Characterization of Human Skin Emanations by Solid Phase Microextraction (SPME) Extraction of Volatiles and Subsequent Analysis by Gas Chromatography-Mass Spectrometry (GC-MS)

by

James Akin

B. S., Chemistry
United States Air Force Academy, 2003

SUBMITTED TO THE DEPARTMENT OF MATERIALS SCIENCE AND ENGINEERING
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN MATERIALS SCIENCE AND ENGINEERING
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2005 [June 2005]

Signature of Author. . . . . .
Department of Materials Science and Engineering
May 15, 2005

Certified by. . . . . .
Cristina E. Davis
Charles Stark Draper Laboratory
Thesis Supervisor

Certified by. . . . . .
Darrell J. Irvine
Assistant Professor of Materials Science and Engineering
Thesis Supervisor

Accepted by. . . . . .
Gerbrand Ceder
Professor of Materials Science and Engineering
Chair, Departmental Committee on Graduate Students

Characterization of Human Skin Emanations by Solid Phase Microextraction (SPME) Extraction of Volatiles and Subsequent Analysis by Gas Chromatography-Mass Spectrometry (GC-MS)

by

James Akin

Submitted to the Department of Materials Science and Engineering
on May 6, 2005 in Partial Fulfillment of the
Requirements for the Degree of Master of Science in
Materials Science and Engineering

ABSTRACT

An experimental study was performed to develop and validate a collection and analysis protocol for human skin emanations. The protocol developed included the rubbing of glass beads on the palms and backs of hands for 20 minutes. The volatile headspace above samples were extracted by a solid-phase microextraction fiber which incorporated a composite coating of liquid polymer matrix and solid porous particles. This protocol provided robust and convenient signatures of human skin emanations and was applied to two experiments for validation. In one experiment, a set of twins donated samples and results suggested qualitative differences between samples of twins. The second experiment involved collections from four unrelated individuals over a period of one month. Multivariate analysis was applied to this data set and indicated a stable signature that can be ascribed to the individual, confirming that the protocol developed here can be extended to larger sample sets of MHC typed individuals.

Technical Supervisor: Cristina E. Davis
Title: Member of the Technical Staff

Thesis Supervisor: Darrell Irvine
Title: Assistant Professor of Materials Science and Engineering

[THIS PAGE INTENTIONALLY LEFT BLANK]

Acknowledgments

6 May 2005

This thesis represents the culmination of two years of inquiry, contemplation, and effort. It has been an invaluable experience in my life, and there are many I would like to thank. I am indebted to my parents and family for the unconditional support they provided.

I would also like the thank the Air Force for allowing me this opportunity to pursue advanced study and research. I think I have learned many valuable lessons-- scientific, academic, and in leadership-- that will aid in my future endeavors as an Air Force officer. In particular, I would like to mention the efforts and mentoring of the graduate programs office at the United States Air Force Academy who made this opportunity first available to me. Also, since coming to Cambridge, the ROTC detachment 365 and its staff has been an invaluable resource.

My experience at Draper has been blessed with outstanding supervisors and colleagues. Thanks to George Schmidt and Loretta Mitrano in the education office for their excellent management of the fellows program with a personal touch. I am deeply indebted to Cristina Davis for selecting me to join this fascinating project and for her mentorship, support, and leadership throughout the last two years. Thank you to everyone in the Draper Bio-engineering group for always being patient, accessible, and helpful. Special thanks to Julie Zeskind, Melissa Krebs, and Maria Holmboe who provided continual support, suggestions, and ideas specifically on this project.

My time at MIT has been an incredible experience. I was continually amazed at the caliber of people I was fortunate enough to work with. Thanks to the all the faculty and staff of DMSE and especially to Darrell Irvine for his advice and support in preparing this thesis.

This thesis was prepared at The Charles Stark Draper Laboratory, Inc., under ARO Contract DAAD19-03-C-0069. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Publication of this thesis does not constitute approval by Draper or the sponsoring agency of the findings or conclusions contained herein. It is published for the exchange and stimulation of ideas.

This document was cleared** by DARPA on [5/11/2005]. All copies should carry the Distribution Statement A(Approved for Public Release, Distribution Unlimited). If you have any questions, please contact the Technical Information Office.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

This thesis focuses on developing a robust skin emanation collection and analysis protocol for use in studying volatile compounds found in human skin emanations. A collection method for this rarely-sampled body effluent was established to capture a robust biological signature using convenient techniques amenable to collecting large numbers of samples from many individuals. Also, the final sample needed to retain most of the biological chemicals present on the skin while still meeting constraints on collection time. In short, the analytical procedure needed be effective and efficient, making an engineering approach to this biological research problem very helpful. The final collection and analysis protocol for skin emanations was then tested on a large sample set of emanations collected from multiple individuals over time. We tested the hypothesis that the chemicals found in skin emanations were unique enough and also invariant over time to allow for separation of individuals by a chemical "odor fingerprint." This was shown using both experimental computational models and a conventional multivariate analysis of the compounds present. In the following sections, relevant background material from immunology, dermatology, and computational data analysis is covered. In the next section, the protocol development is discussed to shed light on the decision making process resulting in the final protocol. In the final section, the results from two experiments, a monozygotic set of twins and four unrelated individuals, are reported and used to validate the usefulness of this protocol. Future applications and work are briefly discussed.

In 1974 it was first suggested that human odor might be linked to specific polymorphic genes such as the major histocompatibility locus[1], a set of genes that control and enable the adaptive immune response. Since then a variety of studies—behavioral, chemical, and genetic in nature—have been conducted, and several hypotheses for a possible link have been proposed. With the advent of powerful computational techniques for data analysis and its application to proteomics, or broad-spectrum genome expression, entirely new kinds of research are now possible to elucidate the origin of human odor. A complex, multivariate "odor phenotype" or intermediate along the path to expression can be computationally modeled and analyzed for patterns of chemical components. Based on experimental design and selection of samples, models can be correlated with underlying factors, such as genetic sequence. In contrast to traditional studies where a large number of variables are rigorously defined and controlled from the outset and sample numbers are often limited, these new discovery based techniques leverage the power of computational methods with sensitive analytical instrumentation and rely on large sample numbers which capture as much biological variability as possible. While some of these studies are seeing application in clinical medicine, expansion into other areas of biological research is possible.

1.1 Adaptive Immune System

1.1.1 Major Histocompatibilty Complex

The human major histocompatibility complex (MHC) is a region of approximately 4 Mb of DNA located on the 6^{th} chromosome, comprised of some 200 genes[2]. The importance of this locus was first realized when tissue transplants were becoming a standard medical procedure, and compatibility between donors and recipients became an important concern. Interestingly, this issue arose specifically when physicians were trying to treat burned aircrew in World War II, and skin grafts were frequently rejected[3]. The original suggestion of self-presentation antigens occurred near the turn of the century, though, in the context of tumor transplantation research in mice[3]. However, this research was largely unnoticed, and it was not until the war that the MHC region was regarded as important. Since that time, more details of the MHC have been elucidated and discovered in a variety of species[4]. Of particular note, the human MHC is named HLA, also known as Human Leukocyte Antigen which is based on its intertwined discovery with blood group antigens. This region was also observed in other mammalian species, and in the mouse genome the MHC region is named H-2. Known functions for the locus include coding for cell-surface proteins present in every cell of the body which, under normal conditions, identify the cell as "self" to T-cells of the immune system. These MHC products (MHC molecules) are anchored in the cell membrane and present short peptides from the interior of the cell. In the event that a cell is infected, the MHC molecules would present a foreign antigen that would be recognized by the T-cell, targeting the cell for the immune response. In contrast, antibodies, located on the B-cells of the immune system and in circulation, can recognize and bind foreign antigens readily, without the process of presentation by MHC molecules.

MHC molecules are divided into two categories, Class I and Class II. Class I molecules are present in all nucleated cells of the body, and their purpose along with other MHC products is to process and present foreign antigens for recognition by cytotoxic T-cells (CD8 T-cells). Class II molecules are present only on certain lymphocytes, i.e. dendritic cells, macrophages, and B-cells. Their purpose is to present antigens for recognition by other effector cells, e.g. helper T-cells, which amplify the immune response through release of cytokines and activation of B cells. Non-nucleated cells in the body, e.g. red blood cells, express little or no Class I or Class II molecules. This is presumably due to the fact that non-nucleated cells cannot be a host for invading viruses or intercellular pathogens. The malaria plasmodium being a notable exception, which spends part of its life cycle within erythrocytes, and its evasion of detection during this phase is, in part, made possible by the lack of MHC presentation molecules.

In addition, the processing and presentation pathway within the cell is different for antigens that will be presented on Class I versus Class II MHC molecules. Antigens destined for Class I presentation start out in the cytosol of the cell, whereas antigens destined for Class II presentation are picked up by endocytosis by the specialized immune surveillance cells from extracellular space. They are conjugated to the MHC Class II molecules through a different route in the endosomes of the cell.

Class I MHC molecules are composed of an alpha-chain consisting of 3 domains[5]. The first two domains form the peptide binding cleft which consists of two alpha-helices on top of an anti-parallel beta-sheet. MHC Class I molecules can accommodate peptides ranging from 8 to 10 residues in length[5]. Compared with the MHC Class II molecules, C-terminal and N-terminal ends of antigens bound in the MHC Class I cleft are significantly buried[5]. The third alpha-chain domain resides closest to the cell-membrane and is attached to the membrane spanning sequence. A fourth domain, the β_2-microglobulin, is non-covalently associated with the alpha-chain, and it is an invariant subunit not encoded within the MHC.

Class II MHC molecules are composed of two chains (α and β) both of which span the membrane. These two chains are non-covalently associated with each other[5]. Again, a similar peptide binding cleft is formed by the first domain in each chain ($\alpha 1$ and $\beta 1$), and the cleft consists of two alpha-helices on top of an anti-parallel beta-sheet. It is important to note that the two halves of the peptide binding cleft are formed by parts of two chains which are non-covalently associated, whereas in the Class I case the cleft is formed by different regions of one chain that are covalently linked. Because of these structural differences, the Class II peptide binding cleft is more open at the ends and, therefore can accommodate peptide segments of much greater length, typically 10-20 residues in length[5].

When an antigen peptide is bound in the binding cleft both the peptide and the bordering alpha-helices of the cleft are in extended conformations. Therefore, the antigen peptide and the MHC molecule mutually stabilize each other through binding. This feature is necessary to ensure stable and tight association between the peptide and the MHC molecule. If the peptides were able to disassociate and bind freely with the MHC cleft, than exogenous peptides could be picked up from extra-cellular spaces and result in targeting of a non-infected cell by the immune response.

These known functions and structures of the MHC molecules have been the starting point for possible theories connecting human odor to the MHC locus.

1.1.2 Haplotypes and Conserved Extended Haplotypes

The genes of the human MHC are quite diverse, with some genes having hundreds of possible alleles. The numbers of alleles (identified as of January 2004) for human MHC genes are given in Figure 1 below [1]. This polymorphism is in accordance with the function of the immune system. Such variety is necessary to ensure an ability to adapt to numerous pathogens and provide for the continuance of the species. Interestingly, despite the polymorphism, it has been observed that certain combinations of alleles have been well conserved in large populations across generations. While most of these studies have been conducted in developed, western populations these conserved combinations of alleles, or conserved extended haplotypes (CEHs), seem to be specific for ethnicity and

nationality. For instance, in the American-Caucasian population, seven specific haplotypes show up with a frequency greater than 1%[6].

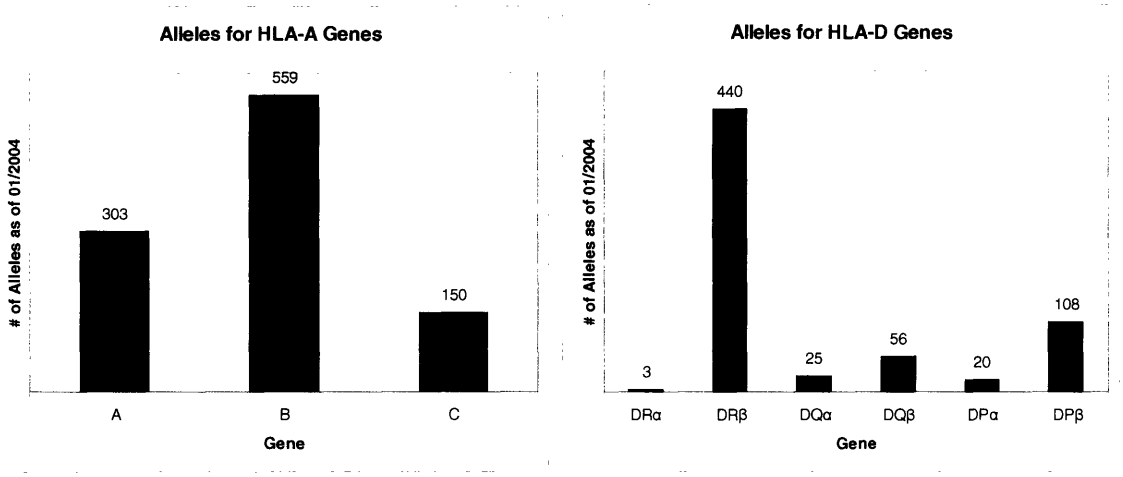


Figure 1: Number Alleles Identified for HLA-A (left) and HLA-D (right) MHC Genes

The alleles typically differ at select residues along the peptide groove. While most of the residues are located in the alpha-helices bordering the groove, some polymorphic residues are also located on the beta-sheet forming the floor between the helices. The residues generally form charged pockets that associate with amino acid side chains in the antigen peptide. Certain specific patterns of amino acids are recognized by any one groove, so the MHC molecule can bind a family of antigens, related by a pattern in their sequence (for example, ...N-N-P-N-N-N-P-N-N-A..., could be a pattern where N stands for non-polar residue, P for polar residue, and A for acidic residue).

1.1.3 Origin of MHC Related Volatiles

The role of olfactory cues in the animal world is well established and recognized. Animals use odors to communicate a broad range of characteristics such as sex, age group, reproductive status, and identity[7, 8]. More generally, an "odortype" has been defined as a secondary genetic trait, comparable to other modes, such as visual recognition (human), used to convey important societal information[9]. Several hypotheses have been proposed to explain the pathway from the HLA and MHC polymorphic genes to the odor phenotype. Volatile signatures of blood, urine, and skin emanations have been correlated to MHC types, mainly in mice, but also in rats and, more tentatively, humans[7, 9]. Initially, studies of H-2 inbred mice suggested that odortypes might be related to the MHC. These studies involved trained mice in a Y-maze that could distinguish odors of other mice, who differed only at the H-2 locus. Later, studies were conducted using mice that had not been trained. Instead, the habituation and dishabituation behaviors of the mice to a series of odor signals were monitored. In these murine studies, filtered and derivatized urine samples were the biological fluid of interest. Mixtures of volatile carboxylic acids have been found to occur in urine and contribute to odor cues in mice[10], as well as in skin secretions of mongooses[8]. These results have been extended to protease-treated serum, under the assumption that volatiles are conjugated in the serum and need to be liberated to form the volatile signature[9].

Another interesting idea is that volatile signatures are related to human MHC genes, but not necessarily to the classic A, B, C, and D loci with known antigen presentation roles. Another set of class I genes exist, E, F, and G, the allelic distribution and function of

which have not been nearly as well-studied[11]. These genes would be inherited along with and in the same manner as the classical MHC genes. Once more is known about these genes and their products, possible mechanistic details between MHC and odor could be further illuminated, or ruled out.

The "carrier" hypothesis of the origin of odor was proposed based on the known function of MHC products, to bind and present small peptide fragments (i.e. antigens in an infected cell). It was shown that MHC molecules were present not only in cell membranes, but also free in circulating plasma[12]. This seems reasonable since active immune surveillance requires continual turn-over of MHC-ligand complexes. In the next step of the pathway, these MHC molecules would be filtered out by the kidneys and released into the urine. However, these large proteins have a low vapor pressure, so it is unlikely they contribute directly to the volatile "odor" signal. On this basis, it has been proposed that the MHC molecules associate with small molecules while in circulation, e.g. small metabolites, or other wastes, and carry these through the kidneys and into the urine. Once in the urine, and after excretion from the body, the molecules denature, or otherwise change conformation, releasing the smaller volatiles into the atmosphere, creating a unique volatile odor. The nature and relative proportions of these volatile components would be dependent on the MHC molecules and their production and assembly, therefore ultimately derived from the MHC genetic code. The direct conjugation of volatile carboxylic acids (the one set of compounds that has been correlated to MHC type in mice) with the MHC molecule is unlikely, since the MHC

peptide groove is intended for larger, conformation-specific peptides. However, some carboxylic acids, such as phenyl acetic acid, have been identified[10] and are known to conjugate with amino acids, such as glycine and taurine, and so including this additional conjugation step in the mechanism is necessary.

1.2 Background Literature for Skin Emanations

1.2.1 Anatomy of Human Skin

The skin is the largest organ in the human body and serves a variety of critical functions. Throughout this study, the skin has been approached as a boundary between the organism and the environment, which facilitates maintenance of the organism's internal equilibrium. Most obviously, it is a barrier to the outside world, and the first line of defense against infection and colonization by pathogens. Also, it can serve as a reservoir for excess water, salt, and metabolic byproducts. Specific to humans, the skin serves an efficient thermoregulatory function through perspiration, i.e. water is excreted onto the surface and absorbs energy through heat as it evaporates. Finally, the skin can often be an indicator of internal status, e.g. the pale color often assumed during illness, reduced turgor indicating dehydration, and odors related to a spectrum of conditions from typhoid to schizophrenia [7, 13, 14]. A variety of cellular, sub-cellular, and supra-cellular structures combine to meet the disparate requirements of this organ and vary according to anatomical region. The following sections will highlight the pertinent details of these structures.

1.2.1.1 Extracellular and Cellular Structures

Human skin consists of three major layers of tissues, each of which can be further divided into various sub-layers. The three main divisions are epidermis, dermis, and basal lamina. The epidermis is the uppermost layer (nearest the exterior environment) and is comprised of several strata (5 de facto layers) with differing compositions of keratin-filled, dead cells, or keratinocytes. The stratum corneum (SC) is the outermost of these layers. It contains the furthest differentiated, squamous keratinocytes, called corneocytes, or horny cells. According to the current "bricks and mortar" model, these corneocytes (bricks) are seeded into a lipid matrix (mortar) exhibiting complex phase behavior[15]. The current description of this phase behavior and the composition of the lipid matrix, called the domain mosaic model, proposes regions of crystalline phase surrounded by a fluid, liquid crystalline phase[16]. This organization of the matrix putatively accounts for the semi-permeable nature of the skin. Limited diffusion can occur in the liquid crystalline phase, and excessive water loss and absorption are avoided while retaining the ability for mechanical flexibility[16]. This is presumably one reason the palms of the hands and soles of the feet, where lipid secreting sebaceous glands are sparse, become "waterlogged" after extended immersion in water while other regions of the skin do not. In this sense the, albeit thicker, SC located on the soles and palms can also be thought of as a more permeable region to the interior of the organism. While these lipids are important, the skin tissue contains and excretes other classes of bio-molecules also, and these will be discussed in a subsequent section.

The dermis is the thickest layer, lying just beneath the epidermis and provides many support functions for the epidermis. All of the glands and hair follicles are rooted in this tissue. In addition, capillaries are imbedded in this tissue and supply nutrients to the surrounding cells through diffusion from the capillary wall and into the tissue. The dermis is also innervated and hosts cells of the immune system. In particular, Langerhaus cells are multinucleated, dendritic cells thought to provide necessary antigen collecting and processing functions. It has also been recently discovered that MHC Class I-ligand complexes can be transferred to Langerhaus cells by surrounding keratinocytes through gap junctions[17, 18]. The basal lamina underlies the two aforementioned layers. Consisting of mainly collagen and laminin, this thin layer provides the boundary of the organ and the structural support from which cells are further differentiated into dermal and epidermal cells.

