METHODS FOR COLLECTION AND PROCESSING OF GENE EXPRESSION DATA

Thesis by

John F. Murphy

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2004

(Defended April 21, 2004)

Success is the ability to go from one failure to another with no loss of enthusiasm.
-Sir Winston Churchill

To my mother and the memory of my father

ACKNOWLEDGMENTS

I feel lucky to have spent five years at Caltech. It's unlikely that I will ever again be surrounded by such a dense mass of thoughtful people engaged in such honest pursuits. Although there are more people that flapped their wings to influence the outcome of my graduate career than I could list, I will attempt to mention some of the larger butterflies.

I am very grateful for the relationship I've had with my advisor, Mark Davis. His humble curiosity was inspirational, and his expansive approach to science allowed us to venture into fascinating new areas. The initial seed that grew into this thesis was planted at a meeting among Mark, Dave Tirrell, Barbara Wold, then-postdoctoral scholar, now-NYU Chemistry Professor Kent Kirshenbaum, and I. Mark, Dave, and Barbara have been with my project from its inception to the final thesis defense, and I thank them all for their careful consideration and helpful advice. I also thank Anand Asthagiri for serving on my committee.

Following that first meeting, Kent had the amusing pleasure of helping me through my first few syntheses. Surely, I would have incurred some grievous injury upon myself if it hadn't been for Kent. I also thank Mona Shahgholi of the Crellin Mass Spectrometry Facility, who shared all of her knowledge with me and helped me understand a field that I had no training in. Brian Williams of the Wold group has been a source of invaluable expertise with all things biological.

I gratefully acknowledge the support of several sponsors for my research. The first is a National Science Foundation Graduate Research Fellowship that provided the support for the first three years of my work here. Second is a Sun Microsystems Academic Equipment Grant that provided the workstation used for the majority of my analytical work. Finally, I am grateful for a grant from the Beckman Institute that supported my work in my final two years.

I must thank the members of the Davis lab, who have been both talented colleagues and great friends, especially the other four 1999 recruits. Without the help of Jeremy Heidel, I would not have survived the learning curve of molecular biology, or fantasy baseball for that matter. Jeremy is the steady hand in the laboratory, and was my backup advisor and Friday-night Atheneaum drinking buddy along with Swaroop Mishra. Swaroop's patience with me is remarkable; he was always around to listen to my complaints of the day, and with remarkable frequency offer simple solutions to my problems. As lucky as I feel to count Steve Popielarski among my friends, I feel exceptionally fortunate to have survived 5 years of friendship with Steve without losing any limbs or otherwise valuable appendages. Although intellectually I recognize the value of our collaboration in the lab, the memories that will stay with me are of kayaking, camping, skiing, and so on. And as trusted officemate Andrea Wight and I pack up and prepare to abandon our basement home of many years, we both wonder—will anyone be so foolish as to open the windows in the coming years? Many others made my time at Caltech enjoyable as well, so thank you Jon Galownia, Joe Holles, Theresa Reineke, and Efrain Hernandez. My friends outside of Caltech have always been there for me—Mike Colonno, Pete Schweitzer, Chris Thunberg, and Tom Aichele.

I would not have made it to Caltech if it hadn't been for the cast of zany ex-Caltechers who largely guided my education at Cornell University: Mike Duncan, Brad Anton, Kelvin Lee, and Bill Olbricht. I am equally grateful to my grandmother, Claire Betz, who supported my education at Cornell, as has given me the freedom to explore new paths in my life after Caltech.

I thank my mother and father, who really did believe in me, even when I was bringing home 'C's in high school math, and didn't believe in myself. Finally, perhaps the greatest of all the butterflies is my wife Kate, who talked me through the toughest times at Caltech, and has unfailingly supported me in every way.

ABSTRACT

Examination of the transcriptional messages encoded in the manifold of mRNA molecules within a cell is a central task of molecular biology and functional genomics. This examination can be broken down into two parts: collection of gene expression data, and analyses of those data. Here, a new method for collecting gene expression data, and two new methods for analyzing those data are presented.

A new method for quantifying gene expression denoted as the Mass-spectrometric Analysis of Gene Expression (MAGE) is developed. MAGE relies on novel conjugates of DNA oligonucleotide 30-mers; each unique sequence is conjugated via photolabile linker to an N-substituted glycine oligomer (peptoid) of unique mass. Deuterated bromoacetic acid is incorporated into some peptoids yielding two chemically identical probe conjugates of different molecular weights for each nucleic acid sequence of interest. Mixtures of these probes, along with 3' adjacent biotin-labeled oligonucleotides, are used to interrogate a target mixture of cDNA. Following hybridization, the two adjacent probes are ligated to enhance the specificity of the identification, and to enable the use of a biotin-affinity column for removal of confounding peptoid tags. The resulting mixture is exposed to longwave ultraviolet light to release the peptoid tags, that are quantified using MALDI-TOF mass spectrometry using the isotopically labeled peptoids as internal standards. These individual components of MAGE are demonstrated.

