, mass spectrometry (MS), amperometric, fluorescence (intrinsic or laser induced (LIF)). These components come together into an instrument that is represented by the diagram of a CE-UV in Figure 2 below.

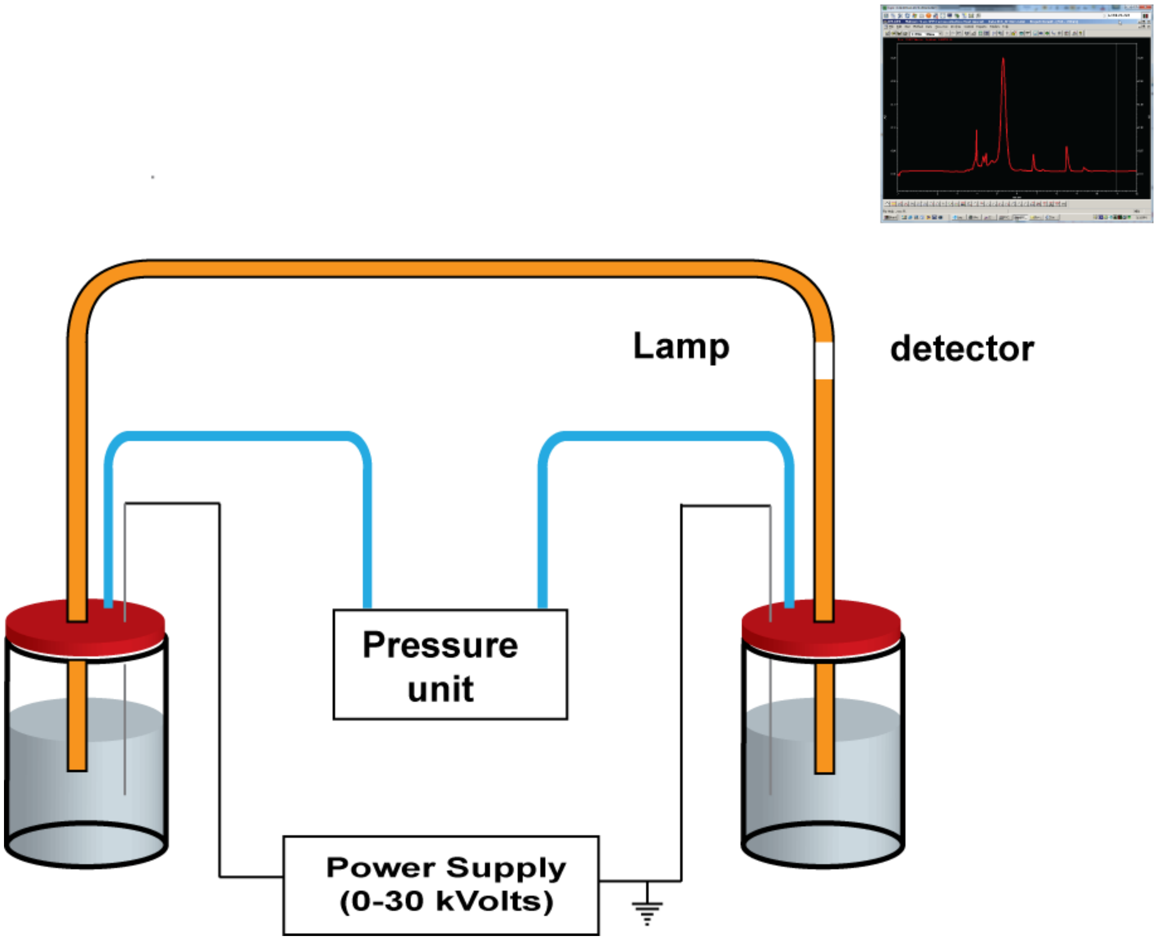


Figure 2: Typical components of a capillary electrophoresis instrument. The ends of a capillary are submerged into buffer containing vials. A potential can be applied across the capillary with a high voltage power supply. The protective coating can be removed to produce a detection window. In some cases, a pressure unit can be used for pressure injections, to pressurize vials during separation and/or to generate a pressure driven flow.

1.1.1.2 Sample introduction

Sample can be introduced into the capillary through electrokinetic injection or by hydrodynamic means. In electrokinetic injection, the inlet of the capillary is placed in the sample vial and an injection voltage is applied allowing the EOF, with a smaller analyte mobility contribution, to dictate the mass of sample introduced into the capillary. Electrokinetic injections find routine use in CE analysis. However, the volume of sample introduced into the capillary critically depends on the EOF and the physio-chemical properties of the sample and are thus subject significant variations with biological samples. One of the more popular methods of introducing a less biased sample is to use hydrodynamic injection either by applying a pressure to the sample vial or by using a siphoning process. The instrument used for this work operates by applying a small pressure to the sealed sample vial. For hydrodynamic injection, Poiseuille's equation relates the injected volume V to the sample`s physical parameters and the pressure differential

$$V = \frac{\Delta P \; \pi \; R^4 t}{8 \, \eta \; l} \qquad \textbf{Equation 1}$$

ΔP is the pressure difference across the capillary, R is the radius of the capillary, t is the time, η is the viscosity and l is the length of the capillary. The reproducibility of the injection can be compromised by poor reproducibility of the pressure generating mechanism and the quality of the components that seal the pressurized sample vial. Viscosity differences between samples and concentration changes as the sample evaporates can prove to be significant source of injection volume variation. In the particular case of protein containing samples, viscosity can increase significantly with increasing protein content and thus cause a decrease in the volume of sample introduced into the capillary. This obviously affects the amount of analyte injected (peak area), but also the starting position of those analytes in the capillary (migration time). Pressure injections are generally preferred to electrokinetic injection when dealing with samples of variable ionic strength and protein concentration because of the better precision of the former.

1.1.1.3 Capillary Zone Electrophoresis and Electrophoretic Mobility

Once injected onto the capillary, ionic species will migrate to the appropriate electrode when the separation potential is applied across the capillary. The migration of ionic species is due to a process called electrophoresis where the ion's migration velocity is referred to as the electrophoretic mobility (μ). The Debye-Huckel-Henry theory provides an acceptable approximation of the electrophoretic mobilities for species in CE:

$$\mu = \frac{q}{6\pi\eta r} \qquad \textbf{Equation 2}$$

where q is the charge of the ionic species, η is the viscosity of the buffer and r is the Stokes' radius of the analyte defined as,

$$r = \frac{k_B T}{6\pi D_i \eta} \qquad \textbf{Equation 3}$$

here D_i is the diffusion coefficient for the analyte, T is the temperature and k_B is Boltzmann's constant. As the size and structural complexity of an analyte molecule increases, the use of the Stokes' radius becomes inappropriate since the analyte molecule does not necessarily adopt a simple spherical shape. Additive and counterion effects can also cause structural changes to the analyte and thus make the spherical approximation deviate enough to invalidate Equation 2. Nevertheless, the approximation allows for a representative depiction of what occurs during a separation in simple CE systems.

In CE, convention holds that the normal mode has the polarity of the inlet electrode as positive and that of the outlet as negative. This is represented above in Figure 2 and below in Figure 3. When a potential is applied across the capillary, the analytes migrate according to their individual electrophoretic mobilities. With normal polarity, we would expect to detect only the cations, but in practice all species (negative, neutral and positive) pass the detector when high pH buffers are used as BGE. In addition to the electrophoretic mobility of the analyte, there is another process called electroosmotic flow (EOF) that occurs. The EOF causes a bulk flow of solvent towards the detector under the conditions stated above. The next section will provide a more detailed account of the EOF.

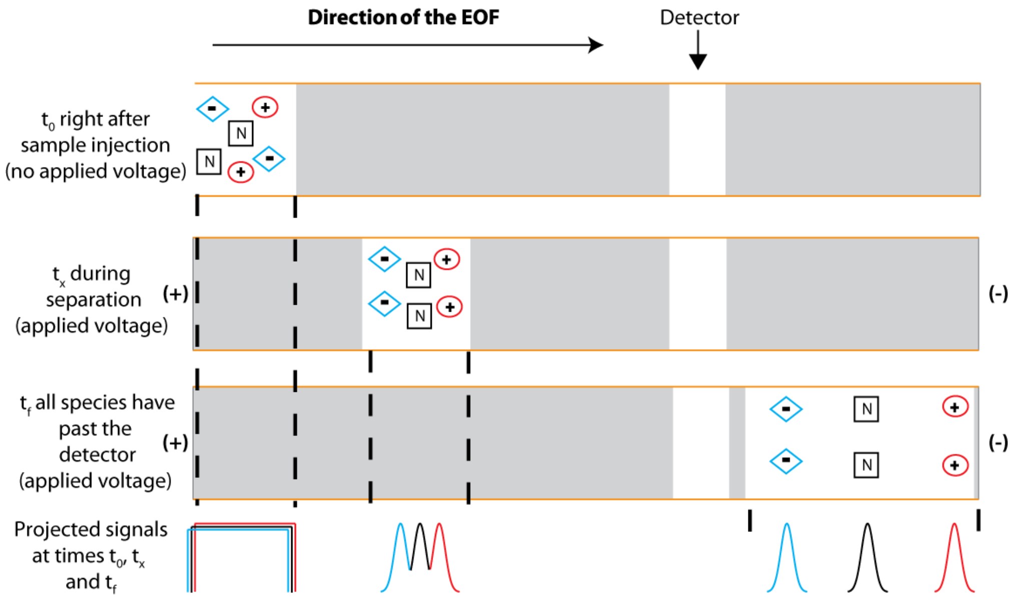

Direction of the EOF
Detector
t0 right after sample injection (no applied voltage)
tx during separation (applied voltage)
tf all species have past the detector (applied voltage)
(+)
(-)
(+)
(-)
N
Projected signals at times t0, tx and tf

Figure 3: Ideal CE separation in normal mode. At t0, the sample mixture is introduced into the capillary. During the separation (t_x), the components of the mixture become more and more resolved. They will be ordered according to charge-to-volume ratios. Once all the components have passed the detector they are well resolved from one another.

1.1.1.4 The Electroosmotic Flow

Helmholtz was first to describe EOF in a series of experiments conducted in the 1870s [24]. His interest revolved around the ionic properties of silica, more specifically the anionic nature of silica at the interface between the silica and an aqueous solution. When he applied an electric field he observed a net flow of solution towards the cathode. Since Helmholtz's discovery, EOF has been extensively investigated. Figure 4 shows the arrangement of ions about the capillary wall that gives rise to the EOF. The silanol groups at the silica/buffer interface are readily deprotonated in high pH BGE because the pKa is around 5.3-6.3 [25] and results in a fixed anionic surface. The charged capillary wall then attracts cations to counter-balance the excess negative charge in several layers. The first layer formed by the cations is the fixed Stern layer beyond which the charge density of cations decreases exponentially as the distance from the capillary wall increases. Beyond the Stern layer, a second, more diffuse layer called the Gouy layer is formed. In the diffuse layer, the cations (and their hydration shell) are free to migrate towards the cathode under the applied electric field. Friction between the mobile cations and the bulk solvent causes a net flow of the entire contents of the capillary towards the cathode. It is critical to note that the "pumping action" originates at the wall which results in plug-like flow that minimizes band broadening.

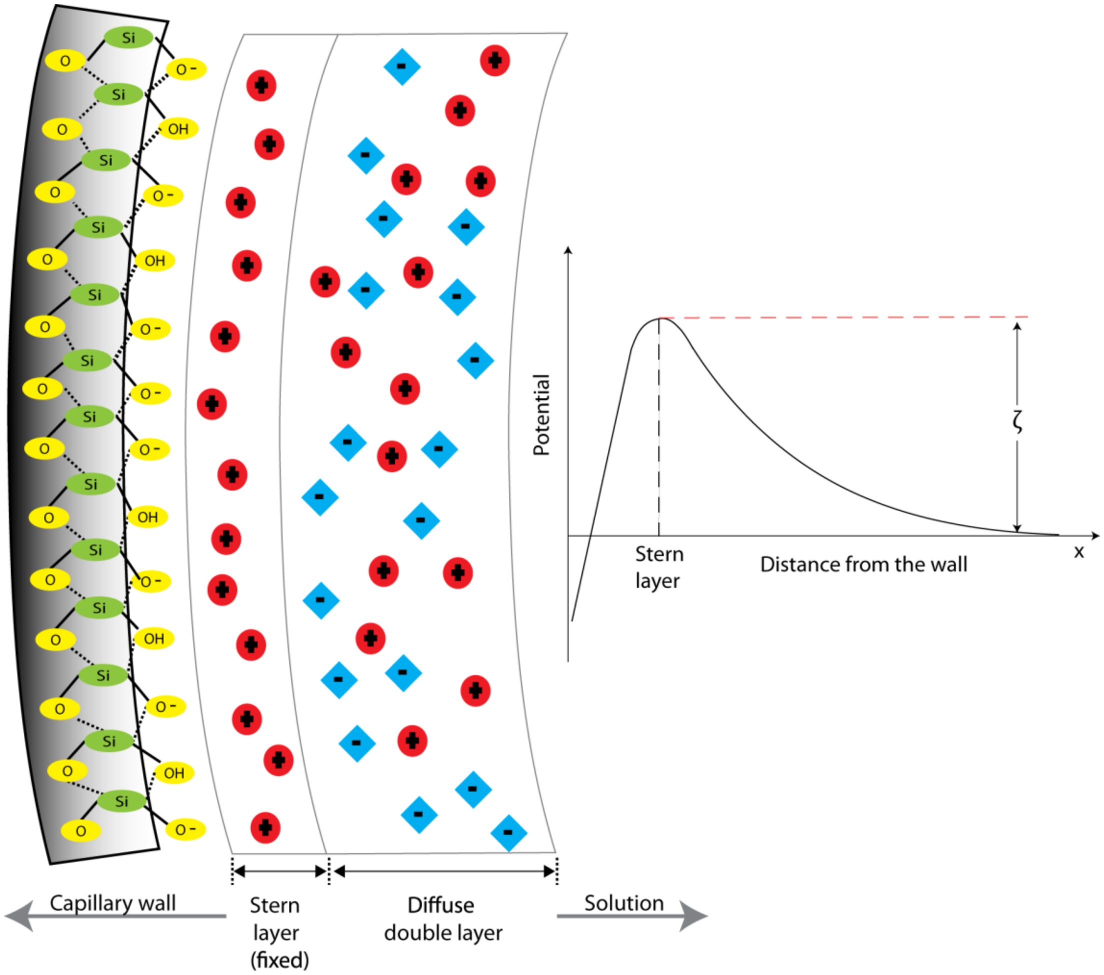


Figure 4: Representation of the charge distribution on the capillary wall. The capillary wall will have protonated and deprotonated silanols depending on the pH of the buffer used. A fixed layer of counter ions (cations here) will form the Stern layer. Beyond that a second layer will form composed of both anions and cations, this layer is called the diffuse double layer.

The magnitude of the EOF increases as the pH increases, since the silica has a broad pKa, but plateaus at pH 8-9 [26]. The magnitude of EOF is also affected by the ionic strength of the BGE; as the charge density at the capillary wall increases the EOF increases. The zeta potential, ζ, can be used to relate the effects of charge density (σ) at the boundary between fixed and mobile ions (where the relative motion occurs), and the EOF

$$\zeta = \frac{\sigma d}{\epsilon_0 \epsilon}$$

Equation 4

here d is the distance from the capillary wall where the double layer ends, *i.e.* the double layer thickness, ϵ_0 is the permittivity of a vacuum and ϵ the dielectric constant of the solution. It is important to realize that σ is not the ionic strength of the solution and in fact σ is inversely proportional to the ionic strength. The zeta potential can be used to define the μ_{EOF}.

$$\mu_{EOF} = \frac{\epsilon_0 \epsilon \zeta}{\eta}$$ **Equation 5**

Equation 5 shows that the magnitude of the EOF is proportional to the ζ and thus σ, and inversely proportional to the viscosity of the solution. Now that EOF has been introduced, it can be accounted for in the calculation of the effective velocity for analytes in CE,

$$v_i = \mu_{app}E = (\mu_{EP} + \mu_{EOF})E$$ **Equation 6**

in this expression μ_{EP} is the electrophoretic mobility and μ_{EOF} is the mobility due to the EOF. To make an analogy with HPLC, the electroosmotic flow can be seen as the equivalent of an electric field-driven pump and replaces the pressure-driven flow that all species in the capillary experience equally; the electrophoretic mobility is the species specific property that allows the separation of different molecules whereas it is the partitioning mechanisms that govern the extent with which different molecules interact with, and are retained onto, the various stationary phases available in HPLC.

Barring more complex interactions between the BGE and other additives in the CE run buffer, the migration order in normal mode CE can be explained by the *q/V* ratio. At one end of the continuum species that have one or several positive charges and a small volume will migrate to the detector the first since both forces, the μ_{EP} and μ_{EOF}, are being applied in the same direction. These will be followed by the detection of species with smaller and smaller *q/V* ratios. Finally the other end of the continuum is reached when negatively charged species pass the detector last as the two forces (μ_{EP} and μ_{EOF}) are now working in opposite in directions. It is worth noting that if a negative molecule has $\mu_{EP} > \mu_{EOF}$, then it will never pass the detector. Luckily, μ_{EOF} is significantly larger than most μ_{EP}. Controlling the EOF becomes critical since it not only has an impact on the separation time, but also on whether or not a species of interest will ever make it to the detector. Unfortunately, many factors can influence the EOF and, as a result, variation in the EOF accounts for a large portion of the variation inherent to CE separations with respect to migration time as will be shown below.

1.1.1.5 Sources of Variation in CE

Generating quality CE data requires a good understanding of the major sources of variation associated with not only the technique, but the sample analyzed as well. Peak width variations and migrations time variations will be explored next since they can have the most important impact on the quality of the data. The experimental conditions can be optimized to minimize the effect on the data quality or to make it possible to use data preprocessing strategies to correct (see COW section) for the unwanted variations.

1.1.1.5.1 Peak width variations

It is important to identify and understand the major sources of variance relevant to CE and how these can be controlled and minimized. To achieve this goal the most appropriate parameter to consider is the Height Equivalent of a Theoretical Plate (HETP).

$$HETP = \frac{L}{N} = \frac{\sigma_{tot}^2}{L}$$ **Equation 7**

This equation conveniently allows for various components contributing to the total variance to be expressed as a series of additive error terms provided that all the dispersive phenomena affecting the total variance act independently from each other. The more important dispersive phenomena operating in CE are longitudinal diffusion, temperature effects (Joule heating), variation in sample introduction and analyte interaction with the capillary wall such as adsorption. All of these can be summed to account for the total variance:

$$\sigma_{tot}^2 = \sigma_{diff}^2 + \sigma_T^2 + \sigma_{int}^2 + \sigma_{wall}^2$$ **Equation 8**

Usually, the band broadening due to diffusion, temperature and sample introduction are insignificant compared to the contribution due to solute-wall interaction in the context of biological samples due to the propensity of proteins to interact with the capillary surface. Therefore, only the contribution of the last term will be discussed.

Analyte-wall interactions can have a significant impact on the observed peak shape and its width (variance). The specific interactions can be hard to predict and identify as it is not always possible to know every species present in a sample and also as adsorption arises from complex interactions between the multitudes of species in solution and the capillary wall [27, 28]. In general, the contribution to the total variance from wall interaction involving small or anionic molecules are not significant but when proteins are analyzed by CE the variance introduced by the analyte-wall interactions can be quite important and can easily rival or supersede the variance coming from all other sources of bandbroadening. It is possible to modify the surface of the capillary to minimize analyte interaction with the wall as described in an extensive literature [27, 28]. However, since in this study no wall coating was employed, only the interaction of proteins with uncoated capillary will be discussed.

Proteins are complex molecules composed of amino acid chains of varying size and conformation. The protein exposed backbone and pendant side chains, all have the potential for interaction with the capillary wall. The interaction can be governed by electrostatic interactions, van der Waals forces and hydrophobic effects depending on the species present in sample, the background electrolytes, additives, the pH of the solution and the temperature [25]. The pH is of particular interest as it governs the charge state of the protein and the wall. It can also affect the protein conformation which influences the hydrodynamic volume. It is convenient to define three relevant scenarios where: (i) pH = pI; (ii) pH < pI; (iii) pH > pI and we assume that the capillary wall still carries a negative charge. In the first scenario, the (non-electrostatic) adsorption of proteins to hydrophilic surfaces is most pronounced when the pH of the solution is close to the pI of the protein because the protein approaches a zero net charge thus minimizing protein/protein lateral repulsion on the surface. Adjusting the pH above or below the pI can, in some cases, reduce protein adsorption, but it generally does not completely prevent it. In the second scenario, electrostatic attraction between the surface and the protein does increase adsorption, but the lateral electrostatic repulsion between adsorbed proteins explains why this scenario shows a net reduction of adsorbed proteins compared to the first scenario. In the third scenario, silica and protein will both be negatively charged causing reduced protein adsorption due to protein/wall and adsorbed protein/protein electrostatic repulsion. However, it is still possible to have some adsorption as proteins can have localized positive charges that can interact with the surface.

The consequences of protein adsorption to the capillary wall in CE are migration time shifts, peak tailing and, in some cases, loss of protein due to permanent adhesion to the capillary wall, all of which can cause variation in the detected peak area.

1.1.1.5.2 Migration time variations

Migration time fluctuations, both between and within runs, result in misalignments and peak distortions when data from run-to-run is compared. Minimizing this type of variation is critical if sophisticated data analyses are going to be used. To use CE data effectively requires time aligned data. The most important source of migration time variation is EOF fluctuation. It was shown above that many parameters influence the EOF such as pH, ionic strength, electric field strength, viscosity, all of which are a function of temperature. Additionally, capillary surface and geometry, presence of surfactants or organic modifiers can all affect the EOF. It becomes clear why controlling the EOF is not an easy task since many of the parameters that influences the EOF are interrelated. Equation 6 shows that the effective velocity of analytes in CE is due to both the EOF and the specific electrophoretic mobilities of each analyte. Contributors to migration time variations will be presented in order of importance from lowest to highest.

The very first equation introduced dealt with was the electrophoretic mobility. Equation 2 can be deceivingly simple, but in practice many parameters can influence the viscosity of the solution used for the separation as well as the charge and volume of analytes, especially for proteins. In Equation 9 below, the factors that modulate each of the terms are introduced into Equation 2. Thus, temperature (T), solvent (S) used, dissolved analyte (DA; concentration, type (molecular weight and composition)), background electrolyte (BGE) concentrations can all affect the viscosity of the solution. The charge on a protein can be influenced by pH because of the many ionisable sites on proteins with different pKa's. Oxidative modifications (Ox) can also alter the charge on the protein by adding new, or modifying existing, ionisable sites on a protein. Temperature is also a factor, as the ionization constants are a function of temperature. Additionally, under certain conditions, proteins can unfold and become denatured (D), thus exposing new ionisable sites which alters the pI. The extent to which a protein will be denatured depends on temperature, solvents present, pH, ionic strength, presence of additives and separation time.

Similarly, the analyte hydrodynamic volume can be influenced by all the factors influencing protein denaturation that were mentioned above, which further changes the protein's interactions with the solvent to name a few. A more realistic Equation 2 should begin to look like,

$$\mu_{EP} = \frac{q(T, Ox, D)}{6\pi * \eta(T, S, DA, BGE) * r(T, D, Ox, S)}$$ **Equation 9**

Equation 5 shows the parameters influencing the EOF. The viscosity is influenced by the same parameters as μ_{EP}. Equation 4 and Equation 5 related the ionic strength (IS) to the

EOF. As ionic strength increases, the interaction between the capillary wall and the cations weaken, in turn diminishing the charge density and with it ζ. Changes in pH can modulate the charge density adjacent to the capillary wall by increasing or decreasing the charge density of the capillary wall itself, *i.e.* deprotonate or protonate the silanols. Additives, modifiers and type of ions used can all have an important effect on ζ. Additionally, the pKa of silanol groups decrease with increased temperature, thus increasing the zeta potential. For protein containing samples the major concern is with protein adsorption (PA) to the capillary wall as discussed above. This wall coating can mask the capillary wall, introduce new hydrophilic or hydrophobic properties to the wall that can cause analyte retention, but also introduce new and different charge density to the wall. This heterogeneity of the capillary wall causes migration time shifts and band broadening. When all these factors come together, Equation 5 becomes

$$\mu_{EOF} = \frac{\epsilon_0 \epsilon * \zeta(T, IS, pH, pKa, PA)}{\eta(T, S, DA, BGE)} \qquad \textbf{Equation 10}$$

Substituting Equation 9 and Equation 10 into Equation 6 gives an overall portrait of the apparent analyte velocity,

$$v_i = \left(\frac{q(T, Ox, D)}{6\pi * \eta(T, S, DA, BGE) * \ r(T, D, Ox, S)} + \frac{\epsilon_0 \epsilon * \zeta(T, IS, pH, pKa, PA)}{\eta(T, S, DA, BGE)}\right) E \qquad \textbf{Equation 11}$$

Considerable efforts should be put in minimizing variation in both peak area and migration time to maximize the quality of the data acquired. Yet, despite the analyst's best precautions, there will still be some migration time variations from sample-to-sample and from run-to-run. Fortunately, there are mathematical ways to correct for misaligned electropherograms and these will be discussed in the chemometrics part of the

Introduction below. The success of time alignment strategies rely largely on making sure that experimental factors contributing to migration time variations are controlled leaving only small migration time corrections to be done. Generally, this assumption can be met with careful experimental design and execution.

1.2 Data analysis strategies (Chemometrics)

The use of statistics is widespread, from very sophisticated applications in modern physics, economics and epidemiology to trivial applications in sales, such as product placement in local grocery stores. It is hard to imagine a day without some reference to results of some sort of statistical analysis: census data, risk factors for such and such disease, latest polls for an upcoming election. The fact is, data is all around us and accumulating at an overwhelming rate. So much so that the cover of "The Economist" in February 2010 was entitled "The data deluge" [29]. They reported that the amount of data being acquired, transmitted, analyzed is astronomical. In 2005, 150 exabytes of data were generated whereas in 2010, 1 200 exabytes of generated data was projected at the time of publication (February 2010). Finding ways of storing all this data is challenge enough, let alone finding the resources to analyze the produce data.

Analytical chemistry is in no way spared from this trend. The best example of a demanding technique, in terms of data generation, is mass spectrometry. Modern instrumentation can output hundreds of gigabytes of data per day. The need for efficient and accurate methods to sift through raw analytical data and find the relevant information is quite real. In chemistry, the term chemometrics was selected to describe the statistical models and methods inspired from statistics and adapted for this specific field of Science. Chemometrics focuses on three main themes: experimental design, calibration and multivariate analysis. The important concept needed to understand the data processing and analysis tools employed in this thesis will be introduced in the following section with a bias toward CE.

1.2.1 Preprocessing strategies

Data preprocessing can be defined as a means to prepare, correct and/or transform data in such a way that artifacts are removed rationally and parsimoniously to enable efficient and accurate chemometric data modeling. The term artifact here refers to undesired alteration in the data, such as baseline offsets and migration time shifts. In the context of this thesis, three basic preprocessing steps are required: baseline correction, migration time correction and data reduction. Baseline corrections are a routine step in most analytical data treatment whether discussing chromatographic, electrophoretic, spectroscopic or spectrometric data. The different ways to correct for this type of artifact will be considered first. Migration time shifts are also commonly observed in chromatography and capillary electrophoresis, yet mathematical means of correcting this type of artifact are not as well known, so a more in-depth introduction will be presented. Finally, the preprocessing section will conclude with an explanation of the data pre-processing strategies that allow reduction in data complexity.

1.2.1.1 Data alignment prior to multivariate analysis

A pre-requisite of multivariate data modeling is that each element of a data array (in this case the signal from a specific species measured at a particular migration time or index) have corresponding elements at matching index position (here migration times) across respective data arrays from all the samples being analyzed. Many of the chemometric tools for data analysis have been developed for spectroscopic data where it is often safe to assume that all of the data is aligned with respect to wavelength because of the high wavelength reproducibility of spectrometer. For data obtained by CE, that assumption cannot be made in the time axis. In the CE section above, the many parameters that can contribute to variation in the migration time such as pH, ionic strength, electric field strength, sample viscosity, temperature, *etc*. were described. Even after careful experimental control of these parameters, there are still some inevitable variations that will result in migration time shifts. Figure 5 highlights the problem of comparing poorly aligned electropherograms. The same analyte present in both samples and thus in both electropherograms is not giving rise to a signal at the same index value (or migration time).

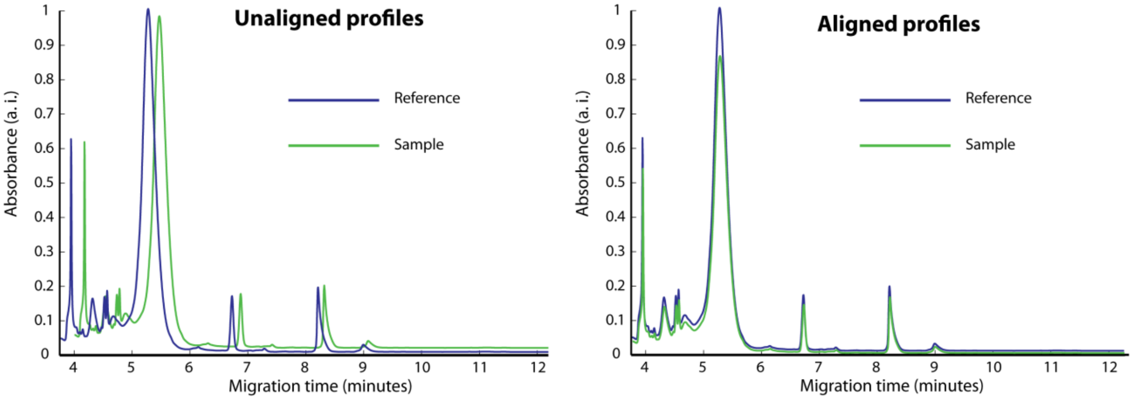


Figure 5: Misaligned and aligned CE profiles. On the left, the green and blue traces are offset from one another causing the signal from the same analyte to show up a different index values. The unaligned samples cannot be compared to each other based on index values. On the right, the blue and green signals have been aligned thus making it possible to compare the signals of each trace at given an index value.

When comparing integrated peak areas in a typical "manual" data analysis there is a sort of "hidden time alignment" that is done, *viz.* the operator makes a qualitative judgement about peak location, integrates the peak and assigns the value to a table corresponding to a particular species independent of the actual migration time. In these situations, where a handful of individual samples are analyzed one at a time, the whole indexing problem is trivial. But when multiple data arrays are combined into higher order arrays for multivariate data analysis, the misalignment problem is far from trivial and often precludes the use of multivariate analyses on the entire data matrix.

1.2.1.1.1 Notation

Before time alignment routines are explained, the notation and the definition of terms used must be clearly stated. Since the data requiring preprocessing is obtained from CE, the electropherogram will be used instead of the more general term measurement vector. As such, *migration time* will be used to refer to the direction along which analyte migrate and upon which alignment is carried out. Similarly, each data point along that axis will have an index value that will be referred to as the *time index.* Scalars are depicted as lowercase italicized letters (*e.g.* x); row vectors and electropherograms are represented as bolded lowercase letters (*e.g.* $\boldsymbol{x}$); bold capital letters are used for data matrices (*e.g.* $\boldsymbol{X}$). To express an individual element of $\boldsymbol{X}$, the following notation will be used, $x(i,j)$, while a range of values for the indices are as such: $i = 1, \dots, I$ and $j = 1, \dots, J$.

Correlation optimized warping involves the alignment of the time axis between a sample electropherogram and reference electropherogram. A sample electropherogram is written as such $x_j = x(t_j)$ and the target/reference is $y_j = y(t_j)$ where $t_j = j$, j being the time index in data points.

Any time alignment routine can be understood as a process that requires the transformation of the time axis of a sample electropherogram in a way that best matches a reference electropherogram. This time axis transformation can be expressed with a warping function $w(t_j)$. Since not all points of sample electropherogram (x_j) are present in the new index obtained from the warping function, the scalar value in the new index can be interpolated from x_j to produce the aligned electropherogram $x(w(t_j))$ in a way that maximizes the similarity between the sample and the reference (y_j). Equation 12 expresses the problem

$$y_j \cong x(w(t_j)) \quad j = 1, \dots, J$$ **Equation 12**

Certainly the simplest way to align electrophoretic data is by making use of an internal standard to normalize migration time. In this case, the analyst would have included some standard that does not interfere with the analytes of interest and also migrates as a clearly resolved peak. The migration time of the standard can be used to rescale the time axis into a migration time ratio, t_R [30]

$$t_R = \frac{t_j}{t_{IS}}$$ **Equation 13**

here t_j is the migration time at a specific index value, t_{IS} is the migration time of the internal standard. Here, instead of using a reference index, all profiles would be projected on a common time axis,

$$y(w(t_j)) \cong x(w(t_j))$$ **Equation 14**

$$y_j\left(\frac{t_j}{t_{IS}}\right) \cong x\left(\frac{t_j}{t_{IS}}\right)$$ **Equation 15**

$$y_j(t_R) \cong x(t_R)$$ **Equation 16**

When used as a warping function, t_R provides a linear correction to generate a warped sample electropherogram. The electropherograms should show a considerable improvement with regards to time alignment when this ratio is used. This type of approach has several advantages. It can account for run-to-run variations in the EOF as long as the EOF remains constant throughout each individual run. It is also simple to implement and understand, but it fails to correct fluctuations of the EOF that occur within a run or any other non-linear source of migration time variation which often occur in analyses of complex samples.

The literature provides a few different solutions for more complex time alignment problems. The algorithms used to time align can be compared and contrasted based on three characteristics: definition of the warping path/function (parametric or non-parametric); the metric used to determine optimum alignment (*e.g.* Euclidean distance [31], correlation coefficient [31, 32], sum of squares of the difference between sample and reference [33]; and optimization algorithm to select optimized warping path. The main time alignment techniques are presented in Table 1, but given the space constraints only Correlation Optimized Warping will be discussed.

Table 1: Selected time alignment techniques successfully applied to chromatography and/or capillary electrophoresis.

Technique name (acronym)	Alignment metric	User input
Correlation Optimized Warping (COW)	Pearson's Correlation Coefficient	Reference profile, segment length, slack size
Dynamic Time Warping (DTW)	Euclidean distance	Reference profile, local continuity constraints, weighing function
Parametric Time Warping (PTW)	Sum of squares of the residuals	Reference profile, Warping function coefficient
Semi-parametric Time Warping (SPW)	Sum of squares of the residuals	Reference profile, Warping function coefficient, number of B-splines, penalty term

All of these approaches have been shown to work with separations data; COW will be explained in more detail since it is the time alignment routine providing the best quality of alignment with a reasonable amount of computational time. Semi-parametric time warping has been shown to be an excellent competitor, especially in terms of alignment speeds, but it is much more complex and does not provide significant advantages over COW [31, 33, 34].

1.2.1.1.2 Correlation Optimized Warping

Nielsen *et al.* [32] first proposed COW as a solution to correct for retention time fluctuations observed in chromatographic data. Correlation Optimized Warping is conceptually simple; the details of the process are presented in the following paragraphs. Time alignment is achieved by breaking down the sample data vector (chromatographic or electrophoretic profile) into segments. The algorithm then produces a new set of segments by linearly stretching, compressing or keeping the original segment. The best combination of these segments is then selected so as to maximize the overall correlation coefficient to the target reference profile. COW requires that the user selects the window size (segment length) and the slack (flexibility) parameters before it searches for the optimal warping path.

To illustrate how COW functions, one potential warping path will be described. Consider the simple synthetic separations data in Figure 6, the first peaks in both profiles are simply offset from one another, while the last peaks are offset but also of slightly different width.

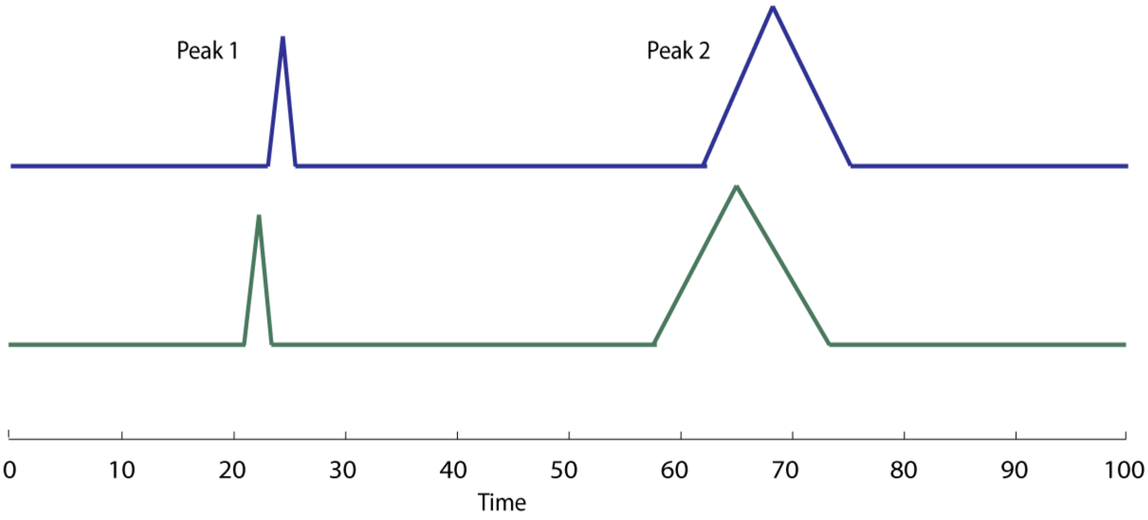


Figure 6: Example of misaligned profiles. On top (blue trace) is the reference profile to which the bottom trace (in green) is to be aligned to.