1.2.1.2 Secretory Glands: Apocrine, Eccrine, and Sebaceous

Three types of secretory glands exist in human skin—eccrine, sebaceous, and apocrine. According to common usage, only the eccrine and apocrine glands are considered "sweat glands", and sometimes even this term refers only to eccrine glands. Sebaceous glands are not considered sweat glands, but can be more generally classed as secretory glands. Since ultimately the substances under consideration in this study are derived from both sweat glands and sebaceous glands, the term "skin emanations" has been adopted, rather than sweat. Other terms in use include skin secretions, or exudates. However the term

emanations was preferred as it more clearly captures the volatile nature of the signal analyzed in this study.

While all three types of glands are found throughout the skin in aggregations that produce secretions, a primary function of eccrine sweat glands is to cool the body during vigorous work. During this perspiration, eccrine glands produce mostly water, which evaporates to cool the skin, and some salts. Eccrine glands are located throughout the body, with the highest concentrations (and the highest concentrations of any secretory gland) located on the soles of the feet and palms of the hands. Eccrine glands are cholinergically stimulated during vigorous work, but also can be stimulated (in the palms) para-sympathetically during emotional stress. By contrast, sebaceous glands are associated with hair follicles and continuously secrete oils, or sebum. This milky-white secretion hydrates and preserves the natural tincture and health of the outermost layers of keratinocytes as discussed above, but also plays a role in creating an odor signal[8, 14]. Sebaceous glands are distributed through the body, and are most concentrated on the face and scalp. The third type of secretory gland found in skin is the apocrine gland. It is commonly believed, due to the distribution of this gland, that it is primarily responsible for scent production in social communication, although no human pheromones have been identified yet[19]. These glands first develop during puberty and are androgenically stimulated. They are located primarily in the axillae and ano-genital regions. Also, in these areas, a hybrid apoeccrine gland has been identified that also develops during

puberty. Now that the overall structure of the skin and the specific secretory glands have been described, the composition of the resulting secretions will be covered.

1.2.1.3 Molecules in Skin Secretions: Lipids, Proteins, and Aqueous

Lipids in and on the skin tissue are unique compared to the vast majority of lipids within the organism[20]. Lipids within the organism can be classified into two basic categories based on their function: structure and storage. Lipids are the major structural constituent of the lipid bi-layer which is critical in cellular and intracellular compartmentalization. Other internal lipids, in the form of triglycerides, form a critical component of metabolic energy storage. Skin lipids, however, have specialized functions, other than structure or storage, and their chemical structures are accordingly unique.

The aqueous portion of skin emanations is primarily derived from the eccrine sweat glands as mentioned above. This secretion consists of mostly water with some dilute salts. The body maintains homeostatic levels of salts in the blood and fluids by releasing excess salts through this aqueous secretion. In the case of the genetic disorder cystic fibrosis, excess chloride ion is transported out of cells lining the lumen of the gland due to a faulty ion transport membrane protein. The proteinaceous component of skin emanations can be divided into two groups of molecules: a set of small, antibiotic proteins (4-7 kDa) and a set of large serum derived proteins (50+ kDa). Soluble portions of MHC Class I and Class II proteins have been identified in dilute eccrine secretions and belong in the latter group.

1.2.2 Collection Methods for Skin Emanations

Human sweat is a rarely-sampled body effluent, and not extensively analyzed for chemical components. Both human urine and human plasma, for example, have relatively straightforward and well-established collection and preparation protocols, and the chemical composition is also well established. Analysis of human blood and urine are important diagnostic tools for many conditions and illnesses, increasingly with the emerging fields of bioinformatics and proteomics. For example, recent studies have shown the viability of analyzing serum proteins for the early diagnosis of ovarian, breast, and prostate cancers[21, 22]. This has been accomplished by analysis of protein components and peptides via time-of-flight mass spectrometry (TOF-MS). Sophisticated computer algorithms are then applied to these large and complex data. In effect, they are mined for "biomarkers", or specific ratios of mass to charge values (m/z values), that segregate the diseased state from the healthy state. Mass to charge values are, of course, related to specific chemical structures and, therefore, proteins. Also of particular note, a recent study was conducted where the mass spectrometer and computer algorithm were replaced by the canine nose in the analysis of urine for detection of bladder cancer. These studies are all the more exciting since early diagnosis of these diseases has a marked effect on survival rates. Analysis of skin emanations may have the potential to be an equally important diagnostic tool. Specifically, the skin is also a reservoir for metabolic wastes, and so it would seem reasonable that small metabolites associated with a disease state would likely be purged as waste. Regarding this waste excretion function,

the skin is also a distributed organ, in contrast to the bladder. In a limited fashion, skin emanations already are used in the clinic, for example, in diagnosing cystic fibrosis in children. Collections methods though are tailored for specific needs—i.e. in the case of cystic fibrosis the analysis specifically quantifies the concentration of chloride ions. No standard collection protocol is present for collecting human skin emanations.

A variety of collection protocols from current literature were reviewed and include the following: a condensed flow of nitrogen over arms and hands[23], washing of the skin with solvents such as ethanol[24], and collection onto glass in a variety of forms[25-28]. As implied above, another set of specialized collection methods has been developed as a means of diagnosing cystic fibrosis. A few of these collection protocols will be discussed below.

1.2.2.1 Flowing Nitrogen over Skin

Several studies used a volatile effluent collection device. This was a "home-made" device into which the donor would place his or her hands. A flow of nitrogen gas was passed through the interior of the chamber, over the donor's skin, and then cryo-trapped in a liquid nitrogen cooled vessel. The advantages of this collection protocol include being able to collect a true volatile signature directly, without the need for an intermediate collection substance or absorbent. The disadvantages of this device include its design and fabrication and the need to wash, or otherwise clean, the collection chamber between donors. In addition, donors would then be required to come to a

specific collection site for analysis of skin emanations, and the ultimate use of this technology would not be portable or useful to a wide scientific audience.

1.2.2.2 Other Collection Methods

Other collections methods used in the literature include sweat droplet collection and droplet collection with the MACRODUCT device[29]. Sweat droplets can be formed as a result of strenuous exercise, or can be stimulated to form by pilocarpine, a cholinergic compound applied to the skin surface. In one study[29], it was found that the MACRODUCT device was able to collect a set of proteins that differed between sexes in humans. Collecting the sweat droplets via direct collection into a vial, instead of using the MACRODUCT device, were insufficient to allow for this discrimination of protein patterns.

1.3 Analytical Methods Used for Skin Emanations

1.3.1 Methods for Sample Preparation and Introduction

The power and usefulness of GC-MS has been extended greatly by improvements in sample preparation techniques. Conventional capillary gas chromatography usually indicates small liquid samples of a few micro liters which are volatilized upon injection, or samples which are in a gaseous state to begin with. However, a variety of techniques have been created or adapted to meet the increasingly diverse needs of a growing user community. Three of the options considered for this study—thermal desorption, static

headspace, and solid-phase microextraction (SPME)—will be discussed below, with a more detailed discussion of the latter, which was ultimately selected for all samples.

1.3.1.1 Thermal Desorption of Deposited Emanations

Thermal desorption is an important route for sample preparation in GC-MS. Solid or liquid deposits on a substrate, such as glass or a SPME fiber are heated beyond their sublimation temperature. The analytes are then transported into the carrier flow of the instrument and become, at this point, similar to a standard liquid injection that has been volatilized in the inlet. It is important to keep the flow path increasing in temperature. If at any point before the detector the temperature dips below the dew point of the analyte, it can condense back out of the vapor phase and be deposited in the instrument.

1.3.1.2 Static Headspace

Static headspace represents often a high-throughput and prevalent technique for sample preparation in GC-MS. Headspace analysis relies on the vapor liquid equilibrium of analytes, usually dissolved in organic solvents, but also water. In order for static headspace to be useful, a sample has to have a significant portion of its components with a high vapor pressure. More molecules can be volatilized by heating of the sample, or mechanical agitation. It is often used in environmental studies and other standard, routine GC-MS analyses. Static headspace was attempted in the initial studies, but skin emanations were of such a low abundance that the instrumentation was not able to detect a significant signal above random noise.

1.3.1.3 Solid Phase Micro Extraction (SPME)

Solid Phase Micro Extraction (SPME) has become a useful and prevalent analytical tool, especially for the analysis of volatiles. SPME usually implies that the extraction device is in a polymer fiber form; however other forms are available such as the Twister® device described in a later section. This section will discuss the history and theory behind SPME, as well as the various applications, advantages, disadvantages, and special considerations in using SPME for extraction of volatiles.

A variety of SPME phases are available commercially, and phases can also be custom made for individual needs. Of particular relevance to this study, most volatile extractions are best conducted with mixed phases that include poly(divinylbenzene), or DVB. The DVB additive is usually present as solid, porous particles, seeded into a matrix of a liquid PDMS phase. Another commercially popular option replaces the PDMS matrix with Carbowax®, a poly(ethyleneglycol), or PEG, with an average molecular weight around 20,000 amu. The increased polarity of the PEG over the PDMS matrix increases slightly the affinity for polar compounds. However, most interactions of volatiles will occur on the large surface area of the porous DVB particles through an adsorption mechanism, rather than through a diffusion mechanism into the liquid matrix. Therefore, these mixed phase fibers are more useful in extractions of volatiles over condensed phases[30]. In addition, because of this adsorption mechanism, the extraction times are significantly shorter as compared to the single phase fibers[31]. On the other

hand, though, this mechanism also leads to a shorter dynamic range for the extraction and increased competitive displacement[31].

1.3.2 Gas Chromatography and Mass Spectrometry (GC-MS)

1.3.2.1 Nature of Data and Sensitivity

The combination of a gas chromatograph for separation followed by mass spectrometry for identification (GC-MS) is a powerful analytical tool that has become increasingly refined and standardized over the last half century. For this study it is a proven technology for detection of volatiles, but also provides a robust amount of data. The mass spectrometer used for this study was an Agilent 5973N (Palo Alto, CA). In quadrupole mass spectrometry (MS), chemical compounds are fragmented and become charged by electron impact (chemical ionization is also possible). The masses are then filtered by a scanning electric field and allowed to contact a detector which registers the mass scanned by the filter and the charge transferred producing a mass-to-charge ratio (m/z). Each compound has a unique fragmentation pattern, and a spectrum of masses can be used for identification of chemical structure. Each spectra is correlated to the time they are retained by the chromatographic column, or retention time (RT). A typical data file consists of a RT axis, an ion count (abundance), and an m/z axis. Due to the large and complicated data acquired, however, efficient and thorough analysis requires methods borrowed from chemometrics, such as principle component analysis (PCA), and genetic algorithms.

1.3.3 Pattern Recognition by Genetic Algorithms

With the revolutionary development of computer processing and its subsequent advancement in power and speed, genetic algorithms have become an increasingly useful tool for multivariate optimization, machine learning, and neural networks[32]. Analysis by genetic algorithms borrows its methods and terminology from biology, and these analyses hinge upon self-evolving numerical models of natural systems, which are inherently complex[33]. A genetic algorithm generates potential solutions to a problem (chromosomes), judges each solution as to how well it solves the problem (determines the fitness), and then generates a new set of solutions preferentially combining and reproducing more fit solutions from the first set (recombination of fit chromosomes). Each iteration of this process is considered a generation.

In this study, the problem is classifying a group of individuals based on the mixture of volatile compounds in their skin emanations. As applied to the GC-MS data, the solution (a chromosome) is a set of RT-m/z coordinate pairs (genes) that, in their relative differences in abundances, are able to segregate the four individuals. A RT-m/z coordinate pair is related to a specific chemical structure, so each gene can be connected with a volatile component. For a simplified qualitative example, let's say that the four individuals were readily classifiable by the presence or absence of a particular compound. That is, person A's skin emits only compound A; person B's skin emits only compound B; and so on. The solution might then be a chromosome whose genes were the identifying RTs and m/z values for compounds A, B, C, and D. The relative abundances

would be as follows in Table 1 for each individual, where a value of 1 is used to indicate 100% abundance since it is absent in samples from all other individuals. This solution also applies to the more complex quantitative case, where all four individuals each emit all four compounds, but in characteristic relative concentrations.

Table 1: Hypothetical Chromosome for Genetic Algorithm Classification

	Compound A	Compound B	Compound C	Compound D
Person A	1	0	0	0
Person B	0	1	0	0
Person C	0	0	1	0
Person D	0	0	0	1

The process of finding the solution is as follows. Data files from each sample are randomly segregated into training data and validation data. The genetic algorithm begins by randomly generating a population of chromosomes. Each chromosome's fitness is determined based on how well its genes are able to classify the four sets of samples (from person A, B, C, and D) used in the training data. The more fit a chromosome, the more likely it will be selected to combine with another chromosome in the population to generate two new chromosomes (offspring) in the next generation. After two new chromosomes are formed, two will be removed (culled) from the entire pool to keep the total population at a constant number. The less fit a chromosome, the more likely it will be removed during this process. In this way, the entire gene pool is constantly being refined to find the most fit chromosome available. As fit chromosomes join to create offspring their genes can recombine so that new gene combinations can be generated.

In the ProteomeQuest® genetic algorithm developed by Correlogic Data Systems®, Bethesda, MD, user defined inputs include the number of genes per chromosome, number of generations, and number of chromosomes. In addition, two more inputs (match and learn) influence the spatial separation and combination of chromosomes in the fitness space. The likelihood of re-combination occurring and the likelihood of culling are finite probabilities derived from the fitness function. Probabilities are used to incorporate an element of randomness, also a hallmark of natural systems, to limit artificial fitting of a model to a set of data.

In 2002 Petricoin and Liotta described a rapid screening process for ovarian cancer and prostate cancer that involved collecting serum samples from patients, analyzing them by SELDI-TOF, and searching the data for biomarkers of cancer with the ProteomeQuest genetic algorithm[21, 22]. Through this procedure, they reported 100% sensitivity and 100% specificity for discriminating a diseased state from a non-diseased state in ovarian cancer[34]. There is some discussion in the literature regarding the confidence in these results[35-37], but the prevailing opinion regards this research as novel and promising, and it can be complementary to well established disease biomarkers, such as prostate specific antigen (PSA)[38-40]. The current discussion in the literature will serve to standardize experimental procedures and analysis methods for discovery based research in proteomics, which requires multivariate, computational approaches. In effect, this represents another joining of disciplines under the broadly applied term of bioengineering.

1.3.4 Data Processing Concerns- Alignment of Retention Times

Pattern recognition techniques such as genetic algorithms and PCA are sensitive to small shifts in RT[41, 42]. These RT shifts can confound building a proper model. For example, a series of chromatograms belong to, in this case, samples from one individual, and a specific component elutes at a RT of 20.00 minutes. However, over the set of samples the RT shifts between 19.00 and 21.00 minutes, and this component will not be used as a marker belonging to this individual due to this variation. The nature of these RT shifts is due to random shifts in partitioning through the column and necessary maintenance of the instrument which leads to trimming of the column is non-linear. In order to account for this an alignment algorithm[43] was modified in MATLAB to include reference files from four individuals. The code for this is given in the appendix.

2. MATERIALS AND METHODS

2.1 Initial Studies for Protocol Development

The majority of the experiments during protocol development were intended to reproduce the results and verify the compatibility of the glass bead collection protocol as described in the aforementioned studies of mosquito attractants[26]. However, a few other ideas were also explored, and some of the results obtained will be highlighted. This will help describe the developmental process resulting in the final protocol. Table 2 below gives a summary of the experiments conducted during protocol development. The numbers refer to the amount of samples collected and analyzed under the indicated protocol. The

column on the left gives the types of substances considered for collecting the skin emanations. The columns on the right give the various combinations of sample preparation (SPME, headspace, or thermal desorption in the inlet) and analytical instrument—FAIMS, flame ionization detector (FID), or mass selective detector (MSD). The grayed out boxes indicate a collection-analysis combination that was incompatible.

Table 2: Summary of Experiments Conducted During Protocol Development

Collection Substance	Sample Preparation and Analytical Instrument				
	SPME		Headspace		Inlet
	FAIMS	FID/MSD	FAIMS	FID/MSD	MSD
Glass Beads		14	20	33*	17*
Silica Beads				4	
Glass Beads, Heavy Sweat		4		2	
Beads with Gloves		1		8	
Filter Paper, Heavy Sweat				2	
Inverted Vial		7		1	
Sock Odors		6			
Twister Stir Bars					8

2.1.1 Anatomical Locations for Sweat

As outlined in a previous section, secretory glands in the skin have variable distributions throughout the body. In order to explore the nature of the volatile signature produced by the body, experiments were conducted to focus on certain anatomical regions. Some of these experiments involved adapting a collection onto glass beads to regions other than the backs and palms of hands, e.g. forehead, face, and back of neck. These regions were chosen because of the abundance of sebaceous glands found there. In addition, a few

experiments were conducted using an inverted headspace vial, with its opening placed snuggly against the skin, in a variety of locations. This idea was intended to directly capture volatiles escaping from the skin surface, small volatiles that were perhaps being missed by the other techniques. Most of these experiments proved inconclusive, but a few are worth noting because of key considerations they provoked, i.e. sock odors, induced perspiration, Twister® extraction, and glass beads rubbed on hands (which was ultimately adopted as the final protocol). The experimental designs had to balance the constraints of time and access to equipment with the most beneficial scientific payoffs. Therefore, the following sections and initial results are not intended as a thorough and rigorous characterization of variable body odors, but rather a logical and expedient search for a robust, useful, and convenient volatile signature for skin emanations.

2.1.2 Collection Substances

2.1.2.1 Socks Worn on Feet for Three to Four Hours

To test the viability of collecting odors from fragments of clothing, an experiment was designed and executed to analyze the odors produced from worn socks. Nylon socks were selected since this material is less absorbent than cotton and, being synthetic, would presumably have a less abundant organic volatile signature. The socks also needed to be thin, in order to minimize the relative amount of foreign material and to facilitate insertion into the vials used for analysis. Three donors volunteered to wear the socks for three to four hours. Two sections were cut from each sock, and each segment was placed

into a 10 mL headspace vial for analysis. A circumscribed midsection of the sock was selected in order to include both eccrine glands located on the sole and sebaceous glands more abundant on the top of the foot. The toe section of the sock was selected in order to collect from a region conventionally regarded as odiferous. Unworn socks were prepared by the same method and analyzed to provide a background or control signature of volatiles. In summary, the sock odors produced a complex signal of many overlapping peaks. This achieved the desired robustness of the signal, but the background signal from the socks themselves turned out to be significant. Due to potential confounding effects of this high abundance background signal, a more inert substrate, such as glass beads or the Twister SPME device was preferred. These two approaches will be described shortly.

2.1.2.2 Induced Perspiration by Exercise

Induced perspiration is a logical and obvious method for sweat collection, and a few experiments were attempted to explore its viability. Three donors volunteered to collect sweat after rigorous exercise. The donors were provided with two vials containing a piece of filter paper, or a set of cleaned glass beads. The donors were instructed to wash their hands after exercise, open the vial marked control and expose the filter paper (or glass beads) contained within to the ambient air. The second vial was then to be opened. The filter paper (or glass beads) was to be applied to the skin in a region with abundant sweat and then returned to the vial which was then sealed. The vial containing the background signature was particularly necessary since samples were being collected offsite. In summary, samples from these studies provided a weak signal, presumably due

to the dilute nature of induced perspiration. Most of this fluid is water, with small amounts of electrolytes, and this accordingly enables its evaporative and thermal regulatory function. From these experiments, it was determined that a collection protocol which did not induce heavy perspiration, but rather passively collected the natural secretions and emanations was preferred. This would minimize the relative amount of water, which can be thought of as a stable matrix that would dilute and trap volatiles. In addition, including exercise in the sample collection protocol would ultimately reduce the number of samples that could be collected due to donor volition.