A strategy for simplification and visualizing of high-dimensional gene expression data, as well as a strategy for inferring the presence of clusters within those data, is

formulated and implemented. In order to visualize high-dimensional gene expression data, principle components analysis is used with subsequent mapping of the data onto an orthogonal set of basis functions known as Andrews curves. This analysis method is demonstrated by visualizing of breast cancer tumor data and yeast sporulation data. In order to cluster gene expression data, the expectation-maximization algorithm is employed to optimize the parameters of a mixture model of Lorentzian distributions. The difference between Lorentzian and Gaussian mixture models is first demonstrated with artificial data, and then applied to yeast sporulation data. The results indicate that mixtures of Lorentzian distributions may have significant utility for gene expression analysis.

The tools demonstrated here offer unique advantages when compared to the current suite of experimental and analytical tools employed by investigators of functional genomics.

TABLE OF CONTENTS

Acknowledgments iv
Abstract vii
Table of Contents viii
List of Tables xii
List of Figures xiv

Chapter One: Preface

1.1 Introduction 2
1.2 Objectives 9
1.2.1 Collecting Gene Expression Data 10
1.2.2 Analyzing Gene Expression Data 11
1.3 References 14

Chapter Two: Expanding the Toolkit for Peptoid Synthesis

2.1 Introduction 22
2.2 Experimental 26
2.2.1 General Peptoid Synthesis 26
2.2.2 Specialized Syntheses 28
2.2.3 Analytical Procedures 31
2.3 Results and Discussion 32
2.3.1 Branched Peptoids 32
2.3.2 Peptoid Capping 35
2.3.3 N-terminal Modifications 37
2.3.4 Macrocyclic Peptoids 41
2.3.5 Oligo-Adamantane Peptoids 46
2.3.6 Oligodeoxynucleotide-Peptoid Conjugates 49
2.4 Summary 53

2.5 References.....54

Chapter Three: MAGE: Mass-spectrometric Analysis of Gene Expression

3.1 Introduction.....61
3.2 Experimental.....67
3.2.1 Oligodeoxynucleotide Probes.....67
3.2.2 MAGE Methodology.....68
3.2.3 Analytical Procedures.....69
3.3 Results and Discussion.....70
3.3.1 Oligodeoxynucleotide Probes.....70
3.3.2 Mass-Spectrometric Quantification.....76
3.3.3 MAGE Methodology.....79
3.4 Summary.....88
3.5 References.....89

Chapter Four: Visualization and Analysis of Gene Expression Data with Model-Based Clustering and Andrews Curves

4.1 Introduction.....93
4.2 Experimental.....103
4.2.1 Algorithm for Clustering by Expectation-Maximization.....103
4.2.2 Computational Methods.....106
4.2.1 Sources of Data.....106
4.3 Results and Discussion.....107
4.3.1 Andrews Curves.....107
4.3.2 EM Clustering of Synthetic Data.....118
4.3.3 EM Clustering of Gene Expression Data.....129
4.4 Summary.....136
4.5 References.....138

Chapter Five: Conclusions and Recommendations

5.1 Peptoid Synthesis 145

5.2 MAGE Methodology 148

5.3 Visualization and Model-Based Clustering 152

5.4 Overall 154

5.5 References 155

Appendix A: Derivation of a Lorentzian distribution from the ratio of two independent normally distributed random variables

A.1 Derivation 159

Appendix B: MATLAB code for visualization and clustering algorithm*s*

B.1 Master algorithm function for generating Andrews curves 162

B.2 Utility for preprocessing data by PCA 164

B.3 Utility for implementing PCA algorithm with SVD 165

B.4 EM adaptive mixture of Lorentzians 166

B.5 EM adaptive mixture of Gaussians 168

B.6 Mixture of Lorentzians with plotting for 2D data sets 170

B.7 Utility for running multiple clustering experiments 173

B.8 Utility for plotting T-distributions 174

B.9 Utility for plotting histograms of cluster values 175

B.10 Utility for plotting cluster means and standard deviations 176

Appendix C: Additional peptoid syntheses

C.1 Peptoids incorporating fluorescent label and iodoacetamide 178

C.2 Peptoids incorporating orthonitrobenzyl moiety 182

C.3 Peptoids incorporating polar, aromatic side groups 185

LIST OF TABLES

Chapter One

Table 1.1 **Distribution of mRNA in a cell.** In a typical cell, the majority of genes are expressed as scarce transcripts. A single scarce gene might only make up .001% of the total transcript population.....6

Chapter Two

Table 2.1 **Peptoid synthesis single linking chemistry scheme of Figliozzi *et al.*, for 100 mg of resin.** These steps are repeated for each monomer addition. 28

Table 2.2 **Primary amine submonomers used for syntheses of branched peptoids BR1, BR2, and BR3.** Branches are introduced by diaminobutane incorporation.....32

Table 2.3 **Primary amine submonomers used for syntheses of capped peptoids CP1 and CP2**. 35

Table 2.4 **Acetylating submonomers used to modify the N-terminus of peptoids NT1-NT6**.....38

Table 2.5 **Primary amine submonomers used for syntheses of macrocyclic precursor peptoid CY1**.....42

Table 2.6 **Primary amine submonomers used for syntheses of adamantyl-peptoids AD1 and AD2.**.....47