The COW algorithm divides the profile in several segments (S1, S2, S3...) and only allows the segment to be stretched/contracted by a maximum amount of “slack”. In this case the segment length is set to 20 and the slack parameter is set to 2. In the example, the top profile is the reference and the bottom is the sample profile.

The first peak in the sample profile (Peak 1, Figure 6) is offset by -2 time units. Correction of this misalignment can be obtained by stretching/contracting the baseline

(SP_{S1}) ahead of the peak of interest. This leaves the peak unaltered, but "slides" it along the baseline. In order to obtain an offset of 2 points to the right, the first 18 point segment in the sample profile (SP_{S1}) must be stretched by 2 points. This is accomplished by setting the boundaries of this segment to 1 and 18 (i.e. using a -2 slack). This 18 point segment is then stretched to match the segment size of the reference, *i.e.* 20. The second segment in the unaligned sample profile starts at data point 19 and ends at 38. It does not need to be stretched or contracted since the optimum solution is to offset this segment. When this segment is positioned after the first warped segment, its index value has now changed from 19 to 21 bringing Peak 1into alignment with the reference.

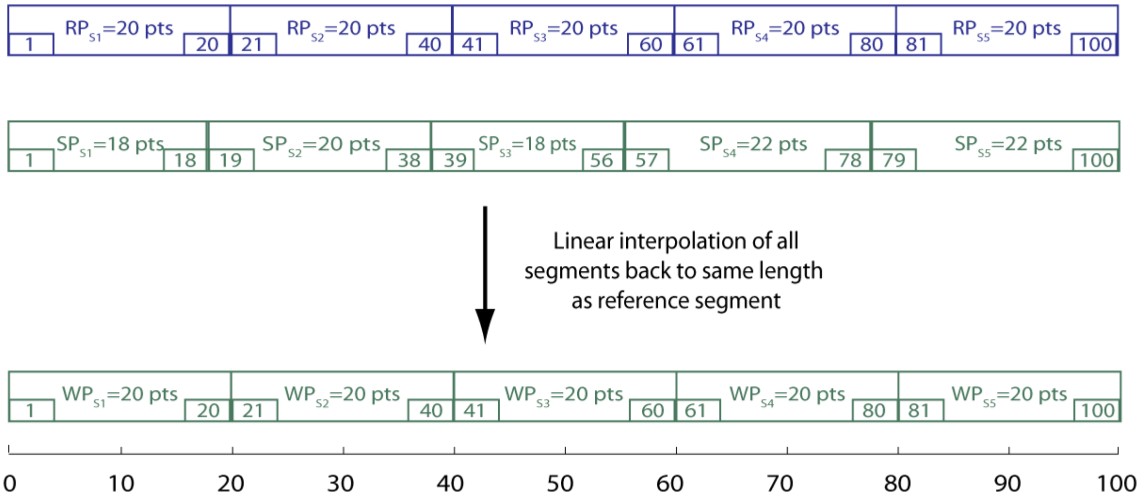


Figure 7: Representation of how COW aligns a profile to a reference by stretching and contracting segments of the entire electropherogram shown in Figure 6. Here the electropherograms are composed of 100 data points, the window size is 20, the slack parameter is 2.

The misalignment of Peak 2 is closer to an offset of 3 data points. Additionally, in this example, Peak 2 in the sample profile is slightly wider compared to Peak 2 in the reference. This makes the alignment of Peak 2 a bit more complex. Conceptually, the *overall* process can be broken into the effects of the changes that occur in each segment. When Segment 1 is stretched, it slides Peak 2 to the right (by 2 points). When Segment 3 is stretched, it increases the translation to the right (an additional 2 points). When Segment 4 is contracted, it narrows the peak, improving the correlation to the reference peak, and partially shifts the peak to the left. Finally, when the last segment is contracted the new *warped* profile is now the correct overall length.

The example above depicts one possible warping path that allows for an improvement in the overall alignment of a profile with respect to a reference. In reality, with these parameters, COW would evaluate a total of 381 possible warping paths and select the optimal warping path on the basis of maximizing the correlation of the warped path to the reference path. For each segment, the correlation coefficient (CC) is calculated as shown in Equation 17 for the second segment from Figure 7 as an example,

$$\mathbf{CC} = \frac{\mathbf{cov}\left(\boxed{\boxed{21}\ RP_{S2}{=}20\ pts\ \boxed{40}}\ ,\ \boxed{\boxed{21}\ WP_{S2}{=}20\ pts\ \boxed{40}}\right)}{\sqrt{\mathbf{var}\left(\boxed{\boxed{21}\ RP_{S2}{=}20\ pts\ \boxed{40}}\right) \mathbf{x}\ \mathbf{var}\left(\boxed{\boxed{21}\ WP_{S2}{=}20\ pts\ \boxed{40}}\right)}} \qquad \textbf{Equation 17}$$

The correlation coefficient is a good measure of the alignment between two profiles as its value is maximized when the traces have both the same shape and are aligned in time. In COW, a correlation coefficient is calculated for each warped sample/reference segment pair. The warping path that has the greatest sum of correlation coefficients is selected as the best warping path.

Dynamic programming is utilized to find this optimal warping path. A detailed description of this optimization routine is outside the scope of this thesis. More information about optimization using dynamic programming can be found in the literature [31, 32]. Briefly, dynamic programming finds the optimal warping path by calculating every possible combination of allowed warped segments based on the two parameters: segment size and slack. After COW alignment, the entire electropherogram can be further preprocessed or directly compared to each other with chemometrics data analysis methods to search for patterns in the data set.

1.2.1.2 Data reduction (i.e. Haar wavelet transform)

After time alignment, the data can be investigated to find the best measured variables that can estimate the dependent variables (in this thesis, birth outcomes/reproductive potential). It is possible to use the signals from single migration times as variables, but using integrated migration time windows is more appropriate since each species migrates as a peak over the course of several seconds. Using a window provides a general improvement in the signal-to-noise ratio, this has been shown in the spectroscopy literature [35]. Furthermore, the number of measured variables to sample number ratio is greatly diminished, reducing the complexity of the computational task and risk of overfitting the data.

Wavelet transforms can be used as preprocessing techniques to reduce the number of variables incorporated in the analysis by simplifying the data. This data reduction also often improves the model generated and reduces the time required to generate the model. The simplest wavelet, the Haar wavelet, is frequently employed to avoid the troublesome, and sometimes, arbitrary selection of an appropriate wavelet family for a particular problem. Moreover, this wavelet transform offers the added benefit of simple implementation.

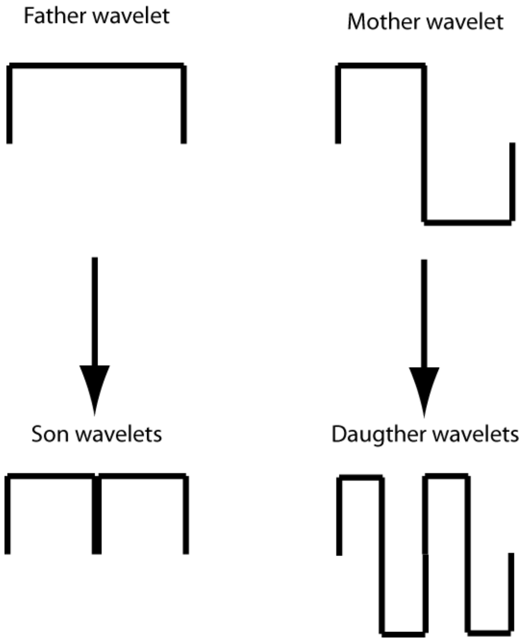


Figure 8: Examples of the Haar Wavelets. On the top left is the father wavelet, below are the son wavelets which are scaled down and offset versions of the father wavelet. On the right is the mother wavelet and similarly below the daughter wavelets.

Wavelet transforms are similar to Fourier transforms in that they transform the data and project it onto a given basis set. An important difference is that wavelet transforms use wavelet functions as opposed to the sine and cosine functions that the Fourier transform employs. The Haar wavelet is a simple mathematical construct and choosing a discrete wavelet transform (DWT) over a continuous one simplifies the implementation even further. The DWT can be defined for data defined over a range 0 ≤ x < 1 by:

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{cases} \mathbf{1}\ \boldsymbol{if}\ \mathbf{0} \leq \boldsymbol{x} < 1 \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

Equation 18

$$\boldsymbol{\psi}(\boldsymbol{x}) = \begin{cases} \mathbf{1}\ \boldsymbol{if}\ \mathbf{0} \leq \boldsymbol{x} < \frac{\mathbf{1}}{\mathbf{2}} \\ -\mathbf{1}\ \boldsymbol{if} \frac{\mathbf{1}}{\mathbf{2}} \leq \boldsymbol{x} < 1 \\ \mathbf{0}\ \boldsymbol{otherwise} \end{cases}$$

Equation 19

$$\boldsymbol{\psi}_{n,k}(x) = \boldsymbol{\psi}(2^n x - k), \text{with } n \text{ and } k \in \overline{\overline{N}}, 0 \leq k < 2^n - 1 \qquad \text{Equation 20}$$

where the subscript $\boldsymbol{n}$ and $\boldsymbol{k}$ represent the scaling and the translation respectively. The essence of the Haar transform is decomposing data into a weighted sum of $\boldsymbol{\psi}_{n,k}$, $\boldsymbol{\psi}$ and $\boldsymbol{\phi}$. The weightings are referred to as the "wavelet coefficients". For the father wavelet $\boldsymbol{\phi}$, the coefficient is obtained by integrating the signal over the entire span of the data. Similarly, for the mother wavelet $\boldsymbol{\psi}$, the first half of the data is integrated and subtracted from the integrated value of the second half of the data span. Daughter and son wavelets are simply scaled down versions of the parent wavelet with offsets. It becomes apparent that the son wavelets behave as a low-pass filter and contain the approximated data while the daughter wavelets behave has high-pass filters and contain the fine details of the data. Figure 9 shows algorithmically how the HT can be implemented. The initial data, $\boldsymbol{a_0}$**,** is divided in into two blocks of coefficients. On the left block, $\boldsymbol{a_1}$**,** contains the coefficient for the son wavelets which constitute the first level approximation of the data, whereas the right block, $\boldsymbol{d_1}$**,** contains the coefficient for the daughter wavelets. The process is then repeated with the $\boldsymbol{a_1}$ block which is further divided into $\boldsymbol{a_2}$ and $\boldsymbol{d_2}$ until the maximal number of dilations/separations $\boldsymbol{n}$ is reached (from Equation 20**).**

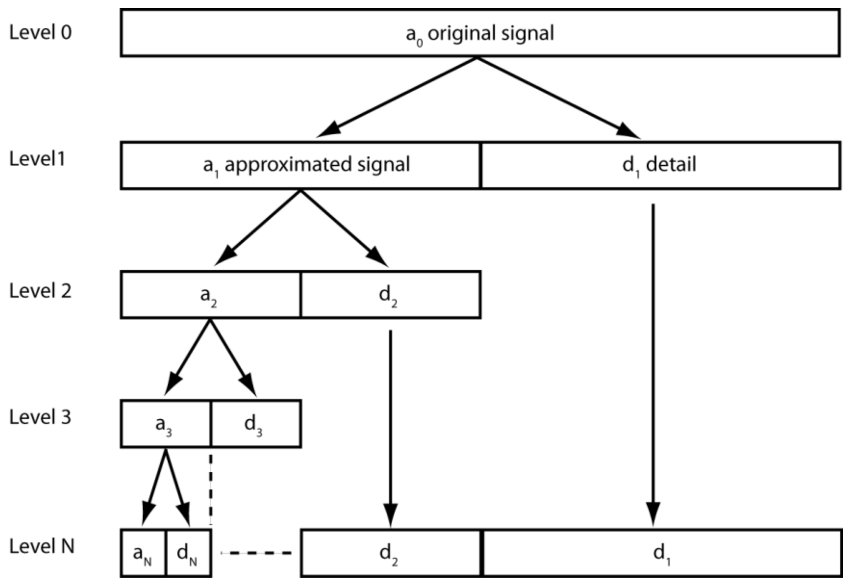


Figure 9: Overview of the discrete wavelet transform by use of the pyramid algorithm [36]. On each level the signal is decomposed into the low- and high-frequency components. On each level, the high frequency component contains the detail and the low frequency component contains the approximated signal. Each level decomposes the approximate signal further into low- and high-frequency components.

It is possible to simplify even further this process by using scaled down and shifted versions of the father wavelet, the previously mentioned son wavelets:

$$\boldsymbol{\phi_{n,k}(x)} = \boldsymbol{\phi(2^n x - k)}, \textbf{with } \boldsymbol{n} \textbf{ and } \boldsymbol{k} \in \overline{\overline{N}}, \boldsymbol{0 \leq k < 2^n - 1}$$

Equation 21

In this case, the HT would decompose a dataset with $\boldsymbol{z}$ data points into weighted sums of a father wavelet and $\boldsymbol{2z-2}$ son wavelets. If a dataset contained 4 data points for example, the resulting HT would be:

$\boldsymbol{[x_1 + x_2 + x_3 + x_4;\ x_1 + x_2;\ x_3 + x_4;\ x_1;\ x_2\ ; x_3\ ; x_4].}$ This is particularly well suited for electrophoretic data since it corresponds to sets of integration windows ranging from the entire profile to individual data points.

1.2.2 Multivariate data analysis (modeling)

1.2.2.1 Bayesian decision theory Classification Strategies

Bayesian decision theory is often used for pattern recognition. Many other areas such as economics, social sciences and artificial intelligence have long been using Bayesian statistics. Only recently have chemical and biological applications been sought for such models [37-41]. It is possible to build a model to properly classify into groups of given data sets. Here, the different Haar wavelets represent features of the system. The best classification solution will be that which finds the most distinct characteristics of each group by selecting the informative Haar wavelets. Similarly to the univariate normal distribution, the multivariate distribution will use the average measurement of a feature and the variance of that measurement as parameters in a Gaussian function to calculate the probability that a sample is part of a group defined by multivariate Gaussian distributions. The resulting probability of belonging to one group or another is used to designate the labelling of a sample into a group. Assignment is based on the highest probability of belonging to a group as opposed to the other. If a number of Haar wavelets are retained as being predictive for the classification, we only need to consider the features of these given wavelets to adequately classify each sample as belonging to their respective classes ultimately giving a parsimonious model.

In this study, Normal distribution is assumed. This means that values for optimal variables (wavelets) will be used to generate the parameters describing Gaussian distributions for each group used in the classification. This will be shown with a univariate example first. To obtain proper classification between two states, *e.g.* normal and abnormal, the variable selected should show a good clustering of values for this variable within a group while showing a good separation of these same values between two groups as depicted in Figure 10.

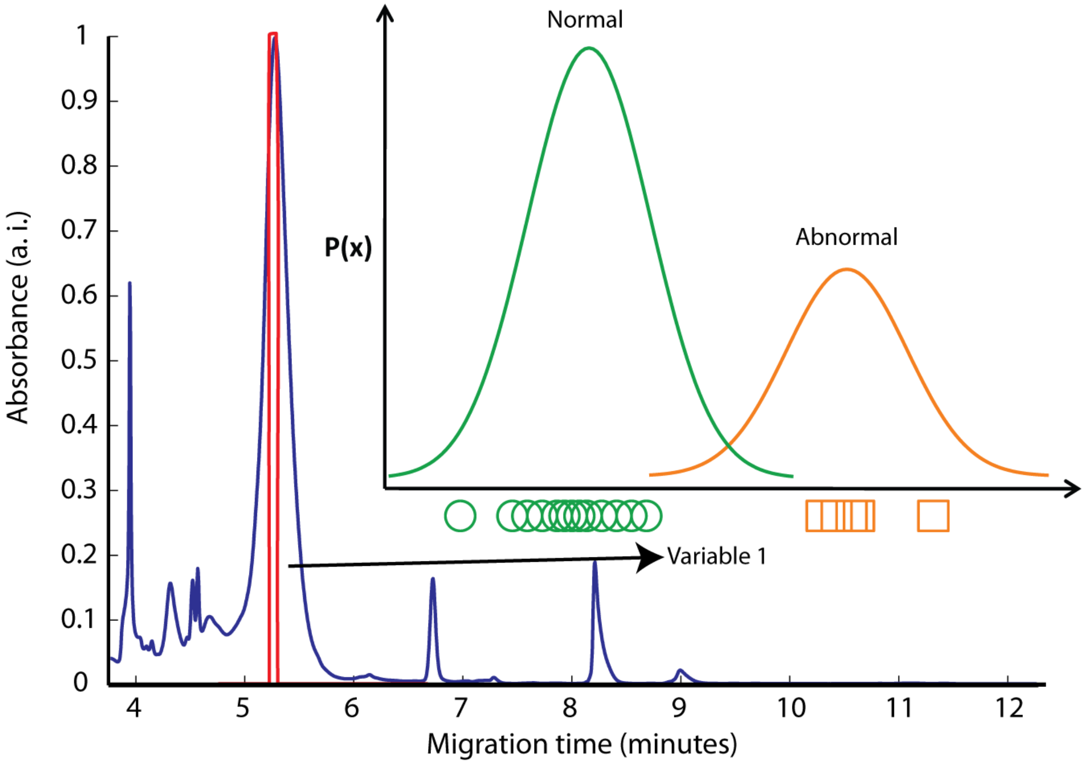


Figure 10: Example of Bayesian decision theory used to classify into two groups with a univariate Gaussian model. A single wavelet (red rectangle) is used to generate two probability distributions, one for each group. The green circles represent Variable 1 values for the normal group and the orange squares represent the same Variable's values for the abnormal group.

It is possible to calculate the means (μ) and standard deviations (σ) for each group with a set of calibration data according to the values of x. Any new analyzed sample can then be compared to those two groups using the Gaussian equation:

$$P_x(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Equation 22

The probability density of a sample being in one group, or the other, can be calculated and the sample classified on the basis of which group has the highest probability. So far univariate models were discussed, yet it is possible to include two or more variables to generate multivariate Gaussian models. In this case a more general form of the Gaussian equation may be used, but the process is essentially the same. Figure 11 shows an example of bivariate distributions.

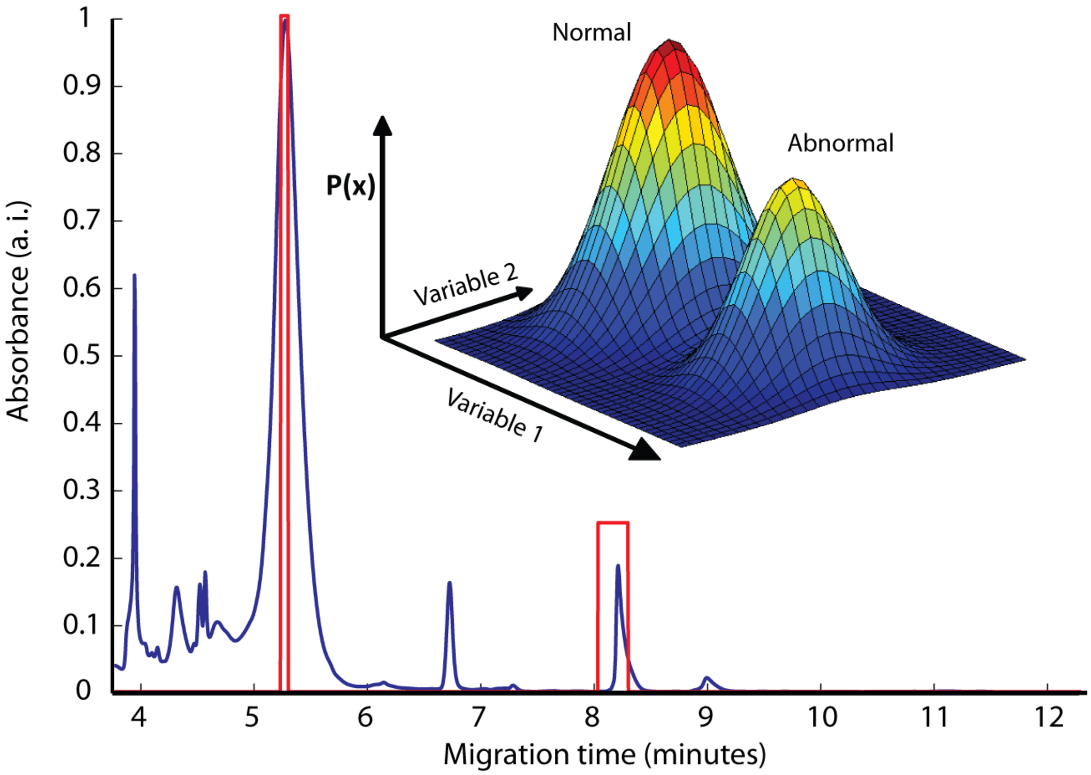


Figure 11: Example of Bayesian decision theory used to classify into two groups with a bivariate Gaussian model. Two wavelets (red rectangles) are used to generate two probability distributions, one for each group.

In a process just like for the univariate models, it is possible to calculate the means ($\boldsymbol{\mu}$), but in this case, a covariance matrix ($\boldsymbol{\Sigma}$) is required. Here again, any new analyzed sample can then be compared to the two groups using the multivariate Gaussian equation:

$$P_x(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$ **Equation 23**

where d is the dimension of the data and $|\Sigma|$ represents the determinant of the covariance matrix. The previously calculated parameters are used to determine the probability of being in one group or the other and here again a sample will be classified into the group for which it has the highest probability density.

1.2.3 Optimization

Traditionally, selecting the best variable(s) to model or monitor processes of any type is done with *a priori* knowledge: a specific wavelength can be selected because it is known to reflect a reaction of interest; the peak area with a specific elution time corresponds to a molecule involved in an biochemical pathway of interest; and the list can go on and on. The obvious benefits of this route are that it is simple to understand and manage, and that the selection of variables of interest resides on sound scientific premises. Biological systems are particularly complex, the processes of interest are dynamic and most often involve multifactorial (multianalyte) contributions to a specific outcome. The consequence of making decisions based solely on the current state of scientific knowledge of a system is that variables of unknown relevance are not included in the modeling, or in the data acquisition step, even though they might be playing a significant role. Unfortunately, for practical reasons, only subsets of variables are considered since including all measured variables can be extremely time consuming when the number of components present in a system is overwhelmingly large. Nevertheless, the significant increases in desktop computer processing power makes it feasible, in some situations, to investigate all possible permutations and combinations of measured variables.

1.2.3.1 Using all permutations and combinations of variables

Evaluating all possible permutations and combinations of variables for a particular system is not particularly hard or complicated; it is, however, very time consuming. The main issue is whether achieving this can be done considering the memory constraints of the computer used and if searching through all the possible scenarios can be completed in a reasonable amount of time.

The COW algorithm, which was introduced above, employs an optimization method that evaluates all possible allowed warping paths using dynamic programming. It can find a solution relatively quickly, in roughly 20 seconds, when working with electropherograms of 4800 points (window size of 20; slack of 2).

When the problem to be optimized involves a large number of samples, above about 100, and a much larger number of variables, above 1000, this type of brute force method is simply no longer feasible. The memory requirements surpass the available memory of the computer and/or solving the problem would take months. Instead, optimization methods that find optimal solution(s) by searching only part of the whole variable space have to be employed.

1.2.3.2 *Genetic algorithm*

Genetic algorithms [42, 43] (GA) are stochastic optimization methods that do not attempt to predict (like derivative based or simplex methods) the direction of the optimum in the variable space during the model creation process. Conceptually, this approach mimics natural selection and evolutionary processes. GA have shown themselves to be excellent choices for parameter optimization in complex data. One of the main advantages is that they avoid being trapped in local minima by randomly generating initial parameters, unlike predictive approaches.

The first step of a GA optimization is the encoding of each variable's index (*i.e.* Haar wavelet number, which in turn represents a specific time window in the electropherogram) into a binary bit string known as the chromosome. A variable, or combinations of variables, is randomly selected and evaluated for its fitness, *i.e.* evaluated by the objective function selected for the optimization. The best results are kept to form a population (or gene pool) and are subjected to genetic operators in the hopes improving the current solution. Crossover and mutation are the commonly used operators to generate the next population. Crossover is similar to mating where the best members of a population can exchange genetic material, *i.e.* exchange good variables to potentially form a better combination of variables. Mutations are rather simple to implement in GA because of the binary encoding. Changing the value of a single bit changes the index value and thus points to a new variable. Genetic operations create a new generation that undergoes the same process to find the best combination(s) of variable. This is repeated until a predefined stopping criterion is met.

1.2.4 Model validation

The risk of over-fitting a data set and finding chance correlations between the variables and the outcomes is relatively high when the number of samples is low and the number of variables is high. Even if the number of variables is reduced to a manageable size with data rank reduction strategies, such as using Haar transform model, validation is important. The ideal situation, when the number of samples analyzed is sufficiently high that the model may be tested with a full validation. This involves generating the model with 2/3 to 3/4 of the sample data set. The remainder of the sample data set is used to evaluate the quality of the model. If the model is good, it will be able to predict adequately the dependent variable of interest using the validation data.

In some situations, it is not possible to do a full validation because the number of samples available to generate the model is too small. In such cases, a leave-one-out cross-validation may be employed to generate the model. In this case, all but one sample is used to generate the model and the remaining sample is used for validation. This process is repeated iteratively through all of the dataset until each sample as undergone the leave one out once. The best model is the one that will have adequately predicted the greatest number of left out samples.

1.3 The sample

One of the great challenges for the analyst is the analysis of biological samples. Intimate knowledge of the sample is important and often dictates which analytical strategies can be employed. Simple samples, with few components or interfering species, might not need very sophisticated analytical approaches. Unfortunately, biological samples seldom fall into this category because of the great concentration dynamic range and bewildering diversity of endogenous and exogenous biochemical species present. For example, genetic material (DNA and RNA) may be present from only a few copies per sample while proteinaceous species (proteins, enzymes and peptides) span concentration ranges from milli- through to zeptomolar. Small molecules (sugars, vitamins, hormones, fatty acids, *etc.*) and ionic species (metals and salts) also cover a broad range of concentrations. These compounds also have extremely diverse physico-chemical properties ranging from ionic, polar to hydrophobic with very different spectrophotometric and mass-spectrometric activities. This makes it virtually impossible to devise a single analytical method that can simultaneously detect every component in a sample. A further complication is that biological samples are often only available in small amounts (*e.g.* cerebrospinal fluid). Thus maximizing the number of species (and information) detected per sample is highly desirable. The majority of this thesis will focus on amniotic fluid as the sample, so the following section presents a review of the limited amniotic fluid biochemical literature. Where possible, literature from other biological fluids, mostly serum and urine, which are believed to be similar to amniotic fluid, will be used to help fill some gaps in the current state of knowledge concerning amniotic fluid.

1.3.1 Amniotic fluid

Amniotic fluid (AF), the fluid that surrounds the fetus during gestation, is a complex and dynamic environment that changes as the pregnancy progresses. AF serves three main functions: nourishment, protection and waste repository. The nutrients and growth factors present in AF facilitate fetal growth. AF provides mechanical cushioning against impact and contains antimicrobial effectors that protect the fetus from infection. Diffusion across fetal membranes and fetal urination largely accounts for the fetal contribution to amniotic fluid, but this varies during gestation. To avoid going into excessive detail, the review of the origins and composition of AF will be limited to the first half of gestation since all of the samples discussed in this thesis are from routine amniocentesis, which is typically carried out around the 15th week of gestation.

1.3.1.1 The origins and development of AF during gestation

Initially, the water component of AF originates from maternal plasma. Hydrostatic and osmotic forces drive diffusion into the amniotic fluid. As the placenta and fetus develop, water and solutes enter the AF primarily via the placenta. Skin keratinisation, beginning at week 19 and completed at week 25, effectively halts water and solute diffusion through the fetal skin. However, the surfaces of the amnion, placenta and umbilical cord remain permeable to water and its solutes. The permeability of the various membranes involved introduces a molecular weight bias towards small molecules compared to serum.

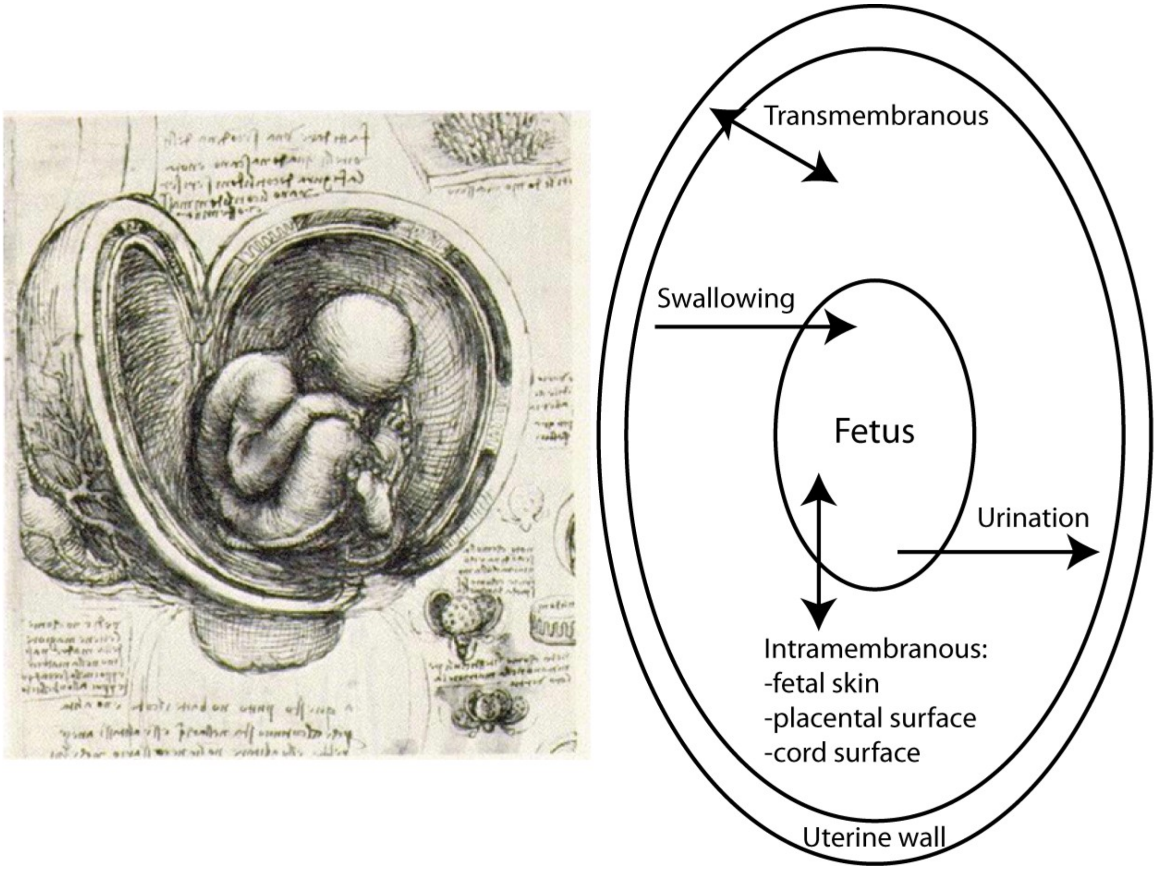


Figure 12: DaVinci's depiction of the womb on the left and amniotic fluid exchange pathways on the right.

Around the 8th week of gestation, the kidneys become functional and produce urine, which also coincides with fetal swallowing. At this early stage of gestation, neither swallowing nor urination affect significantly the content of AF, but as gestation draws to a close, they have a significant effect.

1.3.1.2 Amniotic fluid content

Amniotic fluid has not attracted significant biochemical attention in the past. This lack of interest has left amniotic fluid a poorly characterized and misunderstood biological fluid. This has changed in the last 10 years as published studies have begun to reveal the complex and dynamic nature of amniotic fluid. The previous section has hinted that early to mid-gestation amniotic fluid is not simply urine. In fact, around the 15th week of gestation, fetal urine contributes minimally to the amniotic fluid. Cho *et al.* [44] showed that the amniotic fluid proteome is very similar to the plasma proteome when considering the most abundant proteins (see Table 2).

Table 2: Fifteen highest abundance proteins in amniotic fluid compared to the fifteen most abundant proteins in the plasma proteome [44]. In bold are proteins that are found in high abundance in only one of the two fluids

Amniotic Fluid Proteome	Plasma proteome
Albumin	Albumin
Immunoglobulins	Immunoglobulins
Fribronectin	Serotransferrin
Serotransferrin	Fibrinogen
Complement C3	α_1-microglobulin
α_1-antitrypsin	α_1-antitrypsin
Ceruloplasmin	Complement C3
α-fetoprotein	**Haptoglobulin**
Vitamin D-binding protein	Apolipoprotein A-I
Periostin	**Apolipoprotein B**
Apolipoprotein A-I	α_1-acid glycoprotein
Antithrombin III	Lipoprotein
Transforming growth factor β[i]	Factor H
α_1-microglobulin	Ceruloplasmin
plasminogen	Complement C4

[1] Transforming growth factor β-induced protein ig-h3 precursor

The same study reports the putative identification of 842 distinct proteins in amniotic fluid spanning from low to high abundances using a multidimensional HPLC MS/MS analysis of amniotic fluid protein digests. Several groups [45-49] have reported a similar diversity in the amniotic fluid proteome with some reported having identified close to 1000 distinct proteins depending on the stringency of the protein identification criteria. Most of these proteomic analyses of amniotic fluid have been protein surveys and have not focused on searching for biomarkers in amniotic fluid. Despite the similarities with serum, it is important to mention the discrimination of AF proteins according to molecular weight caused by the passive filtration across various membranes as mention above. This gives rise to a log/linear relationship between the protein concentration in AF compared to serum and the molecular weight as can be seen in Figure 13.

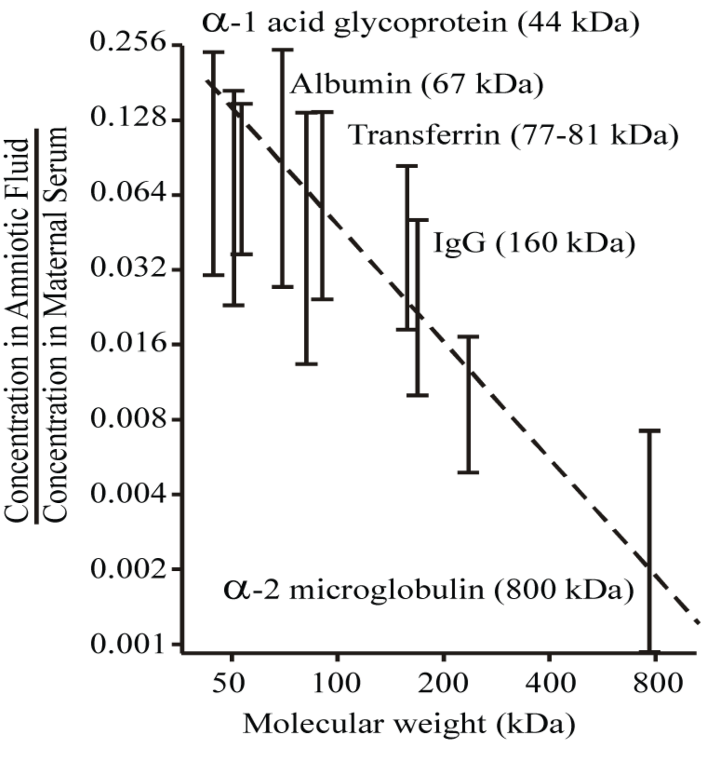


Figure 13: Relationship between the log concentration ratios of protein present in AF over those in serum with respect to protein molecular weight. This figure was adapted from Johnson *et al.* [50].

Proteins are not the only biologically important species present in amniotic fluid. Of considerable interest is establishing the metabolite profiles of biological fluids. This is no simple task as most biological fluids have more than a thousand metabolic species (of endogenous and exogenous sources) that are biologically relevant plus their degradation products. AF is no different. A great variety of biochemical species have been reported to be present, yet the coverage of the AF metabolic profile is incomplete. Here, it is noteworthy to recall that the mother, the placenta and the fetus all contribute to the AF content. The metabonomic analysis of AF, despite the challenges, shows the potential of uncovering novel biomarkers useful for prenatal diagnoses. Establishing a comprehensive list of all the metabolites present in AF is not only outside of the scope of this thesis, but still also outside of the reach of science at the moment.

Modern analytical instrumentation has provided an opportunity to bring us closer to an understanding of the biochemical profile of amniotic fluid. The current state of progress of biomarker discovery in AF, as it pertains to prenatal diagnosis, will be reviewed in the next section.

1.3.2 Prenatal diagnosis

Recently, the gestational window between weeks 11 to 13 as been identified as of great interest towards assessing the initial prenatal care and the required follow-ups [51]. In this time window, most major aneuploidies can be identified by combining maternal characteristics, ultrasonography results and maternal blood tests. Aside from genetic anomalies, the same combination of maternal characteristics and history with biophysical and biochemical testing can help assess the patient-specific risk for a variety of pregnancy complications, such as preterm delivery, preeclampsia, gestational diabetes, fetal growth restriction and macrosomia [51].