2.1.2.3 Twister® SPME device

The Twister® SPME device is a magnetic stir bar which has been coated in an unusually thick layer of poly-dimethyl siloxane (PDMS), traditionally used in standard SPME fibers and, also, chromatographic columns. The advantages of this device over standard SPME is its ability to be immersed in a sample (either liquid or gaseous) and then agitated via a magnetic flux, producing presumably a more robust and thorough extraction. In the case of skin emanations, due to the thick nature of the PDMS phase, the Twister® device could be grasped in the hand without significant damage to the PDMS phase. In the case of a traditional SPME fiber, due to differences in processing and geometry, such handling tends to destroy the polymer phase. Since the Twister® device could then be directly desorbed in the GC inlet, the capability of being handled allowed direct extraction of skin emanations from the hands of the subject, removing the intermediate headspace extraction step involved in the other methods. However, the thick and robust nature of

the PDMS phase is also a drawback in that collection times are significantly increased. The extraction mechanism operates by diffusion into and out of the liquid polymer phase and is much slower than that encountered in the traditional SPME fiber. A labeled picture of the Twister® extraction device is given in Figure 2 below.

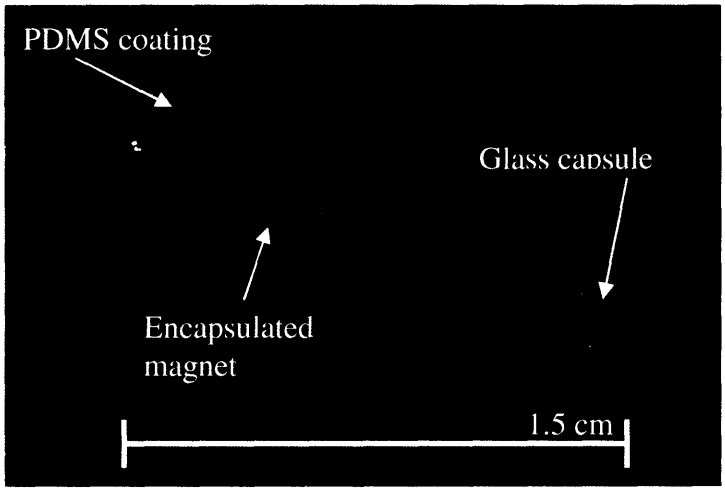


Figure 2: Picture of Twister® PDMS Extraction Device

2.1.2.4 Glass Beads Rubbed on Hands

Rubbing of glass beads on the hands has been used extensively in studies of volatile components which naturally attract mosquitoes to human hosts[26]. Borosilicate glass beads, originally intended for culturing purposes, provide a fairly inert substrate on which to collect human skin emanations. The beads chosen were 3 mm in diameter, which is small enough to fit in a variety of devices for thermal desorption (as discussed in a previous section), and yet big enough for the donor to handle without dropping. For a collection, 30 beads were chosen as a median value. With more beads it becomes

difficult and awkward for the donor to rub the beads in their hands. With fewer beads the surface area on which to collect emanations is reduced. For a sample of 30 beads, roughly 848 square millimeters (or 1.30 in^2) is available to collect skin cells and emanations. A picture of the glass beads used is given in Figure 3 below.

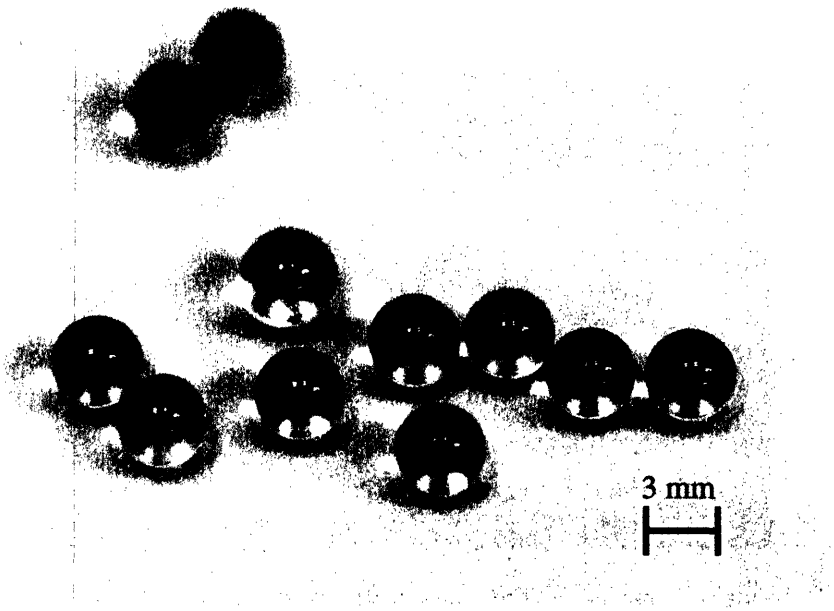


Figure 3: Picture of Glass Beads Used in Collection of Skin Emanations

2.1.3 Cryofocusing of Trace Volatiles and SPME Alternative

Skin emanations deposited on the glass beads presented difficulties for sample preparation. In the aforementioned studies of mosquito attractants, the glass beads were directly desorbed in the GC inlet. This technique was attempted and comparable results to that found in the literature were obtained. However, directly desorbing the beads in the GC inlet raised a few concerns. First, cellular components were present on the beads. This was often easily verified by visual inspection of the beads after a 20 minute

collection. White waxy flakes were found deposited on the beads and appeared to be sloughed corneocytes. Desorbing these coatings of cells directly in the GC inlet raised concerns that non-volatile cellular components were being volatilized, contributing compounds otherwise not present in a volatile signature. In addition such sample desorptions were inherently dirty for the instrumentation. Oftentimes, after retrieving a desorbed sample from the inlet, the previously white coating on the beads was often scorched and browned by the inlet temperature. Inlet liners had to be replaced after a few analyses and could not be reused without first cleaning the glass, and then deactivating it through a complicated and time-consuming process. Given the price of inlet liners, such desorptions were not feasible for large sample sets. In addition, the standard GC inlet installed in the instrument was not intended for this purpose. Temperature cycling of the inlet and repeated opening and closing of the inlet cap would induce accelerated wear. The temperature cycling itself became problematic, as it often took 1 to 1.5 hours for the inlet to cool down from its upper temperature in order to start a new analysis.

Cryofocusing of the column was necessary for direct thermal desorption from the beads in order to provide a high abundance signal for trace volatiles (see the diagram in the Appendix for the instrument setup). Without cryofocusing of the column, the volatiles would enter the column in trace amounts throughout the entire run and be present in the resulting signal as a low broad peak nearly indistinguishable from the baseline[44]. The technique involved immersing the column in a bath of liquid nitrogen during the initial part of the run, i.e. the first 10 minutes. During this time the beads

remained in the closed inlet, and the inlet temperature was ramped to 250 °C. After this initial 10 minutes, the column was removed from the liquid nitrogen bath, wound back around the column basket, and the door to the oven was closed. The temperature programming for the GC analysis began, and the focused slug of analytes began to partition and travel through the column normally. The inlet was operated in a splitless mode, with a low flow of helium carrier gas, throughout the entire analysis.

However, it was discovered that the sub-ambient temperature of the liquid nitrogen bath and the frequent temperature cycling of the GC method was damaging the polar phase of the column. This was apparent by the severe drop in abundances and exclusion of some peaks in the standard (mix of representative ketones) which was analyzed by the GC on a regular basis to ensure proper instrument function. This column damage was not observed on the non-polar HP-5ms column, but since the polar HP-WAXetr column had been selected for all the biological samples to allow for comparison (plasma, urine, skin emanations, and murine samples), a new extraction (or concentration) method had to be explored. Previous experiments had shown the promise of using solid-phase microextraction (SPME), and so an additional experiment was conducted to show that SPME could produce results similar to that obtained by direct desorption of the beads with subsequent cryofocusing. A summary of the results is given in Table 3 below. Both qualitative and quantitative criteria were established to determine the similarity between SPME and cryofocusing preparation techniques.

Table 3: Summary of Results from SPME versus Cryofocusing Experiment

	Person/ Day	Total #	% in Literature	Types of Compounds	Statistical Analysis
CRYO	**1**	22	64% (14)	CA, Aliph, Arom, **Cholest.**	SPME controls were more correlated to SPME samples than CRYO controls were to CRYO samples; both methods were reproducible
	2	36	28% (10)	CA, Ald, Aliph, Arom	
	3	29	38% (11)	CA, Aliph, Arom	
SPME	**1**	16	56% (9)	CA, Ald, Aliph, Arom	
	2	24	17% (4)	CA, Ald, Aliph, Arom, **Vit. E Ace.**	
	3	25	28% (7)	CA, Ald, Aliph, Arom, **EtOH**	

Abbreviations: CA-Carboxylic Acid, Ald-Aldehyde, Aliph-Aliphatic, Arom-Aromatic, Vit. E Ace.-Vitamin E. Acetate, EtOH-Ethanol, Cholest.-Cholesterol

2.1.4 Gas Chromatography Columns and Methods

All skin emanation samples were collected and extracted as described above. After adsorption onto the SPME fiber, the extracted volatiles were then desorbed into the inlet of the gas chromatograph, and the GC/MS method for separation and identification began. This method consisted of a 5 minute desorption of the SPME fiber at 250°C. In addition, the inlet was operated in a splitless mode, during which a relatively low helium carrier gas flow passed through the inlet and into the column. During these five minutes, the GC oven (and column) was set at 50°C. Volatiles desorbing off of the SPME fiber entered this helium gas flow and traveled to the top of the column mounted at the base of the inlet. At the top of the cooler column, volatiles partitioned into the liquid phase and formed a band just below the inlet. At the end of the initial five minutes, the inlet purge valve was opened and the gas flow rate increased 50 fold to sweep any lingering volatiles

to the top of the column, or otherwise out the purge vent. At this point, the temperature programming of the column began. Three temperature ramps were used to help better resolve the complex spectrum of peaks. This method had been optimized for SPME extraction of volatiles in a murine MHC study[45].

2.2 Selection of Donors

2.2.1 Twins and Other Samples Collected at CBR

Three sets of twins volunteered to collect skin emanation samples (according the protocol described above) at the Center for Blood Research (CBR), Boston, MA, where all of the blood and urine samples were also collected. All human samples were collected with informed consent and approval of CBR institutional review board. In addition, CBR determined the MHC-type of these donors. The data collected from these sets of twins represents the initial MHC correlated data collected for skin emanation samples. The MHC typing, as determined by CBR, is given in Table 4 below.

Table 4: MHC Typing of Twins—Skin Emanation Samples Collected at CBR

Donor Nos.	Region	MHC Type
10 & 11	Class I	A*02 (a), A*03 (c), B*27 (a), B*07 (c), Cw*01 (a), Cw*07 (c)
	Class II	DRB1*11 (a), DRB1*15 (c), DQB1*03 (a), DQB1*06 (c)
	Class III	S?31 (a), S?31 (c)
12 & 13	Class I	A*02, A*29, B*35, B*58, Cw*07, Cw*08
	Class II	DRB1*08, DRB1*11, DQB1*03, DQB1*04

	Class III	BF*F, BF*X, C4A*3, C4A*X, C4B*1, C4B*X
17 & 18	Class I	A*31 (a), A*03 (c), B*44 (a), B*07 (c), Cw*03 (a), Cw*07 (c)
	Class II	DRB1*01 (a), DRB1*15 (c), DQB1*05 (a), DQB1*06 (c)
	Class III	F?31 (a), S?31 (c)

It should be noted that three other, MHC-typed individuals (unrelated) have had skin emanations collected from them at CBR as well. One collection has occurred, so far, for each of these donors. This acquired data will be useful in the future, as a larger data set of skin emanation samples is eventually built up and correlated with MHC types.

2.2.2 Unrelated Individuals

Four unrelated individuals volunteered to provide samples of skin emanations according to the collection protocol described above. Collections occurred twice daily, at the same times each day, once in the morning and once in the afternoon. Occasionally, a donor missed a collection, and a convenient time was arranged to conduct a make-up collection. During each collection, a batch of unhandled beads was prepared, just as the batches were prepared for each donor minus the actual collection of emanations, in order to obtain a background signal for that collection. The donor group consisted of two males and two females, ages 22-27, with no known health problems.

3. RESULTS

The goal of this study is to prove that the collection and analysis protocol described above provides a robust biological signature of human skin emanations, and that this process can be incorporated into MHC based studies of skin emanations including large numbers of samples. The results of two experiments, four unrelated individuals and three sets of twins, are presented below and described, qualitatively and quantitatively. Using these results, the efficacy of the collection protocol is judged on the following criteria.

1. A robust signal is present above and beyond that present in the control samples.
2. Qualitatively this signal can be characterized as biological, consisting of compounds reported in the literature and accepted as components of skin emanations.
3. A large sample set can be collected over an extended period of time, and this data can be modeled to remove day to day variations and extract a unique signature that can be ascribed to the individual.
4. The signature of genetically related individuals can be qualitatively described and characterized over multiple collections.

Also, in the process of the study, the interesting effect of competitive adsorption on the SPME phase was observed during some of the multiple extractions conducted on samples and will be highlighted.

3.1 Results from Four Unrelated Individuals Experiment

3.1.1 GC/MS Analysis

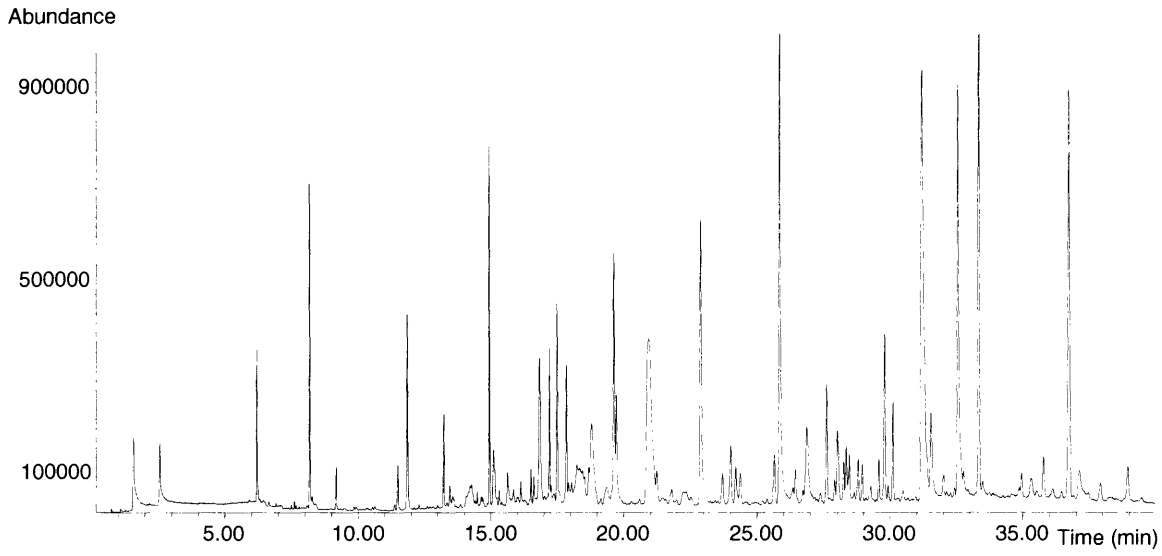


Figure 4: Representative Total Ion Chromatogram (TIC) of Skin Emanations Sample from Four Unrelated Individuals Experiment

Figure 4 above gives a representative response from the GC/MS instrument for a sample of skin emanations collected during the four individuals experiment and extracted by the procedures detailed above. As can be seen in the figure, a complex spectrum of peaks is eluting through the column.

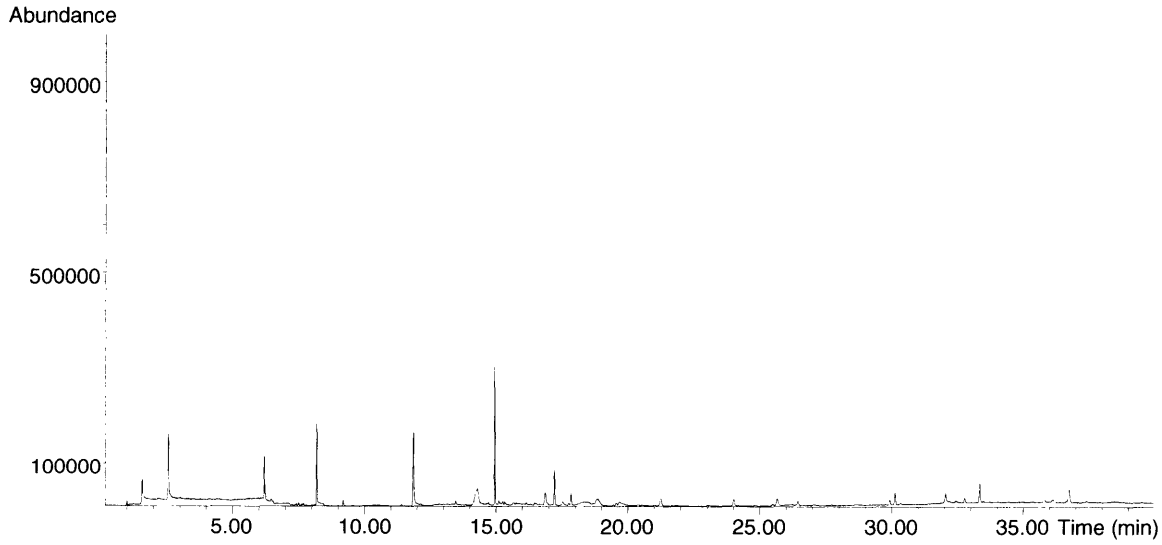


Figure 5: Representative Total Ion Chromatogram (TIC) of Control Sample

Figure 5 above gives a representative volatile signature obtained from a sample of cleaned and un-handled beads, referred to as a control sample. The large peaks eluting at retention times (RTs) up to 15.00 minutes were compared to the NIST library (Version 2, Build 1 July 2002) database of accepted mass spectra. These peaks were identified readily as siloxane compounds characteristic of the SPME fiber, the column phase, or the chromatograph's injector septum and, while undesirable, are difficult to avoid in this analytical procedure. However, the majority of the volatile signature due to skin emanations (comparing to Figure 4) appears to occur after a RT of 15.00 minutes, and so co-elution, or at worst obscuring, of sample peaks by these background siloxanes is limited to the initial period of time.

In order to make the complex spectrum of volatiles present in Figure 4 more manageable and understandable, an ion of particular mass can be extracted from the total ion chromatogram (TIC). Carboxylic acids re-arrange and fragment characteristically between the alpha- and beta- carbons, producing a fragment of mass 60 amu. This is a well studied phenomenon for carboxylic acids and other derivatives of fatty acids known as the McLafferty rearrangement. An m/z value of 60 can be extracted from the complex TIC given in Figure 4, and the resulting extracted ion chromatogram is given in Figure 6 below.