Chapter Three

Table 3.1 **Measured mole fractions compared to predicted mole fractions.** Peptoid ID1 and 10-fold deuterated peptoid ID2 were mixed in various proportions, and analyzed by MALDI-TOF. The areas under the curves respective to each peptoid were used to estimate the relative mole fractions and compare those estimates to the volumetrically measured mole fractions.....79

Table 3.2 **Relative amounts of T1 and T2 in targets**. The targets for experiment 2 are mixtures of anti-sense oligonucleotides representing mouse myogenin and mouse paraoxonase in various proportions.....83

Chapter Four

Table 4.1 **Genes used by Chu *et al.* to create average expression patterns for each of seven classifications** 107

Table 4.2 **The reduced data of Sorlie *et al.* is processed by PCA implemented with SVD.** The largest number of principal components (PCs) that can realistically be visualized with Andrews curves accounts for approximately 21.9% of the variance in the data.....116

Table 4.3 **Likelihoods of fit for the full data set of Chu *et al* analyzed by EM-MoG and EM-MoL with 1 to 12 randomly initialized clusters, with the inclusion of the MDL penalty term.** For each choice of cluster size, each algorithm was executed 30 times to generate the statistics shown. 134

LIST OF FIGURES

Chapter One

Figure 1.1 **The central dogma of molecular biology.** This schematic representation of the central dogma suggests the importance of studying the transcriptome. 2

Chapter Two

Figure 2.1 **Schematic illustrations of peptoids and peptides.** Compared to peptides such as oligo-alanine, upper, peptoids, such as oligo-sarcosine, lower, lack stereochemistry and sites for hydrogen bonding, but retain a similar spacing and overall structure. 22

Figure 2.2 **The solid-phase submonomer method of peptoid synthesis**. This method is executed by repeated, alternating rounds of bromoacetylation and primary amine substitution. Commonly, the finished chain is cleaved by trifluoroacetic acid to form a C-terminal amide cap. 24

Figure 2.3 **Primary amines used for peptoid synthesis.** Peptoid diversity is generated through the choice of primary amines ($R\text{-}NH_2$ in Figure 4.2). Some commonly used classes include aromatic, aliphatic, heterocyclic, cationic, anionic, bulky, small, hydrophobic and hydrophilic. 24

Figure 2.4 **Schematic of how adamantane is incorporated into peptoids by first synthesizing (1)**. 30

Figure 2.5 **Schematic illustrations of three branched structures of increasing size.** Synthesis methods are detailed in Table 2.2.....33

Figure 2.6 **MALDI-TOF spectra of three branched structures of increasing size, BR1 (a), BR2 (b), and BR3 (c).** The masses of the three peptoids are visible as H^+ adducts.....34

Figure 2.7 **RP-HPLC separation and detection at 220 nm of as-made branched peptoid BR2.** Analysis indicated approximately 90% yield of the desired product 35

Figure 2.8 **Schematic illustrations of two capped peptoids, CP1 and CP2.** 36

Figure 2.9 **MALDI-TOF spectra of product of two attempts to cap the peptoid growing chain.** In (a), CP1 largely blocks chain extension, where as in (b), CP2 fails to block chain extension. The masses of the peptoids are visible as H^+ adducts.....37

Figure 2.10 **schematic illustrations of six peptoids with alternative N-termini**. These were formed by terminating the growing peptoid chain with an acetylation step instead of the usual substitution step 38

Figure 2.11 **MALDI-TOF spectra of six N-terminal-modified peptoids, NT1-NT6.** For NT1, (a), two groups of peaks represent the Na^+ and K^+ adducts, and incorporation of bromine accounts for the peak splitting within each of the two groups. For NT2, (b), NT3, (c), NT4, (d), and NT5 (e), the Na^+ and K^+ adducts are evident. For NT6, (f), the H^+ adduct is visible as well as the salt adducts.....39

Figure 2.12 **RP-HPLC separation with detection at 220nm of N-iodoacetyl peptoid NT2 indicating over 95% yield** 41

Figure 2.13 **Schematic illustrations of the synthesis of CY1 free-acid free-amine peptoid for cyclization** 43

Figure 2.14 **Schematic illustration of cyclization of CY1 into CY2 using peptide coupling reagents PyBOP and DIPEA.**.....44

Figure 2.15 **MALDI-TOF spectra of free-acid free-amine peptoid, CY1, (a), and the cyclized product, CY2, (b).** In (a), the H^+ adduct is evident, while in (b), the H^+, Na^+ and K^+ adducts are evident. 45

Figure 2.16 **RP-HPLC separation with detection at 220 nm of pre- and post-cyclization peptoids CY1 (a) and CY2 (b)** 46

Figure 2.17 **Schematic illustrations of AD1 and AD2 tetra-adamantyl peptoids** 47

Figure 2.18 **Schematic illustration of the process for deprotecting pendant amines and conjugating adamantane carboxylic acid to AD2**.....48

Figure 2.19 **MALDI-TOF spectra of tetra-adamantyl peptoids**. AD1, (a), is evidence by the H^+ adduct. In (b), the spectrum indicates H^+ and Na^+ adducts that are somewhat different from the theoretical mass of AD2 49