In practice, genetic testing, sonography and risk factors based on maternal characteristics constitute the main prenatal diagnostic tools. Good biomarkers of abnormal pregnancy are, to say the least, scarce. The next section will review biochemical markers of various abnormal fetal outcomes. The obvious places to search for these markers are biological fluids, such as maternal blood, AF and fetal blood. Maternal blood can be seen as the least invasive biofluid sample to obtain. Unfortunately, it is the farthest removed from the fetal compartment and, as such, might not contain as precise and accurate information about the health status of the fetus. Amniocentesis or fetal blood sampling are considered invasive to very invasive and are carried out when risk factors warrant these more risky procedures. Amniocentesis is carried on a relatively routine basis as more and more women are having children at later stage of their lives. Although certainly not risk free, it is a sampling method much less risky than fetal blood sampling and can still provide important genetic information regarding the fetal health status. Before the week 11 to 13 gestational window becomes an integral part of prenatal care practice, significant breakthroughs into biochemical markers of abnormal fetal development and gestational outcome have to occur.

1.3.2.1 Birth outcomes and abnormal fetal/maternal health statuses

Amniocentesis has typically been performed to screen for genetic disorders such as Down’s syndrome. Yet there is a great potential to search for markers of other types of disorders, such as metabolic disorders or birth outcomes associated with complications such as low birth weights, macrosomia, pre- and post-mature births. These are represented in Figure 14 and will be defined briefly below.

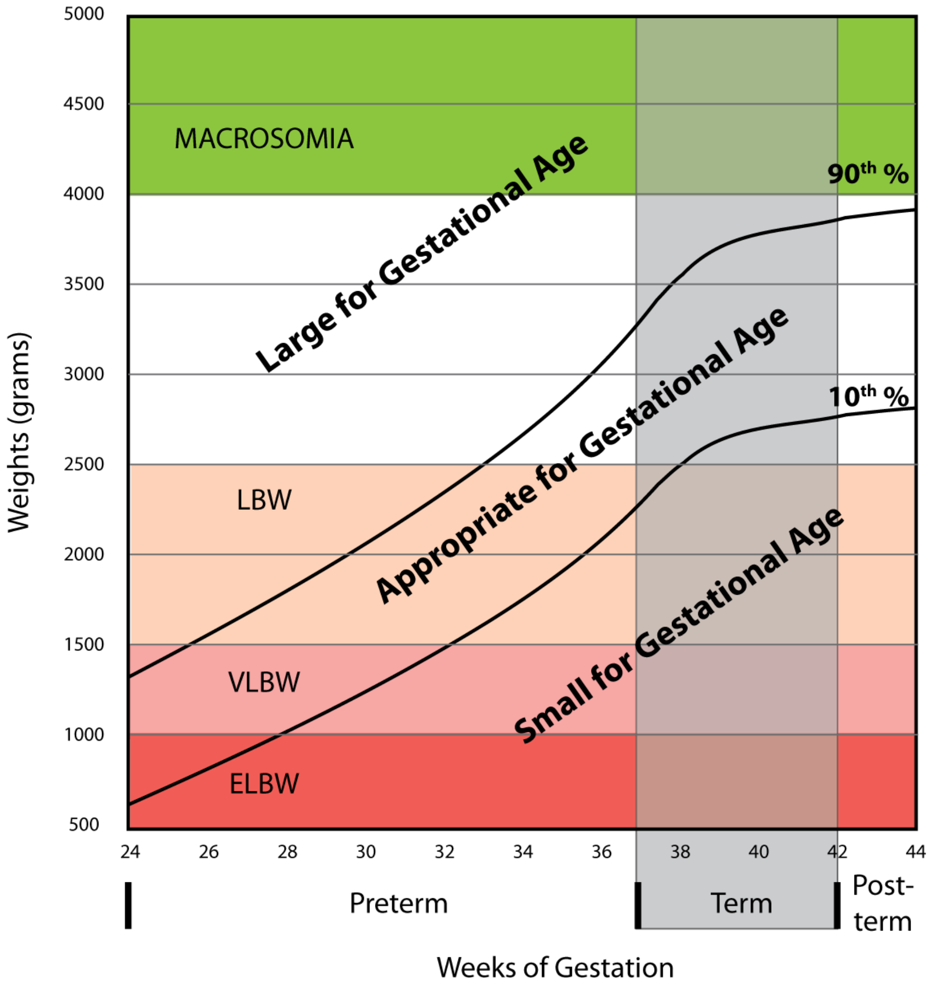


Figure 14: Birth weights categories and birth weight corrected for gestational age categories. Note that percentile lines are very approximate and only serve to give a sense of the SGA, AGA and LGA groupings. This figure was adapted [52].

Premature births (preterm) are defined as birth occurring prior the 37th week of gestation and are the leading cause of perinatal death. Post-mature births, birth after the 42nd week of gestation, are also of concern, but are typically less preoccupying since labor may be induced to prevent such late births. While problematic, advancements in neonatal medicine have improved dramatically the survival rate of either problematic gestational age at birth.

Birth weight categories, at birth, are also of interests. Low birth weights (LBW; < 2500 g) are generally associated with retarded fetal growth. Using the LBW classification leads to problems since birth weights are dependent on the gestational age at birth. For example, a preterm birth will very likely be LBW, but might not be as concerning as a LBW born on the 41st week of gestation. Birth weight categories corrected for gestational age are more appropriate to use as they account for the continuing growth during gestation. Typically, three important categories emerge from this type of classification. Small-for-gestational-age (SGA) neonates are defined as having a birth weight below the 10th percentile for a given gestational age at birth. SGA births are problematic as they are associated with greater risks of perinatal death and handicaps [53]. These risks can be reduced when SGA is identified antepartum [54].

Macrosomia traditionally refers to a birth weight greater than 4000 or 4500 g but, more recently has been defined as large-for-gestational age (LGA) [8]. LGA is similarly defined to SGA, but includes neonates with a birth weight above the 90th percentile for the gestational age at birth. This fetal condition increases the risks of fetal injuries at birth, such as shoulder dystocia, brachial plexus or facial nerve injuries, fractures of the humerus or the clavicle and even birth asphyxia [8]. The mother is at increased risk of trauma to the birth canal and LGA births often warrant a Caesarean section. Additionally, LGA pregnancies are often associated with gestational diabetes mellitus (GDM) pregnancies.

GDM is a maternal condition similar to diabetes albeit slightly less severe, or rather less advanced than diabetes. It is defined as “any degree of glucose intolerance with onset, or first recognition, during pregnancy” [55, 56]. GDM is associated with risks to the fetus and to the mother. Stillbirth, congenital malformations, macrosomia, birth injury, perinatal mortality and postnatal adaptation problems (*e.g.* hypoglycaemia) are all more common as a result of GDM pregnancies. Mothers show higher risk of developing diabetes if GDM is diagnosed during the pregnancy.

There are other abnormal pregnancy outcomes which affect maternal or fetal health statuses, but these are not strictly relevant to this study. The focus of the following section will be on, for the most part, biomarkers of SGA, LGA and GDM.

1.3.2.2 Biomarkers indicative of SGA, LGA or GDM

Above, it was mentioned that “good” biomarkers of abnormal pregnancy outcome are scarce, but this does not mean that the literature is devoid of papers on the subject. The correlation between fetal outcome and biochemical measurements has been demonstrated in the literature. Markers indicative of the onset of SGA, LGA or GDM have been sought in maternal serum, in umbilical cord blood and in AF however, studies have shown a predilection towards SGA.

Kwiterovich *et al.* reported that higher triglyceride and apolipoprotein B levels and decreased high density lipoprotein cholesterol and apolipoprotein A-I levels in cord blood were correlated with infants born as SGA compared to a control group of AGA infants [57]. Concentrations of mitochondrial DNA from umbilical cord tissue normalized to maternal white blood cell nuclear DNA were significantly lower in cases of SGA and LGA [58]. First trimester maternal serum showed that low levels of pregnancy-associated plasma protein A (PAPP-A) and high levels of α-fetoprotein [59] are associated with SGA. Corroborating this study, low levels of PAPP-A were also shown to increase the risk of SGA [59, 60]. Low levels of metastin found in first trimester maternal serum also point to a SGA perinatal outcome [61]. Another study used a composite metric combining fetal nuchal translucency thickness with maternal serum concentration of free β-human chrionic gonadotrophin (β-hCG) and PAPP-A to predict SGA births [62]. Placental growth factor (PLGF), placental protein 13 (PP13) and A Disintegrin And Metalloprotease (ADAM12) are all present in first-trimester maternal serum in lower concentrations [53].

Fasting levels of insulin and C-peptide in maternal serum at 24 to 30 weeks gestation could differentiate LGA and AGA outcomes [63]. Decreased maternal serum α-fetoprotein [64] and elevated vitamin E [65] levels in early gestation were shown to be associated with LGA outcomes. Lower cord blood adiponectin levels combined with elevated insulin and leptin levels were found for LGA infants versus AGA control infants [66]. Recently, first-trimester maternal serum adiponectin levels were found to be lower for cases of macrosomia.

Markers of GDM often overlap with those of LGA since many GDM pregnancies produce LGA neonates. Plasma glucose and HbA1c, markers of diabetes, also serve as markers of GDM during gestation [67]. Maternal serum levels of insulin regulatory species, such as decreased adiponectin [68] or increased leptin [69], have been shown to coincide with GDM. Sex hormone binding globulin [70], C-reactive protein [71] and placental amino acid transporters [72] also showed altered levels in GDM and are potential biomarkers. Here also, composite biochemical and biophysical metrics offers a way to detect GDM by combining maternal serum sex-hormone-binding globulin (SHBG), adeponectin and maternal characteristics [7].

AF biomarkers have received much less attention, but IGF BP1 [73], homocysteine [74] and methionine [75] have shown to be associated with infant birth weight. Elevated glucose in second trimester amniotic fluid was shown to be indicative of GDM [76]. These markers have all been discovered using hypothesis driven strategies.

Despite the few studies involving the search of biomarkers in AF, this biological fluid shows a high potential for biomarker discovery because of its proximity to the fetal compartment as well as the maternal compartment. Additionally, the AF proteome and metabolome are becoming increasingly studied and known.

1.4 Thesis structure

The Introduction was designed to provide sufficient background knowledge to understand the core thesis work that is presented in the next four chapters. The following chapters are in manuscript form and each will be preceded with a short foreword that explains the manuscript's status as well as supplementary information to prepare the reader for the chapter. For the purposes of the overall thesis work, all of the AF samples were analyzed by CE prior to any chemometric data analysis and so, chronologically the GDM (Chapter 2) and LGA (Chapter 4) represent experimental work carried out at the same time. However, Chapter 3 presents the logical extension to the GDM work which is the follow-up experiments where the oxidation status of the human serum albumin recovered from AF is presented. Chapter 4 involves work carried out on *in vitro* fertilization media. This chapter shows another successful application of the chemometric analysis of CE data and speaks to the transferability of the tools developed to other biological sample/outcome type problems. Additional material is presented in Appendices I, II, III and IV to support and more fully describe the work presented in the manuscript chapters as they are written with a clinical audience in mind.

Statement of contributions

Michel Boisvert carried-out the mass spectrometric analysis and processed the electropherogram data using the genetic algorithm for both AF and IVF studies. The algorithm used for the data analysis was written by Michel Boisvert, except for COW and the algorithm provided by David H. Burns (see below). All the efforts to identify the unknown small molecule predictive of LGA were also done by Michel Boisvert.

Tao Gao (M.Sc. research assistant), Nadine Zablith (M.Sc. research assistant) and Celine Lacroix (undergraduate research assistant) collected the capillary electrophoresis data and performed all electrophoresis experiments for the AF study.

Wei Lin (M.Sc. research assistant) collected the capillary electrophoresis data and performed all electrophoresis experiments for the IVF study.

David H. Burns (professor in department of chemistry at McGill University) provided the genetic algorithm and Bayesian classification program, as well as intellectual input to experiment design and data interpretation.

Foreword to Chapter 2

Chapter 2 and Chapter 4 pertain to the analysis of AF by CE whereby the data is analyzed with chemometric tools to link the sample content to maternal or fetal disease/condition. These two chapters are to be submitted for publication as companion papers to Biomarkers in Medicine, a journal designed for a broad audience from researchers to clinicians. Because of this, the data analysis is not described in great details within the papers. The Introduction chapter gave mathematical and conceptual explanations of the chemometric data analysis routines used for preprocessing and processing of the electrophoretic data. Appendix I gives more details about the experimental parameters used.

For more detailed explanations of individual chemometric steps should read the following references:

- COW: Nielsen 1998[32] and Tomasi 2004 [31]
- HAAR and Genetic algorithm employed: Gributs 2006 [43] and Jang 1997 [42]
- Bayesian statistics in medicine: Ashby 2006 [77]

.

Chapter 2

Prediction of Gestational Diabetes Mellitus Based on an Analysis of Amniotic Fluid by Capillary Electrophoresis

Michel R. Boisvert [3], *Kristine G. Koski* [1], *David H. Burns* [2], *Cameron D. Skinner* [3]

[1]School of Dietetics and Human Nutrition, McGill University (Macdonald Campus),
Montreal, Canada H9X-3V9

[2]Department of Chemistry, McGill University, Montreal, Canada H3A-2K6

[3]*Department of Chemistry and Biochemistry, Concordia University,
Montreal, Quebec, Canada H4B 1R6

Email: CSkinner@alcor.concordia.ca, Tel: 1 (514) 848-2424 x 3341, Fax: 1 (514) 848-2868

2.1 Abstract

Aims: To detect gestational diabetes mellitus biomarkers in human amniotic fluid collected for age-related genetic testing at 15 weeks gestation using capillary electrophoresis and a sophisticated data analysis methodology**.**

Materials & Methods: Amniotic fluid samples obtained from mothers undergoing routine amniocentesis were separated by capillary electrophoresis. Data were aligned using correlation optimized warping, reduced by Haar wavelet transformation and samples were classified using a genetic algorithm. The best model maximized the sensitivity and specificity by evaluating a Bayesian statistical model of the data and employed a leave-one-out cross-validation strategy.

Results Gestational diabetes mellitus (GDM, n=14) was distinguished from non-GDM (n=95) with 86% sensitivity and 99% specificity using two wavelets. These wavelets were located in the unresolved protein region and on the edge of the maternally derived albumin peak.

Conclusions: Gestational diabetes is a maternal pathology however it was shown that it alters the biochemical profile of amniotic fluid. These changes can be detected at 15 weeks gestation whereas testing for gestational diabetes is normally carried out at 24-28 weeks and suggests that GDM onset occurs early in gestation.

Keywords: amniotic fluid, gestational diabetes mellitus, capillary electrophoresis, biomarkers of abnormal pregnancy.

2.2 Introduction

The Barker hypothesis, also known as fetal programming, posits that the environment of the growing fetus has direct impacts not only on fetal development but also has lifetime repercussions [78-83]. Studies show that abnormal fetal growth can be linked to future health complications such as increased risk of cardiovascular diseases [79-81], type 2 diabetes [82], dyslipidaemia, obesity and metabolic syndrome [80, 83]. There is also growing evidence that epigenetic modifications may be carried through to subsequent generations with the possibility of increased intergenerational health risks [84]. Early discovery of abnormal fetal development may also be commensurate with treatment and mitigation of the perinatal [85] and long term negative health impacts [86-88].

An important complication that can occur during pregnancy is gestational diabetes mellitus which is defined as a glucose intolerance during pregnancy [89]. Generally it affects 3-8% of all pregnancies in North-America and its prevalence is on the rise [90]. The risk of GDM is increased for women that are 25 years or older, above normal weight, a history of abnormal glucose tolerance, members of ethnic groups with a high prevalence of diabetes or have close relatives with diabetes [14]. Short term complications include macrosomia, hypoglycemia and possible respiratory distress syndrome [91]. Infants born of GDM pregnancies are at elevated risk of developing obesity and type 2 diabetes and their associated complications [92]. The mother with GDM also faces increased risk of developing type 2 diabetes [93] and is at greater risk of cardiovascular disease [92]. Hyperglycemia also is known to increase oxidative stress and cause oxidative damage to proteins and lipids primarily [94].

GDM is routinely diagnosed through a combination of risk assessments and the oral glucose tolerance test (OGTT) [95] at 24-28 weeks. The test involves fasting for at least 8 hours, ingesting 100 g of glucose and tracking plasma glucose values. A GDM diagnosis is made when at least two of the following values are found: fasting ≥95 mg/dl, 1h ≥180 mg/dl, 2h ≥155 mg/dl, 3h ≥140 mg/dl. Using the fasting plasma glucose levels criterion (≥95 mg/dl) the sensitivity and specificity are 58.02% and 68.91% [96], respectively. If perturbations in the glucose metabolism can be observed in weeks 24-28 other metabolic perturbations may be present in AF not only in this time window but also prior to week 24.

At this time the majority of reported markers of GDM involve altered protein levels with only a few small molecules being associated with GDM. Plasma glucose and HbA1c, markers of diabetes, also serve as markers of GDM during gestation [67]. Maternal serum levels of insulin regulatory species, such as decreased adiponectin [68] or increased leptin [69], have been shown to coincide with GDM. Sex hormone binding globulin [70], C-reactive protein [71] and placental amino acid transporters [72] also showed altered levels in GDM and are potential biomarkers. However, there are no generally accepted AF biomarkers for the early detection of GDM. During fetal development, the biochemical state of both mother and fetus is continuously changing, which complicates detecting abnormal biochemical profiles. Despite this increased variation, abnormal development should result in a different growth trajectory and, with the right approach, a differentiable biochemical profile with respect to normal fetal development.

This paper will show that analysis of early second trimester whole AF by capillary electrophoresis coupled with Haar transformed data and Bayesian analysis provides a simple means to survey the AF for underlying differences in AF composition associated with GDM.

2.3 Materials & Methods

2.3.1 Samples

Pregnant women undergoing routine amniocentesis between 12-20 gestational weeks at St. Mary's Hospital Center in Montreal, Canada were invited to participate in this study. Ethical approval was obtained from McGill University and St Mary's Hospital Centre, and signed consents allowed collection of amniotic fluid at Montreal Children's Hospital once genetic testing was completed. The AF samples were stored at –85ºC until analyzed. Application of inclusion criteria (singleton pregnancy) and exclusion criteria (multiple births, genetic anomalies) resulted in 109 mother-infant pairs for whom amniotic fluid samples were analyzed and for whom birth outcomes were available.

From questionnaires and maternal obstetrical chart review, maternal characteristics including maternal age, height, prepregnancy weight, smoking status, parity, amniocentesis week and infant characteristics such as birthing method, gender, birth weight and gestational age were obtained. Gestational age was uniformly calculated on the basis of physicians' estimates using last menstrual period. Birth outcome was determined using a new birth-weight-for-gestational-age-categorization that is based on gestational age and gender [97].

2.3.2 Data collection and processing

The details of the data collection and processing are provided in the following paragraphs after this brief outline. The AF samples were separated by capillary electrophoresis.

Data preprocessing required the time alignment of the electrophoretic data with correlation optimized warping and data reduction with a Haar wavelet transform. Data processing was accomplished with a Bayesian classification strategy.

Capillary electrophoresis was retained as the most appropriate method for the separation of AF's major components because of its potential for fast, highly efficient and automated sample analysis that can accommodate minute amounts of sample. All CE separations were performed on a Beckman Coulter P/ACE Series MDQ capillary electrophoresis system (Beckman, Fullerton, CA) using Beckman 32 Karat Software Version 5.0 for instrument control, data acquisition and analysis. Data were collected at 4 Hz with a photodiode array (PDA) detector covering the 190-350 nm spectral range. Untreated fused silica capillary (75 μm i.d., 360 μm o.d.) purchased from Polymicro Technologies (Phoenix, AZ, US) was cut to 60 cm in length with a window at 50 cm from the inlet. The capillary was conditioned between runs by sequentially flushing for 3 minutes at 1.4 bar followed by a 1 minute wait at 0 bar with 5 mM SDS then 100 mM NaOH. The capillary was then rinsed and filled with the separation buffer (2 minutes at 1.4 bar) and conditioned under 25 kV for 1 minute (0 bar). At the beginning of each day a similar conditioning step was performed except the pressure rinses were 5 minutes and the waits were 2 minutes. All prepared solutions were filtered through 0.45 μm syringe filters and degassed before use. Prior to separation, randomly selected frozen (-85°C) AF samples were thawed in an ice-water bath and diluted 1:1 (v/v) with 0.5 mg/ml thiamine in water as an internal standard [98]. Samples were injected hydrodynamically (10 s, 34.5 mbar) and separated using 75 mM borate, 0.8 mM EDTA pH 9.27 buffer at 25 kV for 20 minutes with the temperature set at 28°C.

Data preprocessing is a critical step where raw data are prepared into a suitable format for subsequent data analysis. It can refer to simple steps such as normalization and baseline correction or to more sophisticated data transformation steps. Guidelines exist for data preprocessing steps that should be applied to separation data prior to chemometric analysis [21, 31, 32, 99]. In this study, the data were baseline corrected by baseline subtraction and normalized to the peak height of albumin. One of the established limitations of CE is migration time variations due to fluctuations in the electroosmotic flow velocity that result in peak area distortions. These issues were minimized by using the accepted practice of dividing the CE signal (at any given time) by its corresponding migration time [21, 99].

To further prepare the data, the electropherograms were time aligned using correlation optimized warping (COW) [31, 32]. The COW algorithm decomposes an electropherogram into small windows that can undergo a constrained stretching or contraction in such a way that the time aligned electropherograms show a major improvement in the alignment of their main features compared to a model target electropherogram. In this case, the electropherograms were aligned to the first sample in the data set as a reference electropherogram since all of the AF electropherograms had a high degree of similarity in terms of presence of peaks [31]. The COW window and slack (stretching/contracting) parameters were set to 20 and 2 data time points (5 and 0.5 s worth of data) respectively based on using a window of at least ½ the width of the smallest peak, which was 15 data time points and the smallest slack to achieve proper alignment. Visually, the alignment of prominent peaks (not shown) was much better and the global correlation coefficients of the sample electropherogram to the target electropherogram typically improved from 0.78 to 0.96 [100].

The final preprocessing step involved a Haar wavelet transformation of the data that denoised and compressed/simplified the data. The son Haar wavelet is a square wave that can be used to decompose the data into integration windows of allowed widths (2, 4, 8, … , 2^n). For the purpose of the Haar transformation, an 8.5 min (2048 data point) slice of the electropherogram was subjected to the Haar transform where with the first 256 wavelets being retained.

The goal of this data analysis was to build a statistical model that successfully differentiated the GDM from non-GDM samples using a minimum number of variables (wavelets). Bayesian statistics, as applied to classification problems [101], are well-established and have already been applied to a variety of fields. For an in depth review of Bayesian statistics in medicine readers are directed to a review by Ashby in 2006 [77]. For this study, a Bayesian classification strategy was used to classify samples into one of two outcomes: GDM or non-GDM.

A common strategy employed in Bayesian classification problems, to minimize the time required to find the best classification model, involves the use of optimization methods. The computational time required to systematically build models with all possible wavelet combinations would have been considerable. Instead, a genetic algorithm, which avoids human variable selection bias, was used to optimize wavelet selection while reducing computational time [102]. Implementation of the genetic algorithm is detailed elsewhere [43].

In the Bayesian strategy, each test sample was classified on the basis of proximity to the known class means via a *posteriori* calculation and the class probabilities from *a priori* information (12.7% for GDM and 87.3% for non-GDM). Inclusion of additional variables (wavelets) in the model was based on the general principle that a parsimonious model that uses fewer variables is more robust and should be chosen unless the addition of another variable increases the success of the model significantly. The best model maximized the sum of the sensitivity (true positive rate) and the specificity (true negative rate) for the classification based on a full leave-one-out cross-validation strategy [101]. Based on the calculated means and standard deviations for each group P values were calculated using student's t-test. A random permutation test was done to determine if the classification model, using the identified variables, is statistically different than models obtained with random permutation of outcomes using the student t-test as the test statistic. A total of 75000 models with randomized outcomes were generated to determine a population mean and standard deviation.

2.4 Results

In this study, the average age (37.8 ± 2.3 years) of the participants was 8.5 years older than the Canadian national average maternal age. Out of the 109 newborns, 55 were male and 54 were female. The prevalence of GDM pregnancy (GDM, n=14; non-GDM, n= 95) in the group was 12.7% *vs.* 2-9% [103] for the population at large. Of those GDM pregnancies, 14% of the women were obese (BMI > 30) and an additional 28% were overweight (BMI >25) but not obese, therefore 58% were of normal weight.

The ethnic composition of the study group consisted of 62% Caucasian, 21% Asian, 7% African, 3% Middle Eastern, 6% Hispanic and 1% other. The population contained 15%

smokers. Amniocentesis was carried-out at 15.1 ± 0.9 weeks with an average gestational age at birth of 39.1 ± 1.9 weeks and average birth weight of 3439 ± 671g.

The CE separation provided a fast, easy method of measuring the AF protein profile with all useful peaks detected in less than 10 minutes [104]. Repeated measurements of a pooled AF sample yielded 11 to 16 % relative standard deviation (RSD) in peak areas and 2.1 to 3.0 % RSD in migration times which necessitated the corrections for electroosmotic flow and area normalization. However, the AF in the study population had about 37-43% RSD on peak areas suggesting that the inherent biological variations were much larger than the error associated with the separation process.

Initially, the CE data were examined without normalization but no significant differences between GDM and non-GDM samples were found. Only when the data were as normalized was it possible to differentiate these maternal states suggesting that it is the relative distributions of the proteins and other biochemical species that give rise to the differences detected here.

By applying the genetic algorithm and Bayesian classification to the transformed data, the best model allowed for the differentiation of GDM samples from non-gestational diabetes samples using two wavelets with 86% sensitivity (2 false negatives) and 99% specificity (1 false positive) with a calculated P value of 0.001 as illustrated in Figure 15. The first of the two wavelets selected is located between the transferrin peak and an unknown partially resolved peak doublet [98]. In this region, multiple proteins co-migrated therefore conclusive identification was not possible without on-line mass spectrometric detection. The second selected wavelet corresponded to the albumin peak.

For GDM, the numerical magnitude of the first wavelet increased while the magnitude of the second wavelet decreased. This model was validated using a full leave-one out strategy was equivalent to a blinded study design. Based on the random permutation test the null hypothesis was rejected for α=0.05, meaning the classification model obtained above is statistically different from the population of models generated, using the same variables, with random permutation of outcomes suggesting that the classification model obtained is not likely to be due to chance alone.

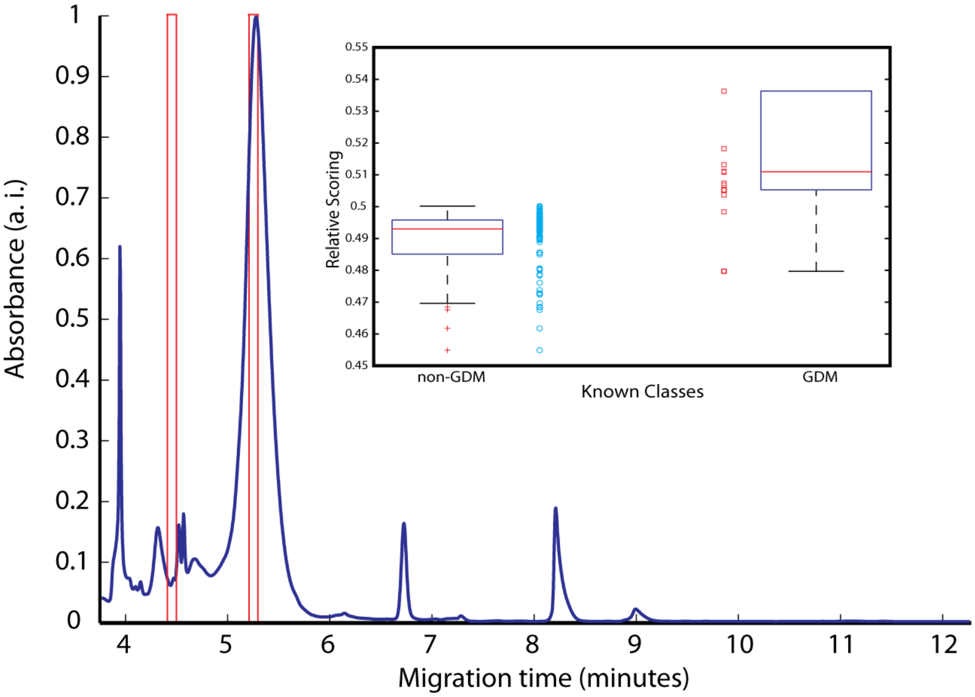


Figure 15: Using 2 variables (wavelets) on the albumin peak, the Bayesian algorithm can correctly classify all cases of GDM (14) and the non-GDM (95). Selected wavelets on electropherogram at 195±5 nm of AMF

2.5 Discussion

Amniotic fluid constitutes an important part of the growing fetus' environment during gestation. The dynamic nature of pregnancy (*e.g.* AF volume, maternal metabolism, body size, health status) and fetal growth (*e.g.* fetal swallowing and urination, skin keratinization, fetal weight etc.) all affect both the biochemical profile and overall concentrations of biochemical species. In early gestation, the majority of AF proteins are maternal in origin. However, the extent of the exchange between amniotic fluid and the fetus is considerable until the skin keratinizes at week 22-26 [105]. At the 15th week of gestation, the AF is a dynamic biochemical system, as such the data processing and analysis strategies must be flexible and robust.

The first selected wavelet corresponds to a valley point between the transferrin and an unresolved doublet peak in the protein band. This region may have been selected since any subtle changes in the migration times of any species surrounding that migration time window would have considerable impact on the value of this wavelet. Peaks on either side of the valley may influence the amplitude of the wavelet but, it is also possible that an unresolved species that underlies the electropherogram directly modulates the amplitude of the wavelet. Given that the first wavelet region corresponds to a mixture of co-migrating proteins no MS analysis was attempted. However, the fact that it was detectable by UV absorbance, suggests that it is a high abundance protein rather than the more difficult to detect minor, or trace, level species.

The second selected wavelet is in the leading edge of the electrophoretic peak associated with albumin, the protein with the highest concentration in AF (see Figure 15.) and serum. Albumin performs many important functions in the circulatory system [106] with control of osmotic pressure, transportation of small molecules and drugs and buffering of oxidative stress being amongst the most important. Diabetes is known to be a disease of both hyperglycemia and increased oxidative stress which results in protein modification.

These modifications could shift the pI of HSA and also induce conformational changes in the protein structure [107] both of which will modulate the electrophoretic mobility and result in an altered peak profile, as detected here.

The main *in-vivo* hyperglycemia associated protein modification is non-enzymatic glycation primarily at a lysine residue. On the other hand, oxidative stress associated modifications are more diverse but in albumin are primarily the oxidation of cysteine and disulfide oxidation. The literature [107-111] and measurements in AF [112] have identified that albumin's cysteine 34, an *in-vivo* redox active thiol that is responsible for albumin's oxidative buffering capacity, is the most susceptible to oxidation. Identified oxidative modifications include, cysteinylation, oxidation to sulfenic, sulfinic and finally sulfonic acid [113]. In 2005, a paper by Bar-Or *et al.* showed that in cases of intrauterine growth restriction (IUGR) oxidative cysteinylation of maternal serum albumin was approximately double that of normal pregnancies. They argued that the low placental blood flow of IUGR establishes a state of elevated oxidative stress that results in increased oxidation of maternal serum albumin to the cysteinylated form [108]. Other modifications have been observed *in-vitro* but have yet to be confirmed *in-vivo* or in AF. There are many unidentified modifications that we have observed in AF but the increasing power of mass spectrometry should provide greater insight into these albumin isoforms and their biological relevancy.

In the case of diabetes, induced protein modifications are significant because these changes can alter albumin's structure and function [109-111]. Glycation of serum albumin is certainly expected to be higher in cases of GDM but our recent MS analysis of AF HSA did not reveal any significant differences between GDM and non-GDM [112].

In that study, we found that increased irreversible oxidation of HSA's cysteine 34 was present for GDM while levels of cysteinylated cysteine 34 were lower. Increased HSA glycation may still be occurring under GDM pregnancies, however glycation is a slower biological reaction [114] than oxidation of albumin's free thiol. Additionally, the extent to which proteins are glycated is controlled by the glucose concentration and the protein's half-life [115]. Glucose concentration in AF is 4-5 mg/mL [116], lower than the 7-8 mg/mL found in serum, so the extent of glycation should be less. As for HSA's half-life in AF, very little is known and thus it is hard to speculate further on the possibility of increased glycation of AF HSA for GDM pregnancies but at the 15th week of gestation AF albumin is believed to originate predominantly from the maternal circulation [117]. This implies that the changes to albumin that we have observed may actually be reflecting changes of maternal serum albumin caused by GDM prior to the 15th week of gestation.

2.6 Conclusions

Using capillary electrophoresis, appropriate data normalization, multivariate analysis and classification allowed GDM to be differentiated from non-GDM on the basis of selected AF electrophoretic regions. In this study, we obtained an 86% sensitivity and 99% specificity with AF collected at 15 weeks gestation. Albumin was identified and is a likely target of oxidative modifications known to occur in GDM AF [112] but may actually be of maternal origin.

Realization that significant biochemical changes are already underway at 15 weeks gestation may also imply that fetal programming is occurring well in advance of the second trimester. Early detection of GDM offers the possibility of remedial action and detection 8-12 weeks earlier using the current strategy should provide additional time to provide a more effective treatment such as change of diet or insulin treatment, both shown to be effective controls of GDM [118].

2.7 Future Perspectives

Amniotic fluid has significant potential as a pool of both birth outcome and maternal biomarkers and requires further investigation. However, the implication of maternal albumin in differentiating GDM from non-GDM pregnancies suggests that maternal proteins may be a more direct and better source of GDM biomarkers. Sampling maternal serum proteins would also have the significant advantage of being much less invasive than amniocentesis sampling and thus much easier to obtain. It can also be drawn at any time during the pregnancy and may allow for earlier prediction, and monitoring, of abnormal pregnancy status.

2.8 Summary points

- Specific changes to the electrophoretic profile of AF allowed differentiation of GDM from non-GDM pregnancies during the early second trimester.
- A simple CE method and sophisticated data analysis and processing methods allowed these species to be located in the electropherogram with high sensitivities and specificities.
- A two wavelet model located on HSA and in the unresolved protein region classified GDM and non-GDM with 86% sensitivity and 99% specificity.
- Due to sample to sample biological variations, and variations in the experimental procedure, the procedure required area normalization and time alignment.
- In this study, Haar transformation and a genetic algorithm allowed the biomarker species to be efficiently located in the data set.

2.9 Acknowledgements

Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR) provided funding for this research (CDS, DHB and KGK) and (MRB) gratefully acknowledges a scholarship from Fonds Québécois de la Recherche sur la Nature et les Technologies.

2.10 Financial disclosure/Acknowledgements

Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR) provided funding for this research (CDS, DHB and KGK) and (MRB) gratefully acknowledges a scholarship from Fonds Québécois de la Recherche sur la Nature et les Technologies.

Foreword to Chapter 3: Albumin and the oxidation of albumin

One of the important results from Chapter 2 is that HSA is linked to GDM. More specifically a region of the HSA peak is predictive of GDM. Unfortunately, the CE-UV method employed was unable to provide sufficient insight about why this specific region of HSA is selected and not another. What complicates the matter is that some proteins are prone to chemical modifications that can alter their migration behaviour in CE. The specific chemical modification of a common protein backbone produces what is commonly referred to as an isoform of this protein. Thus one protein can have a whole family of isoforms that have very similar physico-chemical properties but with sufficient differences to cause alterations in the peak shape, the peak area and the migration time of a protein peak in CE.

In parallel to our CE work on AF, one of our collaborators (DH Burns) was also carrying-out studies of AF using NIR and Raman spectroscopy (unpublished work). He observed that GDM samples showed signs of increased oxidation. Based on the CE and spectroscopic results and the fact that HSA is particularly sensitive to oxidation, it was conjectured that AF HSA may show differences between GDM and non-GDM samples particularly with respect to the distribution of albumin oxidation isoforms.

To establish if there are GDM related differences in the AF HSA required a better understanding of HSA, and its modifications, in AF in comparison with the reported HSA isoforms present in serum. Unlike serum HSA, the HSA isoforms found in AF are not well characterized or described in the literature. This likely due to the assumption that they are one and the same, yet HSA in serum and HSA in AF are subject to environments with different chemical potential.

Appendix II reports the details of this preliminary HSA characterization work and set the stage for the work that resulted in the publication presented in Chapter 3. To the best of our knowledge, this is the first detailed study of AF HSA and was presented at several international conferences. The synopsis of Appendix II is that AF HSA displays a distribution of isoforms that is different than serum HSA. The most striking difference is that AF HSA showed signs of severe oxidative modification when compared to serum HSA. This encouraged testing the hypothesis that AF HSA would show a different pattern of oxidation in GDM pregnancies compared to normal pregnancies.

Chapter 3 presents the study comparing the specific differences in isoform distributions between GDM and non-GDM HSA and has been published in Analytical Chemistry in 2010 [112].

Chapter 3

Increased Oxidative Modifications of Amniotic Fluid Albumin in Pregnancies Associated with Gestational Diabetes Mellitus

Michel R. Boisvert [2], Kristine G. Koski [1], Cameron D. Skinner [2]

[1]School of Dietetics and Human Nutrition, McGill University (Macdonald Campus),

Montreal, Canada H9X-3V9

[2]*Department of Chemistry and Biochemistry, Concordia University,

Montreal, Quebec, Canada H4B 1R6

Email: CSkinner@alcor.concordia.ca

3.1 Abstract

Gestational Diabetes Mellitus (GDM) is a state of hyperglycaemia and increased oxidative stress with onset during pregnancy. Human serum albumin (HSA) was extracted from 26 GDM and 26 nonGDM amniotic fluid samples collected at 15 weeks gestation and analysed by mass spectrometry. The majority of all albumin isoforms were oxidized with the cysteinylated HSA as the base peak in the deconvoluted spectrum. The HSA peak areas, from a control sample, had 36% RSD across the six experimental days, but using the relative isoform distribution improved the precision to 3-6%. The results show that the relative contribution of permanently oxidized HSA was greater (P=0.002) and reversibly oxidized HSA was lower (P=0.006) for GDM compared to nonGDM in the samples measured. This implies that the path towards GDM has been set prior to 15 weeks gestation and results in increased protein oxidation.