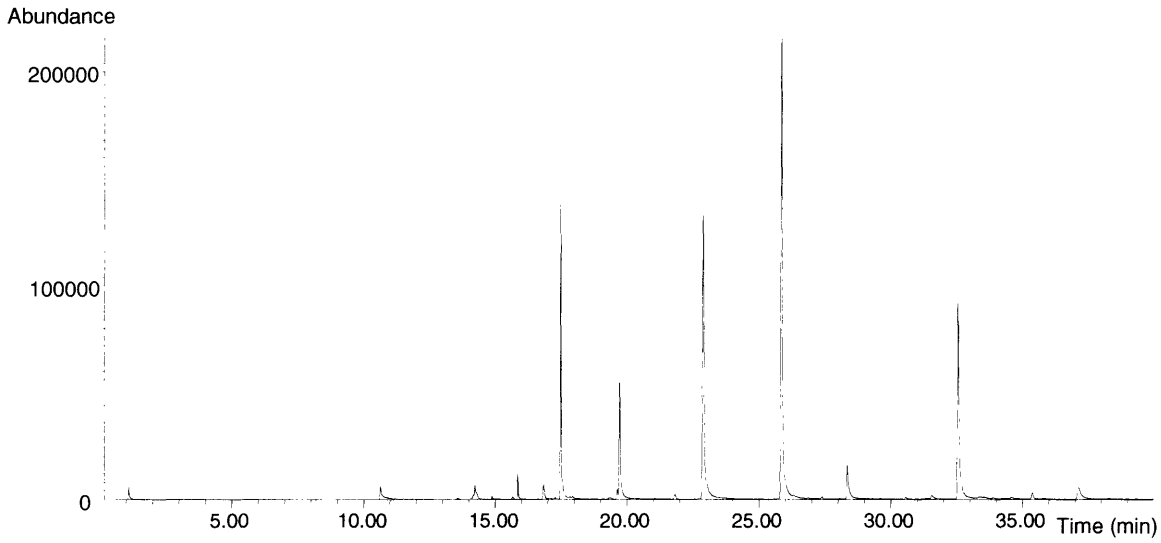


Figure 6: Extracted Ion Chromatogram for Organic Acids in Skin Emanations Sample

As can be seen in Figure 6, a mixture of carboxylic acids is present in the headspace above the beads, and the acids are eluting through the chromatographic column in a specific sequence. As expected, the retention time is increasing with molecular weight, since the molecules with the larger carbon backbone (all compounds having otherwise the same carboxylate group) partition more slowly into the column phase. This trend is summarized in Table 5 below, which gives the average values of the RTs for four of the acids which appear in all 30 samples for each donor. Assuming a normal distribution of retention times, the standard deviation across all 30 samples is given with the mean value. In almost all cases, except for heptanoic acid, the RT is only varying by 0.004-0.006 seconds. This corresponds to the length of time for one scan of the mass range (50-550 amu in this case) by the quadrupole mass ion selector. Therefore, this is really a discrete unit, and obtaining a more precise value is not feasible with the

same acquisition parameters. That is, fragment ions for the particular carboxylic acid could in fact be traveling through the quadrupole, but if this mass selector is not exactly tuned to the relevant mass (which it passes only once per scan), then this species will not reach the detector and be included in the ion count, even though it is in fact eluting during this short time period. Heptanoic acid seems to have a standard deviation in its RT closer to two scans, and this is perhaps due to its later RT and the increased opportunity for dispersive effects to operate.

Table 5: Retention Times for Four Carboxylic Acids in Samples for Each Donor

	Donor 4	Donor 5	Donor 7	Donor 9
Butanoic Acid	14.23 +/- 0.006	14.23 +/- 0.006	14.23 +/- 0.005	14.23 +/- 0.005
Pentanoic Acid	15.86 +/- 0.005	15.86 +/- 0.004	15.87 +/- 0.005	15.86 +/- 0.005
Hexanoic Acid	17.50 +/- 0.005	17.50 +/- 0.005	17.50 +/- 0.006	17.50 +/- 0.005
Heptanoic Acid	19.73 +/- 0.007	19.73 +/- 0.010	19.74 +/- 0.012	19.74 +/- 0.009

Such carboxylic acid mixtures (backbones ranging from C-5 to C-14) are characteristic of biological samples and, when compared to the background signal from the un-handled beads (Figure 5), provides confirmation that biological volatiles are, in fact, being collected on the beads and subsequently extracted by the SPME fiber. The target response (ion count abundance) for each of these carboxylic acids is summarized in Table 6 below. The mean values, across all 30 samples for each donor, are given, but the variability between each sample is quite significant. Assuming a random, normal distribution, the percent variance ranges from 31% to 105%. This variability has a number of sources, to include the inherent nature of biological samples. Organisms are dynamically interacting with their environment at all times. Equilibrium is hard to define

in these complex, multivariate, and ill-defined systems. Other sources of variation include the experimental procedures.

Table 6: Target Response for Four Carboxylic Acids in Samples for Each Donor

	Donor 4	Donor 5	Donor 7	Donor 9
Butanoic Acid	102144.07	46508.27	26560.33	44637.10
Pentanoic Acid	66422.73	42720.73	18240.30	38151.60
Hexanoic Acid	299504.50	486707.03	156532.50	422648.10
Heptanoic Acid	107472.83	219628.83	54044.00	172009.73

A slight tailing trend can also be seen for each peak in Figure 6, i.e. the back side of each peak tails off asymmetrically. This peak tailing is characteristic of carboxylic acids due to the resonance available to the carboxylate group and multiple available interactions with the column phase. Conventionally, when analyzing carboxylic acids by GC, they are first converted (derivatized) to their corresponding methyl esters. These esters are less reactive with the polar phase of the column, and this derivatization leads to narrower peaks and more reproducible abundances. However, derivatization was not practical for the skin emanations samples, which were present as a thin coating on the glass beads, and in addition attempting a chemical reaction on these heterogeneous samples could have altered other, unidentified components and led to loss of sample, or introduction of experimental artifacts. It is important to re-iterate that the results for only a small subset of carboxylic acids has been extracted and presented here in order to obtain a familiarity and understanding of the otherwise complex spectra.

3.1.2 Models Produced from Pattern Recognition

As mentioned previously the large and complex data sets require an efficient and effective analysis procedure. This current study is a fairly new application of such GC/MS data (although, MS in general is becoming increasingly popular in proteomics studies), and it was desirable to explore both standard proven analytical techniques, such as PCA, and also more experimental techniques, such as the Correlogic genetic algorithm.

Correlogic applied the genetic algorithms to the data set obtained from the experiment on the four unrelated individuals described in preceding sections. In all, thirty data files acquired from samples for each donor were prepared and submitted. All files were aligned to a reference file as described above. Data files consisted of a three dimensional matrix with retention time (RT), abundance (or ion count), and mass over charge ratio (m/z) coordinates. Each data file, once converted to binary format and aligned to the reference, was 14.6 MB, and a total of 1.75 GB of data needed to be processed by the Correlogic system. This large amount of data was difficult for Correlogic to process, and the comparison of 4 individuals simultaneously was unprecedented. Therefore, Correlogic installed a new computing cluster and added the functionality of simultaneous comparison of 4 states. Still models could only be built and tested on limited subsets of data, truncating the RT range from the full 40 minutes.

Table 7: Correlogic Model #3121—98.3% Overall Accuracy

Count of Predicted	Actual				
Predicted	0	1	2	3 (blank)	Grand Total
0	14				14
1	1	15			16
2			14		14
3				16	16
(blank)					
Grand Total	15	15	14	16	60

Table 7 above gives one of the most accurate models returned by Correlogic so far. The basic method of model creation with genetic algorithms, in general, was covered in a previous section. Specifically, though, in the Correlogic models, data files are randomly segregated into training and testing sets. The training files are used to build the model, while the testing files are used to validate the model once it has been built. In Table 7, the validation results for the training files for each donor are given. In only one case did the model miss-assign a file to donor 1 (donor B) which actually belonged to donor 0 (donor A). Out of all 60 files, used for validation, this one wrong assignment correlates to the high overall accuracy of 98.3 percent. The biomarkers, consisting of a RT and m/z coordinate pair, were searched within the NIST library of target spectra in order to identify the chemical compounds underlying these factors and will be discussed in the following section.

Principal component analysis was conducted on this data set to verify the classification produced by the genetic algorithm, and the resultant scores and loadings plot is given in Figure 7 below. The figure gives the plot of the data as classified by the two largest principal components. Each component is a linear combination of features (chromatographic peaks) in the original data set that account for the greatest amount of

variation between donors. Each marker (either circle, triangle, cross, or square) represents a data file that was classified according to a model. By visual inspection, one can obtain an idea of how well classified the data is. Similar data files are clustered amongst themselves, while still reasonably segregated from other sets of data files. The crosses correspond to donor A, the squares to donor B, the triangles to donor C, and the circles to donor D. Upon further review of these PCA results, an artifact from the signal processing seemed to be present. The clustering of data for donor B was unusually tight, compared to the other donors. It was necessary to make adjustments to the alignment algorithm in this case. This experience underscores the need to carefully handle the data when using the genetic algorithm for analysis and the need to confirm exploratory data analysis with more standardized and proven techniques. It should be noted, though, even with the artifact introduced during the alignment, a high degree of separation is still noticeable in the PCA model. It is merely the clustering that was affected by the alignment algorithm.

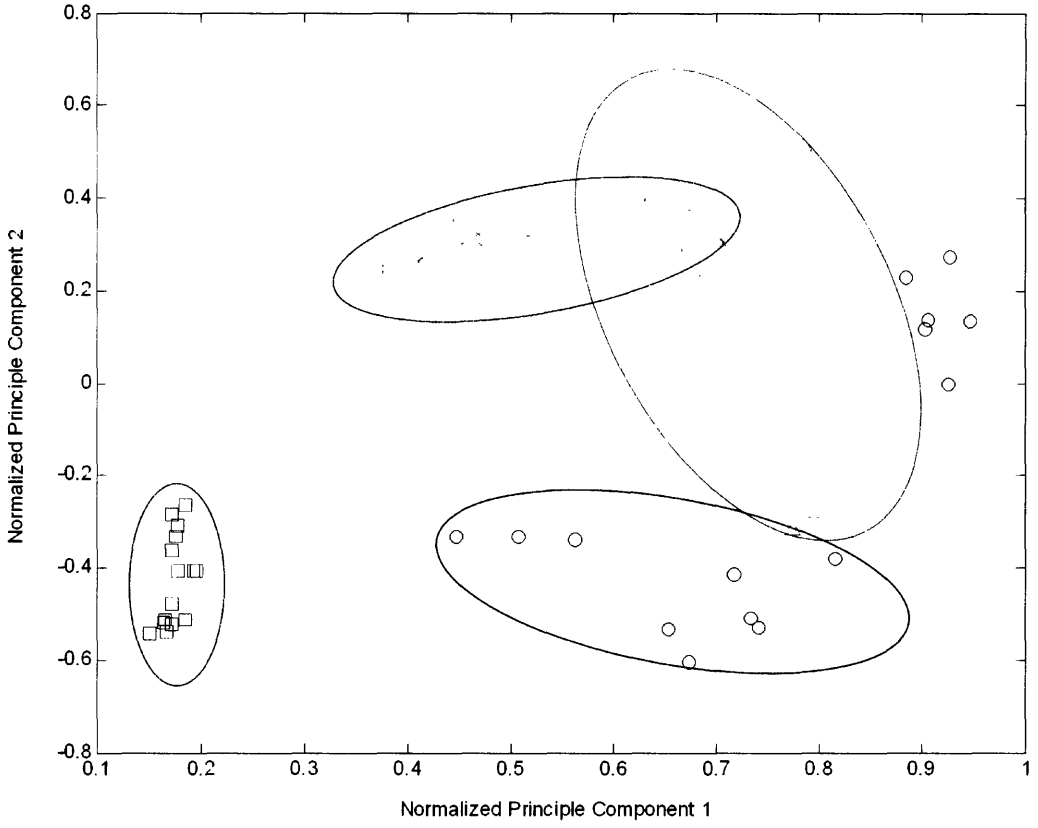


Figure 7: Scores and Loadings Plot from Original PCA

The alignment algorithm created in MATLAB© was modified to account for reference files from 4 individuals. All the files were then re-aligned, and the PCA was repeated for this experiment. The scores and loadings plot is given in Figure 8 below. As previously, the cross markers represent files classified according to the model for donor A. The squares correspond to donor B—the triangles to donor C, and the circles to donor D. The ellipses drawn around the markers are visual aides only and do not represent any formal boundaries.

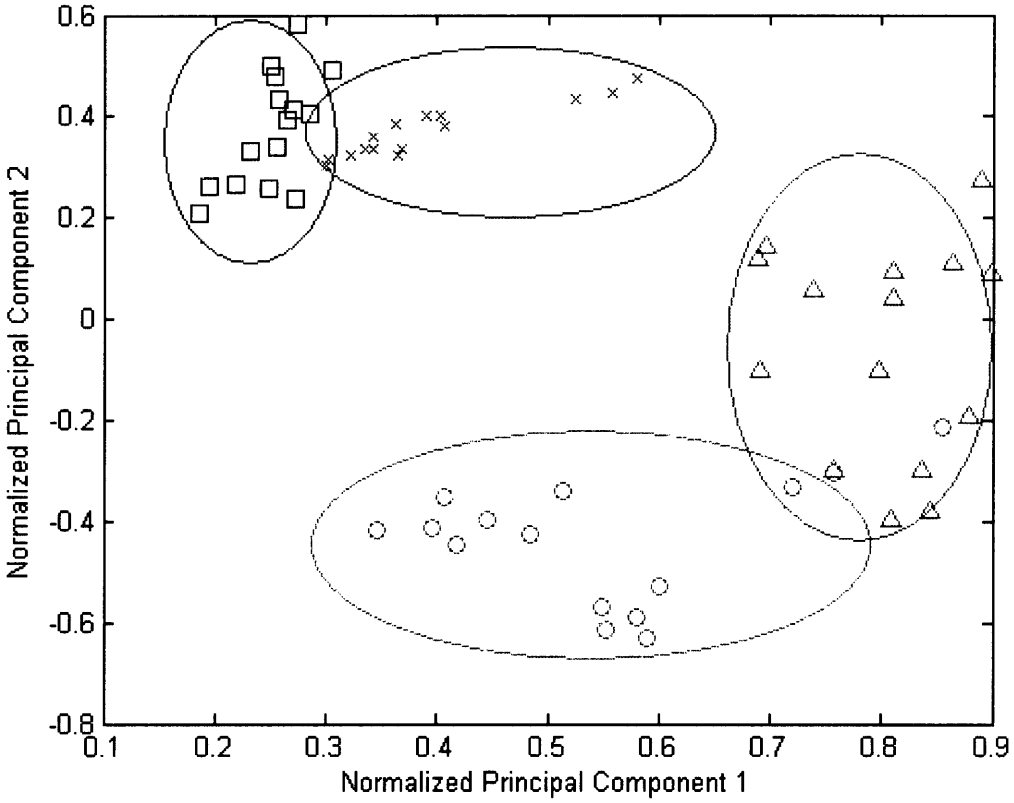


Figure 8: Plot of Two Largest Principal Components from Revised PCA Analysis

To avoid the aforementioned truncation of the RT, an alternative approach was explored. The m/z component of the data was removed for the model building. Models were generated then on the ion abundance and RT data only (basically the TIC out of each data file). Once specific biomarkers had been identified, occurring at a specific RT, the corresponding m/z spectrum could then be investigated for chemical identification. This alternative led to a more straightforward method for chemical identification, as well as, the capability to include the entire RT abscissa in the construction, testing, and validation of the models.

3.1.3 Lists of Compounds Identified

In the Correlogic model that achieved 98.3% classification accuracy, nine biomarkers were identified. The RT and m/z coordinate pairs for each biomarker were extracted from each data file and compared to the NIST library database. The following compounds, in Table 8 below, were identified at each of the nine biomarkers across donor files.

Table 8: Compounds Identified at 9 Biomarkers in Correlogic Model #3121

Compound Name	RT (min)
1,2,3,4-Tetrahydroisoquinolin, 2-acetyl-6,7-dimethoxy-1-phenmethylene-	4.864
Cyclopentasiloxane, decamethyl-	6.593
Cyclopentasiloxane, decamethyl-	6.882
Rhodopin	7.591
(5á)Pregnane-3,20á-diol, 14à,18à-[4-methyl-3-oxo-(1-oxa-4-azabutane-1,4-diyl)]-, diacetate	8.033
5H-Cyclopropa(3,4)benz(1,2-e)azulen-5-one derivative	8.390
Hexadecanoic acid derivative	16.281
Hexadecane, 1,1-bis(dodecyloxy)-	19.802
4a-Phorbol 12,13-didecanoate	28.084

The two siloxanes eluting at 6.593 and 6.882 minutes are components of the column phase or SPME fiber matrix. It is unlikely two biomarkers are originating from the column phase, the SPME fiber, or the vial septa which would have random variations across samples. This is more than likely a mis-identification within the NIST library for these biomarkers, or the background is obscuring the real compound for the library identification. The other 7 compounds in the list have biologically related functions. However, the matches of these compounds with the NIST library targets were not conclusive. Therefore, the potential exists that these are biomarkers of unknown chemical identity.

3.2 Results from Twins Experiment

A series of multiple extractions was conducted on each sample from each twin. This involved extracting the sample three times in immediate succession with a new SPME fiber. The sample vial was maintained at the extraction temperature (65 °C) as one fiber was removed and a freshly conditioned fiber was inserted for the next extraction. Three extractions were conducted on each sample, and representative total ion chromatograms (TICs) for both a control sample of un-handled beads (Figure 9) and a skin emanations sample (Figure 10) are given below.

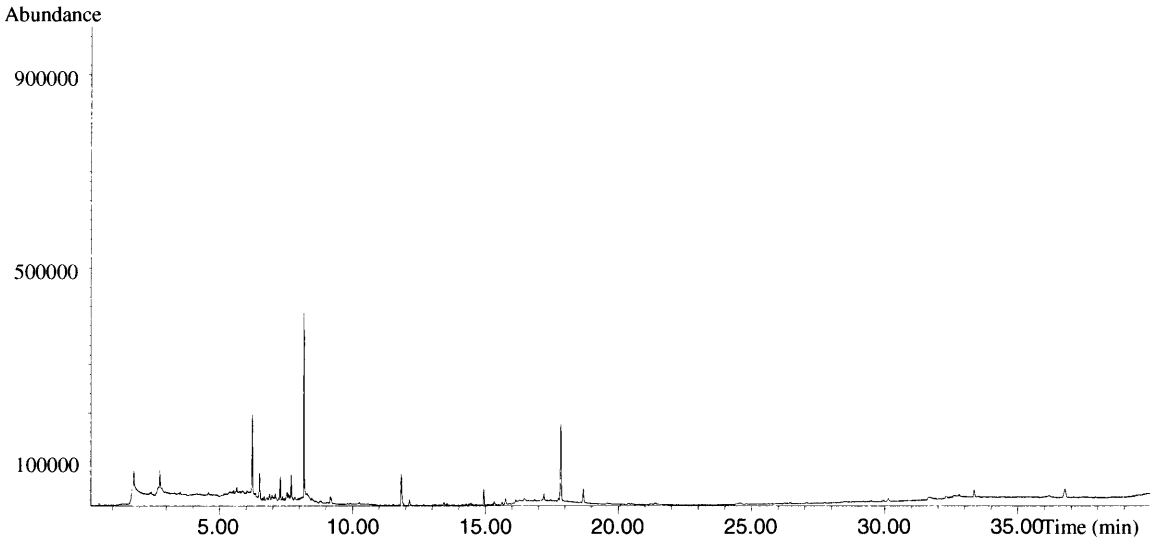


Figure 9: Representative Total Ion Chromatogram (TIC) from Control Sample- Twins Experiment

As can be seen in Figure 9 above, the control beads possess a low level signal. The initial peaks at a RT less than 15.00 minutes were readily identified in a NIST library (Version 2, 1 July 2002) search as siloxanes derived from the SPME fiber, the chromatographic column, and the septum in the sealed 10 mL headspace vial. The larger peak at 17.8

minutes is an unidentified contaminant that was present in most samples from CBR. This is either a component of the CBR clinical environment, or a contaminant deposited on the beads inadvertently while the beads were being prepared for collection. Due to the ubiquitous nature of residue from the hands some level of contamination is unavoidable in preparing samples, even though gloves were worn by the administrator.