Figure 2.20 **Schematic illustration of process for conjugating N-iodoacetyl peptoids to 5'-thiol ODNs**.....50

Figure 2.21 **RP-HPLC separation and detection of commercially prepared 5' disulfide ODN IC1, (a), TCEP-reduced ODN IC2, (b), and peptoid-ODN conjugate IC3, (c)** 51

Figure 2.22 **MALDI-TOF spectra of commercially prepared 5' disulfide ODN IC1, (a), TCEP-reduced ODN IC2, (b), and peptoid-ODN conjugate IC3, (c)**.....52

Chapter Three

Figure 3.1 **The MAGE methodology.** The method uses a ligation step to create molecules with both peptoid mass tag and biotin moieties in one-to-one proportion with sequences of interest. The peptoids are subsequently cleaved, and quantified using isotopic dilution mass spectrometry. 63

Figure 3.2 **Schematic illustrated of the preparation of conjugate PC3.** The peptoid-ODN conjugate PC3 is synthesized by first reducing the 5' disulfide of PC1 and then conjugating an N-iodoacetyl peptoid to PC2 70

Figure 3.3 **RP-HPLC chromatogram of crude PC3 product.** RP-HPLC separation with detection at 260 nm of the crude PC3 product indicates approximately 50% yield. The conjugate PC3 elutes earlier than the starting material PC1. 71

Figure 3.4 **MALDI-TOF analysis of the two main peaks of the RP-HPLC separation of crude PC3**. For each fraction, two MALDI-TOF peaks were detected because the photocleavable bond is fragmented by the MALDI laser. The masses of the two main peaks agreed with the expected products of (a), Figure 3.4, and (b), Figure 3.5. The smaller peaks in (a) and (b) are identical because only the 5' side of the PC1 material was modified in the reaction.....72

Figure 3.5 **Schematic illustration of how exposure to longwave UV light by a lamp or MALDI laser resulted in the creation of two fragments PC4 and PC5 from the PC3 material** 73

Figure 3.6 **Schematic illustration of how exposure to longwave UV light by a lamp or MALDI laser resulted in the creation of two fragments PC6 and PC5 from the PC1 material** 73

Figure 3.7 **RP-HPLC chromatogram of cleavage process.** The purified PC3 peptoid-ODN conjugate was analyzed by RP-HPLC with detection at 260 nm (a, top) and 450 nm (a, bottom). Exposure to longwave UV light resulted in cleavage of PC3 into PC4 and PC5. The DNA fragment PC5 post cleavage is visible at 260 nm (b, top), while the peptoid fragment PC4 is visible at 450 nm (b, bottom) because of the dabcyl label.....75

Figure 3.8 **MALDI-TOF analyses of cleavage process.** After exposure to longwave UV light, the cleavage products of PC3 were collected and analyzed by MALDI-TOF. The fragment visible in 260 nm (a) had a mass that agreed with PC5, and the fragment visible in 450 nm (b) had a mass that agreed with PC4.....76

Figure 3.9 **Schematic illustrations of ID1 and ID2 syntheses.** Isotopically-shifted "heavy" peptoids such as ID2 are synthesized by incorporating D_3-bromoacetic acid 77

Figure 3.10 **MALDI-TOF spectrum of 1:1 mixture of ID1 and ID2** 78

Figure 3.11 **Measured mole fractions compared to predicted mole fractions.** Peptoid ID1 and 10-fold deuterated peptoid ID2 were mixed in various proportions, and analyzed by MALDI-TOF. The areas under the curves respective to each peptoid were used to estimate the relative mole fractions and compare those estimates to the volumetrically measured mole fractions.....78

Figure 3.12 **Probes M1 and M2.** Probe M1 is a 30mer ODN 5' reversibly conjugated to a peptoid. Probe M2 is a 30-mer ODN that is 5' phosphorylated and 3' biotinylated. When arranged 5' M1 M2 3', they form a 60-mer probe for the mouse gene myogenin.....80

Figure 3.13 **Release of M1P tag.** During MAGE, the successfully ligated M1-M2 probes are captured by Neutravidin resin. The peptoid fragments M1P are freed by exposure to longwave UV light 81

Figure 3.14 **MALDI-TOF analysis of M1P tag.** The MAGE methodology was employed to detect a 60-nucleotide segment of the myogenin gene in four mixtures. Mixture A contained 500 pmols of a synthetic ODN myogenin target, Mixtures B and C contained 1 pmol of myogenin cDNA, and mixture D contained only cDNA from the APETELA gene. MAGE detected the target in mixture A, (a), but did not in mixtures B, C or D (representative spectrum, b).....82

Figure 3.15 **Probes M3 and M4.** Probe M3 is a 30mer ODN 5' reversibly conjugated to a peptoid. Probe M4 is a 30-mer ODN that is 5' phosphorylated and 3' biotinylated. When arranged 5' M3 M4 3', they form a 60mer probe for the mouse gene Myogenin 84