3.2 Introduction

Gestational Diabetes Mellitus (GDM) is a state of hyperglycaemia arising during gestation that generally affects 3-8% of all pregnancies in North-America. Its prevalence is on the rise [90]. Left untreated, GDM can lead to diabetes for the mother and also increases the risks of fetal morbidity and perinatal complications [91]. Offspring from a GDM pregnancy are also at increased risk of developing glucose intolerance and diabetes [92]. Both hyperglycaemia and the increased oxidative stress associated with hyperglycaemia can result in excessive post-translational protein modification that can impair protein function [119] and alter biochemical signalling [120]. It is important to appreciate that even relatively small chemical modifications (*e.g.* phosphorylation, glycation, oxidation, alkylation) can cause biochemical cascades that result in important biological consequences [121]. Abnormal distributions of protein isoforms are known to occur in a variety of pathological states and can be used for the diagnosis, or prognosis of Alzheimer's, cardiovascular and autoimmune diseases as well as cancers [122-125]. There is some ambiguity as to the precise definition of isoform [124], but in the context of this paper we will define isoforms as a protein and its various post-translational modifications.

One of the difficulties in isoform analysis is that the post-translational modifications often result in subtle physio-chemical changes to the protein (*e.g.* tertiary structure, hydrophobicity, charge, molecular weight *etc.*) that may be difficult to characterize analytically. Common instrumental methods of isoform analysis include HPLC [126], SDS PAGE, capillary electrophoresis (CE) [127] and capillary isoelectric focusing (cIEF) [128]. Separations based methods such as HPLC, SDS PAGE and CE typically lack the resolving power to separate multiple isoforms simultaneously, especially when the modifications result in minute changes in hydrophobicity, charge and molecular weight. Capillary isoelectric focusing provides high resolution separations of isoforms based on small pI differences but reproducibility and conclusive identification of the modifications remain problematic [128]. The analytical technology with the greatest isoform identification potential is mass spectrometry (MS) [129, 130], especially when coupled with reversed-phase LC. Both CE and cIEF MS couplings are possible but difficult since common CE background electrolytes, many CE additives and cIEF ampholytes cause ion suppression and are problematic with electrospray ionization (ESI) [128]. Isoform analysis with SDS PAGE is used, but the low speed, semi-quantitative nature, high level of labour and low reproducibility of SDS PAGE make this route unappealing [131].

Modern MS is a versatile and powerful analytical technique that routinely achieves the mass accuracies required for isoform analysis. For example, the mass accuracies for time-of-flight MS are below 5-20 ppm and Orbitrap instruments are capable of better than one ppm error, while ion cyclotron resonance mass analysers achieve sub-ppm levels [132]. However, analytical limitations for some of these types of mass analysers include restricted dynamic range and, more importantly, poor quantitative accuracy due to lack of signal reproducibility. Quantitative accuracy can be improved by various strategies that revolve around isotope-based quantification such as absolute quantification by adding isotope-labeled peptides (AQUA) [133], absolute quantification by isotope labelled protein standards (PSAQ) [134] and isobaric tagging for relative and absolute quantitation (iTRAQ) [135, 136]. These methods use labelled analyte protein or peptide as an internal reference to overcome variations in ionization efficiency and the analytical methodology. In some cases, partial least squares calibration can be used to compensate for instrumental non-linearity, as reported for the absolute quantification of bovine serum albumin on a single quadrupole MS (LOD 17 ng, root mean square error in prediction of 127 ng) [137].

However, when attempting to establish a correlation between isoforms and a pathological state, quantitative isoform analysis may not be strictly required. Instead, the relative distribution of the MS signals can be a useful metric of the actual biological distribution. In order for a normalized approach to work, the ionization efficiencies should the same for the various protein isoforms under study. For large proteins, and their respective isoforms, we can reasonably expect that the ionization efficiencies are the same. Human serum albumin (HSA) is the most abundant protein in amniotic fluid [98] and can undergo several oxidative modifications in response to oxidative stress [138, 139]. In the present work, the relative distributions of human serum albumin (HSA) isoform signals from human amniotic fluid were compared between diagnosed GDM and non-GDM.

3.3 Experimental Section

Amniotic Fluid Samples. Pregnant women undergoing routine amniocentesis between 12-20 gestational weeks (mean 15 wks) at St. Mary's Hospital Center in Montreal, Canada were invited to participate in this study. Ethical approval was obtained and signed consents allowed collection of amniotic fluid from Montreal Children's Hospital once genetic testing was completed. The AF samples were stored at –85ºC until analyzed. Application of inclusion criteria (singleton pregnancy) and exclusion criteria (multiple births, genetic anomalies) resulted in 26 AF samples with GDM that were matched to 26 samples without GDM on the basis of gender and birth weight corrected for gestational age for a total of 52 samples that were analyzed. A pooled amniotic fluid sample was prepared and used as a control by mixing ten AF samples followed by aliquotting and freezing.

HSA Isolation and Desalting. Amniotic fluid samples were thawed on wet ice prior to HSA isolation with AlbuminOUT™ (Genotech Biosciences, Maryland Heights, MO, USA) albumin affinity spin columns. To ensure maximum recovery, 400 µL of AF (≈1.6 mg of HSA) was loaded onto the spin column which the manufacturer claims has a 2 mg HSA capacity. The HSA was eluted with 200 µL of the manufacturer's NaCl elution buffer (1-1.5 M) according to the protocol provided with the AlbuminOUT kit. Samples were desalted and washed using 30 kDa NMWCO Ultrafree centrifugal filters (Millipore) using five 500 µL DDW washes and recovered in ≈50 µL. The desalted HSA was then diluted to 100 µL with DDW and stored at -85ºC until MS analysis.

Mass Spectrometry. The HSA samples were thawed on wet ice. One µL of the HSA sample (≈16 µg) was injected, desalted and separated on a Waters CapLC system at 2 µl/min using a NanoEase C18 trap column with a 16 minute linear gradient from 100% A (97%H_2O/3%ACN/0.1%FA) to 90% B (3%H_2O/97%ACN/0.1%FA). A Waters Q-Tof2 system was used for MS detection using nano-electrospray ionization with a capillary voltage of 3.5 kV and a cone energy of 35 V. A mass accuracy of <20 ppm was determined by measuring by mass calibration using Glu-Fib peptide.

Data Analysis. Peaks corresponding to various HSA isoforms were integrated using MassLynx v.4.0 after deconvolution of the raw spectra over the m/z range of 1200 to

1650 with the MaxEnt1 algorithm. Data were analyzed in both Excel and MATLAB 7 (The Mathworks Inc., Natick, MA, USA). The fractional composition was calculated by dividing the isoform's integrated peak area by the total peak area. Student's t tests with a significance threshold set at P = 0.05 were used to compare GDM and nonGDM results.

3.4 Results and Discussion

Modern MS is a powerful analytical technique capable of rapid on-line protein isoform analysis, but sample preparation is critical to fulfilling MS's potential. We found that residual sample salts introduced excessive ionization suppression that led to poor reproducibility and low signal to noise ratios. To prevent this, it was necessary to desalt and wash the HSA sample both on the centrifugal filter and by using the NanoEase trap column. Using the trap column as the final sample preparation stage also allowed effective solvent exchange to improve ESI efficiency. These measures produced high quality spectra and Figure 16 shows a deconvoluted spectrum of a pooled AF HSA sample where several isoforms of HSA have been identified. Quantitative protein MS of biological samples is difficult due to significant variations in sample preparation, ionization efficiency and instrumental sensitivity. When the integrated albumin area, for all isoforms, was calculated from a single pooled sample measured in duplicate on the six experimental days a reproducibility of 36% was found. From this same sample data, reproducibilities of 4.9% and 2.7% were found for Ratio 1 and Ratio 3 respectively, as explained below. Pooled samples prepared each day, and measured in duplicate, yielded reproducibilities of 6.0% & 2.7% RSD for these same measures. The improved precision associated with using the relative measurements illustrates the benefits of this strategy and minimizes both instrumental and sample preparation error.

Figure 16 shows that the HSA in 2nd trimester pooled amniotic fluid is extensively modified; 11-13 prominent peaks were observed as baseline resolved, or as partly resolved peaks depending on the sample. However, the most striking feature of Figure 16 is that the base peak was not from reduced, or mercapto-albumin (66,437 Da, labelled as HSA–SH in Figure 16) but rather from an oxidized species: cysteinylated HSA. This modification occurs from the oxidation of HSA's free thiol [113] (cysteine 34) by S-cysteinylation giving a new isoform (HSA–cys) with a molecular weight of 66,556Da (Δ m/z +119).

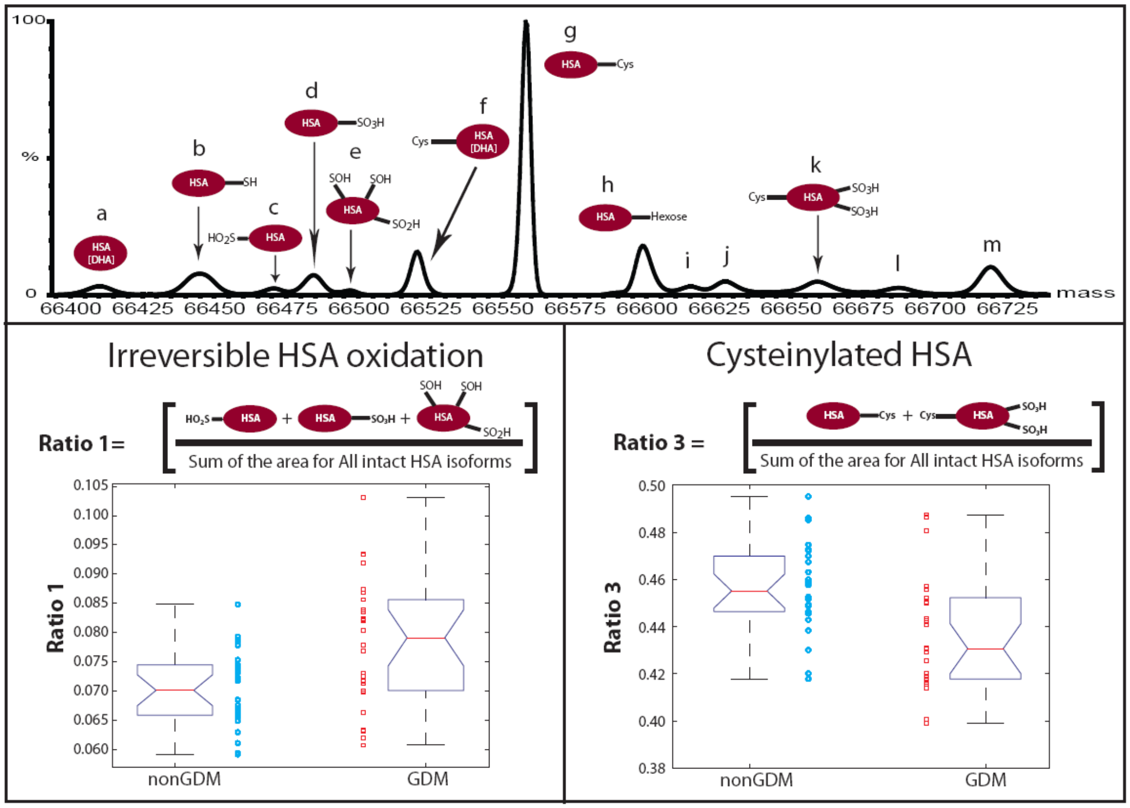


Figure 16: Deconvoluted ESI-MS spectrum of a pooled AF HSA sample showing the main protein modifications observed (top). Each peak has been lettered and the fractional areas are reported in Table 3. In the bottom left panel is the equation for Ratio 1 and a boxplot representation of the results comparing GDM and nonGDM with respect to Irreversible HSA oxidation. In the bottom right panel is the equation for Ratio 3 and a boxplot representation of the results comparing GDM and nonGDM with respect to cysteinylated HSA.

Other typical, *in-vivo*, modifications to the cysteine 34 thiol include oxidation to sulfenic acid (Δ m/z +16, HSA–SOH, not observed), to sulfinic acid (Δ m/z +32, HSA–SO_2H, observed), to sulfonic acid (Δ m/z +48, HSA–SO_3H, observed) and the S-cysteinylation (Δ m/z +119, HSA–cys, observed). In addition to cysteine 34 modification, one or more of HSA's disulfide bonds were oxidized resulting in di-sulfenic (Δ m/z +34, not observed), sulfinic with di-sulfenic acid (Δ m/z +66, $(HOS)_2$–HSA–SO_2H observed) and di-sulfonic acids (Δ m/z +98, not observed). Bar-Or recently identified that after disulfide oxidation, cysteine (cys487) can undergo selective and irreversible oxidation to dehydroalanine due to the proximity of basic arginine residues (arg484 and arg485) *in-*

vivo independent of the oxidation status of Cys34 (Δ m/z -34, HSA[DHA], observed and Δ m/z +85 HSA[DHA]-cys, observed) [140, 141]. Glycation by hexoses [142] (Δ m/z +161, HSA–hexose, observed) is also a common modification that was observed on AF HSA. To the best of our knowledge, these data provides the first detailed examination of intact amniotic fluid HSA protein isoforms by MS.

Mercapto-albumin is normally the predominant isoform (70-80%) in serum HSA [129, 130, 143, 144], but this was not the case for AF HSA where the mercapto-albumin was a minor peak compared to the oxidized isoforms and only contributed 7.5% to the total area (Table 3). There is some evidence to suggest that amniotic fluid is oxic at 15 weeks gestation [145], and normal pregnancy is associated with elevated oxidative stress [146] for both mother and the fetus and may result in increased oxidative protein modification. In maternal serum there are several mechanisms/pathways (e.g. glutaredoxin and thioredoxin) [147, 148] that reduce some forms of oxidized albumin (*i.e.* HSA-cys) to maintain the predominance of the reduced isoform. The activity of these two enzymes has not been reported in AF but thioredoxin was detected in an AF proteome analysis [149]. The placenta is a significant source of AF albumin and elevated levels of placental thioredoxin were measured in pre-eclamptic pregnancies but its effect on AF albumin is unknown [150]. The more highly oxidized HSA thiols (*e.g.* HSA-SO_2H, HSA-SO_3H etc.) are considered irreversibly oxidized, even in serum.

Given that GDM is a condition of elevated oxidative stress we investigated the correlations between oxidized HSA, non-oxidized HSA and GDM. Initially we believed that it would be necessary to use a paired t-test and selected our samples accordingly but this was unnecessary and all values correspond to conventional t-tests with a P of 0.05 for significance. Statistical analysis of individual HSA isoform peaks did not reveal any statistically significant differences between GDM and nonGDM mothers (Table 3) with the exception of the potentially reversibly oxidized cysteinylated HSA (peak g, Table 3). Peaks c, d and e are only partially resolved and the choice of the limits for peak integration are subjective. From a biochemical point of view, these three HSA isoforms contain irreversibly oxidized cys34 and can be justifiably grouped together. In this case the statistical analysis corresponding to all irreversibly oxidized cys34 was significantly greater for GDM than for nonGDM mothers as illustrated by Ratios 1 and 2 (Table 4). In a similar fashion, the potentially reversibly oxidized HSA, where at least one thiol was cysteinylated were significantly lower in AF of GDM compared to nonGDM mothers (Ratios 3 and 4, Table 4). In general, including the dehydroalanine species did not improve discrimination of GDM from nonGDM (Ratios 2 and 4) but the biological significance of this observation is unknown. At the very least, the loss of a disulfide bridge will alter HSA's conformation in a non-reversible way. The opposite trends of the irreversibly and reversibly oxidized HSA warranted investigation using linear combinations of these values which also proved to be significantly different between the groups (Ratios 5, 6 and 7). Interestingly, using just the mercapto-albumin (HSA-SH) did not produce any significant discrimination between GDM and nonGDM ($P = 0.15$).

Table 3: Calculation of relative isoform distributions of HSA in AF and corresponding P value for GDM (n=26) compared to nonGDM (n=26).

Peak	HSA isoforms	Mass difference from HSA-SH (66437 Da)	P value	Fraction of total nonGDM mean (SD)	Fraction of total GDM mean (SD)
a	HSA[DHA]	-34	0.88	0.030 (0.003)	0.030 (0.004)
b	HSA-SH	0	0.15	0.075 (0.016)	0.083 (0.019)
c	HSA–SO_2H	+32	0.33	0.009 (0.004)	0.011 (0.006)
d	HSA–SO_3H	+48	0.08	0.053 (0.007)	0.058 (0.012)
e	$(HOS)_2$–HSA–SO_2H	+66	0.13	0.005 (0.003)	0.069 (0.004)
f	HSA[DHA]-cys	+85	0.81	0.080 (0.009)	0.081 (0.007)
g	HSA–cys	+119	0.01	0.393 (0.019)	0.377 (0.02)
h	HSA–hexose	+161	0.41	0.117 (0.006)	0.120 (0.009)
i	Unknown	+177	0.69	0.0058 (0.006)	0.0066 (0.008)
j	Unknown	+187	0.97	0.051 (0.012)	0.051 (0.011)
k	cys-HSA–$(SO_3H)_2$	+222	0.22	0.050 (0.011)	0.047 (0.008)
l	Unknown	+250	0.91	0.033 (0.014)	0.033 (0.013)
m	Unknown	+276	0.93	0.078 (0.03)	0.078 (0.03)

Table 4: Calculation of relative isoform distributions and linear combinations of oxidized HSA in AF and corresponding P values for GDM compared to nonGDM.

Irreversibly oxidized HSA			
Linear combinations of peaks	P value	Fraction of total nonGDM mean (SD)	Fraction of total GDM mean (SD)
Ratio 1 = c+d+e	0.002	0.068 (0.006)	0.076 (0.015)
Ratio 2 = a+c+d+e+f	0.05	0.179 (0.015)	0.187 (0.016)
Reversibly oxidized HSA			
Ratio 3 = g+k	0.006	0.443 (0.019)	0.424 (0.026)
Ratio 4 = f+g+k	0.03	0.523 (0.025)	0.506 (0.027)
Other combinations			
Ratio 5 = (g+k)-(c+d+e)	0.001	0.374 (0.019)	0.349 (0.03)
Ratio 6 = (f+g+k)-(c+d+e)	0.004	0.455 (0.025)	0.429 (0.03)
Ratio 7 = (g+k)-(a+c+d+e)	0.004	0.343 (0.021)	0.318 (0.036)

These results show that AF HSA is highly oxidized and that the increased oxidative stress associated with GDM alters AF albumin towards the irreversibly oxidized isoforms. Corroborating this concept is a proteomic study of preeclampsia, another pregnancy condition of elevated oxidative stress, in which significantly elevated levels of oxidized momomeric transthyretin were found in AF at 16 weeks gestation [151]. In another study, Bar-Or found that maternal serum HSA, at 30 weeks gestation, was predominantly oxidized as the cysteinylated isoform in intrauterine growth restricted (IUGR) cases, another disease resulting in elevated oxidative stress [141]. It is significant to note that prior to the second trimester, most AF proteins are primarily maternal in origin [152, 153] so the AF measurements presented here may constitute a proxy measurement for maternal serum HSA oxidation.

Currently, only genetic testing is performed during routine amniocentesis, however, our results suggest that AF is an underutilised source of biochemical information and could be useful for early diagnosis of oxidative stress related conditions and may be specific for detection of GDM. Currently, the accepted method of diagnosing GDM is via the fasting plasma glucose level criterion (≥95 mg/dl) and is carried-out at 24-28 weeks gestation but only provides 58% and 69% sensitivity and specificity [154]. Our work shows that GDM can be differentiated from nonGDM at 15 weeks gestation and implies that the path towards GDM has been set prior to the second trimester. However, the high level of expertise required for the MS analysis, and the limited separation of GDM *vs.* nonGDM (Figure 1, bottom panels) would limit the application of the current approach as a clinically relevant diagnostic test. Alternative approaches to quantitatively probe the oxidation state of HSA’s cys34 are being investigated. As suggested by Bar-Or’s results, direct measurements of maternal serum HSA oxidation might provide an even better, and less invasive, method of assessing pregnancy associated oxidative stress and may have the potential for detection of GDM but this requires further investigation.

We have developed an easy and rapid method of assessing the albumin isoform distribution in AF but this same strategy should be applicable to a wider variety of proteins and sample types. Protein isolation, either through the AlbuminOut cartridges or other immuno purification systems and extensive desalting are critical to signal reproducibility but are easy to perform. The sample preparation costs are modest (≈$8/sample) and many samples can be prepared in parallel. Using the relative distribution of the deconvoluted protein spectrum overcomes the difficulties of quantitative sample preparation and mass spectrometry and is useful for monitoring alterations in isoform distribution in pathological conditions.

3.5 Acknowledgements

Support for this research was from the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (CDS and KGK) and a scholarship from Fonds Québécois de la Recherche sur la Nature et les Technologies (MRB).

Postscript to Chapter 3: Statistical significance of results

At the time that this thesis was examined, Chapter 3 was already a published paper and, according to the Thesis Guidelines, the text should not be modified. A random permutation test was identified as being beneficial in validating the significance of the results. To this end, a total of 1000 randomized outcomes were considered, using the same variables as reported in Table 4, to determine a population mean and standard deviation for each of the entries in Table 4.

Based on the random permutation test the null hypothesis was rejected for $\alpha=0.05$, meaning the classification model obtained above is statistically different from the population of models generated with random permutation of outcomes suggesting that the classification model obtained is not likely to be due to chance alone.

Foreword to Chapter 4: From GDM to LGA

Chronologically, the CE-GDM experimental work was carried-out at the same time as the following CE-LGA research however, logically it has been presented separately since LGA and GDM are different physiological conditions. LGA pertains to fetal growth and development whereas, GDM is a maternal health condition that arises during pregnancy. It has been suggested that there is a link between the two as the incidence of LGA neonates is greater in GDM pregnancies. Several studies have implicated GDM as a risk factor of LGA neonates [155, 156] and as such may lead to similarities/overlap in the predictors of GDM and LGA.

Chapter 4 will show that the same HSA region in the AF electropherogram as in Chapter 2 (albumin) combined with an unknown small molecule can be used to correctly classify LGA and AGA neonates. Furthermore, one of the unpublished outcomes from the MS-GDM work (Chapter 3) was that a difference in the HSA isoform distribution was observed between the LGA and non-LGA samples but the number of samples was not sufficient for publication (LGA n=6, AGA n=46). The isoform distribution was also different between the LGA-GDM (n=3) and the LGA non-GDM (n=3).

Again Chapter 2 and 4 are to be submitted together as companion papers to Biomarkers in Medicine.

Chapter 4

Early Prediction of Macrosomia Based on an Analysis of Second Trimester Amniotic Fluid by Capillary Electrophoresis

Michel R. Boisvert [3], Kristine G. Koski [1], David H. Burns [2], Cameron D. Skinner [3]

[1]School of Dietetics and Human Nutrition, McGill University (Macdonald Campus),
Montreal, Canada H9X-3V9

[2]Department of Chemistry, McGill University, Montreal, Canada H3A-2K6

[3]*Department of Chemistry and Biochemistry, Concordia University,
Montreal, Quebec, Canada H4B 1R6

Email: CSkinner@alcor.concordia.ca, Tel: 1 (514) 848-2424 x 3341, Fax: 1 (514) 848-2868

4.1 Abstract

Aims: To identify using capillary electrophoresis (CE) and chemometrics early biomarkers in human amniotic fluid (AF) of large-for gestational-age (LGA) infants

Materials & Methods: Second trimester AF samples, obtained from mothers undergoing age-related amniocentesis (~15 wks gestation), were analyzed by CE. Electropherogram data were aligned using correlation optimized warping and reduced by Haar wavelet transformation. A genetic algorithm using a Bayesian evaluation function and a leave-one-out cross-validation strategy for two birth outcomes: appropriate- versus large-for-gestational age infants (AGA vs LGA).

Results: Large-for-gestational age (LGA, n=23) was distinguished from appropriate-for-gestational age (AGA, n=86) with a sensitivity of 100% and a specificity of 98% using only two wavelets. The first wavelet associated with albumin and the second wavelet with an unknown small molecule.

Conclusions: The approach developed herein allows LGA fetuses to be metabolically distinguished from AGA fetuses early in pregnancy and indicates that birth of a LGA infant is already associated with an altered biochemical profile by the second trimester.

Keywords: amniotic fluid, large for gestational age, capillary electrophoresis, biomarkers of abnormal fetal development

4.2 Introduction

Studies show that abnormal fetal growth including low birth weight and intrauterine growth retardation, both indicators of intrauterine adversity, can be linked to increased risk of cardiovascular diseases [79, 80], type 2 diabetes [82], dyslipidaemia, obesity and metabolic syndrome [80, 83]. Similarly, reports show that LGA neonates also have increased risk for developing obesity, metabolic syndrome, diabetes and cardiovascular diseases [157]. For both these abnormal perinatal outcomes, evidence is emerging of epigenetic modifications *in utero* that may be intergenerational [84, 158]. With evidence for *in utero* fetal programming, there is growing concern that earlier diagnosis is required if we are to mitigate against the increasing incidence of obesity, type 2 diabetes, hypertension and metabolic syndrome in future generations [86].

Some correlations between infant birth weight and biochemical measurements using maternal serum or cord blood and a few using amniotic fluid have been reported. In first trimester maternal serum, low concentrations of pregnancy-associated plasma protein A (PAPP-A) and high levels of α-fetoprotein [59, 60] as well as low concentrations of metastin [61] have been associated with SGA. For LGA infants, decreased maternal serum α-fetoprotein [64] and elevated vitamin E in early gestation [65] and higher fasting concentrations of C-peptide at 24 to 30 weeks gestation were able to differentiate LGA and AGA outcomes [63]. Using cord blood, investigators have reported lower mitochondrial DNA in cases of SGA and LGA [58] and higher triglyceride and apolipoprotein B and lower HDL-cholesterol and apolipoprotein A-I in SGA compared to AGA infants [57]. In LGA infants, lower cord blood adiponectin combined with elevated insulin and leptin have been reported [66]. Several AF species have also been associated with birth weight outcomes: total protein inversely with SGA [159], insulin positively with LGA [160], higher IGF BP1 and IGFBP 3 with SGA and LGA respectively[73], uric acid [161], homocysteine [74] and methionine [75]. Recently, mass spectrometry has been employed to measure AF composition, but this work has focused primarily on cataloguing the proteome rather than finding biomarkers of abnormal fetal growth. Most reports rely on 2D-PAGE and spot analysis [45, 162] and/or liquid fractionation methodologies [163, 164] to achieve sufficient resolution to survey the AF proteome.

In North America, women at an elevated risk of abnormal pregnancy may undergo amniocentesis for genetic testing early in the second trimester. Once tested the amniotic fluid is often discarded. Amniotic fluid is a complex fluid that originates from both maternal and fetal sources, and has been underutilized in the quest of finding important biomarkers of fetal growth and development. Since amniotic fluid is a complex biological sample with a dynamically changing composition that contains 98% water and metabolic species that are necessary for, and by-products of, important reproductive biological processes (e.g. proteins, salts, glucose, uric acid etc.), it deserves further investigation as a potential pool for biomarkers. Moreover, there is a pressing need in prenatal care for the early identification of large-for-gestational age infants. In many countries, including US, Canada, Denmark and Finland, there has been an increase in the number of infants with birth weights > 4000g and > 90th percentile [165-168]. In this report, we demonstrate that capillary zone electrophoresis, one of the simplest of the separation techniques, coupled with a chemometric approach using Haar transformed data and Bayesian analysis provides a simple means of surveying 2nd trimester AF for underlying differences in AF composition associated with birth of large-for-gestational-age infants.

4.3 Materials & Methods

Amniotic fluid collection and sample storage were the same as in the companion paper (Chapter 3 in this thesis). Briefly, 109 mother-infant pairs were available and infants were classified as LGA (n=23) and AGA (n= 86). To classify LGA and AGA infants, we used birth weights that were greater than the 90% for gender and gestational age; AGA infants were defined as those >10% and <90% [168]. All CE separations, reagents and data processing steps were also identical to those used in the companion paper (Chapter 2 in this thesis). The Haar transformed dataset and known fetal LGA/AGA status were analysed using the genetic algorithm to select for the best combination of variables (wavelets) based on evaluation of a Bayesian benefit function as described earlier.

4.4 Results

The women included in this study mothers were undergoing age-related amniocentesis and therefore were older (37.8 ± 2.3) than the Canadian national average (29.3 yrs) and had a slightly higher prevalence of LGA births (19%) compared with the 10% reported in 2006 [169]. Of the LGA births, 29% also had GDM diagnosed during the pregnancy. Fifteen percent were smokers; 38% had BMIs within the normal range (BMI 20-24.9 kg/m^2); 48% were overweight (BMI >25 kg/m^2) and 14% were obese (BMI>30kg/m^2). Ethnicities were as follows: 62% Caucasian, 21% Asian, 7% African, 3% Middle Eastern, 6% Hispanic and 1% other. Amniocentesis was carried-out at 15.1 ± 0.9 weeks with an average gestational age at birth of 39.1 ± 1.9 weeks and average birth weight of 3439 ± 671g. Average birth weight for the LGA infants was 4248 g and for the AGA infants was 3310 g. Of the 109 newborns, 55 were male and 54 were female.

CE separation provided a fast, easy method of measuring the AF biochemical profile with all useful peaks detected in less than 10 minutes [104]. Repeated measurements of a pooled AF sample yielded 11 to 16 % relative standard deviation (RSD) in peak areas and 2.1 to 3.0 % RSD in migration times and necessitated the corrections for electroosmotic flow and area normalization. The genetic algorithm and Bayesian classification algorithm differentiated LGA (n=23) from AGA (n=86) using two wavelets selected from the 256 wavelets that represented the electropherogram with a sensitivity of 100% and a specificity of 98% (Figure 17, P = 0.0003). The model was validated using a full leave-one out strategy and was equivalent to a blinded study design. Also, based on the random permutation test, the null hypothesis was rejected for α=0.05, meaning the classification model obtained above is statistically different from the population of models generated with random permutation of outcomes and suggesting that the classification model obtained is not likely to be due to chance alone.

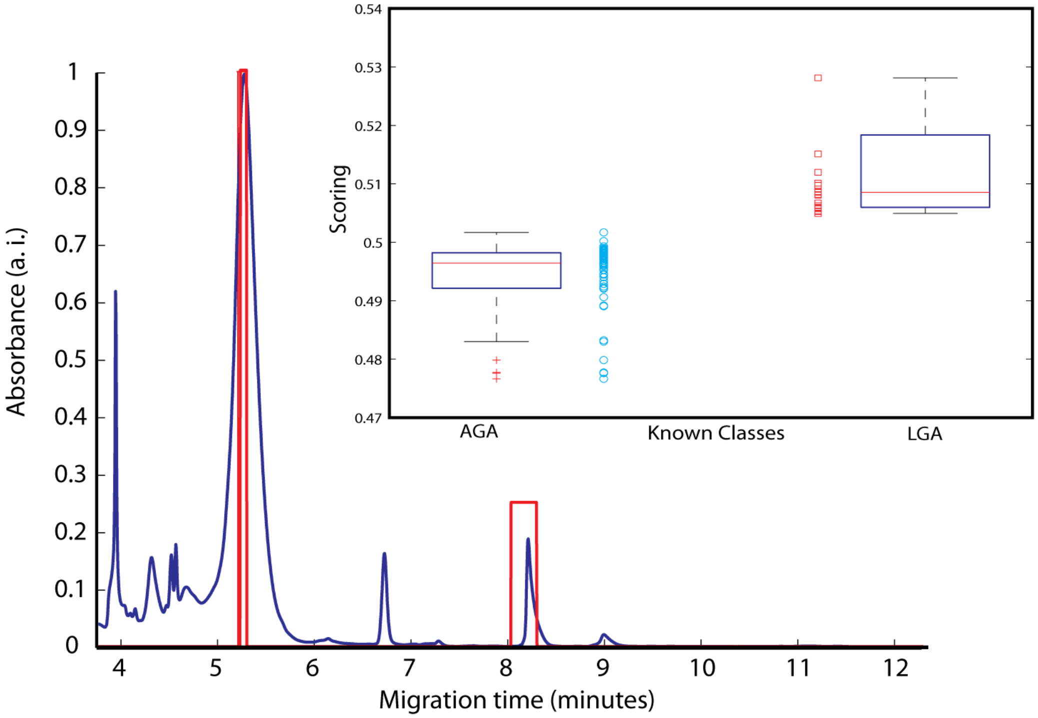


Figure 17 shows an electropherogram (blue trace) and the 2 wavelets (red rectangles) used to build the predictive model for classification of LGA vs. AGA. The inset is a box plot with AGA values in blue circles and LGA in red squares.

The selected wavelets were integrals of the regions in the electropherogram shown in Figure 17. The first wavelet corresponded to a region on the leading side of the albumin peak. The numerical magnitude of the first wavelet decreased for LGA. The second peak corresponded to a small negatively charged species. The magnitude of the second wavelet increased in LGA infants.

4.5 Discussion

Amniotic fluid is a dynamically changing fluid that originates from multiple sources including transudation through the amnion from maternal plasma, through the fetal nasopharyngeal oral and lacrymal secretions, through unkeratinized fetal skin and through the developing fetal kidney [170]. However, pregnancy continually modifies the biochemical profile and the overall concentration of most biochemical species in amniotic fluid. Prior to our investigation, several studies had reported individual differences in biochemical species in cord blood and maternal serum between LGA and AGA infants [58, 63-66, 73-75, 160, 161]. However these were for single analytes at different developmental stages. Our contention had been that a multivariate analysis of a single AF sample could be used reliability to assess the health of the mother-fetus dyad throughout pregnancy. Our results using capillary electophoresis of second trimester amniotic fluid collected at 15.1 ± 0.9 weeks gestation for the differentiation of LGA from AGA infants showed a high degree of selectivity and sensitivity (100% sensitivity and 98% specificity) when the model was built using 2 wavelets. This is much earlier then the current LGA prediction by means of sonographic measurements taken during the 3rd trimester [171-173]. Inclusion of additional variables (wavelets) was rejected since it did not improve the classification significantly. The two wavelets selected correspond to two potential biomarker species for birth of an LGA infant.

The first selected wavelet is part of the electrophoretic peak nominally associated with albumin, the protein with the highest concentration in AF. Albumin is the one of the most abundant proteins in AF and performs multiple functions, including nutrient source through fetal swallowing and as anti-oxidant. In a previous study of HSA in AF we observed 13 distinct masses [112] associated with AF HSA while Bar-Or illustrated 10 HSA isoforms [108] from maternal plasma. The significance of HSA as a potential biomarker is discussed in the companion paper (Chapter 2) where HSA was found to also be linked to GDM. An accelerated metabolic activity is expected for both LGA and

GDM. The decrease in this wavelet's amplitude in LGA indicates a change in the electrophoretic profile of HSA, possibly from protein conformational changes and/or from chemical modifications. Given our previous experience with GDM, a reasonable hypothesis is that the electrophoretic changes may be related to increased HSA oxidation and consequent reduction of the leading half of the albumin peak. However, this needs to be verified via mass spectrometry.

The second selected wavelet corresponds to an unidentified late migrating peak similar to those that have been observed in cerebrospinal fluid and serum CE separations [174]. The exact nature of this small molecule is hard to determine. From the migration behaviour in CE the species has to be a small molecule capable of carrying 2 negative charges. Several potential candidates have been investigated by mass spectrometry combined with CE spiking experiments but none so far correspond to the unknown molecule. Identifying unknown metabolites represent a significant challenge [175, 176] as MS spectral library that are publicly available are few [177, 178] and often contain limited numbers of metabolites in their databases. For instance, the HMDB returns only 19 metabolites (last searched on Jan 13, 2012) present in AF and so far none of them have corresponded to the unknown small molecule.

Proteomic analyses, especially of low abundance proteins, are very popular in biomarker discovery but have only recently been applied to amniotic fluid constituents [45, 162-164]. However, the present study highlights the importance of giving consideration to small molecule metabolites as valuable sources of biomarkers of fetal growth progression. It is worth noting that even relatively high abundance species, such as those that can be detected by UV absorbance, can provide high sensitivities and selectivities.

Obvious LGA metabolites to consider are those involved in production of energy such as carbohydrates and the glycolysis products, fatty acids, and products of amino acid degradation all of which feed into the citric acid cycle and the urea cycle. Establishing the normal concentration range of metabolites present in AF is of high clinical interest for screening, or diagnosis, of metabolic disorders. Only a few initial studies have determined the normal metabolic profile in AF [179-183].