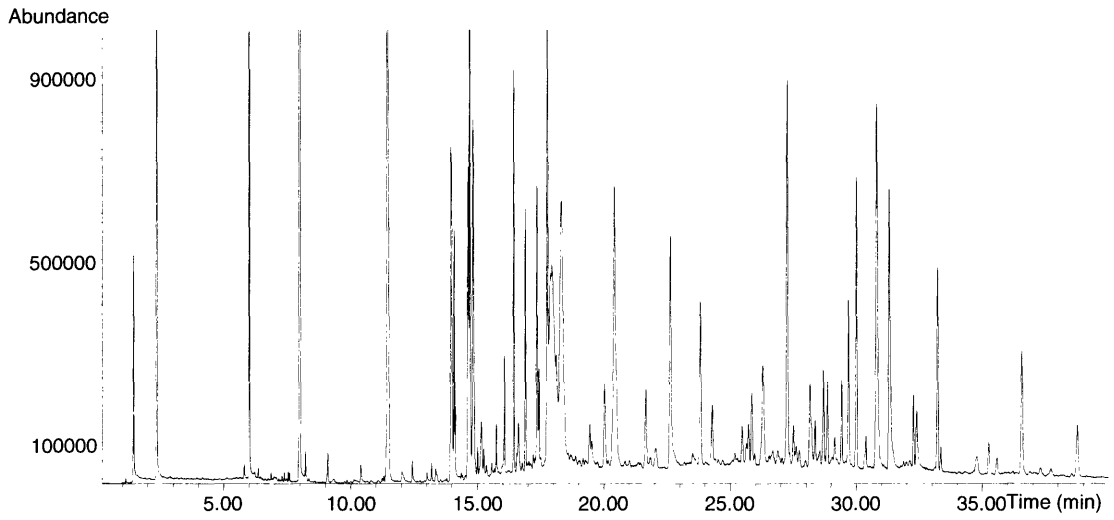


Figure 10: Representative Total Ion Chromatogram (TIC) from Skin Emanations Sample-Twins Experiment

In Figure 10 above, a robust set of peaks is present above that indicated by the control sample. Many of these peaks are co-eluting and overlapping which makes identification via a NIST library search difficult.

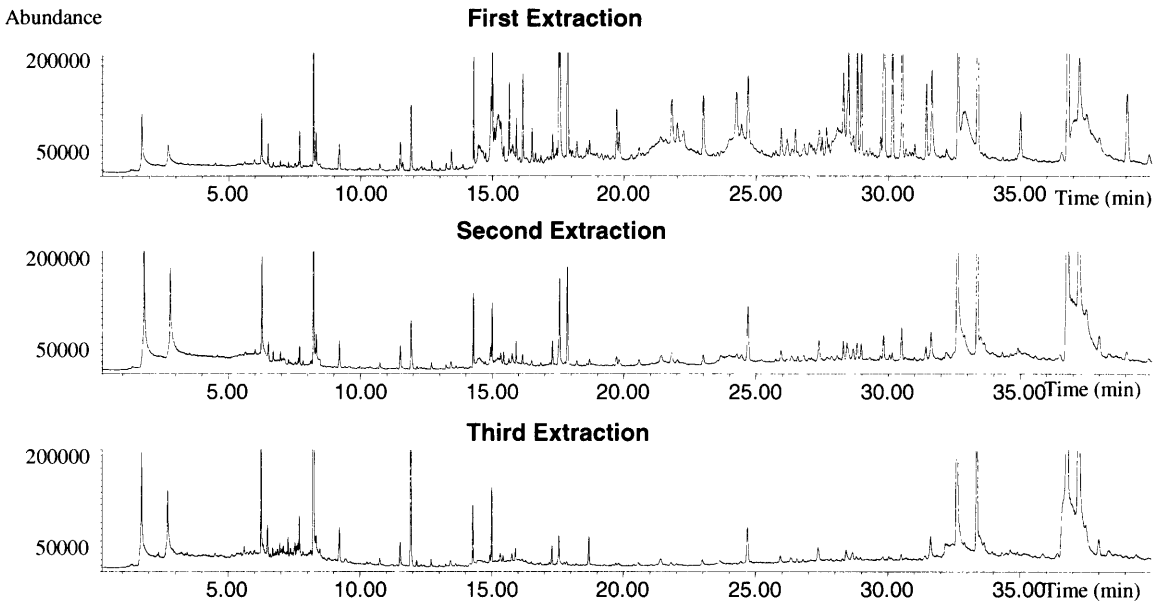


Figure 11: Overlaid TICs for Multiple Extractions- Twins Experiment

As can be seen in Figure 11 above, signal complexity and abundance exhibits a general decrease as more extractions are conducted. Conducting qualitative searches with the NIST library confirms this trend. Some components present in the first extraction are too low abundance to be selected by the integrator and compared to the library spectra database by the time the third extraction is conducted. Upon closer inspection of the chromatograms, the reverse trend was also noticed. By focusing on the extractions for one donor and manually searching the TIC across the RT, two compounds were observed to be missing in the first extraction and showed a stabilizing trend in subsequent extractions. The first compound was eluting at a RT of 15.758 min, and a zoomed in region of the overlaid TIC at this RT is given in Figure 12 below.

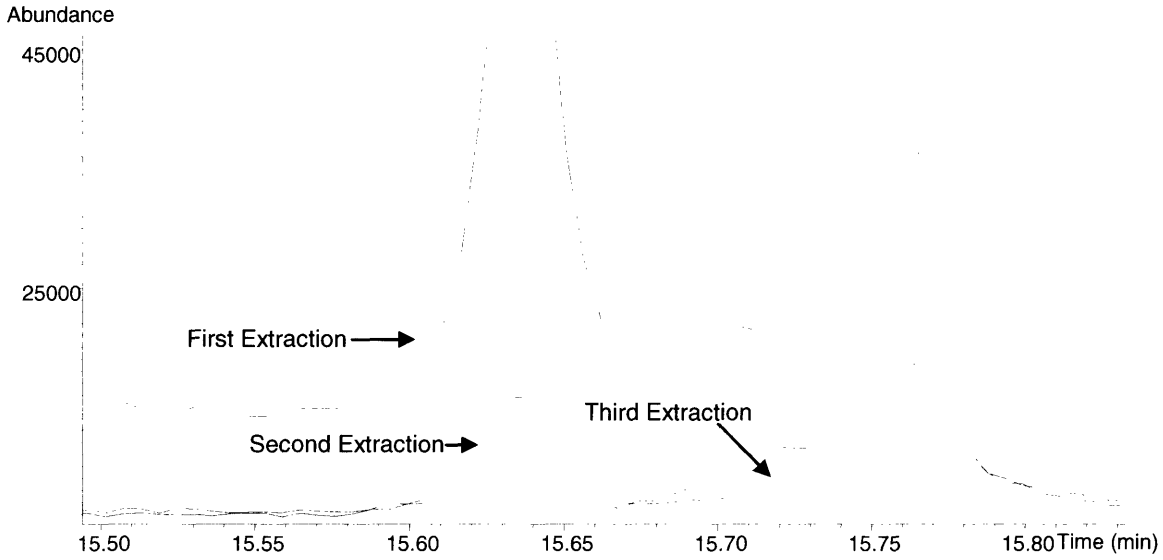


Figure 12: Overlaid TICs of Multiple Extractions for Donor 18

In the first extraction, the broad peak eluting at 15.780 seems to have been concealing another peak to its left at 15.758 which becomes apparent only in the second and third extractions, after the larger peak has been completely extracted. Also, the large peak at 15.636 minutes shows a dramatic decrease between the first and second extraction, but is relatively stable for the second and third extraction. The fact that some peaks are completely extracted while others show a stable response after multiple extractions indicates that equilibrium is being reached between the SPME fiber and the volatiles in the headspace of the vial over the 30 minute extraction time. This is expected based on the nature of the SPME phase. The polymer surface will interact with volatiles in a preferential manner. Similar compounds, with similar polarity and solubility, will adsorb more readily to the surface of the DVB particles and diffuse more easily into the PDMS matrix. Over time as more collisions at the surface occur, the compounds with higher affinity will start to replace those with less affinity. This will occur most readily

for the adsorbed molecules on the surface of the DVB than for the molecules diffusing into the liquid polymer. However, it highlights the issue of competitive adsorption for the SPME-headspace interface, and suggests that a more representative spectrum of volatiles could be captured under transient, non-equilibrium conditions. Such an approach has been recently suggested which proposes that a diffusion based calibration of such extractions while carefully controlling the air flow past the fiber can lead to more quantitatively reproducible results[46]. This idea, considering the short time involved in acquiring the volatile signal, is particularly interesting for designing a device for field use, and such implications will be discussed in a later section.

Figure 13 below compares the two sets of twins for one collection. By following the trace profile of the TICs one can find, in a qualitative manner, peaks that are unique to one of the twin sets. The top two TICs correspond to donors 17 and 18 (one pair of twins) while the bottom two TICs correspond to donors 12 and 13 (another pair of twins). The red lines indicate peaks unique to the latter pair of twins. This was verified by comparing the mass spectra across all four samples. The two peaks identified on the left were not conclusively identified in the NIST spectral library, but the peak on the right, eluting at 37.266 minutes was identified with high confidence as tetradecanoic acid. This is encouraging because it is one of the aforementioned carboxylic acids known to be biologically relevant and able to associate with MHC peptides. The target spectrum compared with the NIST library spectrum is given in Figure 18 in the appendix.

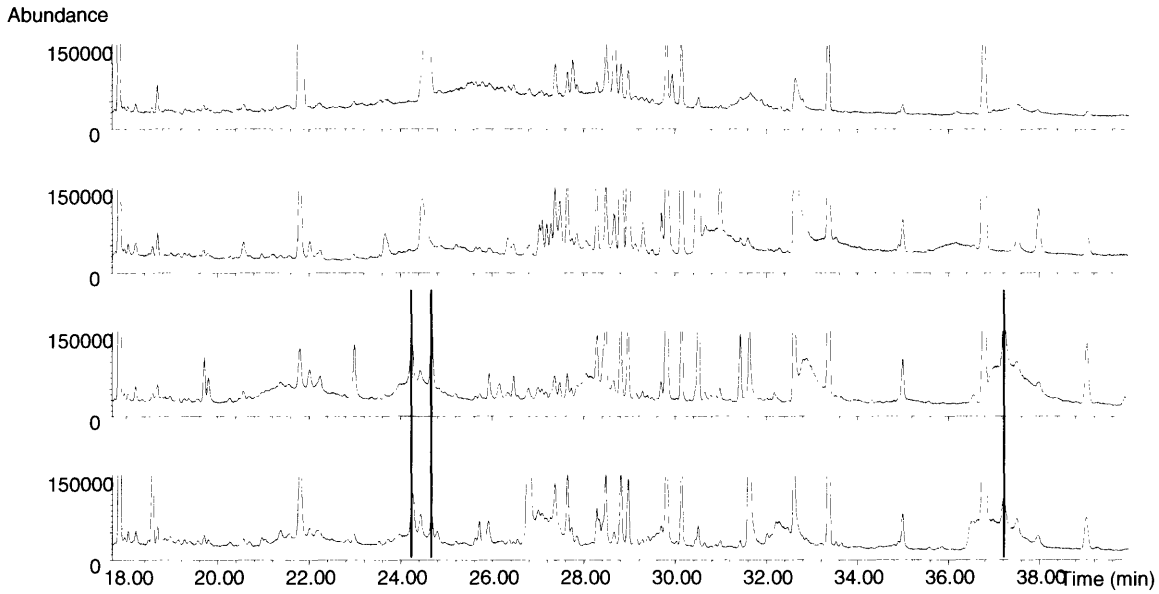


Figure 13: Two Sets of Twins on Same Day

3.1.1 Lists of Compounds Identified

Often positive chemical identifications are a subjective judgment of the analyst. One must look at a variety of factors including RT and signal abundance, in addition to the library search results. The more experience one has with GC/MS and with the particular samples of interest, the more efficient these judgments become. However, compounds were identified with the following criteria:

1. Quality of match with the NIST library target compound was greater than 70
2. Identified in all three extractions at RTs within 0.03 minutes of each other
3. Not identified in any control sample (unhandled beads) at a comparable RT

The quality of match factor is a number generated within the Chemstation® software that compares an unknown spectrum to a library target by taking into account ratios of ion

counts between m/z values, the absence of expected m/z values, and the presence of potentially extraneous m/z values. It is a useful gauge in determining the overall match of an unknown compound to a target in the library, and its range is from 0 to 99, the upper limit indicating an identical match[47, 48]. The value of 70 was selected for this analysis as a balance between opposing concerns. Lower values lead to a larger number of poor matches and became increasingly difficult and unwieldy for analysis. A higher value, though, could potentially rule out critical components. It must be noted that all of these library searches are limited by the completeness of the library used. The NIST library (Version 2, Build 1 July 2002) is a standard, extensive, and high-quality library, but it still contains only 147,198 target spectra. The potential for missing important components of the signal increases by requiring more accurate matches with library targets.

Table 9: Shared Compounds Identified in Skin Emanations

Shared Compounds in Twin Pair 1
1,2-Benzenedicarboxylic acid, bis(2-methylpropyl) ester
Decanal
Dodecanoic acid
Nonanal
Shared Compounds in Twin Pair 2
5,9-Undecadien-2-one, 6,10-dimethyl-, (E)-
Dodecanoic acid
Isopropyl Myristate
Octanal, 2-(phenylmethylene)-

4. DISCUSSION

4.1 Importance of Skin Emanations as a Body Effluent and Metric

Currently, the sweat test is the most reliable method for diagnosing cystic fibrosis. It is also popular because it is non-invasive and quick. Skin emanations in general have potential for many applications in many fields to include homeland security and medicine. In the past, odors emitted by patient have been a surprisingly good tool for disease diagnosis. Furthermore, odor is a non-line-of-sight, latent tool for identification as in the example of bloodhounds.

4.1.1 Confounding Factors

4.1.1.1 Diet, Health, Age, Gender, and Environmental Factors

Human samples (or any biological sample) are inherently complex. Living organisms are constantly adjusting and dynamically interacting with their environment. Analyzing the specific influence of such factors as diet, health, gender, or environment on a skin emanations signature was outside the scope of this study. Rather, in order to reduce the effect of such variables, a large sample set was collected over one month. Donors were not asked to adhere to a certain diet, or make a specific change in their diet. The health status of the donors was not monitored, nor controlled. The four donors were selected from a limited age range (22-27) and consisted of two males and two females. Samples were collected in the mornings and afternoons in order to capture any of the circadian variability in one individual. Environmental factors include living conditions and non-biological contaminants picked up from the surroundings. In order to limit this factor,

skin emanations were collected from a single, confined anatomical area, and donors washed their hands according to the established protocol and rinsed thoroughly prior to every collection. Paper towels were not used, but hands were allowed to hang dry in the air. All collections were conducted in the same laboratory setting with controlled access. The critical factor in the experimental design, however, is the sample number and the time span of collection. It is assumed that over a period of a month, a common denominator will remain amongst all the day to day variability of the signature, and this common denominator can be detected via a multivariate analysis technique. The signature obtained thus can be ascribed to the individual, but not necessarily to any specific "odortype" gene or locus. It is important to note also, applying the model outside the sample group on which it was built would make no sense. This model is only able to classify one individual compared to the other three in the model. The protocol and method can easily be used on larger sample sets and more genetically specific experiments in order to discover more specific underlying factors. Such efforts to extend the generalizability of the model will be discussed in the following section.

4.1.1.2 Experimental Artifacts and Chance

When using experimental data analysis techniques, like genetic algorithms, the following concerns are critical: bias, chance, over-fitting, and generalizability. These issues have been discussed recently in reviewing advances of clinical diagnosis of cancers[38, 39], but they are also applicable to this current study.

Bias, or experimental artifacts, is an important problem in using such data analysis. Bias occurs when an artifact is unintentionally introduced into the data, either by experimental protocol, or data processing. In the event something like this occurs, data could potentially be segregated, not based on the phenomenon of interest, but rather on the spurious differences introduced by changing conditions. Once protocols were finalized in this study no further changes were made during sample collection, or data acquisition. Care was taken to carefully monitor instrument performance and consistency. For example, through routine and inevitable maintenance the chromatography column would need to be trimmed. This had to be done, or otherwise the inlet liner would generate active sites and lead to other potential biases. As the chromatographic column was trimmed (5-10 cm per month), a slight shortening of RT would be expected. Overtime, this shift could become significant, and this is one of the reasons the alignment algorithm was implemented. It must also be noted that samples were run in the order they were collected, therefore a typical sample run would consist of samples from all four donors so that no one donor's data was acquired all at once and at a different time than the other donor's. Also, SPME fiber use had to be carefully monitored. SPME fibers were routinely inspected for oxidation, and after a certain number of uses had to be replaced. This trend also followed sequentially through the data acquisition, and care was taken to ensure that not one SPME fiber was being used exclusively on one donor's set of samples, thereby introducing potential bias. A good example highlighted in the study of potential bias was the effect on clustering that the alignment algorithm had. This was evident once the genetic algorithm results were

compared to more traditional PCA, and underscores the need for comparison to proven methods.

Chance can play a role in model creation in two ways—the model can fit a data set by chance when no underlying markers exist (type I error), or the model can not fit a data set by chance when underlying markers do in fact exist (type II error)[39]. Closely related to false-negative and false-positive conclusions based on chance is the problem of over-fitting. Over fitting occurs when an excess of underlying factors is presumed by the model, and it is able to distinguish between two states based on the combinations of these excessive factors. The more factors, or "biomarkers", there are to build a model, the more likely a chance collection of those factors can classify the two states perfectly, and the factors have nothing to do with the phenomena being analyzed. Therefore, one way to check for the problem of over fitting in a model is to independently validate the model, or have the model classify independent data that were not used in training, or building, of the model. This is similar to using a control group in conventional experimental design[38]. It is also important, in experimental design to use large sample sets and specify few underlying factors in building the model. In this study such an approach was followed. The validation results were presented earlier from the Correlogic genetic algorithm. Validation was performed on sets of data that the model had not used in building of the model, and it was able to successful classify these in several models. Furthermore, since approximately 15 data files were used in the validation, the number of factors was limited to no more than 9 (in the case of the model presented in the text). The

fewer number of factors specified, the less the model is fitting by chance. Rather, the more likely the model has hit upon the underlying factor related to the phenomena of interest, in this case unique odor signature.

Generalizability is the most difficult problem faced by new pattern recognition techniques. Essentially, a model built on data acquired in one study—using one experimental protocol, with one set of subjects, and using one analytical instrument for data acquisition—is not readily extendable when any one of these variables is changed. This is a problem that everyone in this field must confront and is currently struggling with. The solution for now is to continue expanding and collecting immense data sets and boil down the models to the most fundamental, invariant common denominators.

For example, in this study, samples of skin emanations were collected over a period of 30 days. Models were built on this data. If new samples were to be collected after a period of 6 months using the same collection, extraction, and analysis procedures, they could then be validated with the model. If the model was able to maintain the same overall accuracy as the original validation, but with the new set of data collected 6 months in the future, then the generalizability of the model could be proven in part. That is, after a period of extended time, the model was still able to accurately classify skin emanations between four different individuals. This would be the first step of generalizability. However, if the model's accuracy with the new data set was significantly reduced, then the original model was built upon some transient, underlying factors that are no longer effective 6 months later. It is not general enough. In this case, a new model can be

generated from scratch, including the new data. The new validation accuracy could be compared to previous models, to see if any improvements were made, and presumably the model has been created based only on common factors between the original data set and the new data set, effectively excluding the transient factors. A new data set could then be generated at a future time and the process repeated.

4.2 Advantages and Disadvantages of Collection Method

The collection protocol presented here has important advantageous over other collection protocols. First of all, skin emanations are a non-standardized biological fluid, and developing the proper protocol can be time consuming. However, once it is established, such collections are much more convenient, due to their non-invasive nature. Collections from the palms and backs of hands are able to capture both the sebaceous secretions as well as the eccrine secretions, both polar and non-polar components of skin secretions. These secretions are known to produce odors and known to contain metabolic wastes (e.g. excess salts and lipid metabolism). Apocrine secretions are more difficult to collect and are co-located with sebaceous and eccrine glands. Therefore, in order to isolate the apocrine component, one must isolate a single gland and stimulate a secretion. Alternatively, incorporating a collection like the one described in this study, as part of a larger study of skin emanations, is one way of comparing sebaceous and eccrine secretions to combined (with apocrine) secretions. Certain disadvantages must also be addressed. The SPME fiber extraction of the volatile signature will preferential exclude certain compounds, based on the nature of the SPME phase. Therefore, while sample

preparation is greatly simplified, only a subset of volatile components can be extracted at one time. This is partially alleviated by the variety of phases commercially available, and experiment design can include using several phases on one sample. At the same time, SPME may be ideally suited for field applications as discussed in a following section. SPME with automated extraction hardware can also be used for high-throughput analysis with minimal supervision. This is particularly useful when sample sets are large.