Figure 3.16 **Release of M3P tag.** During MAGE, the successfully ligated M1-M2 probes are captured by Neutravidin resin. The peptoid fragments M1P are freed by exposure to longwave UV light 85

Figure 3.17 **MALDI-TOF Detection of M3P tags.** The MAGE methodology was employed to detect a 60-nucleotide anti-sense ODN of the myogenin gene in six mixtures. Each mixture contains 250 pmols of ODN. Mixture A contains entirely T1, mixture F contains entirely T2, and B-E are intermediate amounts listed in Table 3.3. The correct expected mass of the peptoid fragment MP3 was detected in samples A-E (a-e), but not clearly in F (f) 87

Chapter Four

Figure 4.1 **Two example data sets illustrating effective use of PCA.** PCA is capable of extracting a linear pattern such as in (a), where the variance in the data exists almost entirely along one eigenvector, but it is not capable of extracting a nonlinear pattern such as in (b), where PCA would report a roughly equal amount of variance along any two eigenvectors.....98

Figure 4.2 **Univariate T-distributions at several degrees of freedom.** This illustrates that as the number of degrees of freedom increase, the distributions become more Gaussian, and less permissive to outliers 102

Figure 4.3 **The full data of Chu *et al.* is processed by PCA implemented with SVD.** The resulting eigenvalues are used to show the fraction of the total variance explained as more principal components (eigenvectors) are included.....108

Figure 4.4 **After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted.** In black, six genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4, CAT2, ORF YOR100C, and CAR1), and in red, three random genes (ORFs YAR052C, YAR053W, and YAR060C). Here, all seven principal components are used for the Andrews plot.....110

Figure 4.5 **After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted.** In black, six genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4, CAT2, ORF YOR100C, and CAR1), and in red, three random genes (ORFs YAR052C, YAR053W, and YAR060C). Here, only **three** principal components are used for the Andrews plot.....111

Figure 4.6 **After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted.** In black, three genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4), and in red, three genes designated as belonging to the middle class (YSW1, SPR28, SPS2). Here, all seven principal components are used for the Andrews plot 112

Figure 4.7 **After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted.** In black, three genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4), and in red, three genes designated as belonging to the middle class (YSW1, SPR28, SPS2). Here, the data are plotted for four choices of number of principal components 113

Figure 4.8 **After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted.** In black, three genes designated as belonging to the mid-late class (CDC27, DIT2, DIT1), and in red, three genes designated as belonging to the middle class (ORC3, ORF YLL005C, ORF YLL012W). Here, the data are plotted for four choices of number of principal components 114

Figure 4.9 **Plot of analysis of PCA reduction of data of Sorlie *et al.* implemented with SVD.** The resulting eigenvalues are used to show the fraction of the total variance explained as more principal components (eigenvectors) are included.....115

Figure 4.10 **After PCA processing, the data of Sorlie *et al.* is mapped onto the Andrews space and plotted.** In black, three tumors designated as belonging to the Luminal A class (Norway FU15-BE, Norway FU37-BE, Norway FU16-BE), and in red, three tumors from the ERRB2+ class (Northway FU18-BE, Norway FU04-BE, Norway 65-2ndT). Here, the data are plotted for four choices of number of principal components 117

Figure 4.11 **After PCA processing, the data of Sorlie *et al.* is mapped onto the Andrews space and plotted.** In black, three tumors designated as belonging to the Normal class (Benign STF 37, Benign STF 20, Benign STF 11), and in red, three tumors from the Basal class (Norway FU12-BE, Norway FU23-BE, Norway FU39-BE). Here, the data are plotted for four choices of number of principal components 118

Figure 4.12 **Artificial Data Set 1, which contains one persistent cluster and a small number of outliers**.....119

Figure 4.13 **Artificial Data Set 1, after execution of the EM-MoG (a) or EM-MoL (b) algorithm.** In (a), the probability density function for the Gaussian distribution must be quite large in order to explain the distant points, where as the Lorentzian PDF in (b) located the mean of the persistent cluster and ignores the distant points.....120

Figure 4.14 **Artificial Data Set 2, which contains two tight clusters and a small number of outliers**.....121

Figure 4.15 **Artificial Data Set 2, after execution of the EM-MoG (a) or EM-MoL (b) algorithm.** In (a), the MoG algorithm is confounded by the outliers and reaches a maximum likelihood at two clusters with their means (black circles) between the true clusters. In (b), the MoL algorithm correctly identifies the true clusters 122

Figure 4.16 **Artificial Data Set 2, after execution of the EM-MoG (a) or EM-MoL (b) algorithm.** In (a), the PDFs of the two Gaussian clusters overlap along the centerline of the data, whereas in (b) the PDFs are small and centered at the true clusters.....123

Figure 4.17 **Artificial Data Set 3, which contains three faint clusters and a large amount of dense noise**.....124

Figure 4.18 **Artificial Data Set 3, after execution of the EM-MoG (a) or EM-MoL (b) algorithm, with the most probable outcome shown.** In (a), the MoG algorithm is confounded by the noise and provides three clusters with incorrect means (black circles). In (b), the MoL algorithm correctly identifies the true clusters, although the assignment of points near the borders is heuristically arbitrary 125