This paper is the second in a pair of companion papers. In the first paper (presented as Chapter 2 in this thesis), biomarkers of gestational diabetes, a maternal pathology, were determined. In this paper, the same analytical and data analysis methods were used to determine biomarkers of large-for-gestational-age (LGA), a fetal pathology. Many LGA births are believed to be the result of poorly controlled gestational diabetes pregnancies [76, 184, 185] and as such should present a similar biochemical fingerprint as GDM in AF. In this study, only 29% of all neonates born LGA had mothers who had been diagnosed with GDM. This percentage is suggestive that metabolic aberrations associated with GDM do not underscore all cases of LGA. Also, it is interesting that one of the two species associated with LGA (albumin) was also used to characterize GDM (Chapter 2 and 3). Given that LGA and GDM analyses showed some overlap, yet had distinct wavelets, suggested to us that there are both similar and distinct metabolic pathways contributing to both conditions. We conclude, given the results of these two companion papers, that amniotic fluid metabolic profile that is present in pregnancies leading to an LGA infant has a metabolic fingerprint that is distinctive from GDM's fingerprint even though both of these conditions co-occur.

4.6 Conclusions

In conclusion, AF is a complex biological sample that has significant potential as a pool for biomarkers of abnormal fetal development. Using capillary electrophoresis, appropriate data normalization, multivariate analysis and classification allowed specific pregnancy outcomes to be associated with selected electrophoretic peaks or regions. The method developed here allowed LGA fetuses to be distinguished from AGA fetuses early in the second trimester, which is much in advance of current diagnosis using sonography later in pregnancy [171-173] when intervention is impractical or impossible. Further studies could identify definitive AF factors that may provide an important early window into the metabolomics of the developing fetus.

4.7 Future Perspectives

Given our findings, AF collected at the time of routine amniocentesis contains early biomarkers of LGA. However, increasing the resolution of the separation (*e.g.* 2-D chromatography) and combining it with on-line mass spectrometric detection would allow much more biochemical information to be extracted from the AF and expand the potential for biomarker discovery. A paired maternal blood-serum and AF study could provide an excellent means to find biomarkers of abnormal fetal development present in both biological fluids with the ultimate aim of developing non-invasive (to the fetus) assays that could be applied to the broader pregnant population, not just those undergoing routine amniocentesis. This should ultimately lead to earlier intervention strategies that would minimize the incidence of lifetime complications associated with birth of a LGA infant.

4.8 Summary points

- The proximity of AF to the fetus and the direct exchange of biochemical species between AF, mother, placenta and the fetus makes AF a good source of potential biomarkers of abnormal fetal growth due to either fetal or maternal causes. Results suggest that early second trimester AF is an underutilized biofluid for assessing the progression of fetal growth.
- A simple CE method and sophisticated data analysis and processing methods allowed specific biomarkers of LGA to be identified in AF early in the second trimester.
- Due to sample to sample biological variations, and variations in the experimental procedure, the procedure required area normalization and time alignment but Haar transformation and a genetic algorithm allowed the biomarker species to be located in the data set efficiently.
- Two wavelets, one on the leading edge of HSA and one associated with negatively charged small molecule were able to distinguish LGA from AGA with a sensitivity of 100% and specificity of 98%.

4.9 Financial disclosure/Acknowledgements

Natural Sciences and Engineering Research Council of Canada (NSERC) and Canadian Institutes of Health Research (CIHR) provided funding for this research (CDS, DHB and KGK) and (MRB) gratefully acknowledges a scholarship from Fonds Québécois de la Recherche sur la Nature et les Technologies.

Postscript to Chapter 4: The unanswered question.

The previous chapter has left the obvious question of the identity of the small molecule associated with LGA unanswered. The identity of this species was pursued but the difficulty of specific identification was not resolved. Determining the identity of biomarkers of interest in data driven approaches can prove to be very difficult. It is very often the bottle-neck for this type of experimental design. The complexity of biological samples such as AF introduces significant challenges with respect to the isolation and characterization of a putative biomarker. In this case, the CE buffer added another layer of complexity as it interfered with MS analysis of the sample. Additionally, the UV profile (absorbance from 190 to 210 nm) made it difficult to develop isolation methods by HPLC-UV with MS compatible buffers as these almost always require some sort of small volatile organic acid such as formic or acetic acid. These acids absorb strongly in the same UV region.

Appendix III details the approaches and efforts employed to determine the identity of the unknown small molecule present in AF that is predictive of LGA.

Foreword to Chapter 5: Testing the approach on another biological fluid.

While the work with AF was underway a unique opportunity arose to use the techniques that had been developed for AF with culture media that is used for *in-vitro* fertilization (IVF). The volume of IVF media available was limited (~20 µL) and made a CE approach a logical choice. From a sample point of view AF and IVF media show many similarities as they contain the salts, nutrients, amino acids and proteins required for the proper development of the embryo. Yet IVF media is a much “cleaner” and simpler sample as its initial content is known (synthetic biological media) and since the metabolically active embryo may alter its content during culture. Lastly, the incubation of the embryo in the IVF media is done under very well controlled culture conditions that help to minimize sources of variation.

The high cost and low success rate (~33%) generate high interest in research involved with IVF as more and more couples resort to assisted reproductions such as IVF to have a child. The reader may be unfamiliar with IVF so a brief introduction is presented here as it would not be relevant material for publication and is not included in the manuscript that forms Chapter 5.

When a woman is unable to become pregnant through natural conception she can resort to assisted reproduction technology (ART) such as IVF. The process can be broken down into 5 steps.

Step1: Egg production is stimulated with fertility drugs which cause the ovaries to produce several eggs instead of the usual one egg per month. Blood tests are done in this step to monitor hormone levels while transvaginal ultrasound exams are performed regularly to examine the ovaries.

Step 2: Once the eggs have developed, they are retrieved by follicular aspiration which is an outpatient procedure that can be done in the doctor's office. Guided by ultrasound the health care provider introduces a small needle through the vagina into each ovary to retrieve the eggs.

Step 3: The insemination and fertilization of the morphologically best quality eggs is carried by first placing the eggs and the man's sperm together. After a few hours the sperm will enter the egg (fertilization). In some cases the sperm can be injected directly into the egg to enhance the probability of fertilization.

Step 4: Once the egg is fertilized it will begin to divide and is now an embryo. Under normal conditions the embryo will have several cells actively dividing for the next 5 days. During this stage, embryos with high reproductive potential are actively metabolizing and dividing.

Step 5: The best embryo(s) are selected and placed inside the womb by passing a catheter containing the embryo(s) into the vagina and through the cervix. Pregnancy results when an embryo successfully implants to the lining of the womb and begins to grow.

A recent meta-analysis [186] showed that single embryo transfers (SET) result in more at term singleton deliveries compared to double embryo transfers (DET). Although, the initial pregnancy rate is higher in DET, the pregnancy rates for SET and DET become similar if additional frozen single embryo transfer cycles are included. This is important since ART-associated multiple pregnancies often lead to preterm deliveries. This increases the risk of perinatal complication and places an increased burden on health care systems [187].

For a more detailed treatment of the IVF process and ART in general, the reader is directed to a book section on infertility in *Comprehensive Gynecology* by Kaltz [188].

In Chapter 5, the biological sample used is the spent culture media from Step 4. The media is analyzed with a similar approach as for the AF and is used to predict whether or not an embryo from a single embryo transfer will implant successfully or not.

Chapter 5

Analysis of Culture Media Albumin as a Surrogate Marker of Embryo Viability

Running title: Albumin: a Surrogate Marker of Embryo Viability

*M.R. Boisvert[1], W. Lin[1], C.G. Vergouw[2], C.B. Lambalk[2], D.H. Burns[3], C. D. Skinner[1]**

[1]Department of Chemistry and Biochemistry, Concordia University, Montreal, Quebec, Canada

[2]VU University Medical Center, Reproductive Medicine, Amsterdam, the Netherlands

[3]Department of Chemistry, McGill University, Montreal, Canada

[1]Corresponding Author: Tel: 514-848-2424 x 3341; Fax: 514-848-2868;

Email: CSkinner@alcor.concordia.ca

BACKGROUND: Accurate assessment of embryo reproductive potential is critical for assisted reproduction and widespread adoption of single-embryo transfer (SET). Embryo metabolic activity alters the IVF culture media's composition through nutrient consumption but also can modify it via oxidative processes. In this study we investigated if culture media albumin modification, as revealed by sub-micellar capillary zone electrophoresis and mass spectrometry, is correlated to fetal cardiac activity as a measure of embryo reproductive potential.

METHODS: In-vitro cultured embryos were selected on the basis of routine morphological assessment and transferred on Day 3. The spent culture media from 127 SET patients was analysed by sub-micellar SDS capillary zone electrophoresis and 15 additional SET culture media samples were analysed by mass spectrometry. An optimal predictive model of embryo reproductive potential was developed from the electrophoretic data using a genetic algorithm and Bayesian classification program. The relative distribution of albumin isoforms was characterized from the mass spectrometry data.

RESULTS: The sub-micellar capillary zone electrophoresis separation produced partial resolution of three albumin species, possibly of different oxidative states, which were correlated to the embryo's reproductive potential. The genetic algorithm selected two regions in the albumin peaks to generate a model that yielded 91.6% sensitivity and 100% specificity. Mass spectrometry revealed that culture media albumin is distributed into at least 10 isoforms and is appreciably oxidized during culture but there is significantly less albumin oxidation for embryos that had high reproductive potential ($p = 0.04$).

CONCLUSIONS: Albumin profiling, by sub-micellar capillary zone electrophoresis, of spent embryo culture media provides a means to identify the reproductive potential of embryos. Embryos with high reproductive potential were associated with decreased IVF culture media albumin oxidation.

Keywords: in-vitro fertilization, embryo reproductive potential, capillary electrophoresis, oxidative stress, Single Embryo Transfer

5.1 Introduction

Infertility is defined as the failure to conceive after 12 months of properly timed, unprotected intercourse and is estimated to affect up to 15% of reproductive age couples [189, 190]. Several assisted reproductive technologies (ART) have been developed to augment reproduction success, with *in-vitro* fertilization (IVF) being one of the most important. At present 1–3% of children born in developed countries are conceived through ART [191]. However, the World Collaborative Report on ART highlighted that the average pregnancy rate was 28.2%, with high-risk multiple fetus pregnancies occurring 28.3% of the time [192-194]. Single embryo transfers (SET) minimize the risk of multiple fetus pregnancy complications [195] and are becoming increasingly common in the US [192, 194], have become conditionally recommended in Canada [196] and are common practice in some European countries [197]. However, for SET to gain widespread adoption, methodologies to accurately characterize the reproductive potential of individual embryos are needed [198]. Non-invasive methods have been hailed as the path forward towards routine embryo quality assessment while posing minimal risks to the embryo [199, 200].

Evaluation of embryo morphology and cleavage rate [201] has been shown to be related to implantation [202-204] and is widely practiced for selecting the embryos with the best implantation potential. However, persistently low pregnancy rates highlight the limitations of these assessments and have spurred interest in alternative strategies to evaluate the reproductive potential of an embryo. Non-invasive measurements of the culture media are yielding useful predictors of reproductive potential [205]. Currently, much of the focus is directed towards assessment of media nutrient consumption. For example, increased glucose metabolism seems to be an indicator of embryo developmental potential and viability [206, 207]. Amino acid metabolic activity appears to be lower in viable embryos compared to those that arrest [207, 208], while DNA damage elevates blastocyst amino acid metabolism [209]. Pyruvate uptake has been reported by some to increase with embryos that develop to blastocyst stage [206, 210, 211]’ but contradictory evidence [212, 213] leaves the predictive usefulness of this species unclear.

Also important is the role of oxidative stress in the IVF setting which has been comprehensively reviewed by duPlessis [214]. Reactive oxygen species, mainly free radicals, are not highly specific in their reactivity and can modify a wide range of biochemicals. Protein, amino acid and lipid oxidation has been extensively studied and linked to a wide range of adult pathologies [215-217]. The presence of cumulus cells improved blastocyst formation and first cleavage rates in bovine IVF and offered a measure of oocyte protection from exposure to hydrogen peroxide [218]. The activity levels of superoxide dismutase, a powerful anti-oxidant, were seven times higher in porcine follicular fluid than fetal bovine serum. Oocytes cultured in media supplemented with 10% porcine follicular fluid showed little DNA damage, lower intracellular glutathione and continued meiotic progression compared to non-supplemented media. When superoxide dismutase activity was selectively blocked, cell damage and reduced blastocyst formation occurred even in the supplemented media suggesting that oxidative stress and reactive oxygen species can significantly alter the reproductive potential of embryos [219].

For the purposes of this paper, HSA oxidation is of particular interest since this protein is used as an IVF media supplement and is a potent antioxidant due to a free cysteine (cys34). Protein thiols are preferred targets for *in-vivo* and *in-vitro* oxidation, both reversible and irreversible, resulting in multiple protein isoforms [220-223]. Oxidized HSA isoforms have been shown to be altered conformationally with greater hydrophobic region exposure [224, 225] and impaired ligand binding [221] and may be prooxidant under specific *in-vitro* conditions [226].

In IVF culture, supplementing the media with recombinant human albumin has been shown to reduce the rates of apoptosis, nitric oxide (NO) generation and to improve fetal mouse development compared to media with human serum albumin (HSA) or polyvinyl alcohol [227] although, no significant difference was found in human culture [228]. Both bovine serum albumin and reduced glutathione were responsible for increased porcine blastocyst formation suggesting the involvement of a protective antioxidant mechanism [229].

Characterizing the oxidatively modified proteins can be difficult because the modifications often result in subtle physio-chemical changes to the protein (*e.g.* tertiary structure, hydrophobicity, charge, molecular weight, *etc.*). Instrumental strategies to characterise isoforms include HPLC [126], SDS PAGE, capillary isoelectric focusing (cIEF) [128], capillary electrophoresis (CE) [127] and mass spectrometry.

Capillary electrophoresis, a simple, quantitative, high efficiency separation technique that finds extensive use in separations of biological samples [230, 231] features speed of separation, non-destructive analysis and low mass detection limits. Separation in CE occurs via differences in analyte charge to hydrodynamic volume and can provide information on analyte conformation. The minute sample requirements of CE have led to increasing use in single-cell analyses of protein expression, nitric oxide release [232], organelle characterization [233, 234], and nucleic acids (mRNA and DNA) [235, 236]. Given its suitability to such measurements, it is surprising that there have been only a few reports of utilizing CE to investigate embryo protein expression patterns using either single [237] or multi-dimensional strategies [238]. By carrying-out CE separations with sub-micellar concentrations of sodium dodecyl sulfate (SDS), protein unfolding and conformations can be subtly probed [239, 240], but this cannot provide conclusive identification of specific protein modifications. The analytical technology with the greatest isoform identification potential is mass spectrometry (MS) [129, 130] due to the high mass accuracies and resolutions achieved [132]. However, intact protein MS is limited in its ability to quantitatively measure proteins with high precision. Some of these difficulties with quantitative MS can be overcome by normalization of the individual isoform MS signals to the total signal from all isoforms. This in turn enables correlations between changes in isoform distribution and a pathological state to be investigated, as has been demonstrated by Bar-Or [108] and our work with amniotic fluid [241].

In this report, sub-micellar SDS capillary electrophoresis was used as a rapid non-invasive method to investigate modifications of albumin in the spent culture media as a surrogate marker for embryo reproductive potential. Positive results from the CE study prompted further investigation of specific oxidative albumin modifications of IVF media and changes to the albumin isoform distribution by mass spectrometry.

5.2 Experimental

5.2.1 Patient population and stimulation protocol

The ethics review committees of Concordia, McGill and VU University medical center gave approval for this study. Patients (n=127 for the CE study and 14 for the MS study) under 38 years of age, or with positive response to previous IVF or intra-cytoplasmic sperm injection (ICSI) treatment, underwent controlled ovarian hyperstimulation using a 'long' protocol with Decapeptyl (Ferring, Copenhagen, Denmark), a GnRH agonist and Gonal F (Serono, Geneva, Switzerland), Puregon (Schering Plough, Oss, the Netherlands) or Menopur (Ferring) gonadotropins. A short GnRH agonist protocol was administered to women over 38 years or those that had a poor response previously.

Vaginal ultrasonography and serum estradiol was used to monitor ovarian response. 10000 IU of human chorionic gonadotrophine (Pregnyl, Schering Plough, Oss, the Netherlands) was administered sub-cutaneously when there was at least one follicle ≥18 mm and three or more follicles ≥16 mm. Ultrasonographic directed oocyte retrieval was performed 36 hours later.

5.2.2 Embryo culture procedure

Insemination of the oocyte was initiated 40 hours after hCG injection using either IVF and/or ICSI procedures. Fertilization was scored 16-18 hours after insemination. Embryos were cultured individually in 25 μl pre-equilibrated drops of IVF media. Embryo-free control drops of media were also incubated at the same time. Individual embryos with the highest number of blastomeres and the least fragmentation were transferred on Day 3 after oocyte retrieval. After transfer, the spent media drops and control drops, were immediately frozen. Positive pregnancy was defined as fetal cardiac activity (FCA) at 12 weeks gestation. For the CE study human tubal fluid (Lonza, Belgium) plus 10% protein solution, (Sanquin, the Netherlands) was used as the media (HTF+) while for the MS study SAGE cleavage IVF media (sIVF) (SAGE In Vitro Fertilization, Inc., Trumbull, CT 06611 USA) was used.

5.2.3 Capillary electrophoresis

Electropherograms were collected at 190 and 198 nm (5 nm bandwidth) at 4 Hz using a Beckman Coulter P/ACE MDQ capillary electrophoresis system (Fullerton, CA, USA). Fused silica capillary (60 cm in length, window at 50 cm, 100 μm inside diameter, 365 μm outside diameter) was from Polymicro Technologies (Phoenix, AZ, USA). The capillary was prepared for each sample injection by filling with borate buffer (2 minutes at 1.4 bar) and conditioned under 25 kV for 1 minute (0 bar). Frozen HTF+ culture media samples were selected randomly and thawed in ice-water, diluted with 2 volumes of water and injected hydrodynamically (34.5 mbar, 10 s). Samples were separated by applying 25 kV, at 20ºC, with 75mM borate, pH 9.25 and 5 mM sodium dodecyl sulphate (SDS) buffer as the background electrolyte. The capillary was conditioned in-between runs by flushing for 1 minute at 1.4 bar followed by a 0.5 minute wait at 0 bar with 5 mM SDS then 100 mM NaOH. All prepared solutions were filtered (0.45 μm) and degassed.

Commercial IVF media (G-1-3 PLUS, Vitrolife, Englewood, CO, USA) was used for method development, optimization and subjected to ten -85 ºC freeze-thaw cycles to determine effects of freezing.

5.2.4 Mass Spectrometry

Stock sIVF media, control drops and spent media drops were thawed on wet ice and low molecular weight species, including salts, were removed from a 15 µl sample using a 30 kDa nominal molecular weight cut-off Ultrafree centrifugal filter (Millipore, Billerica, MA, USA). The samples were rinsed with water five times and the albumin recovered in 20 µl of water. An aliquot of the HSA isolate (1 µl) was loaded onto a Waters NanoEase C18 trap column and eluted with a 16 minute linear gradient from 100% A (97% H_2O/3% acetonitrile/0.1% formic acid) to 90% B (3% H_2O/97% acetonitrile/0.1% formic acid) at 2 µl/min. Mass spectra were collected on a Waters Q-Tof2 system by nano-electrospray ionization with the capillary voltage set at 3.5 kV and a cone energy of 35V. The MaxEnt1 algorithm (MassLynx v. 4.0), over the 1200 to 1650 m/z range of the raw spectra, was used to calculate the deconvoluted spectrum of HSA.

5.2.5 Data processing

All of the CE data processing procedures were carried-out using Matlab (The MathWorks Inc., MA USA). The separation window, containing the sample related peaks, from 6.5 to 11 minutes was extracted from the electropherogram. All of the electropherograms were normalized to a maximum of 1 and aligned to the first sample in the dataset as a reference electropherogram using correlation optimized warping (COW) with window and slack parameters set to 20 and 2 points (5 and 0.5 seconds worth of data) respectively [32, 242]. The COW aligned data was then Haar wavelet transformed from the 1024 data points that encompassed the eluted species, to 256 wavelets [43]. The transformed dataset and FCA outcomes were analysed using a genetic algorithm and Bayesian statistical model program described in detail elsewhere [43]. Briefly, individual models were built with 2-5 variables created from an initial population of 75 randomly selected variables out of the 256 possible variables (wavelets). The fitness of the individual model was evaluated using a full leave-one-out cross-validation strategy to generate the sensitivity and specificity values. Models with the highest combined sensitivity and specificity were retained and evolved through 300 generations to yield the optimal model.

The MS data was processed by deconvolution of the raw spectra using the MaxEnt1 algorithm and peaks corresponding to various HSA isoforms were integrated. The isoform fractional composition was calculated by dividing each isoform's integrated peak area by the total albumin peak area. Statistical analysis of the isoform distribution was carried out in Matlab with a significance threshold set at p = 0.05. A random permutation test was done to determine if the results of the student t-test obtained are statistically different than student t-test calculated with random permutation of outcomes. A total of 1000 randomized outcomes were considered, for each of the entries in Table 5 below, to determine a population mean and standard deviation.

5.3 Results

The SDS concentration in the CE background electrolyte was the most important variable to the overall separation. Figure 18 shows that at 5 mM SDS, albumin was partially resolved into three species while at lower concentrations of SDS, albumin migrated as a single peak, or a poorly resolved set of peaks (*e.g.* 3 mM in Figure 18).

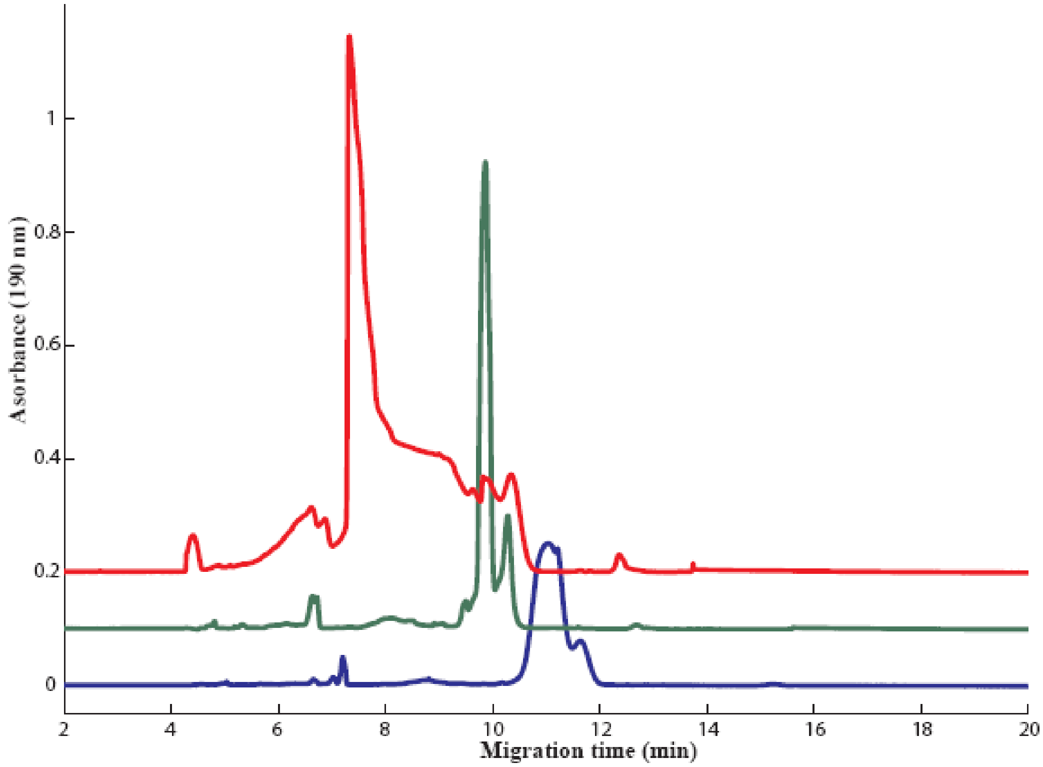


Figure 18: Electropherograms of unused (Vitrolife) media solutions separated with different amounts of SDS in the background electrolyte on a 65 cm capillary. The buffer was 75 mM borate with 3.5 mM (bottom trace), 5 mM (middle), and 10 mM (top) SDS. Traces were offset by 0.1 AU for clarity.

Whereas, at concentrations just above the critical micellar concentration (CMC) of 6-7 mM resulted in irreproducible separations (e.g. 10 mM as shown). The optimal SDS concentration was 5 mM. Borate concentration was also varied, but at concentrations lower than 50 mM peak broadening was found while at concentrations higher the 75 mM there was little improvement in peak shape and the separation was slower (data not shown). The highest resolution separation, in 75 mM borate and 5 mM SDS, was observed with a capillary temperature of 20 ºC. Higher temperatures resulted in faster separations but with significantly lower resolution of the albumin species (data not shown). The optimized sub-micellar CZE separation produced a rapid separation with all species migrating in less than 11 minutes. The most prominent feature in the electropherogram was the albumin but some amino acids and media ingredients were reliably detected and identified in the unused media (see Figure 19) but were below the detection limit in the spent media. Subjecting the media to multiple freeze-thaw cycles produced no detectable effects on the separation profile.

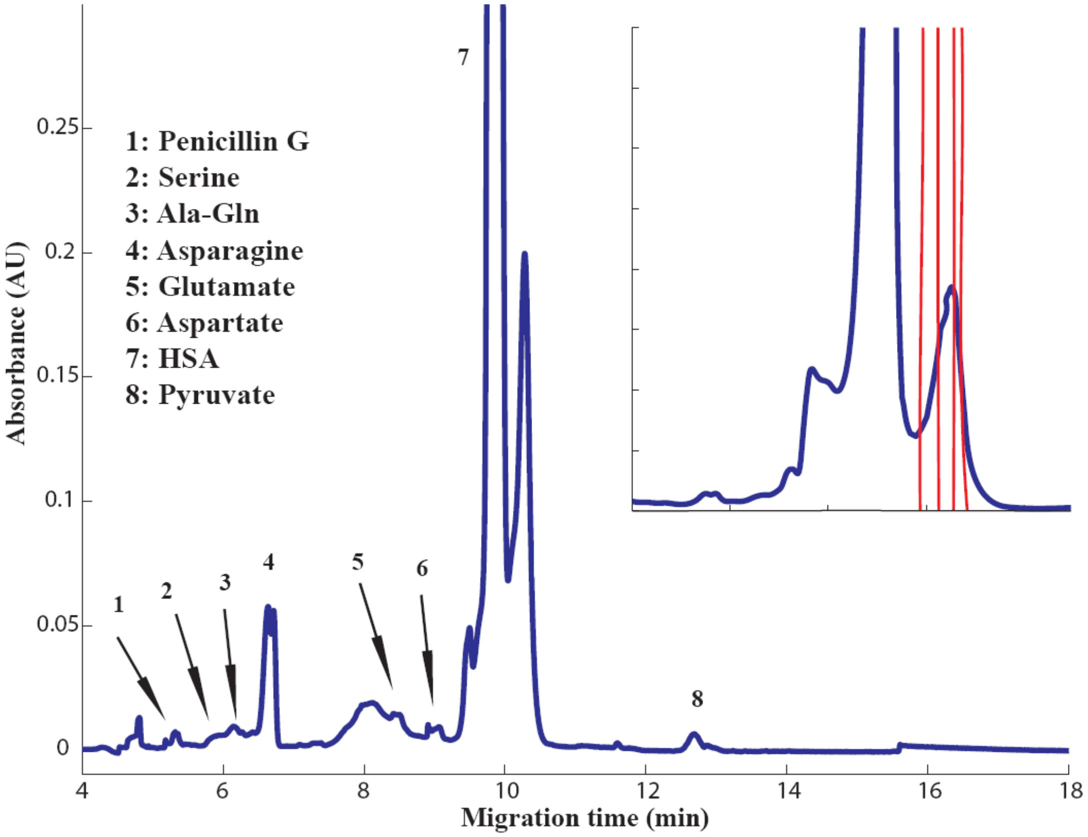


Figure 19: An electropherogram of unused media diluted with two volumes of distilled water and injected hydrodynamically onto the capillary. The separation was carried-out at 25 kV using 75 mM borate + 5 mM SDS at pH 9.25 and measured at 195 +/- 5 nm. Locations of some known media ingredients noted on the electropherogram were determined from spiking experiments. Inset shows part of a spent media electropherogram with the two regions selected by the genetic algorithm, in red, as being predictive of implantation.

In the 127 SET samples analysed by CE, the overall pregnancy rate was 28.4% with no multiple pregnancies. After data alignment, normalization and Haar transformation, the genetic algorithm produced a model, using full leave one out cross validation, with relative embryo viability scores that correlated to the embyro's reproductive potential. A model using two wavelets in albumin's third peak, as shown in Figure 19, was deemed to be the most parsimonious with 33 true positives, 91 true negatives, 0 false positives and 3 false negatives yielding a 91.6% sensitivity and 100% specificity. The mean relative viability scores (and standard deviation) developed by the model were: 0.534 (0.133) for the positive and 0.482 (0.0484) for the negative SET results. Figure 20 illustrates the classification results in a box plot format.

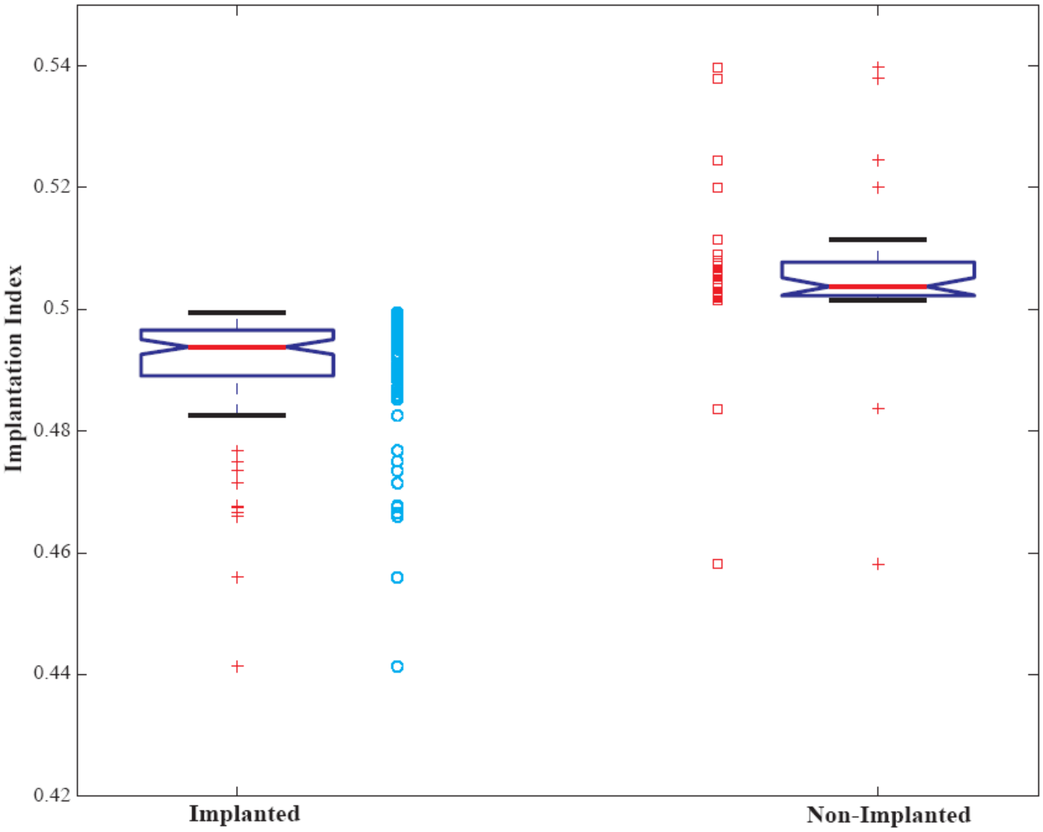


Figure 20: Box and whisker plot of two wavelet model showing differentiation of non-implanted (circles) and implanted (squares) embryos. See text for details of interpretation.

Figure 21 shows the deconvoluted mass spectrum of albumin in sIVF media with known HSA modifications labelled. Similar spectra were obtained from the HTF+ and VitroLife G-1-3 PLUS, (Vitrolife, Englewood, CO, USA) media but with different peak intensities (data not shown). In the sIVF samples, reduced or mercapto-albumin (HSA-SH) was expected at a mass of 66437 Da and was observed at an apparent mass of 66436 Da (peak f Figure 4) well within the 20 ppm mass error of the measurement. Observed, and identified, albumin modifications included oxidation of one of HSA's thiols to sulfinic acid (Δ m/z +32, HSA-SO_2H, peak g) and S-cysteinylation (Δ m/z +119, HSA-cys, peak k) [113]. Also observed was oxidation of a disulfide to produce HSA with a sulfinic acid and with di-sulfenic acids (Δ m/z +66, $(HOS)_2$–HSA–SO_2H, peak h) and S-cysteinylation with di-sulfonic acid (Δ m/z +222, cys-HSA–$(SO_3H)_2$, peak n). Additionally, S-cysteinylation of the dehydroxyanaline isoform of HSA recently reported by Bar-Or *et al.* was observed (Δ m/z +85, HSA[DHA]-cys, peak i) [108]. Glycation of albumin by hexoses (Δ m/z +161, HSA-hexose, peak l) was also observed.

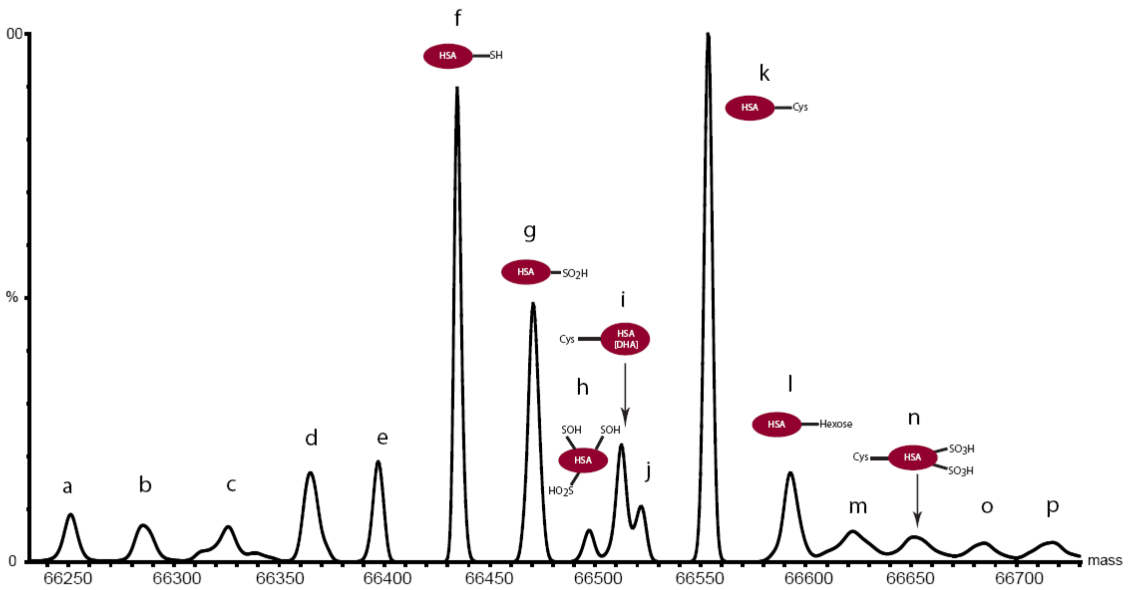


Figure 21: Deconvoluted mass spectrum of albumin isolated from IVF media. Protein masses were calculated using maximum entropy deconvolution. See text for peak identification.

In the 14 SET samples subjected to MS analysis, the pregnancy rate was 43%. Analysis of the percentage isoform distribution of the sIVF stock media (not subjected to culture conditions) showed that only 17.8±0.7% (n=5) was mercapto-albumin (peak f) whereas cysteinylated (peaks k+n) and irreversibly oxidized albumin (peaks g+h) made-up 20.8±0.4% and 13.4±0.3% respectively. One-way ANOVA also revealed a statistically significant difference in the level of irreversible oxidation between the stock media (13.4±0.3%), the control (14.8±0.3%) and spent media (14.3±0.4%) ($p = 0.000005$). We examined if single, or combinations (sums, products, ratios) of isoforms yielded significant differences between embryos that implanted and those that did not. Inverse least-squares (ILS) was also evaluated with the multiple parameter models. The models with the fewest number of peaks required to achieve significance ($p < 0.05$) are shown.

The simplest ILS model that achieved significance is shown only when the peaks differ from those found with non-ILS models.

In the first round of analyses, only peaks from identified HSA species (e.g. peaks f, g, h, i, k, l and n) were used to build the models and are summarized in Table 5. In the second round of analyses, all of the peaks detected in Figure 21 were used to build the models and are summarized in Table 6.

Based on the random permutation test the null hypothesis was rejected for α=0.05 for the peak combination of peaks *g* and *h*, meaning the classification model obtained above is statistically different from the population of models generated with a random permutation of outcomes and the classification model obtained is not likely to be due to chance alone. The other entries in Tables 5 and 6 were not subjected to the permutation due to time constraints and will be the subject of future work prior to publication of these results.

Table 5 : Statistical analysis results of identified peaks based on calculations using the normalized MS peak areas.

Model type	Identified peaks from Figure 21	Parameter calculated from the normalized areas (implant)	Parameter calculated from the normalized areas (non-implant)	Standard deviation (implant)	Standard deviation (non-implant)	p-value
Single peak*	h	0.009	0.012	0.001	0.002	0.06
Sum of peaks and t-test	g + h	0.140	0.145	0.002	0.005	0.04
Two peaks, ILS and t-test	h & n	0.3	0.8	0.2	0.3	0.01
Product and t-test	h * k	0.0020	0.0024	0.0001	0.0004	0.04
Ratio and t-test	h / n	0.48	0.61	0.09	0.10	0.02

* The statistical model using a single peak did not achieve $p < 0.05$ but is included here for comparative purposes only.