4.3 Comparison of Compounds to Those Identified in Literature

Some of the compounds identified in the literature as components of skin secretions are given in the tables below. Similar mixtures of carboxylic acids and other biological compounds have been found in qualitative searches of the data obtained in this study. This indicates that a complex biological signature, specifically skin emanations, is being collected and characterized via this collection and analysis protocol. Such mixtures of compounds are known to produce odors, and can be detected by conventional analytical instrumentation, such as GC/MS.

Table 10: Compounds Identified in Literature

Carboxylic Acids[26]	**Male Axillae[49]**	**Apocrine Secretions[50]**
acetic acid	2-methylhexanoic acid	Hexanoic acid
2-propenoic acid	3-methylhexanoic acid	(E)-3-methyl-2-pentenoic acid
propanoic acid	dimethylsulfone ($C_2H_6SO_2$)	3-methylhexanoic acid
2-butenoic acid	γ-C_8-lactone	(Z)-3-methyl-2-hexenoic acid
2-methyl-2-butenoic	4-ethylpentanoic	Heptanoic acid

acid	acid	
3-methyl-2-pentenoic acid	(Z)-3-methyl-2-hexenoic acid	(E)-3-methyl-2-hexenoic acid
3-methylpentanoic acid	2-ethylhexanoic acid	Octanoic acid
hexanoic acid	*n*-heptanoic acid	7-octenoic acid
heptanoic acid	2-methylheptanoic acid	Nonanoic acid
octanoic acid	(E)-3-methyl-2-hexenoic acid	C_9-unsaturated acid
nonanoic acid	phenol	Decanoic acid
decanoic acid	γ-C_9-lactone	Undecanoic acid
undecanoic acid	*n*-octanoic acid	
dodecanoic acid	2-methyloctanoic acid	
methyldodecanoic acid	4-ethylheptanoic acid	
tridecanoic acid	7-octenoic acid	
tetradecenoic acid	γ-C_{10}-lactone	
methyltridecanoic acid	*n*-tetradecanol	
tetradecanoic acid	*n*-nananoic acid	
pentadecenoic acid	2-methylnonanoic acid	
methyltetradecanoic acid	4-ethyloctanoic acid ("goat acid")	
methyltetradecanoic acid	*n*-hexanoic acid	
methyltetradecanoic acid	*n*-decanoic acid	
pentadecanoic acid	2-methyldecanoic acid	
9-hexadecanoic acid	4-ethylnonanoic acid	
methylpentadecanoic acid	9-decenoic acid	
hexadecanoic acid	*n*-hexadecanol	
heptadecenoic acid	*n*-undecanoic acid	
methylhexadecanoic acid	4-ethyldecanoic acid	
heptadecanoic acid		

11-phenoxyundecanoic acid
9,12-octadecadienoic acid
9-octadecenoic acid
methylheptadecanoic acid
octadecanoic acid
docosanoic acid
lactic acid
hexanedioic acid
heptanedioic acid
benzoic acid
4-hydroxybenzoic acid
4-hydroxy-3-methoxybenzoic acid

5. CONCLUSIONS

Recent studies correlating mating behavior with MHC genetics suggest that odor may be derived from an individual's DNA. In this study, the volatile signatures of skin emanations, as part of a larger study including volatiles derived from both blood and urine, were collected and characterized. A large variety of non-standardized collection procedures for this biological fluid have been described in the literature, and it was necessary to develop a robust, reproducible, and convenient collection protocol tailored to the needs of this study. Several different collection protocols were attempted, but the most useful protocol ended up being the rubbing of glass beads on the backs and palms of hands as described in previous studies identifying mosquito attractants to human hosts[26]. The next issue to address was sample preparation for the GC/MS instrument.

Direct thermal desorption with cryofocusing ended up causing damage to the chromatographic column and incurred additional difficulties with cellular contaminants and instrument maintenance, while static headspace extraction did not provide a significant instrument response distinguishable from noise. Solid phase microextraction (SPME) of the volatile signature of the beads proved to be a better alternative, especially for large sample sets. In addition, standardizing this extraction across all biological samples (urine and blood also) allows for closer comparison of results as data sets are expanded in the future.

An experiment was designed and conducted involving 4 unrelated individuals who each donated approximately 30 samples of skin emanations, over a period of 30 days. The Correlogic genetic algorithm was applied to this data and several models were produced. One model in particular achieved an overall classification accuracy of 98.3%. This indicates that a stable underlying signature is present amongst the otherwise variable, complex emanations produced over time. These results were confirmed with PCA analysis. The PCA analysis, however, brought to light a problem in the alignment algorithm that was effecting the clustering of the samples. The algorithm was modified to include reference files for all four donors, and modeling was conducted again. Results from these models show a reduction in the accuracy of classification, however further adjustments need to be made to the alignment algorithm. In particular, the threshold needs to be set higher so as to avoid an overabundance of landmarks. This abundance of landmarks, usually correlated with a background signal of siloxanes, or another noisy component, causes erratic fitting of the piecewise cubic spline approximation function. It

is encouraging to note that the second round of PCA analysis indicated a more normal and expected clustering pattern, and the artifact had been removed, i.e. clustering was independent of donor in both aligned and unaligned analyses.

Samples from three sets of twins and from three other unrelated individuals have been collected and MHC typed. Results from these samples consist of lists of identified compounds through NIST library searching and suggest potential links with MHC volatiles identified in other studies. In addition, these analyses provided interesting insights into the potential of SPME extraction. Specifically, the issue of competitive adsorption seemed significant. From the lists of compounds, no conclusive connections can necessarily be made to MHC type, but certain volatiles seemed unique to each twin set. In addition, some of the identified components, such as carboxylic acids, have been correlated in other studies to MHC types in mice. Carboxylic acids are ubiquitous products of lipid metabolism in many species and are known to associate with amino acids which in turn can associate with the peptide binding groove in MHC products. These represent the only set of identified compounds, in current literature, reliable traced to MHC genetics in both urine and blood. The collection, extraction, and analysis protocol presented here was shown to provide a robust signal from human skin emanations, including the set of expected carboxylic acids, and was able to uniquely classify four unrelated individuals using both PCA and genetic algorithm modeling. Therefore, this protocol represents a viable method to be extended for more broad-ranging, large sample-size MHC studies.

6. FUTURE DIRECTIONS

6.1 SPME-FAIMS Experiments and Considerations for Potential Device Design

Some headway has been made recently on miniaturizing GC/MS for field use. The FAIMS detector, while unproven in this application, is also a good candidate for miniaturization. The FAIMS detector can detect a wide range of compounds, but sometimes some initial separation is necessary. A SPME fiber may be an ideal separation for this detector. The SPME phase is similar to that found in a chromatographic column (PDMS). If the SPME were desorbed for analysis with a temperature ramp, analytes would selectively leave the SPME phase at specific temperature. All the studies so far have been conducted at high temperature, isothermal desorption, and this idea has not been fully explored. A ramped desorption provides a degree of separation that could be similar, if not as extensive, as a chromatographic column—just the amount of separation the FAIMS detector needs. In addition, using the SPME device in a non-equilibrium, transient extraction as mentioned above allows for quick capture of volatiles and, especially with the DVB mixed phases, avoids competitive adsorption which decreases the robustness of the signature extracted. A non-equilibrium, transient extraction is ideal for field applications because the extraction time is on the order of seconds, not minutes.

6.2 Potential Use in Medical Applications

The set of proteins found in serum, or the "serum proteome", has been the subject of intense research for disease diagnosis. For example, the studies mentioned earlier,

employing the Correlogic algorithm, focused on serum proteins and related metabolites as markers for prostate and ovarian cancer[21, 22]. The identity and relative quantities of "biomarkers" for a disease state has been the desired goal. Often, however, large high-concentration serum proteins that enable the circulatory function of this fluid are problematic for the analyst. They tend to obscure the low-concentration, critical "biomarkers" and require special sample preparation steps, like affinity chromatography. Skin emanations, on the other hand, are a filtrate of the blood to begin with, similar to urine, and lack the high concentrations of serum specific proteins, like albumin, trypsin, and immunoglobulin. Substances secreted on the skin are often metabolic wastes and represent a unique mixture with great potential for disease diagnosis. Furthermore, collection of this fluid, while currently not standardized, is non-invasive and convenient, as its successful implementation in diagnosing cystic fibrosis proves.

REFERENCES:

1. Thomas, L., *The lives of a cell.* 1974, New York: Viking. 16-19.
2. Janeway, C., Jr., *Immunobiology: The Immune System in Health and Disease.* 6 ed. 2005, New York: Garland Science.
3. Klein, J., *Natural History of the Major Histocompatibility Complex.* 1986, New York: John Wiley & Sons, Inc.
4. Trowsdale, J., *"Both man & bird & beast": comparative organization of MHC genes.* Immunogenetics, 1995. **41**: p. 1-17.
5. Janeway, C., Jr., et al., *Immunobiology: The Immune System in Health and Disease.* 6 ed. 2005, New York: Garland Science.
6. Yunis, E.J., et al., *Inheritable variable sizes of DNA stretches in the human MHC: conserved extended haplotypes and their fragments or blocks.* Tissue Antigens, 2003. **62**: p. 1-20.
7. Brown, R.E., *What is the role of the immune system in determining individual distinct body odours?* Int. J. Immunopharmacol., 1995. **17**(8): p. 655-661.
8. Gorman, M.L., *A mechanism for individual recognition by odour in herpestes auropunctatus (carnivora: viverridae).* Animal Behavior, 1976. **24**: p. 141-145.
9. Yamazaki, K., et al., *Odortypes: their origin and composition.* Proc. Natl. Acad. Sci. USA, 1999. **96**: p. 1522-1525.
10. Singer, A.G., G.K. Beauchamp, and K. Yamazaki, *Volatile signals of the major histocompatibility complex in male mouse urine.* Proc. Natl. Acad. Sci. USA, 1997. **94**: p. 2210-2214.
11. Rammensee, H.-G., J. Bachmann, and S. Stevanovic, *MHC ligands and peptide motifs.* 1997, New York: Chapman and Hall.
12. Singh, P.B., *Classical class I transplantation antigens in solution in the body fluids.* 1986.
13. Freedberg, I.M., ed. *Fitzpatrick's dermatology in general medicine.* 5 ed. Vol. 1. 1999, McGraw-Hill: New York.
14. Liddell, K., *Smell as a diagnostic marker.* Journal of Postgraduate Medicine, 1976. **52**: p. 136.
15. Jager, M.W.d., et al., *The phase behavior of skin lipid mixtures based on synthetic ceramides.* Chemistry and Physics of Lipids, 2003. **124**(2): p. 123-134.
16. Forslind, B., *A domain mosaic model of the skin barrier.* Acta Dermato-Venereologica, 1994. **74**(1): p. 1-6.
17. Heath, W.R. and F.R. Carbone, *Coupling and cross-presentation.* Nature, 2005. **434**: p. 27-28.
18. Neijssen, J., et al., *Cross-presentation by intercellular peptide transfer through gap junctions.* Nature, 2005. **434**: p. 83-88.
19. Preti, G. 2005: Cambridge, MA.
20. Nicolaides, N., *Skin lipids: their biochemical uniqueness.* Science, 1974. **186**(4158): p. 19-26.

21. Petricoin, E.F., III, et al., *Use of proteomic patterns in serum to identify ovarian cancer.* The Lancet, 2002. **359**: p. 572-577.
22. Petricoin, E.F., III, et al., *Serum proteomic patterns for detection of prostate cancer.* Journal of the National Cancer Institute, 2002. **94**(20): p. 1576-1578.
23. Naitoh, K., et al., *Direct temperature-controlled trapping system and its use for the gas chromatographic determination of organic vapor released from human skin.* Analytical Chemistry, 2000. **72**: p. 2797-2801.
24. Ramotowski, R.S., *Composition of latent print residue*, in *Advances in Fingerprint Technology*. 2001.
25. Bernier, U.R., et al., *Analysis of human skin emanations by gas chromatography/mass spectrometry. 2. identification of volatile compounds that are candidate attractants for the yellow fever mosquito (aedes aegypti).* Analytical Chemistry, 2000. **72**: p. 747-756.
26. Bernier, U.R., et al., *Chemical analysis of human skin emanations: comparison of volatiles from humans that differ in attraction of aedes aegypti (diptera: culicidae).* Journal of the American Mosquito Control Association, 2002. **18**(3): p. 186-195.
27. Braks, M.A.H. and W. Takken, *Incubated human sweat but not fresh sweat attracts the malaria mosquito anopheles gambiae sensu stricto.* Journal of Chemical Ecology, 1999. **25**(3): p. 663-672.
28. Healy, T.P. and M.J.W. Copland, *Human sweat and 2-oxopentanoic acid elicit a landing response from anopheles gambiae.* Medical and Veterinary Entomology, 2000. **14**: p. 195-200.
29. Sens, D.A., M.A. Simmons, and S.S. Spicer, *The Analysis of Human Sweat Proteins by Isoelectric Focusing. I. Sweat Collection Utilizing the Macroduct System Demostrates the Presence of Previously Unrecognized Sex-Related Proteins.* Pediatric Research, 1985. **19**(8): p. 873-878.
30. Koziel, J., M. Jia, and J. Pawliszyn, *Air sampling with porous solid-phase microextraction fibers.* Analytical Chemistry, 2000. **72**: p. 5178-5186.
31. Pawliszyn, J., *Solid phase microextraction: theory and practice.* 1997, New York: Wiley-VCH.
32. Mitchell, M., *An introduction to genetic algorithms.* 1996, Cambridge, MA: MIT Press.
33. Holland, J.H., *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.* 1992, Cambridge, MA: MIT Press.
34. Conrads, T.P., et al., *High-resolution serum proteomic features for ovarian cancer detection.* Endocrine-Related Cancer, 2004. **11**: p. 163-178.
35. Petricoin, E. and L.A. Liotta, *The vision for a new diagnostic paradigm.* Clinical Chemistry, 2003. **49**(8): p. 1276-1278.
36. Check, E., *Running before we can walk?* Nature, 2004. **429**: p. 496-497.
37. Diamandis, E.P., *Proteomic patterns in biological fluids: do they represent the future of cancer diagnostics?* Clinical Chemistry, 2003. **49**(8): p. 1272-1275.

38. Ransohoff, D.F., *Rules of evidence for cancer molecular-marker discovery and validation.* Nature Reviews: Cancer, 2004. **4**: p. 309-314.
39. Ransohoff, D.F., *Bias as a threat to the validity of cancer molecular-marker research.* Nature Reviews: Cancer, 2005. **5**(2): p. 142-149.
40. Ornstein, D.K., et al., *Serum proteomic profiling can discriminate prostate cancer from benigh prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml.* Urology, 2004. **172**(4): p. 1302-1305.
41. Malmquist, G. and R. Danielsson, *Alignment of chromatographic profiles for principal component analysis: a prerequisite for fingerprinting methods.* Journal of Chromatography A, 1994. **687**: p. 71-88.
42. Willse, A., et al., *Identification of Major Histocompatibility Complex-Regulated Body Odorants by Statistical Analysis of a Comparative Gas Ghromatography/Mass Spectrometry Experiment.* Analytical Chemistry, 2005.
43. Krebs, M.D., et al., *Alignment of analytical sensor data by landmark selection from complex chemical mixtures.* in review, 2005.
44. Bernier, U.R. 2003.
45. Schaefer, M.L., D.A. Young, and D. Restrepo, *Olfactory fingerprints for major histocompatibility complex-determined body odors.* Journal of Neuroscience, 2001. **21**(7): p. 2481-2487.
46. Augusto, F., J. Koziel, and J. Pawliszyn, *Design and validation of portable SPME devices for rapid field air sampling and diffusion-based calibration.* Analytical Chemistry, 2001. **73**(3): p. 481-486.
47. McLafferty, F.W. and D.B. Stauffer, *Retrieval and interpretative computer programs for mass spectrometry.* J. Chem. Inf. Comput. Sci., 1985. **25**: p. 245-252.
48. Pesyna, G.M., et al., *Statistical occurrence of mass and abundance values in mass spectra.* Analytical Chemistry, 1975. **47**(7): p. 1161-1164.
49. Zeng, X.N., et al., *Analysis of charateristic odors from human male axillae.* Journal of Chemical Ecology, 1991. **17**: p. 1469-1491.
50. Zeng, X.N., et al., *An investigation of human apocrine gland secretion for axillary odor.* Journal of Chemical Ecology, 1992. **18**: p. 1039-1055.

Appendix

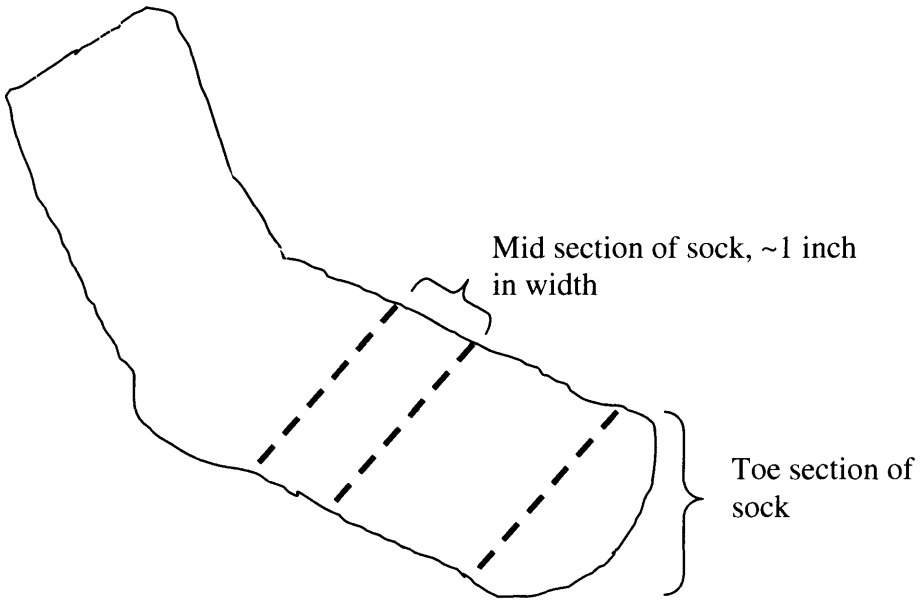


Figure 14: Sectioning of Socks for Sock Odors Experiment

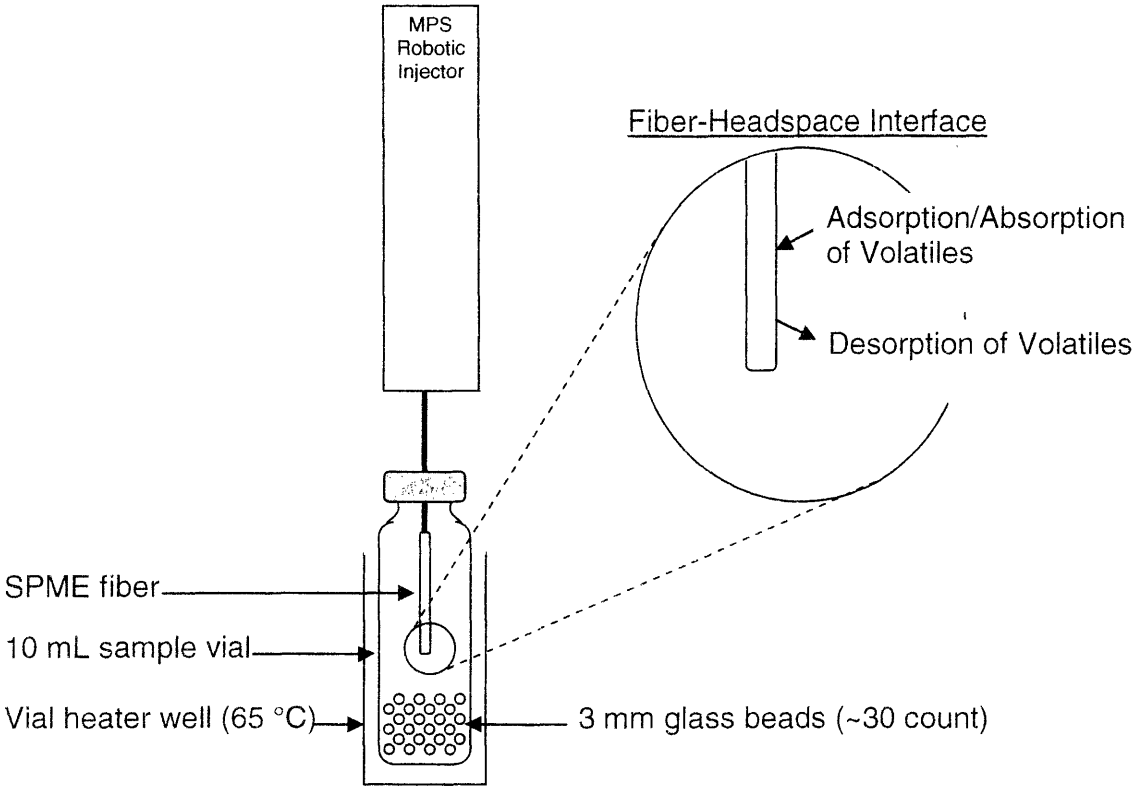
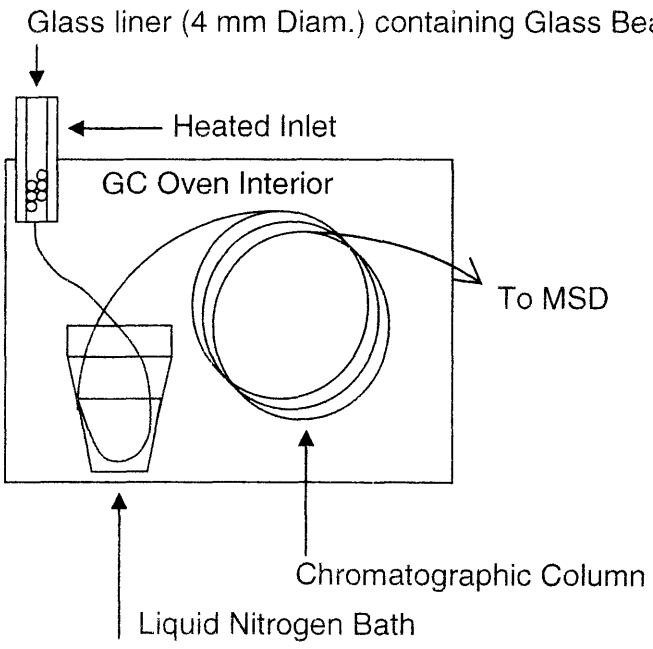