Figure 4.19 **Artificial Data Set 3, after execution of the EM-MoG (a) or EM-MoL (b) algorithm.** In (a), the PDFs of the three Gaussian clusters spread across the data in an unintuitive manner, whereas in (b) the PDFs are small and centered at the true clusters 126

Figure 4.20 **Artificial Data Set 3, after execution of the EM-MoG and EM-MoL algorithm over 100 trials.** The data points are the black dots, the means proposed by MoG are the blue circles, and the means proposed by MoL are the red crosses 127

Figure 4.21 **Artificial Data Set 3 was processed by 100 trials of the EM-MoG (a) and EM-MoL (b) algorithms.** The means of the proposed PDFs were binned into a histogram. Because the y-coordinates of the means of two of the true clusters are very similar, the y-coordinate histogram of (b) has a 200-count bin 128

Figure 4.22 **Confusion matrices for results of 477 pre-classified genes from the data set of Chu *et al.* analyzed by EM-MoG and MoL with 2 randomly initialized clusters.** In (a), MoG proposes two relatively diffuse clusters, whereas in (b), MoL proposes two clusters that more precisely partition the true clusters. 131

Figure 4.23 **Plot of the likelihood of fit of full data set of Chu *et al.* analyzed by EM-MoG (a) and EM-MoL (b) with 1 to 30 randomly initialized clusters, with and without the inclusion of the MDL penalty term.** With the penalty term, both clustering methods find 7-9 clusters to be optimal, which agrees well with the 7 classes of Chu *et al*. 133

Figure 4.24 **Histograms of cluster fits for the full data set of Chu *et al.* analyzed by EM-MoG (a) and EM-MoL (b) with 7 randomly initialized clusters.** One of the clusters was broken down into its seven dimensions, and each dimension was binned into a histogram and fit by the appropriate PDF. 135

Chapter Five

Figure 5.1 **Schematic illustrated of MAGE probes in different orientations.** Probes M1 and M2 are in the configuration used in this thesis. By reversing the probes to the configuration shown in M3 and M4, the peptoid fragment of photocleavage would be a single species, instead of several as in this thesis.....149

Figure 5.2 **Schematic illustrated of alternative MAGE peptoid fragments.** By reversing the configuration of the probes, the peptoid fragment of MAGE is changed from PF1, which was present in multiple products, to PF2 149

Appendix C

Figure C.1 **Schematic illustration of fluoro-iodo peptoid FI1.** The iodoacetamide is added near the C-terminus for conjugation to 5'-thiol ODNs..177

Figure C.2 **Schematic illustration of fluoro-iodo peptoid FI2.** The iodoacetamide is added near the C-terminus for conjugation to 5'-thiol ODNs 178

Figure C.3 **MALDI-TOF analyses of FI1 (a) and FI2 (b).** The iodoacetamide is added near the C-terminus for conjugation to 5'-thiol ODNs 179

Figure C.4 **Schematic illustration of ortho-nitro aniline-incorporating peptoid ON1 and ON2** 180

Figure C.5 **ESI-Q analysis of ON1 peptoid.** The spectrum indicates incomplete yield 181

Figure C.6 **ESI-Q analyses of ortho-nitro aniline incorporating peptoid ON2 before (a) and after (b) UV radiation**.....182

Figure C.7 **Schematic illustration of four common MALDI matrices** 183

Figure C.8 **Schematic illustration of peptoids incorporating polar, aromatic side chains**.....184

CHAPTER 1
PREFACE

1.1 Introduction.....2
1.2 Objectives.....9
1.2.1 Collecting Gene Expression Data.....10
1.2.2 Analyzing Gene Expression Data.....11
1.3 References.....14

1.1 Introduction

Gene expression analysis is the examination of the transition of information encoded in nuclear DNA to the collection of proteins that are the ultimate product of that DNA. The central dogma of molecular biology, Figure 1.1, describes the flow of cellular information in general terms, originating in the genome where information is encoded in DNA molecules.