Table 6: Statistical analysis results of all peaks based on calculations using the normalized MS peak areas.

Model type	Peaks from Figure 21	Parameter calculated from the normalized areas (implant)	Parameter calculated from the normalized areas (non-implant)	Standard deviation (implant)	Standard deviation (non-implant)	p-value
Single peak	d	0.063	0.066	0.001	0.002	0.009
Sum of peaks and t-test	d + h	0.072	0.077	0.002	0.004	0.007
Two peaks, ILS and t-test	d & p	0.3	0.8	0.2	0.3	0.003
Product and t-test	d * g	0.0082	0.0088	0.0002	0.0005	0.01
Ratio and t-test	h / o	0.7	1.0	0.1	0.3	0.02

5.4 Discussion

In this study, capillary electrophoresis, with a sub-micellar detergent background electrolyte, provided a simple and rapid method of characterizing the spent IVF media. Several amino acids and other media ingredients were visible in the electropherogram (Figure 19) of the stock media but many were below detectable levels, especially in the spent media. In particular, the amino acids correlated to reproductive potential (asparagine, leucine, glycine [243], and glutamate [244]) were either not well-resolved from other amino acids or were below detectable levels with these separation conditions. Albumin was the most prominent feature in the electropherogram and was partially resolved into three peaks by the surfactant. Sodium dodecyl sulfate acts on proteins by binding to accessible positive and hydrophobic regions. At high concentrations, SDS denatures proteins but, at sub-micellar concentrations the proteins are not completely denatured and the various protein conformations can be probed by CE [245, 246]. In the present case, albumin-SDS binding gave rise to increased differences in the charge to hydrodynamic volume ratio and improved the electrophoretic selectivity over non-SDS containing electrolytes.

Using the HTF+ media separation data, the genetic algorithm selected two regions of the partially resolved albumin as being correlated to, and hence predictive of, FCA and reproductive potential. Inclusion of additional variables/wavelets in the model did not significantly improve the results (*i.e.* sensitivity and specificity) and were rejected. Normalization of the electropherogram data to the peak maximum (of albumin) discounts the possibility that the observed effect was due to changing albumin concentration but was rather a change in the relative distribution of the albumin species as resolved by submicellular CE. Identification of the exact albumin species selected by the genetic algorithm would require further analysis. One route to such an identification is by CE with on-line mass spectrometric detection but both the SDS and borate buffer are known to cause severe ionization suppression in mass spectrometry [247, 248]. Alternatively, multiple fraction collection, dialysis and MS analysis is in principle a viable route to identification of the specific modification(s) but is very time consuming due to the very small volumes injected. However, the later migrating peaks in CE are associated with more negatively charged species such as those that have been oxidized. In albumin, the most readily oxidized amino acid is Cysteine 34 but oxidation of any of the 17 disulfide bridges could also produce acidic isoforms.

We speculated that an oxidative modification of HSA seemed likely given albumin's high concentration and susceptibility to oxidative modification. Previous work with near infrared spectroscopy also implicated an oxidative mechanism but was unable to identify the specific biochemical species involved [249]. Instead of pursuing a challenging CE-based strategy for identification of albumin's *in-vitro* modification we decided a direct measurement of albumin's isoforms by MS was warranted. However, by the time we had developed the MS technique the IVF culture protocol at VU had changed to using the Sage media [241]. Nonetheless, we found that all of the commercial media (VitroLife G-3, Sage Cleavage media and the HTF+) all showed a similar number of HSA isoforms and an oxidized albumin base peak in the stock media mass spectrum. A similar predominance has been observed in other pharmalogical preparations [223] containing albumin and contrasts with serum where mercapto-albumin is normally the predominant isoform (70-80% of all albumin) [129, 130, 143, 144] Both "reversible" (cysteinylated) and “irreversible” oxidative (*e.g.* sulfinic acid) modifications of albumin were observed, however, in IVF culture media, all oxidation is expected to be irreversible in practice since neither glutaredoxin nor thioredoxin, the enzymes responsible for reducing mixed disulfides *in-vivo*, were present in the media [147, 148, 250].

Media cultured with an embryo showed less irreversible oxidation than the control drops highlighting that embryos employ active antioxidant processes [250-252] but significantly higher levels of irreversible oxidation were found in the samples where implantation did not occur. This suggests that incompetent embryos were less capable of mounting an antioxidant response to the cumulative oxidative insult brought about by the metabolic processes and culture conditions [214]. This may result in inactivation of key enzymes [253], protein oxidation, modification of nucleic acids and depletion of finite stocks of antioxidants [254]. Results from the examination of the best combinations of isoforms to differentiate between embryos with high and low reproductive potential also highlighted the importance of oxidative stress. Table 5 shows that when the statistical analysis was restricted to the peaks that have been identified, only oxidized HSA isoforms appeared in the models (irreversibly oxidized peaks h, n and g and reversibly oxidized peak k). It is also noteworthy that the irreversibly oxidized HSA associated with peak h achieved a $p = 0.056$ when evaluated by the t-test. When the statistical analysis was expanded to include all the peaks shown in Figure 21, Peak d was significant by itself and was included in nearly all the combinations as shown in Table 6. The biological significance of these results is unclear as peaks d, p and o require identification. In general, statistical analyses with greater numbers of parameters and/or using ILS yielded smaller p values than those reported in both tables; however, given the small number of measurements (14) the models with the fewest parameters to achieve significance are shown.

In summary, the commercial media used in IVF have many different formulations but they all attempt to meet the metabolic needs of the developing embryo [255]. Media formulations supplemented with albumin all displayed characteristic patterns of modifications due to oxidation but this was modulated by the embryo. These modifications resulted in differences in the electrophoretic profile, using sub-micellar separation conditions, and provided a simple and objective method of assessing the reproductive potential of an embryo with high sensitivity and specificity. The specific HSA isoforms were evaluated directly by MS and statistical analysis suggested an oxidative stress mechanism that resulted in selective HSA modifications. In both the CE and MS analyses, use of the relative albumin signal rather than the absolute values was necessary and should be robust in the face of varying culture media compositions and experimental variations.

Non-invasive assessment of embryo reproductive potential is crucial for the continued development of ART and minimizing the risks of multiple fetus pregnancy. Only when the embryologist is armed with a reliable method of assessing reproductive potential, can accurate decisions be made about which specific embryos should be transferred.

5.5 Acknowledgements

We would like to thank all members of the IVF team from the VU University Medical Center in Amsterdam, especially the laboratory technicians and embryologists, for their assistance. We are grateful to Molecular Biometrics LLC for facilitating access to the samples.

5.6 Funding

This study was partially funded by Molecular Biometrics LLC, New Jersey, USA.

Chapter 6

Conclusions

A chemometric method was developed to perform untargeted biomarker discovery in biological samples. The underlying goals were to maximize the information extracted from biological samples and search for biomarkers in an unbiased way. The method was designed in such a way that the analysis process required minimal user input to achieve successful classification and potential biomarker discovery. Variables/species of interest were further investigated to determine the nature of the biochemical species giving rise to the corresponding signal. Identification of putative biomarkers found in untargeted experimental design is one of the greatest challenges for this approach. Biological samples are complex by nature, containing genomic, proteomic and metabolic species in great numbers and diversity. This work gives an indication of what can be achieved, but represents a small fraction of the potential of such untargeted approaches towards biomarker discovery.

Modern analytical instrumentation can process large numbers of samples and produce information rich data sets. The traditional data analysis strategies employed by analytical chemists do not allow for efficient data processing. To analyze these large datasets, data preprocessing and processing strategies appropriate for separation science data need to be employed.

Simple peak integration, a mainstay in analytical chemistry, misses a large part of information present in data where multiple species contribute to the signal (peak) such as collected with CE. The problem is that the electrophoretic profile showed characteristics that were not compatible with pattern recognition and data mining chemometrics. The application of COW to the electrophoretic data made it possible to proceed with further chemometric data processing, classification and to locate potential biomarkers.

The Haar transform proved to be a very convenient and efficient way of compressing the electrophoretic data into a manageable number of variables for the genetic algorithm and Bayesian classification routine. Other data simplification or transforms are certainly viable and can accomplish the same task, but the square wave has the significant advantage that interpretation of the selected region is unambiguous.

The most exciting outcome of the Bayesian classification strategy is the successful differentiation of GDM from non-GDM, AGA from LGA and, for the IVF study, between implanted and non-implanted embryos. Furthermore, important biological findings came from the investigation into the chemical nature of retained variables. This tends to show that the same strategy could be applied to a wider variety of sample/outcome pairs to facilitate the design of new diagnostic tests, but also to reveal important biomarkers.

Untargeted biomarker discovery strategies can be seen as a way to cast a much larger net by designing the data generation process to maximize the number of analytically detected species. This allows a maximal number of species to be tested with regards to a possible link with the disease and, as such, includes all detected species as potential biomarkers, regardless of *a priori* biological knowledge of the sample or the investigated disease/outcome. In practice, it is the number of signals that are maximized, signals for which the corresponding biochemical species may not be identified/known.

Characterizing and identifying molecules that are potential biomarkers is the current bottleneck for this type of strategy. This is exemplified by the efforts to identify the unknown small molecule relevant to predicting LGA (Chapter 4 and Appendix III).

Designing separation methods with the idea of maximizing the chances of identifying the source of all signals present appears to be crucial for biomarker discovery. And so, in addition to the classic resolution/separation time optimization, and considerations with regards to sufficient resolution with respect to chemometric data analysis, the optimized separation conditions should be compatible with current molecular characterization tools. Indeed, MS and NMR are excellent tools to identify and characterize molecules, but require sample matrices that can be incompatible with the separation of biological fluids by CE. Thus, the samples, buffers and additives used should not only be evaluated with respect to the quality of the separation, but also with the risks of interfering with identification by MS and/or NMR.

In this study, capillary electrophoresis was successfully employed to search for biomarkers in AF. The separation provided sufficient resolution to separate, or partially separate, the major UV absorbing components in AF and produce quality data. The major disadvantage with CE lies in the above mentioned point involving the identification of unknowns of interest. Yet, it has the potential of being used as a diagnostic tool as it is less expensive than HPLC and can more readily be miniaturized into a small point-of-care type of instrument, a so-call lab-on-a-chip.

Human serum albumin in AF appears to be an important predictor for both GDM and LGA. In GDM, the increased oxidative stress associated with hyperglycemia was reflected by an increased level of irreversible oxidation of HSA and relative decrease in levels of cysteinylated HSA. This may be the reason for the difference in the CE signals for the AF HSA peak for samples with and without GDM. Changes in the distribution of the HSA isoforms present is expected to induce migration shifts and peak shape distortions.

In the IVF spent media analysis, the oxidation levels of HSA were again important. Both the CE-UV and MS analysis of the media indicates that the environment of the embryo is modified during the embryo culture. Additionally, the HSA oxidation profile of the spent media was different for samples where the implantation was successful and produced a positive fetal cardiac activity assessment. This is particularly important as it indicates that signs of successful implantation are already present in the IVF media before the fertilized egg is returned to the womb.

Human serum albumin is a marker of oxidative stress as shown by the MS analysis of AF HSA and of the HSA in IVF media. It does not appear to be specific to GDM as this has been reported elsewhere [108, 256] for other diseases. Nevertheless, determining the relative abundance of HSA isoforms, more specifically those indicative of oxidation, could be a very useful clinical diagnostic measurement, especially when combined with other biochemical measurements and risk factors.

Gestational diabetes mellitus is detectable around the 15th week of gestation using oxidized HSA as a marker. Both GDM and LGA are detectable by week 15 using the CE analysis method. Yet currently, screening for GDM is done around the 26th week of gestation and screening for LGA is done with sonography much later in pregnancy [171-

173]. In both cases, it could be possible to detect and diagnose these conditions earlier and take appropriate action to mitigate the consequences of each condition.

Untargeted biomarker discovery is a powerful strategy to survey biofluids for important biomarkers of disease/abnormal conditions especially when only a few biomarkers have been identified. Yet, for this approach to be accepted, it cannot, and should not, be seen as a standalone solution to biomarker discovery. Signal(s)/species identified as being a biomarker(s) of a certain disease need to be tested and validated by separate corroborating experiments. To provide convincing evidence, biomarker discovery experiment should include not only the untargeted survey portion, but also a complementary hypothesis driven portion to confirm biomarkers identified during the first portion.

Chapter 7

Future Work

The most pressing future work involves the characterization of the unknown small molecule involved with predicting LGA. This will be briefly discussed in the next section. The following sections will cover the long term future works that are directly spawned from this thesis and will be outlined as future projects that others could pursue.

7.1.1 Characterization of unknown small molecule in amniotic fluid

Continued investigation of filtered AF (<5kDa) is preferred to simplify the sample and eliminate protein interferences with the separation. Efforts already taken in this direction are described in Appendix III. Work has begun to optimize the CE separation of the low molecular weight fraction (LMWf) of AF. Collecting peaks with improved resolution could eliminate some of the comigrating peak that can be collected into the same fraction as that of the peak of interest, thus simplifying the MS characterization of all species found in the fraction.

The most viable and useful option at the moment is the development of a HPLC-UV method to separate the LMWf of AF. A HPLC method could be used for the identification of the unknown molecule from the LGA study and would be useful for future experiments on AF (see the following section). Injections in HPLC have approximately 1000 times larger volume than in CE and by consequence have 1000 times more mass of analytes per injection and thus larger mass of analyte collected in each fraction. To locate the fraction in which the unknown is present, collected fractions can be analyzed by CE. Only the fraction that gives rise to a CE signal with the appropriate migration time (such as in Figure 17) would then be analyzed by MS. The increase mass of unknown may also make NMR a possible characterization tool.

In parallel, literature and metabolite databases are continuously revisited for new candidate molecules that show physico-chemical properties that correspond to the CE migration time should be investigated. On the basis of this literature work, CE spiking experiments can be used to confirm or eliminate candidate molecules.

Alternatively, developing a CE-MS method in negative mode (MS) [257, 258] would provide direct detection and characterization of analytes present in the LMWf of AF. This would be highly informative and of great interest, yet CE-MS is not as well established as LC-MS and the current MS facilities at Concordia University are not well equipped for coupling of CE to MS.

7.1.2 Paired serum/AF samples

One of the obvious limitations of the AF experiments is that amniocentesis comes with increased risk to the fetus and, as such, is only performed when the benefits outweigh the risks. Maternal serum offers a potential window into the fetal compartment, as exchanges between the fetal and maternal compartment go both ways. The Bar-Or paper on IUGR [108] is an example of this. It would be extremely interesting to look at paired samples of maternal blood and AF, perhaps even maternal urine, to see if biomarkers present in AF are quantitatively detectable in maternal fluids obtained at the same time by much less invasive methods (blood and urine sampling). This is a considerable task to manage since would it requires handling, acquiring data and analyzing data for a large numbers of paired samples.

Each biofluid should be prepared in a similar fashion, except for blood as it is preferable to work with the serum fraction of blood. There is a net advantage to fractionating samples into categories in terms of molecular weight as metabolites and proteins require different separation conditions. As such each type of biofluid could be filtered into 2 or 3 fractions: LMW (<5 kDa or even <1 kDa), medium molecular weights (MMW; >5 kDa and <60 kDa) and above, high molecular weight (HMW; >60 kDa). Beyond this step, it would be worth considering further sample preparation steps depending on what type of instrumental analysis are to be carried out.

Proper sample storing, labeling and aliquoting are critical. It may seem trivial, but large scale studies with several different instrument operators can get quickly get chaotic. Storing is pretty obvious, although special thought should be put into organizing and centralizing all aliquots and samples. Labeling should follow a basic scheme, such as "principal investigator initials"-"operator initials"-"sample number (at least 4 digits)". Finally, determining the number of aliquots and volumes required per sample is important. Here are some examples of approximate volumes that should be set aside: for CE about 10-20 µL/aliquot; for HPLC 20-40 µL/aliquot; for HSA isolation ~400 µL/aliquot (AF) or ~40 µL/aliquot (serum).

The choices of possible instrumental analysis for such samples are fairly large, especially when considering the possible combinations of separation science instruments and detectors. I would consider CE again as it requires very small volumes and minimal new method optimization. Yet, it would be difficult to choose CE over HPLC as HPLC offers a more robust and reproducible separation and would provide new information. More reproducible data means simpler data preprocessing and less noisy data. Additionally, HPLC is very easily coupled to a MS making it a very appealing choice for direct characterization of unknowns. Selecting an appropriate stationary phase might take some optimization for the small molecules, but for proteins, a C8 stationary phase should provide sufficient retention for the water soluble proteins that compose the MMW and HMW fractions. Efforts should be made in the method development step to interface the HPLC to MS and try to characterize the high abundance molecules before the analysis of samples begins on a large scale.

As for detection, using UV detection offers some important advantages over MS or fluorescence detection in terms of simplicity, costs and reproducibility, but comes at the cost of higher limits of detection. The lower detection limits of MS and fluorescence detection should give rise to increased chemical rank in the data set, which is good in itself for untargeted biomarker discovery, but it can introduce increased noise and a much higher variable to sample ratio.

The data analysis steps do not necessarily need to change. All the algorithms developed and used for this thesis could be applied to any data generated by CE or HPLC. That being said, once quality data is obtained, a great variety of chemometric data analysis routines can be applied to the data set to find interesting patterns or to test new routines against real data. Some possible modification/improvements of the current sets of chemometric data analysis routines are suggested below.

Again, this is a large and demanding undertaking, but the information gained from each type of biofluid on its own is worth the effort, not to mention the potential to learn from the comparison between each fluid. Furthermore, analyzing all these types of fluids as one large project should take less time than having each type of sample as a segregated project.

7.1.3 Longitutinal study on maternal serum

The previous section discussed the possibility of paired analysis of different biological fluids drawn once. Another interesting angle is to consider maternal serum drawn at different times during the pregnancy and try to establish patterns, not only between groups with different conditions, but also in time. All of the considerations in the previous section with respect to sample handling and preparation, data generation, data analysis apply.

7.1.4 HSA isoform analysis

One of the important findings of this thesis work is that HSA is a proxy for the oxidative status of a biological media (AF and IVF media). Future work along this line of investigation can involve: (1) a refinement of the method for the determination of abundance or relative abundance of HSA isoforms; (2) application of the current method to other biofluids, such as cerebrospinal fluid or biofluids of animal origins that contain albumin; and (3) extension of the approach to other proteins that can be used as markers of oxidative stress in a biological sample.

In the early 1990s, Suzuki *et al.* [256] published a paper where they obtained a partially resolved triplet peak for HSA by HPLC on a proprietary Asahipak GS-520H column. They attributed these to the reduced form of HSA as well as to 2 unidentified oxidized forms of HSA. It would be interesting to use some form of separation that can resolve the isoforms in the hopes of getting a better signal-to-noise for the MS detection, possibly iso-electric focusing.

Alternatively, labeling strategies involving HSA's cysteine 34 could be exploited to get a relative measurement of reversibly oxidized HSA. This could be achieved by complete oxidation of the remaining free thiols to ensure that no mercaptoalbumin remains in solution, followed by the displacement of the cysteines on cysteine 34 by a mass or fluorescent label. If a fluorescent label is used, a second label that will non-specifically target HSA could be used to determine the total HSA concentration, thus giving the relative abundance of reversible oxidation over the sum of all the HSA isoforms.

Other proteins are susceptible to oxidation and might prove to be better markers of OS or more specific markers of a condition than HSA can be. Any free thiol containing protein, such as the α1-antitrypsin present in AF and serum, could be considered as a candidate biomarker for OS.

7.1.5 Refining the chemometrics tools

The code that allows for the chemometric data analysis strategies employed in this thesis can be used again with little modification. Nevertheless, there is room for improvement and added functionality. For the moment, it is not user friendly and requires significant understanding of computer programming and chemometrics to be correctly used. Work to simplify and clean up the algorithm employed would benefit present and future users as they may not always have the detailed knowledge of the code to benefit from it.

The algorithm should be expanded to include more complex data preprocessing strategies for normalization, baseline and artifact correction. Additionally, adapting COW to align multidimensional data would be highly beneficial in the context of ever larger multidimensional datasets from ever more sophisticated analytical instrumentation. Very recently (2011), a new algorithm that appears to perform better than COW was recently published and is referred to as icoshift [259]. It would be very interesting to employ this new algorithm and evaluate if it can improve the quality of the alignment.

The Bayesian algorithm should be modified in different ways to account for the nature of the data. There is no guarantee that measurements cluster according to a normal distribution and as such log-normal, logistic, Poisson distribution, *etc.* could be explored in the classification model. The classification itself could be expanded to several groups (not only 2) as diseases are much more complex than present or absent. A case in point is the SGA, AGA and LGA classification of births.

Finally, the genetic algorithm is indifferent to the benefit function it employs (currently Bayesian). There are several other relevant pattern recognition strategies that could be made available to the user and might be more appropriate depending on the classification problem at hand, such as principal component analysis, hierarchical clustering and k-nearest neighbours to name but a few.

It is an exciting time for untargeted biomarker discovery. The rapidly increasing sophistication of analytical instrumentation provides the means to more completely capture the complexity of biological samples and produce information rich data sets. The matching continual increase in computational power allows for these rich data sets to be analyzed by chemometric data analysis routines. Clearly, untargeted biomarker discovery is a field in its infancy, but as it matures, it has the potential to identify novel biomarkers and open up new avenues of investigation in chemistry, biology and medicine.

References

1. Barry, M.J., *Screening for Prostate Cancer — The Controversy That Refuses to Die.* New England Journal of Medicine, 2009. **360**(13): p. 1351-1354.
2. Schröder, F.H., et al., *Screening and Prostate-Cancer Mortality in a Randomized European Study.* New England Journal of Medicine, 2009. **360**(13): p. 1320-1328.
3. Andriole, G.L., et al., *Mortality Results from a Randomized Prostate-Cancer Screening Trial.* New England Journal of Medicine, 2009. **360**(13): p. 1310-1319.
4. Sarrats, A., et al., *Differential percentage of serum prostate-specific antigen subforms suggests a new way to improve prostate cancer diagnosis.* The Prostate, 2010. **70**(1): p. 1-9.
5. Gossain, V.V. and S. Aldasouqi, *The challenge of undiagnosed pre-diabetes, diabetes and associated cardiovascular disease.* International Journal of Diabetes Mellitus, 2010. **2**(1): p. 43-46.
6. Melander, O., et al., *Novel and Conventional Biomarkers for Prediction of Incident Cardiovascular Events in the Community.* JAMA: The Journal of the American Medical Association, 2009. **302**(1): p. 49-57.
7. Nanda, S., et al., *Prediction of gestational diabetes mellitus by maternal factors and biomarkers at 11 to 13 weeks.* Prenatal Diagnosis, 2011. **31**(2): p. 135-141.
8. Nanda, S., et al., *Maternal serum adiponectin at 11 to 13 weeks of gestation in the prediction of macrosomia.* Prenatal Diagnosis, 2011. **31**(5): p. 479-483.
9. Kullo, I.J. and L.T. Cooper, *Early identification of cardiovascular risk using genomics and proteomics.* Nat Rev Cardiol, 2010. **7**(6): p. 309-317.
10. Green, E.D. and M.S. Guyer, *Charting a course for genomic medicine from base pairs to bedside.* Nature, 2011. **470**(7333): p. 204-213.
11. Gerszten, R.E. and T.J. Wang, *The search for new cardiovascular biomarkers.* Nature, 2008. **451**(7181): p. 949-952.
12. Rubakhin, S.S., et al., *Profiling metabolites and peptides in single cells.* Nat Meth, 2011. **8**(4s): p. S20-S29.
13. Baker, M., *Metabolomics: from small molecules to big ideas.* Nat Meth, 2011. **8**(2): p. 117-121.
14. Roux, A., et al., *Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review.* Clinical Biochemistry, 2011. **44**(1): p. 119-135.
15. Brown, S.D., et al., *Chemometrics.* Analytical Chemistry, 1996. **68**(12): p. 21-62.
16. Lavine, B.K., *Chemometrics.* Analytical Chemistry, 1998. **70**(12): p. 209-228.
17. Lavine, B. and J. Workman, *Chemometrics.* Analytical Chemistry, 2006. **78**(12): p. 4137-4145.
18. Lavine, B. and J. Workman, *Chemometrics.* Analytical Chemistry, 2008. **80**(12): p. 4519-4531.
19. Lavine, B. and J. Workman, *Chemometrics.* Analytical Chemistry, 2010. **82**(12): p. 4699-4711.
20. Baumgartner, C., et al., *Bioinformatic-driven search for metabolic biomarkers in disease.* Journal of Clinical Bioinformatics, 2011. **1**(1): p. 2.
21. Landers, J.P., *Handbook of Capillary and Microchip Electrophoresis and Associated Microtechniques*. 3rd ed, ed. J.P. Landers. 2007: CRC Press. 1592.

22. Frost, N.W., M. Jing, and M.T. Bowser, *Capillary Electrophoresis.* Analytical Chemistry, 2010. **82**(12): p. 4682-4698.
23. Geiger, M., A.L. Hogerton, and M.T. Bowser, *Capillary Electrophoresis.* Analytical Chemistry, 2011. **84**(2): p. 577-596.
24. Helmholtz, H., *Studien über electrische Grenzschichten.* Annalen der Physik, 1879. **243**(7): p. 337-382.
25. Stutz, H., *Protein attachment onto silica surfaces – a survey of molecular fundamentals, resulting effects and novel preventive strategies in CE.* ELECTROPHORESIS, 2009. **30**(12): p. 2032-2061.
26. Rosenholm, J.M., et al., *On the Nature of the Brønsted Acidic Groups on Native and Functionalized Mesoporous Siliceous SBA-15 as Studied by Benzylamine Adsorption from Solution.* Langmuir, 2007. **23**(8): p. 4315-4323.
27. Schure, M.R. and A.M. Lenhoff, *Consequences of wall adsorption in capillary electrophoresis: theory and simulation.* Analytical Chemistry, 1993. **65**(21): p. 3024-3037.
28. GHOSAL, S., *The effect of wall interactions in capillary-zone electrophoresis.* Journal of Fluid Mechanics, 2003. **491**: p. 285-300.
29. *The data deluge*, in *The Economist*. 2010.
30. Yang, J., S. Bose, and D.S. Hage, *Improved reproducibility in capillary electrophoresis through the use of mobility and migration time ratios.* Journal of Chromatography A, 1996. **735**(1-2): p. 209-220.
31. Tomasi, G., F. vandenBerg, and C. Andersson, *Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data.* JOURNAL OF CHEMOMETRICS, 2004(18): p. 231-241.
32. Nielsen, N.-P.V., J.M. Carstensen, and J. Smedsgaard, *Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping.* Journal of Chromatography A, 1998. **805**(1-2): p. 17-35.
33. Szymańska, E., et al., *Evaluation of different warping methods for the analysis of CE profiles of urinary nucleosides.* ELECTROPHORESIS, 2007. **28**(16): p. 2861-2873.
34. Skov, T., et al., *Automated alignment of chromatographic data.* JOURNAL OF CHEMOMETRICS, 2006. **20**(11-12): p. 484-497.
35. Brenchley, J.M., U. Horchner, and J.H. Kalivas, *Wavelength Selection Characterization for NIR Spectra.* Applied Spectroscopy, 1997. **51**: p. 689-699.
36. Sundling, C.M., et al., *Wavelets in Chemistry and Cheminformatics*, in *Reviews in Computational Chemistry*, K.B. Lipkowitz, T.R. Cundari, and V.J. Gillet, Editors. 2006, Wiley-VCH. p. 295-329.
37. Li, Y. and R. Anderson-Sprecher, *Facies identification from well logs: A comparison of discriminant analysis and naive Bayes classifier.* Journal of Petroleum Science and Engineering, 2006. **53**(3-4): p. 149-157.
38. Kristen L. Mello, S.D.B., *Novel ?hybrid? classification method employing Bayesian networks.* Journal of Chemometrics, 1999. **13**(6): p. 579-590.
39. Nathaniel A. Woody, S.D.B., *Hybrid Bayesian networks: making the hybrid Bayesian classifier robust to missing training data.* Journal of Chemometrics, 2003. **17**(5): p. 266-273.
40. Myles, A.J., et al., *An introduction to decision tree modeling.* Journal of Chemometrics, 2004. **18**(6): p. 275-285.

41. Yu, J. and X.-W. Chen, *Bayesian neural network approaches to ovarian cancer identification from high-resolution mass spectrometry data.* Bioinformatics, 2005. **21**(suppl_1): p. i487-494.

42. Jang, J.-S.R., C.-T. Sun, and E. Mizutani, *Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence*. 1997, Upper Saddle River: Prentice-Halt, Inc.

43. Gributs, C.E.W. and D.H. Burns, *Parsimonious calibration models for near-infrared spectroscopy using wavelets and scaling functions.* Chemometrics and Intelligent Laboratory Systems, 2006. **83**(1): p. 44-53.

44. Cho, C.-K.J., et al., *Proteomics Analysis of Human Amniotic Fluid.* Molecular & Cellular Proteomics, 2007. **6**(8): p. 1406-1415.

45. Tsangaris, G.T., et al., *The normal human amniotic fluid supernatant proteome.* In Vivo, 2006. **20**(4): p. 479-490.

46. Michel, P.E., et al., *Proteome analysis of human plasma and amniotic fluid by Off-Gel™ isoelectric focusing followed by nano-LC-MS/MS.* ELECTROPHORESIS, 2006. **27**(5-6): p. 1169-1181.

47. Liberatori, S., et al., *A two-dimensional protein map of human amniotic fluid at 17 weeks' gestation.* ELECTROPHORESIS, 1997. **18**(15): p. 2816-2822.

48. Nilsson, S., et al., *Explorative Study of the Protein Composition of Amniotic Fluid by Liquid Chromatography Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrometry.* Journal of Proteome Research, 2004. **3**(4): p. 884-889.

49. Vuadens, F., et al., *Identification of biologic markers of the premature rupture of fetal membranes: Proteomic approach.* PROTEOMICS, 2003. **3**(8): p. 1521-1525.

50. Johnson, A.M., et al., *Amniotic fluid proteins: Maternal and fetal contributions.* The Journal of Pediatrics, 1974. **84**(4): p. 588-593.

51. Nicolaides, K.H., *A model for a new pyramid of prenatal care based on the 11 to 13 weeks' assessment.* Prenat Diagn, 2011. **31**(1): p. 3-6.

52. Malul, Y. 2010 [cited 2012 Jan. 8]; Available from: http://en.wikipedia.org/wiki/File:Weight_vs_gestational_Age.jpg.

53. Karagiannis, G., et al., *Prediction of Small-for-Gestation Neonates from Biophysical and Biochemical Markers at 11–13 Weeks.* Fetal Diagnosis and Therapy, 2011. **29**(2): p. 148-154.

54. Lindqvist, P.G. and J. Molin, *Does antenatal identification of small-for-gestational age fetuses significantly improve their outcome?* Ultrasound in Obstetrics and Gynecology, 2005. **25**(3): p. 258-264.

55. International Association of Diabetes and Pregnancy Study Groups Consensus Panel, *International Association of Diabetes and Pregnancy Study Groups Recommendations on the Diagnosis and Classification of Hyperglycemia in Pregnancy.* Diabetes Care, 2010. **33**(3): p. 676-682.

56. American Diabetes Association, *Diagnosis and Classification of Diabetes Mellitus.* Diabetes Care, 2006. **29**(suppl 1): p. s43-s48.

57. Kwiterovich, P.O., Jr., et al., *A Large High-Density Lipoprotein Enriched in Apolipoprotein C-I: A Novel Biochemical Marker in Infants of Lower Birth Weight and Younger Gestational Age.* JAMA, 2005. **293**(15): p. 1891-1899.

58. Gemma, C., et al., *Mitochondrial DNA Depletion in Small- and Large-for-Gestational-Age Newborns[ast].* Obesity, 2006. **14**(12): p. 2193-2199.

59. Smith, G.C.S., et al., *Pregnancy-Associated Plasma Protein A and Alpha-fetoprotein and Prediction of Adverse Perinatal Outcome.* Obstet Gynecol, 2006. **107**(1): p. 161-166.

60. Peterson, S.E. and H.N. Simhan, *First-trimester pregnancy-associated plasma protein A and subsequent abnormalities of fetal growth.* American Journal of Obstetrics and Gynecology, 2008. **198**(5): p. e43-e45.

61. Eva M. L. Smets, et al., *Decreased plasma levels of metastin in early pregnancy are associated with small for gestational age neonates.* Prenatal Diagnosis, 2008. **28**(4): p. 299-303.

62. Poon, L.C.Y., et al., *Reference range of birth weight with gestation and first-trimester prediction of small-for-gestation neonates.* Prenatal Diagnosis, 2011. **31**(1): p. 58-65.

63. Valensise, H., et al., *C-peptide and insulin levels at 24-30 weeks' gestation: an increased risk of adverse pregnancy outcomes?* European Journal of Obstetrics & Gynecology and Reproductive Biology, 2002. **103**(2): p. 130-135.

64. Baschat, A.A., et al., *Very Low Second-Trimester Maternal Serum Alpha-fetoprotein: Association With High Birth Weight.* Obstet Gynecol, 2002. **99**(4): p. 531-536.

65. Scholl, T.O., et al., *Vitamin E: maternal concentrations are associated with fetal growth.* Am J Clin Nutr, 2006. **84**(6): p. 1442-1448.

66. Mazaki-Tovi, S., et al., *Cord blood adiponectin in large-for-gestational age newborns.* American Journal of Obstetrics and Gynecology, 2005. **193**(3, Supplement 1): p. 1238-1242.

67. Lapolla, A., et al., *Can plasma glucose and HbA1c predict fetal growth in mothers with different glucose tolerance levels?* Diabetes Research and Clinical Practice, 2007. **77**(3): p. 465-470.

68. Williams, M.A., et al., *Plasma Adiponectin Concentrations in Early Pregnancy and Subsequent Risk of Gestational Diabetes Mellitus.* J Clin Endocrinol Metab, 2004. **89**(5): p. 2306-2311.

69. Qiu, C., et al., *Increased Maternal Plasma Leptin in Early Pregnancy and Risk of Gestational Diabetes Mellitus.* Obstet Gynecol, 2004. **103**(3): p. 519-525.

70. Thadhani, R., et al., *First-trimester sex hormone binding globulin and subsequent gestational diabetes mellitus.* American Journal of Obstetrics and Gynecology, 2003. **189**(1): p. 171-176.

71. Wolf, M., et al., *First-Trimester C-Reactive Protein and Subsequent Gestational Diabetes.* Diabetes Care, 2003. **26**(3): p. 819-824.

72. Jansson, T., et al., *Alterations in the Activity of Placental Amino Acid Transporters in Pregnancies Complicated by Diabetes.* Diabetes, 2002. **51**(7): p. 2214-2219.

73. Tisi, D.K., et al., *Insulin-Like Growth Factor II and Binding Proteins 1 and 3 from Second Trimester Human Amniotic Fluid Are Associated with Infant Birth Weight.* J. Nutr., 2005. **135**(7): p. 1667-1672.

74. Grandone, E., et al., *Homocysteine levels in amniotic fluid. Relationship with birth-weight.* Thrombosis and Haemostasis, 2006. **95**(4): p. 625-628.

75. Monsen, A.-L.B., J. Schneede, and P.M. Ueland, *Mid-trimester amniotic fluid methionine concentrations: a predictor of birth weight and length.* Metabolism, 2006. **55**(9): p. 1186-1191.

76. Tisi, D.K., et al., *Fetal Exposure to Altered Amniotic Fluid Glucose, Insulin, and Insulin-Like Growth Factor–Binding Protein 1 Occurs Before Screening for Gestational Diabetes Mellitus.* Diabetes Care, 2011. **34**(1): p. 139-144.

77. Ashby, D., *Bayesian statistics in medicine: a 25 year review.* Statistics in Medicine, 2006. **25**(21): p. 3589-3631.

78. Barker, D.J.P., *The fetal and infant origins of adult disease.* British Medical Journal, 1990. **301**: p. 1111.

79. Tappy, L., *Adiposity in children born small for gestational age.* International Journal of Obesity, 2006. **30**: p. S36-S40.

80. Levy-Marchal, C. and D. Jaquet, *Long-term metabolic consequences of being born small for gestational age.* Pediatric Diabetes, 2004. **5**(3): p. 147-153.

81. O'Regan, D., et al., *Prenatal dexamethasone 'programmes' hypotension, but stress-induced hypertension in adult offspring.* J Endocrinol, 2008. **196**(2): p. 343-352.

82. Barker, D.J.P., *EDITORIAL: The developmental origins of adult disease.* European Journal of Epidemiology, 2003. **18**(8): p. 733-736.

83. Lévy-Marchal, C. and P. Czernichow, *Small for Gestational Age and the Metabolic Syndrome: Which Mechanism Is Suggested by Epidemiological and Clinical Studies?* Hormone Research, 2006. **65**: p. 123-130.

84. Emanuel, I., et al., *The Washington State Intergenerational Study of Birth Outcomes: methodology and some comparisons of maternal birthweight and infant birthweight and gestation in four ethnic groups.* Paediatric & Perinatal Epidemiology, 1999. **13**(3): p. 352-371.

85. Lindqvist, P.G. and J. Molin, *Does antenatal identification of small-for-gestational age fetuses significantly improve their outcome?* Ultrasound in Obstetrics and Gynecology, 2005. **25**: p. 258-264.