Figure 15: Diagram of SPME Extraction



General Method:

1. Beads are sealed in inlet
2. Inlet ramped to 250°C (5 min)
3. Oven remains at ambient (and open) for first 10 minutes
4. LiN_2 bath removed
5. Oven temperature program begins (3 temperature ramps)

Figure 16: Diagram of Instrument Setup for Thermal Desorption of Glass Beads followed by Cryofocusing

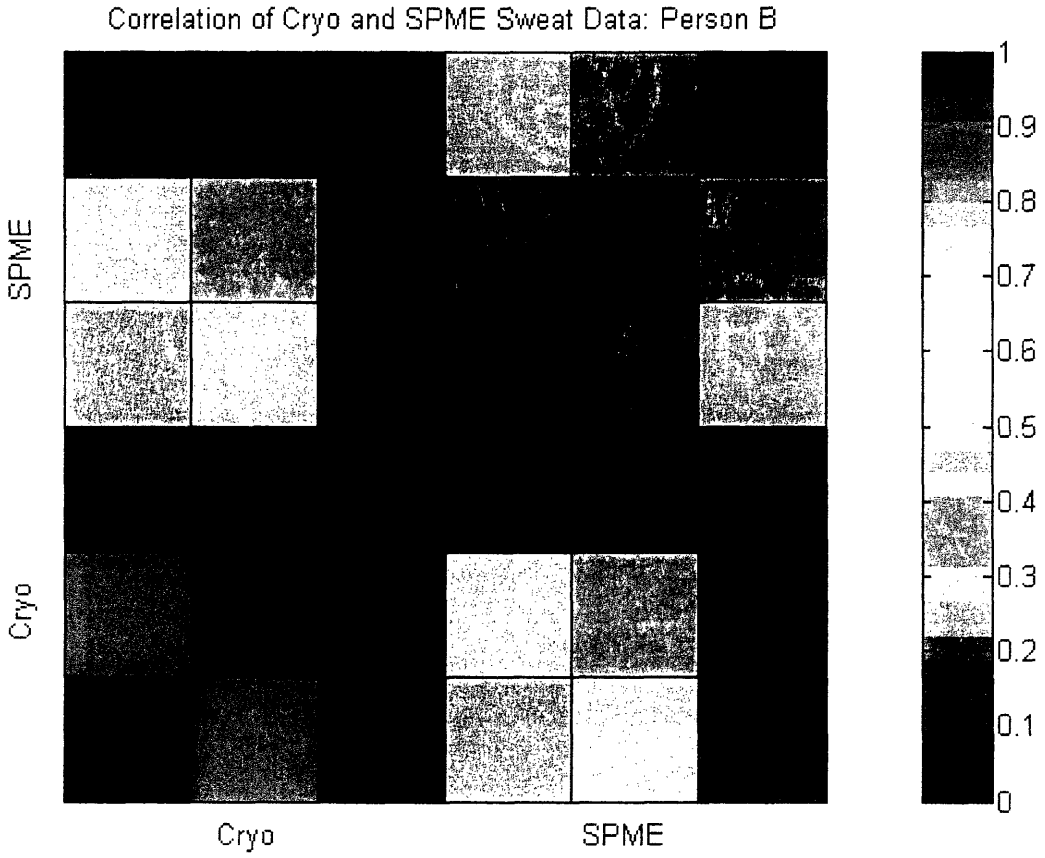


Figure 17: Dot Product Comparisons of Data for SPME versus CRYO Experiment

Table 11: List of Compounds for Donor 12 and Donor 13

Donor 12, Day 1 (07/02/04)

1,2-Benzenedicarboxylic acid, bis(2-methylpropyl) ester
7-Acetyl-6-ethyl-1,1,4,4-tetramethyltetralin
Butanoic acid
Cyclohexadecane

Decanal
Dodecanoic acid
Hexanoic acid
Nonanal
Octanal, 2-(phenylmethylene)-

Donor 13, Day 1 (07/02/04)
1,2-Benzenedicarboxylic acid, bis(2-methylpropyl) ester
5,9-Undecadien-2-one, 6,10-dimethyl-, (E)-
Decanal
Dodecanoic acid
Ethanol, 2-(2-ethoxyethoxy)-
n-Decanoic acid
Nonanal
Nonanoic acid

Table 12: List of Compounds for Donor 17 and Donor 18

Donor 17, Day 1 (07/02/04)
5,9-Undecadien-2-one, 6,10-dimethyl-, (E)-
7-Acetyl-6-ethyl-1,1,4,4-tetramethyltetralin
Cyclopentaneacetic acid, 3-oxo-2-pentyl-, methyl ester
Dodecanoic acid
Ethylene brassylate
Formamide, N,N-dibutyl-
Hexadecanoic acid, methyl ester
Isopropyl Myristate
Octanal, 2-(phenylmethylene)-

Donor 18, Day 1 (07/02/04)
1,2-Benzenedicarboxylic acid, bis(2-methylpropyl) ester
2,6,10,14,18,22-Tetracosahexaene, 2,6,10,15,19,23-hexamethyl-, (all-E)-
5,9-Undecadien-2-one, 6,10-dimethyl-, (E)-
Butane, 1,1'-[oxybis(2,1-ethanediyloxy)]bis-
Cyclohexadecane
Cyclopenta[g]-2-benzopyran, 1,3,4,6,7,8-hexahydro-4,6,6,7,8,8-hexamethyl-
Dodecanoic acid
Hexanoic acid, 2-ethyl-, hexadecyl ester
Isopropyl Myristate
Naphthalene
Octanal, 2-(phenylmethylene)-

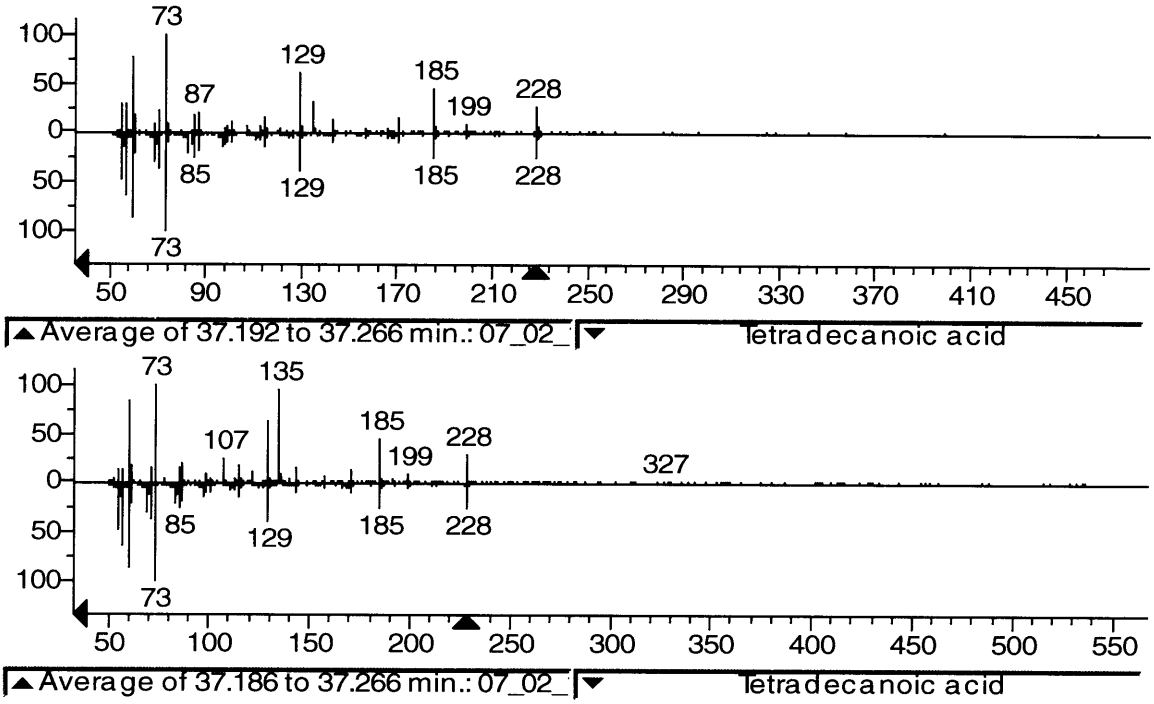


Figure 18: NIST Library Spectral Comparisons with Peak Unique to Donor 12 (top) and Donor 13 (bottom) when Compared to Donors 17 and 18