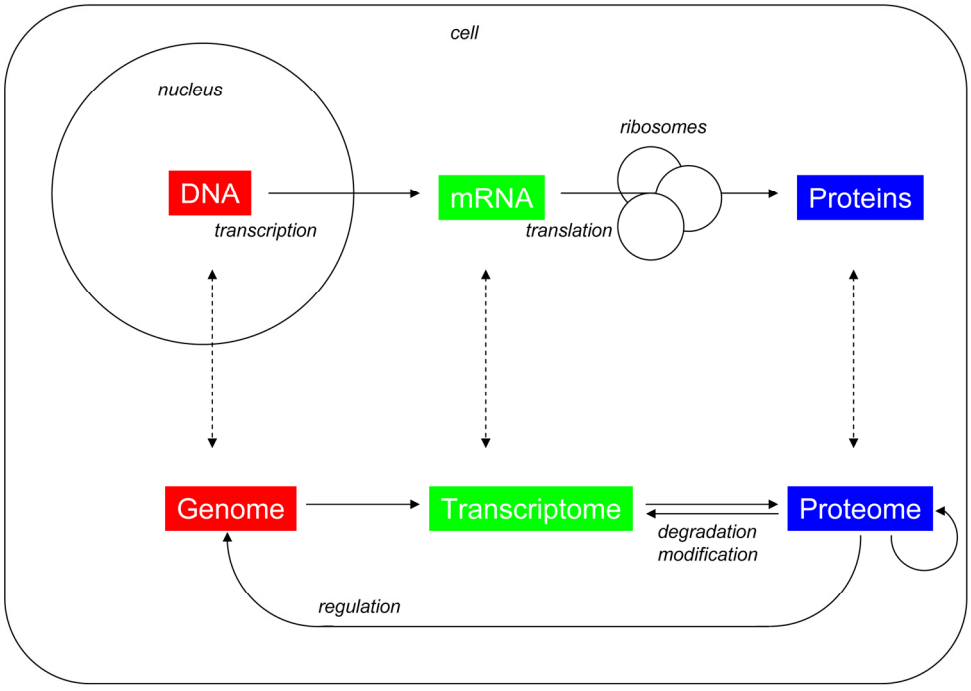


Figure 1.1: **The central dogma of molecular biology.** This schematic representation of the central dogma suggests the importance of studying the transcriptome.

In the nucleus, DNA is transcribed into a complementary manifold of mRNA molecules, known as the transcriptome. These messages are subsequently transported to ribosomes where transcripts are translated into proteins. The transcriptome is not a copy of the genome, because only portions of the genome are transcribed. The selection of which and extent to which genes are transcribed is largely a function of the state of the proteome, which acts in part to regulation of gene expression. The state of the transcriptome is thus a function of the basic information content of the genome, the regulatory action of the proteome, and also the proteome-mediated degradation of transcripts. Finally, the state of the proteome is a function of, among other things, the transcripts that reach the ribosomes. The proteome, which harbors the enzymes that catalyze the reactions inside the cell, is also subject to self-modification and modification from information inputs from outside the cell.

A basic experimental need of functional genomics is the ability to measure the abundance of identifiable sequences of either mRNA or DNA derived from mRNA. Although the relationship between the transcriptome and its product, the more functionally diverse proteome, is not yet fully understood[1-3], it has been repeatedly demonstrated that even in isolation, the transcriptome is an information-rich molecular phenotype.

Recently, transcriptome analysis has been applied to the classification of breast cancers[4], prostate cancers[5], adult acute myeloid leukemias[6], and follicular thyroid tumors[7], where significant clinical factors such as time to distant metastasis and overall survival are correlated to the abundances of a subset of the transcriptome. Another class of studies has sought to infer relationships among genes or proteins by examining

changes in gene expression for a particular organism or tissue across a relevant range of conditions. This has been the aim of studies of yeast sporulation[8], stress response in human cell culture[9], and *C. elegans* development[10]. Another intriguing class of studies seeks to determine the quantitative behavior of subsets of the transcriptome[11-15]. Unlike a general classification or identification analysis, efforts to model gene regulation networks demand highly quantitative data on a gene-by-gene basis. This demand is often met by using large-scale data to initiate a framework for analysis, and then conducting as many lower-level analyses using more precise, but time-consuming, methods as are practical[16].

Sequence-specific nucleic acid detection has been a fundamental technique of molecular biology for decades[17,18]. The current state-of-the-art techniques that are designed to be quantitative can largely be categorized into several classes: PCR-based, sequencing-based, and microarray-based. Because they incorporate exponential amplification, PCR-based methods, such as real-time PCR and competitive PCR, are currently the most sensitive techniques available. The sequencing-based methods, such as serial analysis of gene expression (SAGE) and differential display, are principally advantageous when the genes or sequences of interest are not known before the experiment. Microarrays require advanced knowledge of sequences of interest, but they make practical the simultaneous analysis of thousands of sequences. Of these methods, only SAGE is inherently quantitative, but it cannot be applied to rarer transcripts in a statistically robust manner with today's sequencing technology. Thus, clever methods of normalizing the other technologies have been devised in order to provide quantitative data.

In order to make PCR quantitative, the sequence of interest (SOI) is amplified along with a sequence of known initial abundance. In the case of real-time PCR[19-21], the focus is to eliminate all of the variability in PCR that comes after the exponential phase, by employing fluorescent labels to monitor the PCR reaction kinetics and using an intermediate, exponential-phase abundance for the calculation. In a simple example, the sample to be analyzed is divided into two aliquots, and in one aliquot the SOI is amplified, and in the other aliquot an endogenous comparator sequence is amplified. The comparator sequence is chosen to have a constant abundance across all of the samples to be analyzed. So-called housekeeping genes are generally candidates for comparator sequences.

In competitive PCR[22,23], the comparator sequence is exogenous, and is exactly the same as the SOI with the exception of a one-base mismatch. The SOI and a known amount of comparator are co-amplified by the same primers in the same reaction, and because of their similarity, they are amplified at the same rate as long as their starting concentrations are fairly similar. Here, the difficulty lies in quantifying the relative amounts of two nearly identical sequences. The most sophisticated way to do this yet proposed is the method of Ding and Cantor[23], where a base extension reaction is used to produce two small oligodeoxynucleotides (ODNs) of different masses in known proportion the relative abundances of the two amplified sequences. These ODNs are then quantified by MALDI-TOF mass spectrometry. This technique does not require the expensive fluorescent tagging systems that real-time PCR does, but neither PCR-based technique is suitable for significant multiplexing in a laboratory of typical resources.