86. de Vries, A., et al., *Prenatal dexamethasone exposure induces changes in nonhuman primate offspring cardiometabolic and hypothalamic-pituitary-adrenal axis function.* The Journal of Clinical Investigation, 2007. **117**(4): p. 1058-1067.

87. Oberlander, T.F., et al., *Prenatal exposure to maternal depression, neonatal methylation of human glucocorticoid receptor gene (NR3C1) and infant cortisol stress responses.* Epigenetics, 2008. **3**(2): p. 97-106.

88. Heijmans, B.T., et al., *The epigenome: archive of the prenatal environment.* Epigenetics, 2009. **4**(8): p. 526-31.

89. Buchanan, T.A., et al., *What Is Gestational Diabetes?* Diabetes Care, 2007. **30**(Supplement_2): p. S105-111.

90. Dabelea, D., et al., *Increasing Prevalence of Gestational Diabetes Mellitus (GDM) Over Time and by Birth Cohort: Kaiser Permanente of Colorado GDM Screening Program.* Diabetes Care, 2005. **28**(3): p. 579-584.

91. Georgiou, H., et al., *Screening for biomarkers predictive of gestational diabetes mellitus.* Acta Diabetologica, 2008. **45**(3): p. 157-165.

92. Kitzmiller, J.L., L. Dang-Kilduff, and M.M. Taslimi, *Gestational Diabetes After Delivery: Short-term management and long-term risks.* Diabetes Care, 2007. **30**(Supplement_2): p. S225-235.

93. Kim, C., K.M. Newton, and R.H. Knopp, *Gestational Diabetes and the Incidence of Type 2 Diabetes: A systematic review.* Diabetes Care, 2002. **25**(10): p. 1862-1868.

94. Araki, E. and T. Nishikawa, *Oxidative stress: A cause and therapeutic target of diabetic complications.* Journal of Diabetes Investigation, 2010. **1**(3): p. 90-96.

95. *Standards of Medical Care in Diabetes- 2009.* Diabetes Care, 2009. **32**(Supplement 1): p. S13-S61.

96. Akinci, B., et al., *Is fasting glucose level during oral glucose tolerance test an indicator of the insulin need in gestational diabetes?* Diabetes Research and Clinical Practice, 2008. **82**(2): p. 219-225.

97. Kramer, M.S., et al., *A New and Improved Population-Based Canadian Reference for Birth Weight for Gestational Age.* Pediatrics, 2001. **108**(2): p. e35-.

98. Gao, T., et al., *Identification and quantitation of human amniotic fluid components using capillary zone electrophoresis.* Analytical Biochemistry, 2009. **388**(1): p. 155-157.

99. Mayer, B.X., *How to increase precision in capillary electrophoresis.* Journal of Chromatography A, 2001. **907**(1-2): p. 21-37.

100. van Nederkassel, A.M., et al., *A comparison of three algorithms for chromatograms alignment.* Journal of Chromatography A, 2006. **1118**(2): p. 199-210.

101. MacKay, D.J.C., *Information Theory, Inference, and Learning Algorithms*. 1 ed. 2003, Cambridge: Cambridge University Press. 628.

102. Riccardo, L., *Genetic algorithms in chemometrics and chemistry: a review.* Journal of Chemometrics, 2001. **15**(7): p. 559-569.

103. van Leeuwen, M., et al., *Comparison of Accuracy Measures of Two Screening Tests for Gestational Diabetes Mellitus.* Diabetes Care, 2007. **30**(11): p. 2779-2784.

104. Stewart, C.J., R.K. Iles, and D. Perrett, *The analysis of human amniotic fluid using capillary electrophoresis.* ELECTROPHORESIS, 2001. **22**(6): p. 1136-1142.

105. Hardman, M.J., et al., *Barrier Formation in the Human Fetus is Patterned.* Journal of Investigative Dermatology, 1999. **113**(6): p. 1106-1113.

106. Peters, T., Jr., *All About Albumin* 1995: Elsevier. 432.

107. Gabaldon, M., *Thiol dependent isomerization of bovine albumin.* International Journal of Biological Macromolecules, 2009. **44**(1): p. 43-50.

108. Bar-Or, D., et al., *Cysteinylation of maternal plasma albumin and its association with intrauterine growth restriction.* Prenatal Diagnosis, 2005. **25**(3): p. 245-249.

109. Muravskaya, E.V., A.G. Lapko, and V.A. Muravskii, *Modification of Transport Function of Plasma Albumin during Atherosclerosis and Diabetes Mellitus.* Bulletin of Experimental Biology and Medicine, 2003. **135**(5): p. 433-435.

110. Ahmed, N., et al., *Peptide Mapping Identifies Hotspot Site of Modification in Human Serum Albumin by Methylglyoxal Involved in Ligand Binding and Esterase Activity.* J. Biol. Chem., 2005. **280**(7): p. 5724-5732.

111. Faure, P., et al., *Impairment of the antioxidant properties of serum albumin in patients with diabetes: protective effects of metformin.* Clinical Science, 2008. **114**(3): p. 251-256.

112. Boisvert, M.R., K.G. Koski, and C.D. Skinner, *Increased oxidative modifications of amniotic fluid albumin in pregnancies associated with gestational diabetes mellitus.* Anal Chem, 2010. **82**(3): p. 1133-7.

113. Carballal, S., et al., *Sulfenic acid in human serum albumin.* Amino Acids, 2007. **32**(4): p. 543-551.

114. Capote, F.P. and J.-C. Sanchez, *Strategies for proteomic analysis of non-enzymatically glycated proteins.* Mass Spectrometry Reviews, 2009. **28**(1): p. 135-146.

115. Watkins, N.G., S.R. Thorpe, and J.W. Baynes, *Glycation of amino groups in protein. Studies on the specificity of modification of RNase by glucose.* Journal of Biological Chemistry, 1985. **260**(19): p. 10629-10636.

116. WEISS, P.A.M., et al., *Amniotic Fluid Glucose Values in Normal and Abnormal Pregnancies.* Obstetrics & Gynecology, 1985. **65**(3): p. 333-339.

117. Pasman, S., *Fetal fluid and protein dynamics*, in *Department of Obstetrics*. 2010, Leiden University: Leiden

118. Metzger, B.E., et al., *Summary and Recommendations of the Fifth International Workshop-Conference on Gestational Diabetes Mellitus.* Diabetes Care, 2007. **30**(Supplement 2): p. S251-S260.

119. Martinez-Sanchez, G., et al., *Oxidized proteins and their contribution to redox homeostasis.* Redox Report, 2005. **10**(4): p. 175-185.

120. Fröjdö, S., H. Vidal, and L. Pirola, *Alterations of insulin signaling in type 2 diabetes: A review of the current evidence from humans.* Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease, 2009. **1792**(2): p. 83-92.
121. Dalle-Donne, I., et al., *Protein S-glutathionylation: a regulatory device from bacteria to humans.* Trends in Biochemical Sciences, 2009. **34**(2): p. 85-96.
122. Han-Ling, Y., et al., *Aberrant profiles of native and oxidized glycoproteins in Alzheimer plasma.* PROTEOMICS, 2003. **3**(11): p. 2240-2248.
123. Rich, S. and V.V. McLaughlin, *Endothelin Receptor Blockers in Cardiovascular Disease.* Circulation, 2003. **108**(18): p. 2184-2190.
124. Rodriguez-Pineiro, A.M., et al., *Relevance of Protein Isoforms in Proteomic Studies.* Current Proteomics, 2007. **4**(4): p. 235-252.
125. Subbaramaiah, K. and A.J. Dannenberg, *Cyclooxygenase 2: a molecular target for cancer prevention and treatment.* Trends in Pharmacological Sciences, 2003. **24**(2): p. 96-102.
126. Anraku, M., et al., *Intravenous iron administration induces oxidation of serum albumin in hemodialysis patients.* Kidney International, 2004. **66**(2): p. 841-848.
127. William, W.P.C., et al., *Rapid separation of protein isoforms by capillary zone electrophoresis with new dynamic coatings.* ELECTROPHORESIS, 2005. **26**(11): p. 2179-2186.
128. Ferenc, K., *Recent applications of capillary isoelectric focusing.* ELECTROPHORESIS, 2003. **24**(22-23): p. 3908-3916.
129. Aldini, G., et al., *Mass spectrometric characterization of covalent modification of human serum albumin by 4-hydroxy-trans-2-nonenal.* Journal of Mass Spectrometry, 2006. **41**(9): p. 1149-1161.
130. Kawakami, A., et al., *Identification and characterization of oxidized human serum albumin: A slight structural change impairs its ligand-binding and antioxidant functions.* FEBS Journal, 2006. **273**: p. 3346-3357.
131. Mann, M. and N.L. Kelleher, *Precision proteomics: The case for high resolution and high mass accuracy.* Proceedings of the National Academy of Sciences, 2008. **105**(47): p. 18132-18138.
132. Dass, C., *Fundamentals of contemporary mass spectrometry*. 2007, Hoboken, New Jersey: John Wiley & Sons, Inc. 512.
133. Gerber, S.A., et al., *Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS.* Proceedings of the National Academy of Sciences, 2003. **100**(12): p. 6940-6945.
134. Wilm, M., *Quantitative proteomics in biological research.* PROTEOMICS, 2009. **9**(20): p. 4590-4605.
135. Dawn, P.R., E.S. Luis, and O.K. Bernd, *Quantitative analysis with modern bioanalytical mass spectrometry and stable isotope labeling.* Journal of Labelled Compounds and Radiopharmaceuticals, 2007. **50**(11-12): p. 1124-1136.
136. MacCoss, M.J. and D.E. Matthews, *Quantitative MS for Proteomics: Teaching a New Dog Old Tricks.* Analytical Chemistry, 2005. **77**(15): p. 294 A-302 A.
137. Michaud, F.-T., et al., *Multivariate analysis of single quadrupole LC-MS spectra for routine characterization and quantification of intact proteins.* PROTEOMICS, 2009. **9**(3): p. 512-520.
138. Gianazza, E., J. Crawford, and I. Miller, *Detecting oxidative post-translational modifications in proteins.* Amino Acids, 2007. **33**(1): p. 51-6.

139. Oettl, K. and R.E. Stauber, *Physiological and pathological changes in the redox state of human serum albumin critically influence its binding properties.* Br. J. Pharmacol., 2007. **151**(5): p. 580-90.

140. Bar-Or, R., L.T. Rael, and D. Bar-Or, *Dehydroalanine derived from cysteine is a common post-translational modification in human serum albumin.* Rapid Communications in Mass Spectrometry, 2008. **22**(5): p. 711-6.

141. Bar-Or, D., et al., *Heterogeneity and oxidation status of commercial human albumin preparations in clinical use.* Crit Care Med, 2005. **33**(7): p. 1638-41.

142. Lapolla, A., et al., *The role of mass spectrometry in the study of non-enzymatic protein glycation in diabetes: An update.* Mass Spectrometry Reviews, 2006. **25**(5): p. 775-797.

143. Beck, J.L., et al., *Direct observation of covalent adducts with Cys34 of human serum albumin using mass spectrometry.* Analytical Biochemistry, 2004. **325**(2): p. 326-336.

144. Martina, K., et al., *Characterization of cysteinylation of pharmaceutical-grade human serum albumin by electrospray ionization mass spectrometry and low-energy collision-induced dissociation tandem mass spectrometry.* Rapid Communications in Mass Spectrometry, 2005. **19**(20): p. 2965-2973.

145. Lurie, S., et al., *Different Degrees of Fetal Oxidative Stress in Elective and Emergent Cesarean Section.* Neonatology, 2007. **92**(2): p. 111-115.

146. Myatt, L. and X. Cui, *Oxidative stress in the placenta.* Histochemistry and Cell Biology, 2004. **122**(4): p. 369-382.

147. Liu, L., et al., *A metabolic enzyme for S-nitrosothiol conserved from bacteria to humans.* Nature, 2001. **410**(6827): p. 490-494.

148. Sahoo, R., et al., *Effect of nitrosative stress on Schizosaccharomyces pombe: Inactivation of glutathione reductase by peroxynitrite.* Free Radical Biology and Medicine, 2006. **40**(4): p. 625-631.

149. Cho, C.-K.J., et al., *Proteomics Analysis of Human Amniotic Fluid.* Molecular and Cellular Proteomics, 2007. **6**(8): p. 1406-1415.

150. Shibata, E., et al., *Enhanced Protein Levels of Protein Thiol/Disulphide Oxidoreductases in Placentae from Pre-eclamptic subjects.* Placenta, 2001. **22**(6): p. 566-572.

151. Vascotto, C., et al., *Oxidized Transthyretin in Amniotic Fluid as an Early Marker of Preeclampsia.* Journal of Proteome Research, 2006. **6**(1): p. 160-170.

152. Gitlin, D., et al., *The Selectivity of the Human Placenta in the Transfer of Plasma Proteins from Mother to Fetus.* Journal of Clinical Investigation, 1964. **43**(10): p. 1938-1951.

153. Johnson, A.M., et al., *Amniotic fluid proteins: Maternal and fetal contributions* Journal of Pediatrics, 1974. **84**(4): p. 588-593.

154. American Diabetes Association, *Standards of Medical Care in Diabetes- 2009.* Diabetes Care, 2009. **32**(Supplement 1): p. S13-S61.

155. Donma, M.M., *Macrosomia, top of the iceberg: The charm of underlying factors.* Pediatrics International, 2011. **53**(1): p. 78-84.

156. Kerényi, Z., et al., *Maternal Glycemia and Risk of Large-for-Gestational-Age Babies in a Population-Based Screening.* Diabetes Care, 2009. **32**(12): p. 2200-2205.

157. Ahlsson, F.S.E., et al., *Lipolysis and Insulin Sensitivity at Birth in Infants Who Are Large for Gestational Age.* Pediatrics, 2007. **120**(5): p. 958-965.

158. Agnihotri, B., et al., *Trends in human birth weight across two successive generations.* Indian journal of pediatrics, 2008. **75**(2): p. 111-7.

159. Tisi, D.K., J.J. Emard, and K.G. Koski, *Total Protein Concentration in Human Amniotic Fluid Is Negatively Associated with Infant Birth Weight.* J. Nutr., 2004. **134**(7): p. 1754-1758.

160. Carpenter, M.W., et al., *Amniotic Fluid Insulin at 14–20 Weeks' Gestation: Association with later maternal glucose intolerance and birth macrosomia.* Diabetes Care, 2001. **24**(7): p. 1259-1263.

161. Gao, T., et al., *Second trimester amniotic fluid transferrin and uric acid predict infant birth outcomes.* Prenatal Diagnosis, 2008. **28**(9): p. 810-814.

162. Soo-Jin Park, et al., *Proteome analysis of human amnion and amniotic fluid by two-dimensional electrophoresis and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry.* PROTEOMICS, 2006. **6**(1): p. 349-363.

163. Michaels, J.E.A., et al., *Comprehensive Proteomic Analysis of the Human Amniotic Fluid Proteome: Gestational Age-Dependent Changes.* J. Proteome Res., 2007. **6**(4): p. 1277-1285.

164. Gianazza, E., et al., *Mapping the 5-50-kDa fraction of human amniotic fluid proteins by 2-DE and ESI-MS.* PROTEOMICS - Clinical Applications, 2007. **1**(2): p. 167-175.

165. Boulet, S.L., et al., *Macrosomic births in the united states: Determinants, outcomes, and proposed grades of risk.* American Journal of Obstetrics and Gynecology, 2003. **188**(5): p. 1372-1378.

166. Heiskanen, N., K. Raatikainen, and S. Heinonen, *Fetal Macrosomia – A Continuing Obstetric Challenge.* Biology of the Neonate, 2006. **90**(2): p. 98-103.

167. Ørskou, J., et al., *Maternal Characteristics and Lifestyle Factors and the Risk of Delivering High Birth Weight Infants.* Obstetrics & Gynecology, 2003. **102**(1): p. 115-120.

168. Kramer, M.S., et al., *Why are babies getting bigger? Temporal trends in fetal growth and its determinants.* The Journal of Pediatrics, 2002. **141**(4): p. 538-542.

169. Statistics_Canada, *Births*, Healths_Statistics_Division, Editor. 2008, ISSN 1710-5285. p. 19.

170. Moore, K.L. and T.V.N. Persaud, *The Developing Human: Clinically Oriented Embryology*. 6 ed. 2003: W.B. Saunders Company.

171. Weiner, Z., et al., *Clinical and ultrasonographic weight estimation in large for gestational age fetus.* European Journal of Obstetrics & Gynecology and Reproductive Biology, 2002. **105**(1): p. 20-24.

172. Kernaghan, D., et al., *Fetal size and growth velocity in the prediction of the large for gestational age (LGA) infant in a glucose impaired population.* European Journal of Obstetrics & Gynecology and Reproductive Biology, 2007. **132**(2): p. 189-192.

173. Chauhan, S.P., et al., *Limitations of clinical and sonographic estimates of birth weight: experience with 1034 parturients.* Obstet Gynecol, 1998. **91**(1): p. 72-77.

174. Lloyd, D.K., *Capillary electrophoresis analysis of biofluids with a focus on less commonly analyzed matrices.* Journal of Chromatography B, 2008. **866**(1-2): p. 154-166.

175. Clements, M. and L. Li, *Strategy of using microsome-based metabolite production to facilitate the identification of endogenous metabolites by liquid chromatography mass spectrometry.* Analytica Chimica Acta, 2011. **685**(1): p. 36-44.

176. Want, E.J., et al., *From Exogenous to Endogenous: The Inevitable Imprint of Mass Spectrometry in Metabolomics.* Journal of Proteome Research, 2006. **6**(2): p. 459-468.

177. Smith, C.A., et al., *METLIN: A Metabolite Mass Spectral Database.* Therapeutic Drug Monitoring, 2005. **27**(6): p. 747-751.

178. Wishart, D.S., et al., *HMDB: a knowledgebase for the human metabolome.* Nucleic Acids Research, 2009. **37**(suppl 1): p. D603-D610.

179. Amorini, A., et al., *Metabolic profile of amniotic fluid as a biochemical tool to screen for inborn errors of metabolism and fetal anomalies.* Molecular and Cellular Biochemistry: p. 1-12.

180. Ottolenghi, C., et al., *Gestational age-related reference values for amniotic fluid organic acids.* Prenatal Diagnosis, 2010. **30**(1): p. 43-48.

181. Graça, G.a., et al., *1H NMR Based Metabonomics of Human Amniotic Fluid for the Metabolic Characterization of Fetus Malformations.* Journal of Proteome Research, 2009. **8**(8): p. 4144-4150.

182. Graça, G.a., et al., *Metabolite Profiling of Human Amniotic Fluid by Hyphenated Nuclear Magnetic Resonance Spectroscopy.* Analytical Chemistry, 2008. **80**(15): p. 6085-6092.

183. Cohn, B., et al., *Quantitative metabolic profiles of 2nd and 3rd trimester human amniotic fluid using &lt;sup&gt;1&lt;/sup&gt;H HR-MAS spectroscopy.* Magnetic Resonance Materials in Physics, Biology and Medicine, 2009. **22**(6): p. 343-352.

184. Luoto, R., et al., *Primary Prevention of Gestational Diabetes Mellitus and Large-for-Gestational-Age Newborns by Lifestyle Counseling: A Cluster-Randomized Controlled Trial.* PLoS Med, 2011. **8**(5): p. e1001036.

185. Cok, T., E. Tarim, and T. Bagis, *Isolated abnormal value during the 3-hour glucose tolerance test: which value is associated with macrosomia?* Journal of Maternal-Fetal and Neonatal Medicine, 2011. **24**(8): p. 1039-1041.

186. McLernon, D.J., et al., *Clinical effectiveness of elective single versus double embryo transfer: meta-analysis of individual patient data from randomised trials.* BMJ, 2010. **341**.

187. Bromer, J.G., et al., *Preterm deliveries that result from multiple pregnancies associated with assisted reproductive technologies in the USA: a cost analysis.* Current Opinion in Obstetrics and Gynecology, 2011. **23**(3): p. 168-173 10.1097/GCO.0b013e32834551cd.

188. Katz, V., et al., *Infertility: etiology, diagnostic evaluation, management, prognosis*, in *Comprehensive Gynecology*. 2007, Elsevier: Philadelphia.

189. Guttmacher, A.F., *Factors affecting normal expectancy of conception.* J Am Med Assoc, 1956. **161**(9): p. 855-60.

190. Mosher, W.D. and W.F. Pratt, *Fecundity and infertility in the United States: incidence and trends.* Fertil Steril, 1991. **56**(2): p. 192-3.

191. Sunderam, S., et al., *Assisted Reproductive Technology Surveillance --- United States, 2006*, D.o.R. Health and N.C.f.C.D.P.a.H. Promotion, Editors. 2009. p. 1-25.

192. de Mouzon, J., et al., *World collaborative report on Assisted Reproductive Technology, 2002.* Hum Reprod, 2009. **24**(9): p. 2310-20.

193. Schieve, L.A., et al., *Are Children Born After Assisted Reproductive Technology at Increased Risk for Adverse Health Outcomes?* Obstetrics & Gynecology, 2004. **103**(6): p. 1154-1163.

194. Basatemur, E. and A. Sutcliffe, *Follow-up of children born after ART.* Placenta, 2008. **29 Suppl B**: p. 135-40.

195. Kjellberg, A.T., P. Carlsson, and C. Bergh, *Randomized single versus double embryo transfer: obstetric and paediatric outcome and a cost-effectiveness analysis.* Hum Reprod, 2006. **21**(1): p. 210-6.

196. *Guidelines for the number of embryos to transfer following in vitro fertilization No. 182, September 2006.* Int J Gynaecol Obstet, 2008. **102**(2): p. 203-16.

197. Bergh, C., *Single embryo transfer: a mini-review.* Hum Reprod, 2005. **20**(2): p. 323-7.

198. Brison, D.R., et al., *Predicting human embryo viability: the road to non-invasive analysis of the secretome using metabolic footprinting.* Reprod Biomed Online, 2007. **15**(3): p. 296-302.

199. Sakkas, D. and D.K. Gardner, *Noninvasive methods to assess embryo quality.* Current Opinion in Obstetrics & Gynecology, 2005. **17**(3): p. 283-288.

200. Brison, D.R., et al., *Predicting human embryo viability: the road to non-invasive analysis of the secretome using metabolic footprinting.* Reproductive Biomedicine Online, 2007. **15**(3): p. 296-302.

201. Brezinova, J., et al., *Evaluation of day one embryo quality and IVF outcome--a comparison of two scoring systems.* Reprod Biol Endocrinol, 2009. **7**: p. 9.

202. Ciray, H.N., et al., *Early cleavage morphology affects the quality and implantation potential of day 3 embryos.* Fertil Steril, 2006. **85**(2): p. 358-65.

203. Dennis, S.J., et al., *Embryo morphology score on day 3 is predictive of implantation and live birth rates.* Journal of Assisted Reproduction and Genetics, 2006. **23**(4): p. 171-175.

204. Steer, C.V., et al., *The cumulative embryo score: a predictive embryo scoring technique to select the optimal number of embryos to transfer in an in-vitro fertilization and embryo transfer programme.* Hum Reprod, 1992. **7**(1): p. 117-9.

205. Botros, L., D. Sakkas, and E. Seli, *Metabolomics and its application for non-invasive embryo assessment in IVF.* Molecular Human Reproduction, 2008. **14**(12): p. 679-690.

206. Gardner, D.K., et al., *Noninvasive assessment of human embryo nutrient consumption as a measure of developmental potential.* Fertil Steril, 2001. **76**(6): p. 1175-80.

207. Devreker, F., *Uptake and release of metabolites in human preimplantation embryos*, in *Human Preimplantation Embryo Selection*, K. Elder and J. Cohen, Editors. 2007, Informa: London, UK. p. 325–336.

208. Sturmey, R.G., et al., *DNA damage and metabolic activity in the preimplantation embryo.* Hum Reprod, 2009. **24**(1): p. 81-91.

209. Stokes, P.J., et al., *Metabolism of human embryos following cryopreservation: implications for the safety and selection of embryos for transfer in clinical IVF.* Hum Reprod, 2007. **22**(3): p. 829-35.

210. Hardy, K., et al., *Non-Invasive Measurement of Glucose and Pyruvate Uptake by Individual Human Oocytes and Preimplantation Embryos.* Human Reproduction, 1989. **4**(2): p. 188-191.

211. Gott, A.L., et al., *Non-invasive measurement of pyruvate and glucose uptake and lactate production by single human preimplantation embryos.* Hum Reprod, 1990. **5**(1): p. 104-8.

212. Conaghan, J., et al., *Selection criteria for human embryo transfer: a comparison of pyruvate uptake and morphology.* J Assist Reprod Genet, 1993. **10**(1): p. 21-30.

213. Conaghan, J., et al., *Effects of pyruvate and glucose on the development of human preimplantation embryos in vitro.* J Reprod Fertil, 1993. **99**(1): p. 87-95.

214. du Plessis, S.S., et al., *Impact of oxidative stress on IVF.* Expert Review of Obstretrics and Gynecology, 2008. **3**: p. 539-554.

215. Stadtman, E.R., *Protein oxidation and aging.* Free Radical Research, 2006. **40**(12): p. 1250-1258.

216. Arteel, G.E., *Oxidants and antioxidants in alcohol-induced liver disease.* Gastroenterology, 2003. **124**(3): p. 778-790.

217. Stocks, J. and N.E. Miller, *Capillary electrophoresis to monitor the oxidative modification of low density lipoproteins.* Journal of Lipid Research, 1998. **39**(6): p. 1305-1309.

218. Fatehi, A.N., et al., *Presence of cumulus cells during in vitro fertilization protects the bovine oocyte against oxidative stress and improves first cleavage but does not affect further development.* Zygote, 2005. **13**(02): p. 177-185.

219. Tatemoto, H., et al., *Protection of porcine oocytes against cell damage caused by oxidative stress during in vitro maturation: role of superoxide dismutase activity in porcine follicular fluid.* Biol Reprod, 2004. **71**(4): p. 1150-7.

220. Di Simplicio, P., et al., *Thiolation and nitrosation of cysteines in biological fluids and cells.* Amino Acids, 2003. **25**(3-4): p. 323-39.

221. Bonini, M.G., D.C. Fernandes, and O. Augusto, *Albumin oxidation to diverse radicals by the peroxidase activity of Cu,Zn-superoxide dismutase in the presence of bicarbonate or nitrite: diffusible radicals produce cysteinyl and solvent-exposed and -unexposed tyrosyl radicals.* Biochemistry, 2004. **43**(2): p. 344-51.

222. Himmelfarb, J. and E. McMonagle, *Albumin is the major plasma protein target of oxidant stress in uremia.* Kidney Int, 2001. **60**(1): p. 358-363.

223. Kleinova, M., et al., *Characterization of cysteinylation of pharmaceutical-grade human serum albumin by electrospray ionization mass spectrometry and low-energy collision-induced dissociation tandem mass spectrometry.* Rapid Commun Mass Spectrom, 2005. **19**(20): p. 2965-73.

224. Rasheed, Z. and R. Ali, *Reactive oxygen species damaged human serum albumin in patients with type 1 diabetes mellitus: biochemical and immunological studies.* Life Sci, 2006. **79**(24): p. 2320-8.

225. Anraku, M., et al., *Validation of the chloramine-T induced oxidation of human serum albumin as a model for oxidative damage in vivo.* Pharmaceutical Research, 2003. **20**(4): p. 684-692.

226. Gryzunov, Y.A., et al., *Binding of fatty acids facilitates oxidation of cysteine-34 and converts copper-albumin complexes from antioxidants to prooxidants.* Arch Biochem Biophys, 2003. **413**(1): p. 53-66.

227. TAKENAKA, M. and T. HORIUCHI, *Recombinant human albumin supports mouse blastocyst development, suppresses apoptosis in blastocysts and improves fetal development.* Reproductive Medicine and Biology, 2007. **6**(4): p. 195-201.

228. Bungum, M., P. Humaidan, and L. Bungum, *Recombinant human albumin as protein source in culture media used for in vitro fertilization, a prospective randomized study.* Fertility and Sterility, 2002. **78**(Supplement 1): p. S56-S57.

229. Wang, W.H. and B.N. Day, *Development of porcine embryos produced by IVM/IVF in a medium with or without protein supplementation: effects of extracellular glutathione.* Zygote, 2002. **10**(2): p. 109-115.

230. Kraly, J., et al., *Bioanalytical applications of capillary electrophoresis.* Anal Chem, 2006. **78**(12): p. 4097-110.

231. Huang, Y.F., et al., *Capillary electrophoresis-based separation techniques for the analysis of proteins.* Electrophoresis, 2006. **27**(18): p. 3503-22.

232. Yang, Q., et al., *Single cell determination of nitric oxide release using capillary electrophoresis with laser-induced fluorescence detection.* J Chromatogr A, 2008. **1201**(1): p. 120-7.

233. Chen, Y., G. Xiong, and E.A. Arriaga, *CE analysis of the acidic organelles of a single cell.* Electrophoresis, 2007. **28**(14): p. 2406-15.

234. Johnson, R.D., et al., *Analysis of mitochondria isolated from single cells.* Anal Bioanal Chem, 2007. **387**(1): p. 107-18.

235. Huang, W.H., et al., *Recent advances in single-cell analysis using capillary electrophoresis and microfluidic devices.* J Chromatogr B Analyt Technol Biomed Life Sci, 2008. **866**(1-2): p. 104-22.

236. Chiu, D.T. and R.M. Lorenz, *Chemistry and biology in femtoliter and picoliter volume droplets.* Acc Chem Res, 2009. **42**(5): p. 649-58.

237. Hu, S., et al., *Protein analysis of an individual Caenorhabditis elegans single-cell embryo by capillary electrophoresis.* J Chromatogr B Biomed Sci Appl, 2001. **752**(2): p. 307-10.

238. Harwood, M.M., et al., *Single-cell protein analysis of a single mouse embryo by two-dimensional capillary electrophoresis.* J Chromatogr A, 2006. **1130**(2): p. 190-4.

239. Schneider, G.F., et al., *Pathway for unfolding of ubiquitin in sodium dodecyl sulfate, studied by capillary electrophoresis.* J Am Chem Soc, 2008. **130**(51): p. 17384-93.

240. Stutz, H., et al., *Detection of coexisting protein conformations in capillary zone electrophoresis subsequent to transient contact with sodium dodecyl sulfate solutions.* Electrophoresis, 2005. **26**(6): p. 1089-1105.

241. Boisvert, M.R., K.G. Koski, and C.D. Skinner, *Increased Oxidative Modifications of Amniotic Fluid Albumin in Pregnancies Associated with Gestational Diabetes Mellitus.* Analytical Chemistry, 2010.

242. Tomasi, G., F.v.d. Berg, and C. Andersson, *Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data.* Journal of Chemometrics, 2004. **18**(5): p. 231-241.

243. Brison, D.R., et al., *Identification of viable embryos in IVF by non-invasive measurement of amino acid turnover.* Hum Reprod, 2004. **19**(10): p. 2319-24.

244. Seli, E., et al., *Noninvasive metabolomic profiling of embryo culture media using proton nuclear magnetic resonance correlates with reproductive potential of embryos in women undergoing in vitro fertilization.* Fertility and Sterility, 2008. **90**(6): p. 2183-2189.

245. Oakes, J., *Protein–surfactant interactions. Nuclear magnetic resonance and binding isotherm studies of interactions between bovine serum albumin and sodium dodecyl sulphate.* Journal of the Chemical Society, Faraday Transactions 1: Physical Chemistry in Condensed Phases, 1974. **70**: p. 2200-2209.

246. Gudiksen, K.L., I. Gitlin, and G.M. Whitesides, *Differentiation of proteins based on characteristic patterns of association and denaturation in solutions of SDS.* Proc Natl Acad Sci U S A, 2006. **103**(21): p. 7968-72.

247. García, M.C., *The effect of the mobile phase additives on sensitivity in the analysis of peptides and proteins by high-performance liquid chromatography-electrospray mass spectrometry.* Journal of Chromatography B, 2005. **825**(2): p. 111-123.

248. Eriksson, J.H.C., et al., *Feasibility of nonvolatile buffers in capillary electrophoresis-electrospray ionization-mass spectrometry of proteins.* Electrophoresis, 2004. **25**(1): p. 43-49.

249. Vergouw, C.G., et al., *Metabolomic profiling by near-infrared spectroscopy as a tool to assess embryo viability: a novel, non-invasive method for embryo selection.* Human Reproduction, 2008. **23**(7): p. 1499-1504.

250. Goto, Y., et al., *Oxidative stress on mouse embryo development in vitro.* Free Radic Biol Med, 1992. **13**(1): p. 47-53.

251. Guerin, P., S. El Mouatassim, and Y. Menezo, *Oxidative stress and protection against reactive oxygen species in the pre-implantation embryo and its surroundings.* Hum Reprod Update, 2001. **7**(2): p. 175-89.

252. Menezo, Y. and P. Guerin, *[Gamete and embryo protection against oxidative stress during medically assisted reproduction].* Bull Acad Natl Med, 2005. **189**(4): p. 715-26; discussion 726-8.

253. Weber, H., et al., *Oxidative stress triggers thiol oxidation in the glyceraldehyde-3-phosphate dehydrogenase of Staphylococcus aureus.* Mol Microbiol, 2004. **52**(1): p. 133-40.

254. Gardiner, C.S. and D.J. Reed, *Glutathione redox cycle-driven recovery of reduced glutathione after oxidation by tertiary-butyl hydroperoxide in preimplantation mouse embryos.* Arch Biochem Biophys, 1995. **321**(1): p. 6-12.

255. Gardner, D.K. and M. Lane, *Culture and selection of viable blastocysts: a feasible proposition for human IVF?* Hum Reprod Update, 1997. **3**(4): p. 367-82.

256. Suzuki, E., et al., *Increased oxidized form of human serum albumin in patients with diabetes mellitus.* Diabetes Research and Clinical Practice, 1992. **18**(3): p. 153-158.

257. Kok, M.G.M., G.J. de Jong, and G.W. Somsen, *Sensitivity enhancement in capillary electrophoresis-mass spectrometry of anionic metabolites using a triethylamine-containing background electrolyte and sheath liquid.* ELECTROPHORESIS, 2011. **32**(21): p. 3016-3024.

258. van Pinxteren, D. and H. Herrmann, *Determination of functionalised carboxylic acids in atmospheric particles and cloud water using capillary electrophoresis/mass spectrometry.* Journal of Chromatography A, 2007. **1171**(1-2): p. 112-123.

259. Tomasi, G., F. Savorani, and S.B. Engelsen, *icoshift: An effective tool for the alignment of chromatographic data.* Journal of Chromatography A, 2011. **1218**(43): p. 7832-7840.

260. Altria, K., *Essential peak area normalisation for quantitative impurity content determination by capillary electrophoresis.* Chromatographia, 1993. **35**(3): p. 177-182.

261. Guo, K. and L. Li, *Differential 12C-/13C-Isotope Dansylation Labeling and Fast Liquid Chromatography/Mass Spectrometry for Absolute and Relative Quantification of the Metabolome.* Analytical Chemistry, 2009. **81**(10): p. 3919-3932.

262. Leyva, A., et al., *Clinical and biochemical studies of high-dose thymidine treatment in patients with solid tumors.* Journal of Cancer Research and Clinical Oncology, 1984. **107**(3): p. 211-216.

263. Greter, J. and C. Jacobson, *Urinary organic acids: isolation and quantification for routine metabolic screening.* Clin Chem, 1987. **33**(4): p. 473-480.

264. Shoemaker, J.D. and W.H. Elliott, *Automated screening of urine samples for carbohydrates, organic and amino acids after treatment with urease.* Journal of Chromatography B: Biomedical Sciences and Applications, 1991. **562**(1-2): p. 125-138.

265. Gustafsson, J., et al., *Bile Acid Metabolism in Extrahepatic Biliary Atresia.* Upsala Journal of Medical Sciences, 2006. **111**(1): p. 131-136.

266. Batta, A.K., et al., *Characterization of serum and urinary bile acids in patients with primary biliary cirrhosis by gas-liquid chromatography-mass spectrometry: effect of ursodeoxycholic acid treatment.* Journal of Lipid Research, 1989. **30**(12): p. 1953-62.

267. Tadano, T., et al., *Studies of serum and feces bile acids determination by gas chromatography-mass spectrometry.* Rinsho Byori, 2006. **54**(2): p. 103-10.

Appendix I

Methodological information to supplement Chapters 2 & 4

1 Data import

The PDA data were exported as an ascii tab delimited data file using the Beckman 32 Karat Software's built-in export utility. Each sample data was then imported into MatLab as 161 wavelengths by 4800 migration time points.

2 Basic Preprocessing: baseline, dividing-by-time, normalization.

Once the data were imported into MatLab, the basic preprocessing of the PDA data was applied. The first ten wavelengths (190 to 199 nm) were averaged and stored as a row vector. A simple baseline subtraction then followed using the average signal value from index values 240 to 300 (60 to 75 seconds). Next, each signal data point was divided by the migration time in seconds corresponding to the index for each data point (*e.g.* signal at index 240 = y, then y/60s gives the time corrected signal) [260]. Finally, the signal was normalized to the peak height of albumin.

3 Correlation Optimized Warping

The first 150 data points for each sample electropherogram were replaced with the value of the 151st data point to remove the instrument auto-zero artifacts. Due to the high similarly of the profiles, the first electrogram in the dataset was selected to be the reference profile to which all the others were aligned. The window size was selected to be 20 as per the recommendation made by Nielsen [32] in the original publication that window size should match that of the smallest peak in the profile. The slack parameter was set to 2. Profiles that showed a global correlation coefficient of less than 0.9 when compared to the reference profile were excluded from further analysis.