MATLAB Script for Alignment Algorithm:

```
function [] = AlignBlindLandmarks4pchip(varargin)
%
% Use landmarks derived from 2 reference data sets to time align
another chromatogram to them.
%
% Usage format is:
%
% AlignBlindLandmarks4pchip();

%  The thresholds are the minimum total ion abundance that you want to
use to
%  identify a peak.  The higher the threshold, the fewer peaks
identified.
%  MaxTimeOffset is the maximum allowable misalignment between files,
%  landmarks will not be checked for identical ness between files
unless
%  they are within this time offset value (in seconds).

%  MinCorr is the minimum correlation value of an m/z spectra from two
%  files to be considered identical.
```

```
%  allowtime is the max time (sec) allowed between usable landmarks
when
%  calculating the functional approximation.

%  Example of inputs:
%  reffile1 = 'grey_01interp.bin';
%  reffile2 = 'grey_03interp.bin';
%  reffile3 = 'grey_04interp.bin';
%  reffile4 = 'grey_05interp.bin';
%  filename = 'grey_02interp.bin';
%  threshold1 = 600000;
%  threshold2 = 600000;
%  threshold3 = 600000;
%  threshold4 = 600000;
%  threshold5 = 600000;
%  outputfile5 = 'grey_02align.bin';
%  MaxTimeOffset = 20;
%  MinCorr = 0.99;
%  allowtime = 20;

%%% open the diary file
diaryName = strcat('diary_AlignBlindLandmarks2pchip_',
datestr(datevec(now),30), '.txt');
diary (diaryName);
fprintf('\n\n\n\n****** Beginning AlignBlindLandmarks2pchip at %s
******\n\n', datestr(datevec(now),0));

%%% define the number of parameters in each array
numParams = 14;

%%% check number of input arguments
if (nargin == 0)
    fprintf('ERROR: AlignBlindLandmarks2() must have inputs.
Goodbye.\n\n');
    diary off;
    return;
end

%%% check validity of all input array arguments
for (ii = 1:nargin)
    if (length(varargin{ii}) ~= numParams)
        fprintf('ERROR: each bpskTest() input argument array must have
%d elements. Goodbye.\n\n', numParams);
        diary off;
        return;
    end
end

%%% sanitize all the input arguments
for (inpArgIdx = 1:nargin)

    reffile1 = varargin{inpArgIdx}{1};
```

```
    reffile2 = varargin{inpArgIdx}{2};
    reffile3 = varargin{inpArgIdx}{3};
    reffile4 = varargin{inpArgIdx}{4};
    filename = varargin{inpArgIdx}{5};
    threshold1 = varargin{inpArgIdx}{6};
    threshold2 = varargin{inpArgIdx}{7};
    threshold3 = varargin{inpArgIdx}{8};
    threshold4 = varargin{inpArgIdx}{9};
    threshold5 = varargin{inpArgIdx}{10};
    outputfile5 = varargin{inpArgIdx}{11};
    MaxTimeOffset = varargin{inpArgIdx}{12};
    MinCorr = varargin{inpArgIdx}{13};
    allowtime = varargin{inpArgIdx}{14};

    fprintf('\nProcessing task %d of %d with reffile1=%s, reffile2=%s,
reffile3=%s, reffile4=%s, filename=%s, threshold1=%d, threshold2=%d,
threshold3=%d, threshold4=%d, threshold5=%5, outputfile5=%s\n',
inpArgIdx, nargin, reffile1, reffile2, reffile3, reffile4, filename,
threshold1, threshold2, threshold3, threshold4, threshold5,
outputfile5);

%%  reference file #1
        fidm = fopen(reffile1, 'r');
        RTPtsm = fread(fidm,1,'long');
        MZPtsm = fread(fidm,1,'long');
        RTm = fread(fidm,RTPtsm,'single');
        MZm = fread(fidm,MZPtsm,'single');
        InterpIntensityMatrixm = fread(fidm,RTPtsm*MZPtsm,'single');
        fclose(fidm);

InterpIntensityMatrixm=reshape(InterpIntensityMatrixm,RTPtsm,MZPtsm);
        CGram1 = sum(InterpIntensityMatrixm');

    %     plot(RTm,CGram), xlabel('RT, Seconds'), ylabel('Abundance'),
title(infilem)
    %     pause

    %     InterpIntensityMatrixm = sqrt(abs(InterpIntensityMatrixm));
    %     imagesc(MZm,RTm, InterpIntensityMatrixm), ylabel('RT,
Seconds'), xlabel('M/Z'), title(filename1)
    %     pause

    IMatrix1 = InterpIntensityMatrixm;
    MZ1 = MZm';
    time1 = RTm;
    TIC1 = CGram1';

    clear RTm MZm InterpIntensityMatrixm RTPtsm MZPtsm fidm

%%  reference file #2
        fidm = fopen(reffile2, 'r');
        RTPtsm = fread(fidm,1,'long');
```

```
        MZPtsm = fread(fidm,1,'long');
        RTm = fread(fidm,RTPtsm,'single');
        MZm = fread(fidm,MZPtsm,'single');
        InterpIntensityMatrixm = fread(fidm,RTPtsm*MZPtsm,'single');
        fclose(fidm);

InterpIntensityMatrixm=reshape(InterpIntensityMatrixm,RTPtsm,MZPtsm);
        CGram2 = sum(InterpIntensityMatrixm');

    %     plot(RTm,CGram), xlabel('RT, Seconds'), ylabel('Abundance'),
title(infilem)
    %     pause

    %     InterpIntensityMatrixm = sqrt(abs(InterpIntensityMatrixm));
    %     imagesc(MZm,RTm, InterpIntensityMatrixm), ylabel('RT,
Seconds'), xlabel('M/Z'), title(infilem)
    %     pause

    IMatrix2 = InterpIntensityMatrixm;
    MZ2 = MZm';
    time2 = RTm;
    TIC2 = CGram2';

    clear RTm MZm InterpIntensityMatrixm RTPtsm MZPtsm fidm

 %%  reference file #3
        fidm = fopen(reffile3, 'r');
        RTPtsm = fread(fidm,1,'long');
        MZPtsm = fread(fidm,1,'long');
        RTm = fread(fidm,RTPtsm,'single');
        MZm = fread(fidm,MZPtsm,'single');
        InterpIntensityMatrixm = fread(fidm,RTPtsm*MZPtsm,'single');
        fclose(fidm);

InterpIntensityMatrixm=reshape(InterpIntensityMatrixm,RTPtsm,MZPtsm);
        CGram3 = sum(InterpIntensityMatrixm');

    %     plot(RTm,CGram), xlabel('RT, Seconds'), ylabel('Abundance'),
title(infilem)
    %     pause

    %     InterpIntensityMatrixm = sqrt(abs(InterpIntensityMatrixm));
    %     imagesc(MZm,RTm, InterpIntensityMatrixm), ylabel('RT,
Seconds'), xlabel('M/Z'), title(infilem)
    %     pause

    IMatrix3 = InterpIntensityMatrixm;
    MZ3 = MZm';
    time3 = RTm;
    TIC3 = CGram3';

    clear RTm MZm InterpIntensityMatrixm RTPtsm MZPtsm fidm
```

```
%%  reference file #4
        fidm = fopen(reffile4, 'r');
        RTPtsm = fread(fidm,1,'long');
        MZPtsm = fread(fidm,1,'long');
        RTm = fread(fidm,RTPtsm,'single');
        MZm = fread(fidm,MZPtsm,'single');
        InterpIntensityMatrixm = fread(fidm,RTPtsm*MZPtsm,'single');
        fclose(fidm);

InterpIntensityMatrixm=reshape(InterpIntensityMatrixm,RTPtsm,MZPtsm);
        CGram4 = sum(InterpIntensityMatrixm');

    %     plot(RTm,CGram), xlabel('RT, Seconds'), ylabel('Abundance'),
title(infilem)
    %     pause

    %     InterpIntensityMatrixm = sqrt(abs(InterpIntensityMatrixm));
    %     imagesc(MZm,RTm, InterpIntensityMatrixm), ylabel('RT,
Seconds'), xlabel('M/Z'), title(infilem)
    %     pause

    IMatrix4 = InterpIntensityMatrixm;
    MZ4 = MZm';
    time4 = RTm;
    TIC4 = CGram4';

    clear RTm MZm InterpIntensityMatrixm RTPtsm MZPtsm fidm

%%  file to be aligned
        fidm = fopen(filename, 'r');
        RTPtsm = fread(fidm,1,'long');
        MZPtsm = fread(fidm,1,'long');
        RTm = fread(fidm,RTPtsm,'single');
        MZm = fread(fidm,MZPtsm,'single');
        InterpIntensityMatrixm = fread(fidm,RTPtsm*MZPtsm,'single');
        fclose(fidm);

InterpIntensityMatrixm=reshape(InterpIntensityMatrixm,RTPtsm,MZPtsm);
        CGram5 = sum(InterpIntensityMatrixm');

    %     plot(RTm,CGram), xlabel('RT, Seconds'), ylabel('Abundance'),
title(infilem)
    %     pause

    %     InterpIntensityMatrixm = sqrt(abs(InterpIntensityMatrixm));
    %     imagesc(MZm,RTm, InterpIntensityMatrixm), ylabel('RT,
Seconds'), xlabel('M/Z'), title(infilem)
    %     pause

    IMatrix5 = InterpIntensityMatrixm;
    MZ5 = MZm';
```

```
    time5 = RTm;
    TIC5 = CGram5';

    IMatrix1(:,208) = zeros(RTPtsm,1);
    IMatrix1(:,282) = zeros(RTPtsm,1);
    IMatrix2(:,208) = zeros(RTPtsm,1);
    IMatrix2(:,282) = zeros(RTPtsm,1);
    IMatrix3(:,208) = zeros(RTPtsm,1);
    IMatrix3(:,282) = zeros(RTPtsm,1);
    IMatrix4(:,208) = zeros(RTPtsm,1);
    IMatrix4(:,282) = zeros(RTPtsm,1);
    IMatrix5(:,208) = zeros(RTPtsm,1);
    IMatrix5(:,282) = zeros(RTPtsm,1);

    CGram1 = sum(IMatrix1');
    CGram2 = sum(IMatrix2');
    CGram3 = sum(IMatrix3');
    CGram4 = sum(IMatrix4');
    CGram5 = sum(IMatrix5');

    [PeakAmp1, PeakRT1, PeakInds1] = FindPeaksV2(time1, TIC1,
threshold1);
    [PeakAmp2, PeakRT2, PeakInds2] = FindPeaksV2(time2, TIC2,
threshold2);
    [PeakAmp3, PeakRT3, PeakInds3] = FindPeaksV2(time3, TIC3,
threshold3);
    [PeakAmp4, PeakRT4, PeakInds4] = FindPeaksV2(time4, TIC4,
threshold4);
    [PeakAmp5, PeakRT5, PeakInds5] = FindPeaksV2(time5, TIC5,
threshold5);

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Plot unaligned TICs, frame-by-frame
    Frames = fix(length(time2)/10);
        for n = 1:10
        Inds = ((n-1)*Frames+1):(n*Frames);
        figure (1); subplot(5,2,n);
plot(time1(Inds),TIC1(Inds),time2(Inds),TIC2(Inds),time3(Inds),TIC3(Ind
s),time4(Inds),TIC4(Inds),time5(Inds),TIC5(Inds)), xlabel('Retention
Time (sec)'), ylabel('Total Ion Abundance')
        grid on%, legend('Sample 1','Sample 2', 'Sample 3','Sample
4','Sample 5')
    end
%     figure (1);
plot(time1,TIC1,time2,TIC2,time3,TIC3,time4,TIC4,time5,TIC5),
xlabel('Retention Time (sec)'), ylabel('Total Ion Abundance')
%          legend('Sample 1','Sample 2','Sample 3','Sample 4','Sample
5'), grid on
%     %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %Plot chromatogram with landmarks circled
    figure (2); plot(time1, TIC1, PeakRT1, PeakAmp1,'o'),
title(['Sample 1 - ',num2str(length(PeakInds1)),' Landmarks'])
```

```
        grid on, xlabel('Retention Time (sec)'), ylabel('Total Ion
Abundance'), legend('Chromatogram','Landmarks')
    figure (3); plot(time2, TIC2, PeakRT2, PeakAmp2,'o'),
title(['Sample 2 - ',num2str(length(PeakInds2)),' Landmarks'])
        grid on, xlabel('Retention Time (sec)'), ylabel('Total Ion
Abundance'), legend('Chromatogram','Landmarks')
    figure (4); plot(time3, TIC3, PeakRT3, PeakAmp3,'o'),
title(['Sample 3 - ',num2str(length(PeakInds3)),' Landmarks'])
        grid on, xlabel('Retention Time (sec)'), ylabel('Total Ion
Abundance'), legend('Chromatogram','Landmarks')
    figure (5); plot(time4, TIC4, PeakRT4, PeakAmp4,'o'),
title(['Sample 4 - ',num2str(length(PeakInds4)),' Landmarks'])
        grid on, xlabel('Retention Time (sec)'), ylabel('Total Ion
Abundance'), legend('Chromatogram','Landmarks')
    figure (6); plot(time5, TIC5, PeakRT5, PeakAmp5,'o'),
title(['Sample 5 - ',num2str(length(PeakInds5)),' Landmarks'])
        grid on, xlabel('Retention Time (sec)'), ylabel('Total Ion
Abundance'), legend('Chromatogram','Landmarks')

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
    % Take the landmarks in Sample 1 as the alignment reference, and
find
    % offset to identical features in Sample 2

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
%    MaxTimeOffset = 20;                                    %%  max
tolerable misalignment, seconds
%    MinCorr = 0.99;
%     LandmarkMatrix = zeros((length(PeakInds1)+length(PeakInds2)),7);
%     LandmarkMatrix(:,1) = [PeakInds1'; PeakInds2'];
%     LandmarkMatrix(:,2) = [PeakRT1'; PeakRT2'];
%     LandmarkMatrix(:,3) = [PeakAmp1'; PeakAmp2'];
%     LandmarkMatrix = sortrows(LandmarkMatrix,[1]);
%     PeakInds = LandmarkMatrix(:,1)';
%     PeakRTs = LandmarkMatrix(:,2)';
%     PeakAmps = LandmarkMatrix(:,3)';
%
%     for n = 1:length(PeakRTs)
%         for m = 1:length(PeakInds3)
%             if abs(PeakRTs(n) - PeakRT3(m)) < MaxTimeOffset
%                % Trace1 = IMatrix1(PeakInds1(n),:);
%                % Trace2 = IMatrix2(PeakInds2(m),:);
%                % Corrcoef =
(Trace1*Trace2')/sqrt((Trace1*Trace1')*(Trace2*Trace2'));
%                 Trace1 = IMatrix1((PeakInds(n)-9):(PeakInds(n)+9),:);
%                 Trace3 = IMatrix2((PeakInds3(m)-
9):(PeakInds3(m)+9),:);
%                % Trace1 = Trace1 - mean(mean(Trace1));
%                % Trace2 = Trace2 - mean(mean(Trace2));
%                 Corrcoef =
sum(sum(Trace1.*Trace3))/sqrt(sum(sum(Trace1.*Trace1))*sum(sum(Trace3.*
Trace3))),
```

```
%                   if Corrcoef > MinCorr
%                      %plot(MZ1,sum(Trace1),MZ2,sum(Trace2)),
title([num2str(Corrcoef), '  ', num2str(PeakRT1(n)),'  ',
num2str(PeakRT2(m))])
%                      %pause
%                      %figure (13); plot(MZ1,Trace1,MZ2,Trace2);
%                      %pause
%                       LandmarkMatrix(n,4)=PeakInds3(m);
%                       LandmarkMatrix(n,5)=PeakRT3(m);
%                       LandmarkMatrix(n,6)=PeakAmp3(m);
%                       LandmarkMatrix(n,7)=Corrcoef;
%                   end
%               end
%           end
%       end
%       LandmarkMatrix21 = LandmarkMatrix;
    LandmarkMatrix = zeros(length(PeakInds1),7);
    LandmarkMatrix(:,1) = PeakInds1';
    LandmarkMatrix(:,2) = PeakRT1';
    LandmarkMatrix(:,3) = PeakAmp1';
    for n = 1:length(PeakInds1)
        for m = 1:length(PeakInds5)
            if abs(PeakRT1(n) - PeakRT5(m)) < MaxTimeOffset
               % Trace1 = IMatrix1(PeakInds1(n),:);
               % Trace2 = IMatrix2(PeakInds2(m),:);
               % Corrcoef =
(Trace1*Trace2')/sqrt((Trace1*Trace1')*(Trace2*Trace2'));
                Trace1 = IMatrix1((PeakInds1(n)-9):(PeakInds1(n)+9),:);
                Trace5 = IMatrix5((PeakInds5(m)-9):(PeakInds5(m)+9),:);
               % Trace1 = Trace1 - mean(mean(Trace1));
               % Trace2 = Trace2 - mean(mean(Trace2));
                Corrcoef =
sum(sum(Trace1.*Trace5))/sqrt(sum(sum(Trace1.*Trace1))*sum(sum(Trace5.*
Trace5)));
                if Corrcoef > MinCorr
                   %plot(MZ1,sum(Trace1),MZ2,sum(Trace2)),
title([num2str(Corrcoef), '  ', num2str(PeakRT1(n)),'  ',
num2str(PeakRT2(m))])
                   %pause
                   %figure (13); plot(MZ1,Trace1,MZ2,Trace2);
                   %pause
                    LandmarkMatrix(n,4)=PeakInds5(m);
                    LandmarkMatrix(n,5)=PeakRT5(m);
                    LandmarkMatrix(n,6)=PeakAmp5(m);
                    LandmarkMatrix(n,7)=Corrcoef;
               end
           end
        end
    end
    LandmarkMatrix51 = LandmarkMatrix;
    %%%%%
    LandmarkMatrix = zeros(length(PeakInds2),7);
    LandmarkMatrix(:,1) = PeakInds2';
    LandmarkMatrix(:,2) = PeakRT2';
```

```
    LandmarkMatrix(:,3) = PeakAmp2';
    for n = 1:length(PeakInds2)
        for m = 1:length(PeakInds5)
            if abs(PeakRT2(n) - PeakRT5(m)) < MaxTimeOffset
               % Trace1 = IMatrix1(PeakInds1(n),:);
               % Trace2 = IMatrix2(PeakInds2(m),:);
               % Corrcoef =
(Trace1*Trace2')/sqrt((Trace1*Trace1')*(Trace2*Trace2'));
                Trace2 = IMatrix2((PeakInds2(n)-9):(PeakInds2(n)+9),:);
                Trace5 = IMatrix5((PeakInds5(m)-9):(PeakInds5(m)+9),:);
               % Trace1 = Trace1 - mean(mean(Trace1));
               % Trace2 = Trace2 - mean(mean(Trace2));
                Corrcoef =
sum(sum(Trace2.*Trace5))/sqrt(sum(sum(Trace2.*Trace2))*sum(sum(Trace5.*
Trace5)));
                if Corrcoef > MinCorr
                   %plot(MZ1,sum(Trace1),MZ2,sum(Trace2)),
title([num2str(Corrcoef), '  ', num2str(PeakRT1(n)),'  ',
num2str(PeakRT2(m))])
                   %pause
                   %figure (13); plot(MZ1,Trace1,MZ2,Trace2);
                   %pause
                    LandmarkMatrix(n,4)=PeakInds5(m);
                    LandmarkMatrix(n,5)=PeakRT5(m);
                    LandmarkMatrix(n,6)=PeakAmp5(m);
                    LandmarkMatrix(n,7)=Corrcoef;
               end
           end
        end
    end
    LandmarkMatrix52 = LandmarkMatrix;
    %%%%%%%%
    LandmarkMatrix = zeros(length(PeakInds3),7);
    LandmarkMatrix(:,1) = PeakInds3';
    LandmarkMatrix(:,2) = PeakRT3';
    LandmarkMatrix(:,3) = PeakAmp3';
    for n = 1:length(PeakInds3)
        for m = 1:length(PeakInds5)
            if abs(PeakRT3(n) - PeakRT5(m)) < MaxTimeOffset
               % Trace1 = IMatrix1(PeakInds1(n),:);
               % Trace2 = IMatrix2(PeakInds2(m),:);
               % Corrcoef =
(Trace1*Trace2')/sqrt((Trace1*Trace1')*(Trace2*Trace2'));
                Trace3 = IMatrix3((PeakInds3(n)-9):(PeakInds3(n)+9),:);
                Trace5 = IMatrix5((PeakInds5(m)-9):(PeakInds5(m)+9),:);
               % Trace1 = Trace1 - mean(mean(Trace1));
               % Trace2 = Trace2 - mean(mean(Trace2));
                Corrcoef =
sum(sum(Trace3.*Trace5))/sqrt(sum(sum(Trace3.*Trace3))*sum(sum(Trace5.*
Trace5)));
                if Corrcoef > MinCorr
                   %plot(MZ1,sum(Trace1),MZ2,sum(Trace2)),
title([num2str(Corrcoef), '  ', num2str(PeakRT1(n)),'  ',
num2str(PeakRT2(m))])
```

```
                    %pause
                    %figure (13); plot(MZ1,Trace1,MZ2,Trace2);
                    %pause
                     LandmarkMatrix(n,4)=PeakInds5(m);
                     LandmarkMatrix(n,5)=PeakRT5(m);
                     LandmarkMatrix(n,6)=PeakAmp5(m);
                     LandmarkMatrix(n,7)=Corrcoef;
                end
            end
        end
    end
    LandmarkMatrix53 = LandmarkMatrix;
    %%%%%%%%%
    LandmarkMatrix = zeros(length(PeakInds4),7);
    LandmarkMatrix(:,1) = PeakInds4';
    LandmarkMatrix(:,2) = PeakRT4';
    LandmarkMatrix(:,3) = PeakAmp4';
    for n = 1:length(PeakInds4)
        for m = 1:length(PeakInds5)
            if abs(PeakRT4(n) - PeakRT5(m)) < MaxTimeOffset
               % Trace1 = IMatrix1(PeakInds1(n),:);
               % Trace2 = IMatrix2(PeakInds2(m),:);
               % Corrcoef =
(Trace1*Trace2')/sqrt((Trace1*Trace1')*(Trace2*Trace2'));
                Trace4 = IMatrix4((PeakInds4(n)-9):(PeakInds4(n)+9),:);
                Trace5 = IMatrix5((PeakInds5(m)-9):(PeakInds5(m)+9),:);
               % Trace1 = Trace1 - mean(mean(Trace1));
               % Trace2 = Trace2 - mean(mean(Trace2));
                Corrcoef =
sum(sum(Trace4.*Trace5))/sqrt(sum(sum(Trace4.*Trace4))*sum(sum(Trace5.*
Trace5)));
                if Corrcoef > MinCorr
                   %plot(MZ1,sum(Trace1),MZ2,sum(Trace2)),
title([num2str(Corrcoef), '  ', num2str(PeakRT1(n)),'  ',
num2str(PeakRT2(m))])
                    %pause
                    %figure (13); plot(MZ1,Trace1,MZ2,Trace2);
                    %pause
                     LandmarkMatrix(n,4)=PeakInds5(m);
                     LandmarkMatrix(n,5)=PeakRT5(m);
                     LandmarkMatrix(n,6)=PeakAmp5(m);
                     LandmarkMatrix(n,7)=Corrcoef;
                end
            end
        end
    end
    LandmarkMatrix54 = LandmarkMatrix;
    clear LandmarkMatrix;
    LandmarkMatrix = [LandmarkMatrix51; LandmarkMatrix52;
LandmarkMatrix53; LandmarkMatrix54];
    LandmarkMatrix = sortrows(LandmarkMatrix,[2,1]);
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
    %   Sample-derived time warping function
    %
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
    I = find(LandmarkMatrix(:,4));  %   identify usable landmarks
    Map = [LandmarkMatrix(I, 5) LandmarkMatrix(I, 2)-LandmarkMatrix(I,
5)];
    Map = [min(time3) Map(1,2); Map; max(time3) Map(length(I),2)];
    Map = [Map(1:(length(I)+1),:);
mean(Map((length(I)+1):(length(I)+2),:)) ;Map((length(I)+2),:)];

    i=1;
    while i<=(length(Map)-1);
        if Map(i,:)==Map(i+1,:);
            Map(i+1,:) = [];
        end
        if Map(i,1)==Map(i+1,1);
            Map(i+1,1)=(Map(i+1,1)+0.01);
        end
        i=i+1;
    end

% the below command will take Map (which contains the common landmarks
and time
% offset) and add "dummy" landmarks between two points that are far
% apart.  These points will have the same offset as the second of the
two
% spread landmarks... This helps constrain the cubic spline function.
 %   allowtime = 10;              %max time allowed between usable
landmarks
    r = 1;
    while r < (length(Map)-1)
        if (Map(r+1,1)-Map(r,1)) > allowtime
            Map = [Map(1:(r-1),:); (Map(r,1):(allowtime):Map(r+1,1))',
Map(r,2)*ones(ceil((Map(r+1)-Map(r))/(allowtime)),1); Map(r+1:end,:)];
        end
        r=r+1;
    end


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
    %   Adjust time axis using warping function
    %   Plot functional approximations
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%
    Xunique=Map(:,1); % added by JJA, 9 February 2005, because pchip
was giving error -> the data abscissae should be distinct; the two
```

```
comments below are from a suggest solution from the mathworks website,
searching for this error
    Yunique=Map(:,2);        %[b,i,j]=unique(x); % Remove duplicates
from x
    [Bb,Yy,Zz]=unique(Xunique); %x55=interp1(b, y(i), 5.5) % Obtain the
interpolated value at x=5.5 - this succeeds

    pp = pchip(Bb,Yunique(Yy));

    Delta = ppval(pp,time5);

%     recommented the following out because I ran into the same pchip
error
%     again even with the higher threshold; this leads me to believe
the
%     compound matrix of four ref files includes double entries
%     pp = pchip(Map(:,1),Map(:,2));      % original code from
*2pchip.m; removed 'unique' call above because it seemed to be a
threshold too low problem
%     Delta = ppval(pp,time5);            % original code from
*2pchip.m with time3 --> time 5

    %figure (5);
plot(Map(:,1),Map(:,2),'o',(0:max(Map(:,1))./(length(RTm)-
1):max(Map(:,1))), pp), xlabel('Retention Time(sec)'), ylabel('RT
Sample 1 - RT Sample 4 (sec)')
    %    legend('Landmarks','Functional Approximation'), grid on
%     pp=pp';
%     RT3p = time5 + pp;

%     pp = csape(Map(:,1),Map(:,2));        %cubic spline
%     Delta = ppval(pp,time5);
    figure (7); plot(Map(:,1),Map(:,2),'o',time5, Delta),
xlabel('Retention Time(sec)'), ylabel('RT Sample 1-4 - RT Sample 5
(sec)')
       grid on, legend('Landmarks','Functional Approximation')
     RT5p = time5 + Delta;

    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    %   Plot aligned TICs frame-by-frame
    %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
    Frames = fix(length(time2)/10);
    for n = 1:10
        Inds = ((n-1)*Frames+1):(n*Frames);
        figure (8); subplot(5,2,n);
plot(time1(Inds),TIC1(Inds),time2(Inds),TIC2(Inds),time3(Inds),TIC3(Ind
s),time4(Inds),TIC4(Inds),RT5p(Inds),TIC5(Inds)), xlabel('Retention
Time (sec)'), ylabel('Total Ion Abundance')
        grid on%, legend('Sample 1','Sample 2', 'Sample 3','Sample
4','Sample 5')
    end
```

```
%     figure (6);
plot(time1,TIC1,time2,TIC2,time3,TIC3,time4,TIC4,RT3p,TIC3),
xlabel('Aligned Retention Time (sec)'), ylabel('Total Ion Abundance')
%          legend('Sample 1','Sample 2', 'Sample 3','Sample 4','Sample
5'), grid on
%Check to make sure all looks good before outputting to file
%pause

    %%% save results to binary file (outputfile)
    outputFid5 = fopen(outputfile5, 'w');

    COUNT = fwrite(outputFid5,length(RT5p),'long');
    COUNT = fwrite(outputFid5,length(MZ5),'long');

    count = fwrite(outputFid5, RT5p, 'single');
    fprintf('Wrote %d single-precision values to file %s\n', count,
outputfile5);

    count = fwrite(outputFid5, MZ5, 'single');
    fprintf('Wrote %d single-precision values to file %s\n', count,
outputfile5);

    count = fwrite(outputFid5, IMatrix5, 'single');
    fprintf('Wrote %d single-precision values to file %s\n', count,
outputfile5);

    fclose(outputFid5);


end % for (inpArgIdx = 1:nargin)

%%% wrap it up
fclose('all');
fprintf('\n\n****** End AlignBlindLandmarks2 at %s ******\n\n\n\n\n\n',
datestr(datevec(now),0));
diary off;
```