4 Haar Transform of the Data

From the original 4800 migration time data points, a window of 2048 data points (900 to 2947; 225s to 736.75s) was selected to undergo the discrete Haar wavelet transform (son wavelets). The first 256 wavelets were retained for further analysis.

5 Genetic Algorithm Parameters

The population size for each generation was set to 100. The number of variables that were used to build the model was set to 2, the cross-over rate was set to 1.0, the mutation rate was set at 0.15 and the variable index values were encoded with 12 bits. The GA was allowed to evolve over 300 generations. The benefit function using Bayesian decision theory is described next.

6 Benefit Function of the Genetic Algorithm

The fitness of variables selected by the GA was evaluated using accuracy of the leave-one out classification into normal and abnormal groups using Bayesian decision theory. The accuracy was calculated by summing true positive and true negative over all samples. Two Gaussian distributions were constructed out of the calibration data by determining the centres and covariance matrices corresponding to each group's distribution. The resulting parameters were used to determine the probability density of the left out sample to belong to one group or the other.

7 Random Permutation Test

The data set was comprised of 109 samples and after the Haar wavelet transform, 256 variables. This leaves a high potential for over fitting and generating models that can provide adequate classification of the data by chance. A random permutation test was employed to determine if the models obtained showed significant differences from models obtained from the random permutation of class labels. This is done by comparing the value of a test statistic (T) for the model generated with the correct class label to the distribution of this same test statistic calculated on a large number of random permutations of the class labels. If the number calculated T-values of random permutation is sufficiently large it will allow the use of a Z-test to evaluate if the T-value (T*) from the T-values generated from the random permutation of class labels.

In this particular case the T-test employed was the student's t-test. For both the LGA and GDM classifications a T* was calculated. To obtain the parameters for the Z-test, the genetic algorithm was modified to evaluate models produced from the random permutation of the class labels. The genetic algorithm would produce a new random permutation of class labels after each iteration. The algorithm was allowed to go through 500 iterations, in each iterations 150 combinations of 2 variables were used to determine the T-values.

For the GDM classification, 75000 T-values were obtained and the resulting $\boldsymbol{\mu}$ and σ were 0.0895 and 0.6349 respectively. The results of the Z-test for those parameters and with a T* of 2.3207 lead to the rejection of the null hypothesis ($\alpha = 0.05$) and gave a p-value of 0.017.

For the LGA classification, 75000 T-values were obtained and the resulting $\boldsymbol{\mu}$ and σ were 0.2087 and 0.8366 respectively. The results of the Z-test for those parameters and with a T* of 2.2868 lead to the rejection of the null hypothesis ($\alpha = 0.05$) and gave a p-value of 0.013.

In both cases, the models with the correct class labels were shown to be statistically different from the distribution generated with the random permutation of class labels.

Appendix II

Additional Results

Chapter 2 and Chapter 4 show the results from the chemometric data analysis of AF separated by CE. Some additional results and information are included in this Appendix to complement results discussed in the above mentioned Chapters. These are the single variables results and bivariates plots for both GDM and LGA outcomes.

1 Single variable result for GDM

Using the same approach described in Chapter 2 and in Appendix I but with only one variable to generate the classification model. The results can be seen in Figure 22 below. In this case the retained wavelet is located on the transferring peak. The sensitivity is 79% (3 false negatives) and the specificity is 100%.

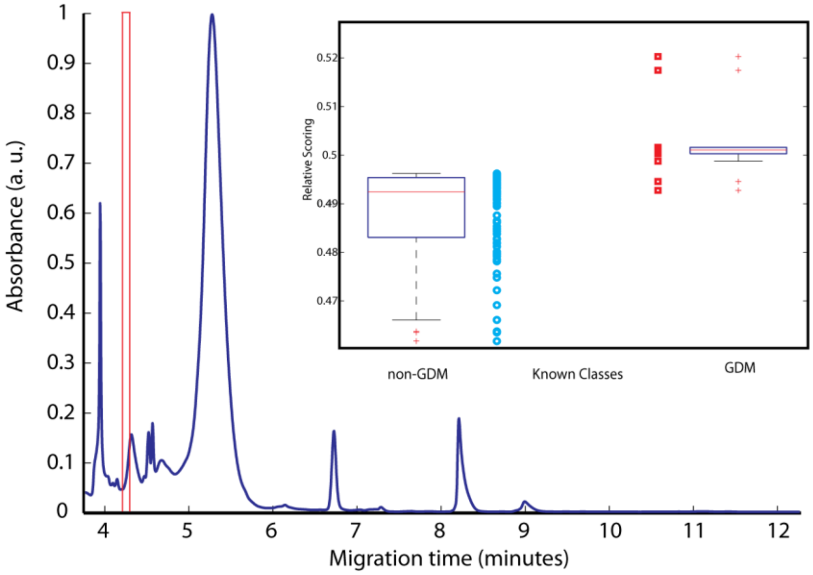


Figure 22: Using 1 variable (wavelet) on the transferrin peak, the Bayesian algorithm can classify with a sensitivity of 79% and specificity of 100%. Selected wavelets on electropherogram at 195±5 nm of AMF

2 Single variable result for LGA

Using the same approach described in Chapter 2 and in Appendix I but with only one variable to generate the classification model. The results can be seen in Figure 23 below. In this case the retained wavelet is located on the transferring peak. The sensitivity is 100% and the specificity is 93% (6 false positives).

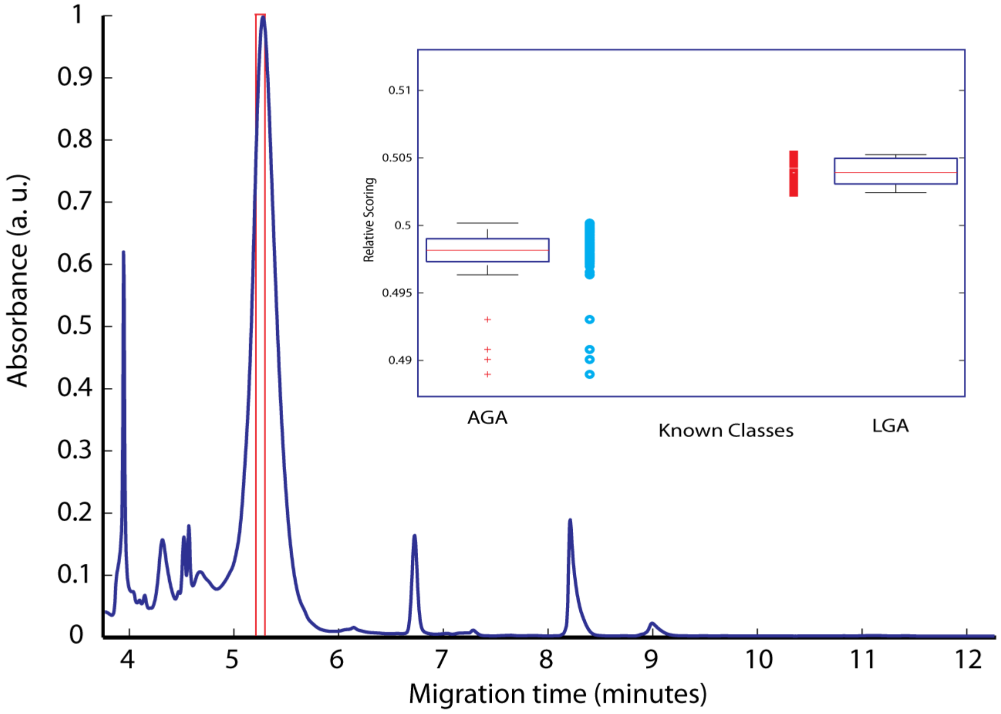


Figure 23: Using 1 variable (wavelet) on the HSA peak, the Bayesian algorithm can classify with a sensitivity of 100% and specificity of 93%. Selected wavelets on electropherogram at 195±5 nm of AMF

3 Bivariate plot of the selected wavelet for GDM

Figure 15 shows the selected wavelets and the resulting classification of GDM samples obtained when using those wavelets. To compliment Figure 15, a bivariate plot of the coefficient for the 2 wavelets used for the classification is shown below (Figure 24).

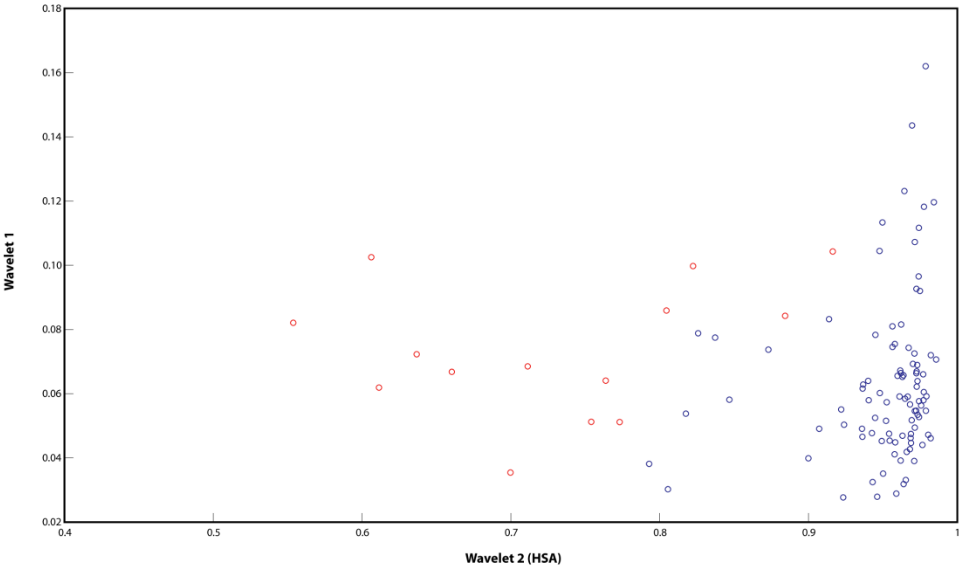


Figure 24: Scatter plot of GDM (in red) and non-GDM (in blue) AF samples. The x-axis corresponds to the coefficient for Wavelet 2 (HSA) and the y-axis corresponds to the coefficient for Wavelet 1 in the unresolved protein region.

4 Bivariate plot of the selected wavelet for LGA

Figure 15 shows the selected wavelets and the resulting classification of LGA samples obtained when using those wavelets. To compliment Figure 17, a bivariate plot of the coefficient for the 2 wavelets used for the classification is shown below (Figure 25).

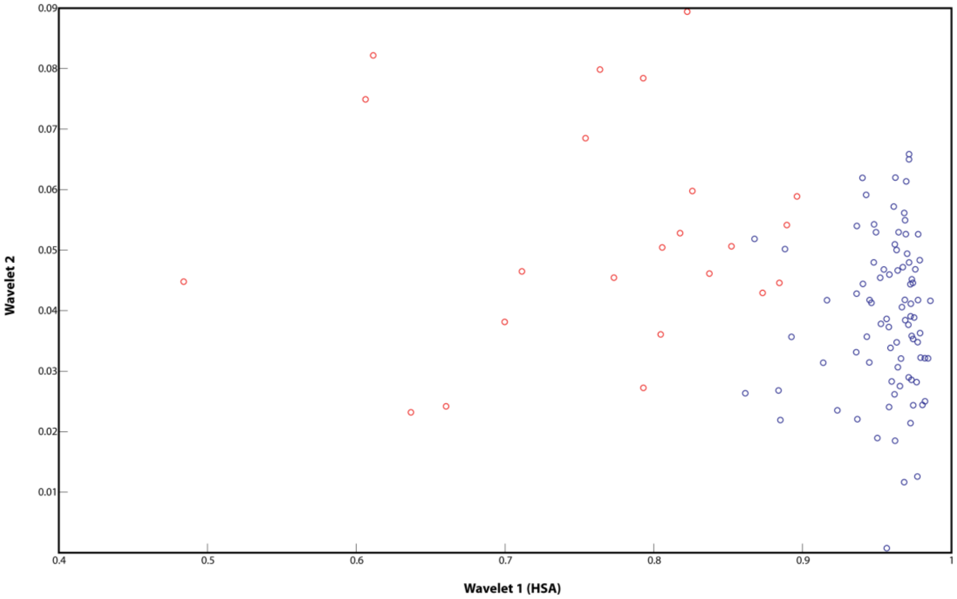


Figure 25: Scatter plot of LGA (in red) and AGA (in blue) AF samples. The x-axis corresponds to the coefficient for Wavelet 1 (HSA) and the y-axis corresponds to the coefficient for Wavelet 2 (late migrating peak).

Appendix III

Preliminary work on HSA

1 Albumin isolation and trypsin digestion of albumin

Frozen amniotic fluid samples were thawed on wet ice. Albumin (~ 4 mg/ml) was isolated from 400 µL AF of with AlbuminOUT™ (Genotech Biosciences) albumin affinity spin columns. The retained albumin (~ 1.6 mg) was released from the affinity column with 200 µl the buffer provided with the kit (NaCl solution buffered at pH 7.2). The albumin fractions were desalted with ultrafree 5 kDa centrifugal filters (Millipore). The resulting solution was diluted to 100 µl with ddH_2O water.

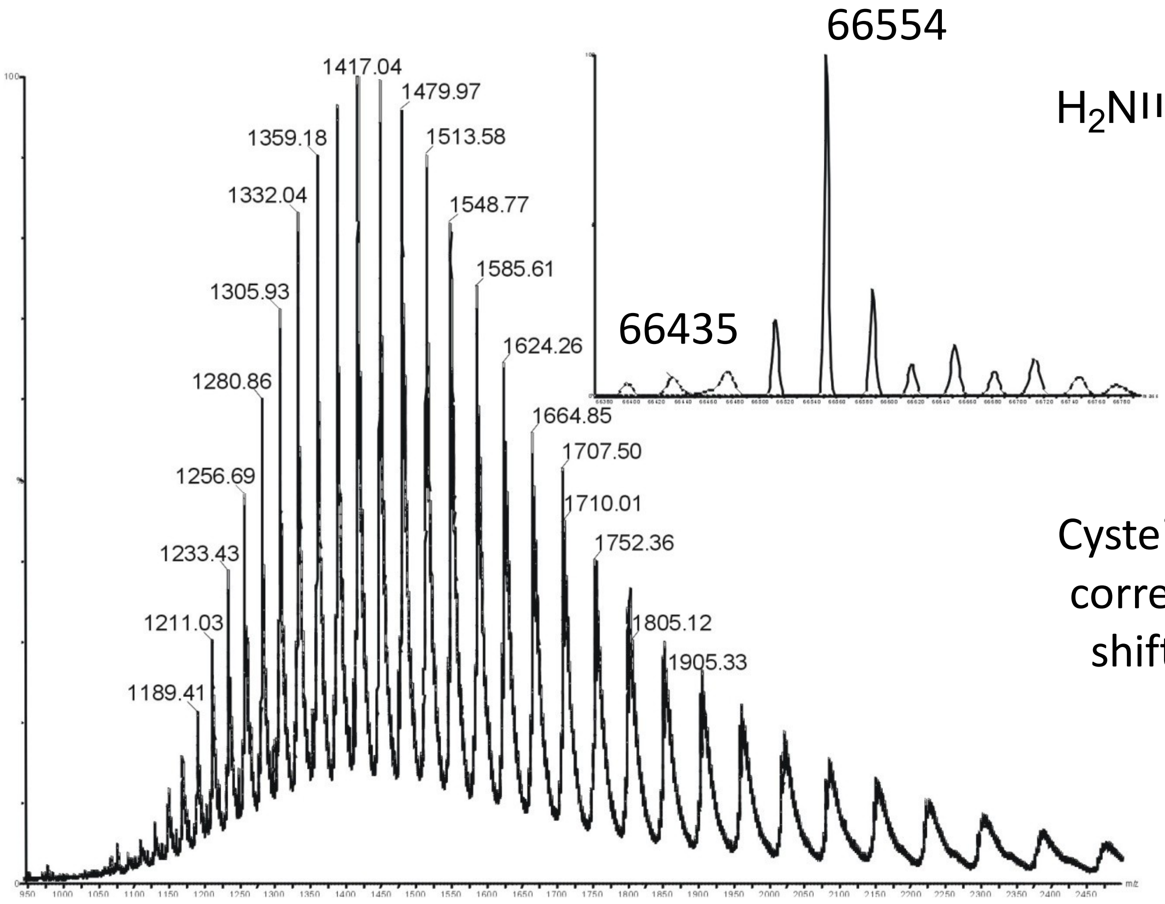
Albumin (160 µg) was digested with trypsin (3.4 µg) in 100 µl NH_4HCO_3–NH_4OH at pH 8.5. The digestion was allowed to proceed overnight at 37°C. After digestion the solution was lyophilized for 2 h at 43°C.

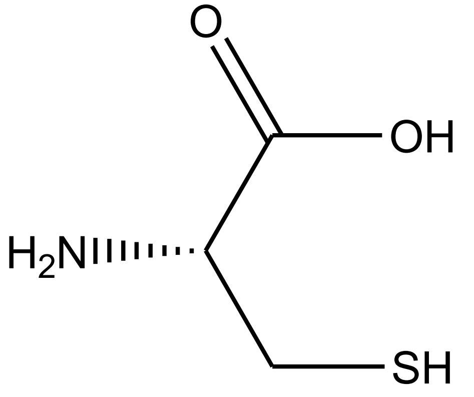
2 Intact proteins

Undigested proteins were analyzed by direct infusion into ESI-MS at a flow rate of 2 µl/min. The isolated proteins were diluted into 70% ddH_2O/30% ACN/0.1% FA (v/v/v) to give a protein concentration of at most 40 pmol/µl. The spectra were obtained by scanning over the mass range: m/z = 600 to 2500. The instrumental conditions were: positive ion mode; cone 40 V; capillary 3.5 kV; collision energy 10 V; desolvation temperature 150°C.

A crude protein separation was done using Micromass capLC with a NanoEase C18 guard column. Flow rate was set at 2 ul/min with linear gradient of 3%/min increase of solvent B (97% ACN/3%water/0.1%FA) with respect to solvent A (97% water/3%ACN/0.1%FA) from starting at 5% after 5 minutes to 90% after 35 minutes.

The resulting raw spectrum was deconvoluted to obtain protein molecular weights (MW) that correspond to the various HSA isoform present shown in Figure 26. The base peak corresponds to the expected MW of cysteinylated HSA.





cysteine

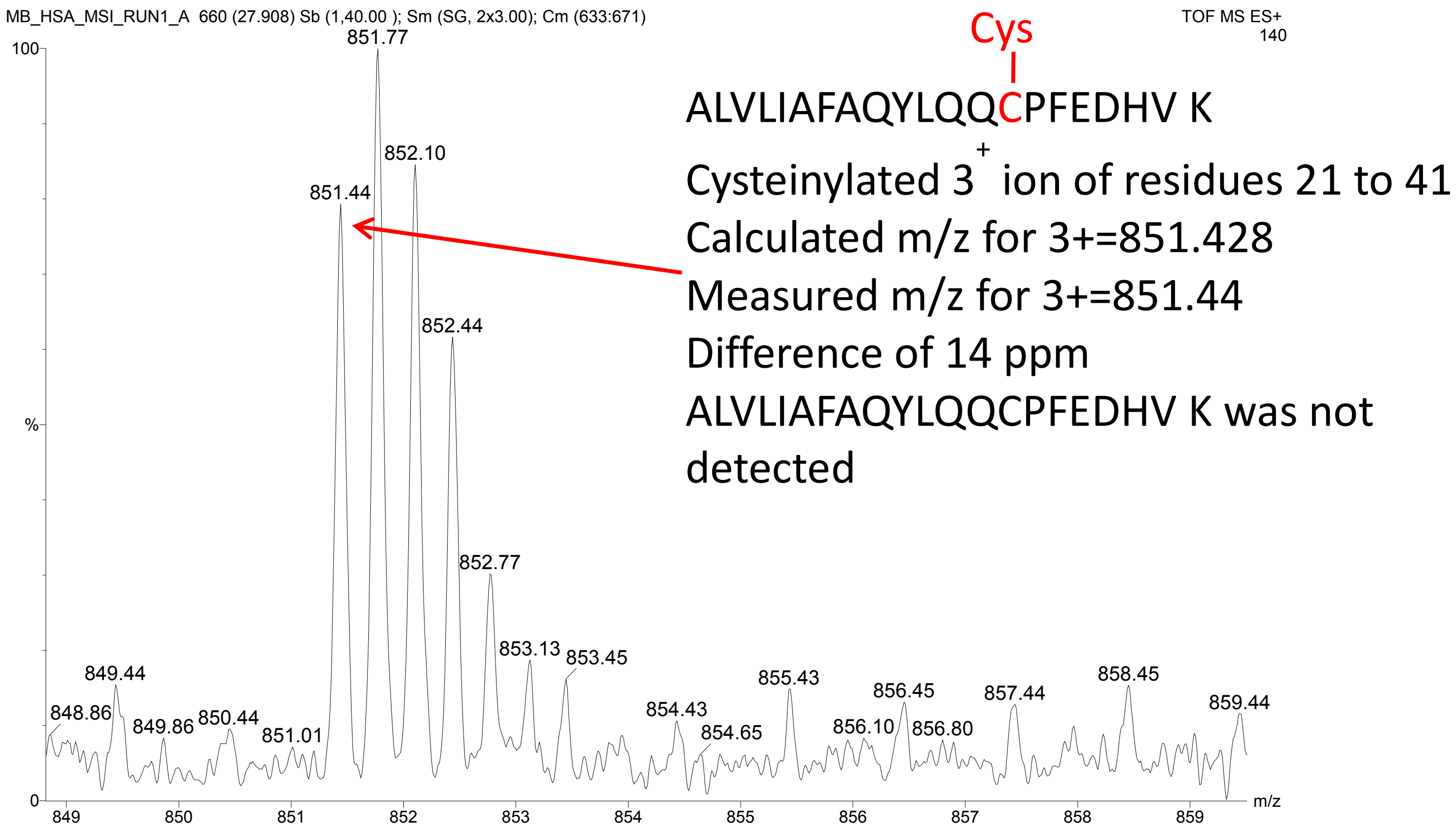
Cysteinylation as a PTM corresponds to a mass shift of +119.004 Da (121.02-2H)

Figure 26: Raw and deconvoluted spectrum of HSA where cysteinylated HSA is the base peak in the spectrum.

3 HSA digestion

Prior to MS analysis the digests were resuspended in 0.1% formic acid and 10 µl of this solution was diluted to 60 µl (4.8 pmol/µl). The HSA tryptic digests were analyzed with the same mass analyzer and CapLC system mentioned in the previous section. The diluted digests (4.8 pmol/µl) solutions were injected (5 µl) on a Waters NanoEase column 3 µm Atlantis dC18, 75 µm x 100 mm. The flow rate was set at 6 µl/min with a 60 minute linear gradient from 95% solvent A (same as above) to 65% solvent B (same as above) followed by a 10 minute wash with 90% solvent B. The column was equilibrated with starting conditions for 20 minutes. LC-MS/MS experiments were carried under the same conditions except for the collision energy which was raised from 10 V to 35 V when fragmentation was desired.

The peptide fragment containing the cysteine 34 should contain the peptide segment with residues 21 to 41. The calculate mass for this peptide segment if cysteinlylated is 851.43 for the triply charged fragment. This mass was indeed present in the spectrum in low abundance within 14 ppm of its calculated value (see Figure 27). Figure 27 shows the initial spectrum.



ALVLIAFAQYLQQCPFEDHV K

Cysteinylated 3^+ ion of residues 21 to 41

Calculated m/z for 3+=851.428

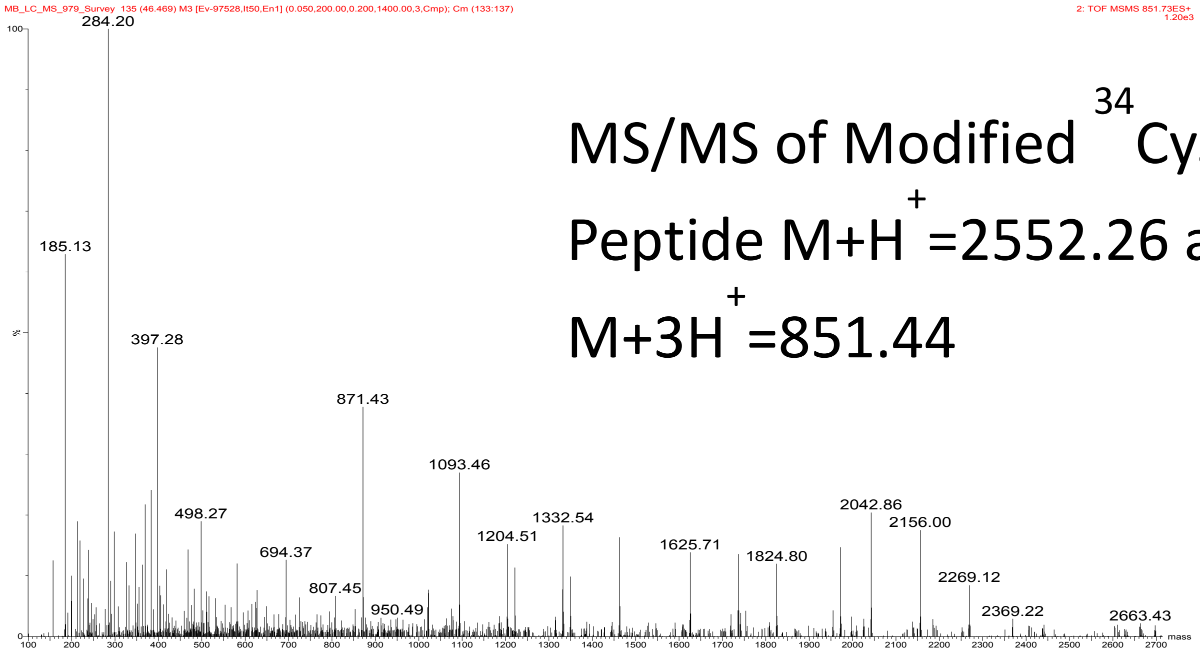
Measured m/z for 3+=851.44

Difference of 14 ppm

ALVLIAFAQYLQQCPFEDHV K was not detected

Figure 27: HSA peptide fragment 21-41 with cysteinylated cysteine 34.

Further confirmation that the observed mass was that of the triply charge cysteinylated version of the HSA peptide fragment (21-41), MS/MS fragmentation was conducted on the 851.44 peak. The triply charged 851.44 generated the spectrum in Figure 28. The sequencing of the peptide fragment was carried out and yielded convincing evidence that the 851.44 peak was indeed that of cysteinylated cysteine 34 peptide segment of HSA. The sequencing of the peptide was done with the highest observed mass error of 20 ppm (see Table 7).



MS/MS of Modified ^{34}Cys Peptide $M+H^+$=2552.26 and $M+3H^+$=851.44

Peptide $M+H^+$=2552.26 and $M+3H^+$=851.44, Sequence ALVLIAFAQYLQQCPFEDHV K

Figure 28: MS/MS spectrum of the triply charged 851.44 peak.

Table 7: Sequencing of the HSA fragment 21 to 41 containing cysteinylated cysteine 34.

m/z	Δ m/z	AAs	Calculated	Δm/z (in ppm)	Ion type
2269.12	283.14	A+L+V	283.18	16	y19 +y20
2156.00	113.12	I/L	113.08	19	y18
2042.86	113.14	I/L	113.08	29	y17
1824.80	218.06	A+F	218.09	16	y16
1625.71	199.09	A+Q	199.10	3	y14
1093.46	532.25	Y+L+Q+Q	532.26	8	y12
871.43	222.03	C+Cys	222.01	20	y8
498.27	373.16	P+F+E	373.16	4	y5
		D+H+V+K	479.25	6	y1+19.02u

Appendix IV

Identification of unknown predictive of LGA

Determining the identity of the molecule that gives rise to the second peak involved in the prediction of LGA is of great interest. A variety of strategies were investigated to obtain more information to identify the unknown molecule. The methods employed are outlined in this appendix. Amniotic fluid samples were separated by CE in a similar fashion as described previously except that a fraction collection step was included to collect the molecule to be identified. Fractions were then analyzed by MS under a variety of different conditions. Efforts to develop HPLC separation of AF were also pursued to isolate and identify the unknown. These will be described with more detail in the following sections.

1 Fraction collection by capillary electrophoresis

The AF sample was thawed on an ice water bath prior to separation and diluted 1:1 in ddH$_2$O. In some cases, the AF was filtered through a 5 kDa molecular weight cut off centrifugal filter to collect the filtrate and eliminate the proteins from the sample.

The separation conditions employed for the fraction collection are essentially the same as for the separation and are described in Chapter 2. Briefly, a 75 mM borate buffer with 0.8 mM EDTA with a pH adjusted to 9.2 was used for the separation. The separation voltage was set to 25 kV for a 20 minute separation. The capillary was conditioned before each run by flushing a 5 mM SDS solution, followed by 100 mM NaOH and finally the run buffer was introduced. For each step, the solutions were flushed through the capillary for

2 minutes at 1.4 bar with a 1 minute wait in between. Prior to injection and separation, the capillary was equilibrated with the run buffer under a 25 kV voltage for one minute.

The appropriate parameters to collect the CE fraction were calculated based on the apparent mobility of the unknown peak and the capillary dimensions. The migration time of the unknown is the time taken to reach the detector and not the time taken to exit the capillary. Therefore, the apparent mobility is used to determine the expected time when the unknown will exit the capillary as shown by the example below.

The calculated time of exit from the capillary is used in the timed program to stop the CE separation and change the outlet vial to a collection vial containing 20 μL of ddH_2O. Once the vial was in place, 1 psi was applied to the inlet to push the portion of the capillary content in which the unknown is present into the collection vial. Repeated AF separations (10 to 20 runs/vial) and thus CE fractions were collected together into the same vial to obtain sufficient mass of unknown for subsequent MS characterization steps. These steps were carried out for both the filtered and non-filtered samples.

2 First approach: C18 trap column and positive mode ESI-QToF detection

The pooled unknown containing fractions (5 μL) where injected onto a NanoEase C18 trap column with a Waters CapLC system and directed into an ESI-QToF2 (Waters) mass spectrometer. The MS settings were: capillary voltage 3.3 kV, cone 35 V, collision energy 10 V, source temperature at 80 ºC and the desolvation temperature 150 ºC. The instrument was set to scan from 50 to 950 m/z in positive mode. The mobile phases used where A 0.1% formic acid in water and B 0.1% formic acid in acetonitrile. The gradient was set to start at 20% and go to 60% B in 10 minutes at 1 μL/min followed by a steep increase to 90% B in 3 minutes at 2 μL/min.

The resulting spectra were compared to MS solvent blank and to a CE solvent blank to eliminate peaks due to contaminants in any of those two potential sources. Only two masses appeared in the fractions, with m/z values of 432.28 and 226.18 coming out late and therefore showing fairly high hydrophobicity at this low pH. MS/MS of those two masses were conducted but did not provide sufficient information to characterize the molecules giving rise to those peaks.

The initial decision to run in positive mode was based on the fact that the instrument as an order of magnitude more sensitivity in positive mode than negative mode. Yet the unknown is an anion under basic conditions, so it was important to consider running the MS investigations in negative mode with an anion exchange HPLC column.

3 Second approach: AEX HPLC column and negative mode ESI-QToF detection

The pooled fractions were analyzed on the same QToF2 instrument previously mentioned. A anion exchange column [Phenomenex Luna NH_2, 5 μm, 15 cm, 4.6 cm ID] was used with an Agilent HP 1050 HPLC system. The MS settings were the following: capillary voltage 3.0 kV, cone 25 V, collision energy 10 V, source temperature 90 ºC and the desolvation temperature 250 ºC. The instrument was set to scan from 50 to 500 m/z in negative mode. The mobile phase used was 70% water with 5 mM ammonium acetate buffer at pH 5.3 and 30% acetonitrile.

In this case, collected CE fractions, LMW AF fractions and blanks were compared to determine the peaks that were present in both the CE and LMW AF fractions, both not in the blanks. Peaks of interest were then further analyzed to determine the exact mass and chemical formulae. Exact masses were used to search metabolite databases (HMDB [178] and METLIN [177]) and find the potential metabolites that had reasonable UV profiles and pKa's while being likely present in AF. Those that were deemed relevant were purchased and spiked into AF to determine whether or not there migration time corresponded with that of the unknown species. The most relevant database hits are summarized in Table 8.

Table 8: Anions found in the collected AF fraction analyzed by HPLC-MS.

$[M-H]^-$ (Th)	Keep/reject reason	Candidates	Biological location	reference
141.02	UV profile does not match	5-Hydroxymethyluracil	Urine Blood	Guo *et al.* [261] Leyva *et al.* [262]
195.00	UV profile and charge to volume ratio appears correct	Series of isomers: Galactonic acid, mannonate, gluconic acid and gulonic acid	Urine	Greter *et al.* [263]
198.08	UV profile ok but should carry 3 negative charges and migrate slower than unknown	O-phosporyl-L-homoserine		
209.02	UV profile and charge to volume ratio appears correct	Series of isomers: Saccharic acid and mucic acid	Urine	Shoemaker *et al.* [264]
262.68	N/A	No candidate fit the CE and UV data	N/A	N/A
276.99	N/A	No known relevant metabolite at this value	N/A	N/A
330.97	N/A	No known relevant metabolite at this value	N/A	N/A
344.97	3-4 negative charges and would migrate too late	5-O-(1-carboxyvinyl)-3-phosphoshikimate		
407.28	Not very soluble in water and at best at ~2 µM concentration including all isomers	Several possible cholic acid isomers	Newborn blood Urine Bile	Gustafsson *et al.* [265] Batta *et al.* [266] Tadano *et al.* [267]

Mannoate, gluconic, gulonic and galactonic acids, as well as saccharic and mucic acid, were purchased from Sigma-Aldrich and used in spiking experiments to determine if any of them were the species corresponding to the unknown. Although the migration times were close, none of these candidates matched the migration time of the unknown.

4 Other spiking experiments

With the MS experiment not yielding any definitive results and the fact that dioic acid did appear to have migration time that corresponds well with molecules having such functionalities, the main acids involved in the citric acid cycle were considered. They are expected to be present in AF and at sufficient concentrations to be detected by UV. Thus oxalic, oxalacetic, fumaric, 2-ketoglutaric, malic, methylmalonic, glutaric and citric acids were purchased from Sigma-Aldrich. These were used in spiking experiments to see if any of them had a corresponding migration time with the peak of interest. Unfortunately none of them showed a migration time that matched with the migration time of the unknown.

Appendix V

Oxygenation study on MS profile of HSA

In an effort to assess the effect of saturating levels of oxygen on HSA in IVF media used in Chapter 5, samples were exposed to O_2. Additionally, to verify if small molecular weight species may have an impact on the media's ability to buffer oxidative stress, aliquots of the media were filtered through 30 kDa centrifugal molecular weight cut-off filters to remove species below 30 kDa. Both filtered and unfiltered media were divided into 3 aliquots. These were exposed to O_2 for 0 h, 1 h and overnight. All six samples were then analyzed by LC-MS to determine if any change in the ratio of HSA isoforms could be detected.

Analysis of peak integration values normalized to the total area of HSA isoforms showed minimal, if any, differences between samples exposed to O_2 for varied periods of time (Table 9). Only the freshly thawed sample showed any difference. Fresh media was injected at a HSA concentration of 90 μM whereas the others were at a HSA concentration of 150 μM. A concentration effect could account for the variation seen between fresh and O_2 exposed samples.

Table 9: Effect of oxygen saturation of unfiltered media on relative abundance of HSA isoforms.

HSA PTMs	rdHSA	HSA-SO_2H	HSA-$2SO_2H$	Sulfation	Cys-HSA	deoxi-HSA
Molecular Weight (Da)	66437.3	66467.5	66496	66520.8	66553.8	66587.2
Δ MW	0	30.2	58.7	83.5	116.5	149.9
Untreated	23.9%	16.7%	5.0%	10.7%	35.4%	8.2%
Incubation 24h no O_2	26.0%	13.2%	5.7%	9.6%	38.7%	6.8%
1h O_2+23h incubation	25.9%	12.7%	5.4%	9.0%	38.4%	8.7%
24h O_2	26.1%	13.2%	4.7%	8.1%	38.8%	9.2%
Average	25.5%	13.9%	5.2%	9.3%	37.8%	8.2%
Standard Deviation	1.0%	1.9%	0.4%	1.1%	1.6%	1.0%
Relative error	4%	13%	9%	11%	4%	12%

Similarly, the intact protein profile (spectra not shown) for the filtered sample do not show any differences with respect to the exposure time to O_2. The data are summarized in Table 10.

Table 10: Effect of oxygen saturation of filtered media on relative abundance of HSA isoforms.

HSA PTMs	rdHSA	HSA-SO_2H	HSA-$2SO_2H$	Sulfation	Cys-HSA	deoxi-HSA
Molecular Weight (Da)	66437.3	66467.5	66496	66520.8	66553.8	66587.2
Δ MW	0	30.2	58.7	83.5	116.5	149.9
Untreated	24.2%	15.2%	3.7%	10.4%	37%	9.0%
Incubation 24h no O_2	23.9%	12.5%	3.4%	9.5%	43%	8.1%
1h O_2+23h incubation	22.1%	13.2%	3.5%	10.3%	41%	9.9%
24h O_2	25.7%	13.2%	3.1%	9.3%	41%	7.9%
Average	24.0%	13.5%	3.4%	9.9%	40.%	8.7%
Standard Deviation	1.5%	1.2%	0.2%	0.6%	2%	0.9%
Relative error	6%	8%	7%	6%	5%	11%