Serial analysis of gene expression (SAGE) employs enzymatic techniques to create short tags from a pool of cDNA that are subsequently concatenated, cloned into plasmids, and sequenced[24]. SAGE is inherently quantitative, and thus very well suited for multi-laboratory collaboration. However, due to the fundamental statistics of gene expression, Table 3.1, even if 10^5 tags are sequenced, rare transcripts are not reliably quantified. Differential display, unlike the other methods discussed here, does not require expensive consumables or equipment. Differential display functions by using PCR primers designed to hybridize to a small fraction of the sequences in a typical cDNA sequence, and amplify only those[25,26]. The typically 50-100 products are displayed using gel electrophoresis, and sequences that are differentially displayed between two samples can be isolated and sequenced. Depending on how rigorously the labeling process is handled, this method can determine relative abundances of SOIs, but it cannot approach the reliability of the low-throughput PCR methods.

	Copies per Cell of Each mRNA sequence		Number of Different mRNA Sequences in Each Class		Total Number of mRNA Molecules in Each Class
Abundant class	12,000	X	4	=	48,000
Intermediate class	300	X	500	=	150,000
Scarce class	15	X	11,000	=	165,000

Table 1.1: Distribution of mRNA in a cell. In a typical cell, the majority of genes are expressed as scarce transcripts. A single scarce gene might only make up .001% of the total transcript population[27].

Microarrays are surfaces onto which probe sequences are spatially arranged at high density. Labels are incorporated into the target sequences, and then the targets and probes are contacted and allowed to hybridize. After washing, the microarray is

visualized, and the resulting display in combination with the spatial map of sequences on the surface indicates which sequences were present in the target. A carefully controlled system of fluorescent labeling can confer quantitativeness to microarray methodologies. Three types of microarrays account for the majority of studies: short-ODN, long-ODN, and cDNA. Short-ODN microarrays[28,29] are composed of multiple different probe ODN sequences for each target SOI. Because the length of the probe ODNs is typically no more than 25 nucleotides, single-base mismatch controls and sophisticated statistical techniques are employed to produce aggregate quantitative figures. Long-ODN microarrays rely on probe ODNs that are composed of at least 50 nucleotides, and thus have fewer problems with cross-hybridization compared to short-ODN microarrays. The longer ODNs can be synthesized in situ[30], non-specifically immobilized[31], or covalently attached[32] to the surface. Most microarray studies rely on robotically spotted cDNA microarrays[8,33], where the probes are either full-length cloned cDNAs or large PCR-amplified fragments that are robotically deposited on the surface in a non-specific manner. Typically, the target mRNA pool from one of two samples being compared is labeled during reverse transcription with the dye Cy3, and the other with Cy5[34]. The two target samples are simultaneously hybridized to the same probe array, and the intensity of each spot is measured at wavelengths appropriate for each dye. The relative abundance of the target SOIs is inferred from these intensities. Recently, a method has been developed to combine stringent labeling procedures with printed dye calibration spots in order to produce absolute abundances from microarray data[35].

Microarray experiments do not include an inherent amplification step, however a variety of global amplification schemes have been tested with varying degrees of success.

Two of the most popular methods are linear amplification schemes that offer significantly more reproducibility than global PCR-based schemes: *in vitro* transcription[36] and the aRNA-based method of Eberwine[37-39]. Using these methods, the minimum starting material requirement for microarrays can be lowered to about one μg of total RNA, or the amount found in 10^5 to 10^6 cells, which is still far more than is required for PCR-based methods.

Microarrays are most often employed as screening tools. Successful studies have sought to use the aggregate data for broad classification[40] or to identify genes with behavior worthy of further investigation[41]. In either case, false information will generally not confound the overall result of a meaningful classification or a collection of genes of interest. Improving microarray methodology in order to maximize reproducibility[42-44], and formulating statistical models to extract the maximum amount of relevant information from each experiment[45-47] are major areas of research. This work is hindered by an incomplete understanding of the physics of hybridization between free ODNs and tethered ODNs[48-50], as well as the sources of noise in gene expression analysis[51-54]. Because of this, it is common for studies to verify particular microarray results with low-throughput, high-fidelity methods, most commonly real-time PCR[55]. These subsequent studies are often critical because many significant biological processes are affected by relatively small changes in abundance of relatively scarce transcripts. Another reason that PCR methods are important supplements to microarray studies is that although fluorescent detection systems in principle operate over 4-5 orders of dynamic range, in practice microarrays experience signal compression, signal deterioration, and floor effects when operated beyond 2-3 orders of dynamics range[56,57].

There remains a need for new methods for gene expression analysis, or the more general problem of sequence-specific nucleic acid quantification, especially for addressing the problem of efficiently collecting very reliable, unambiguously quantitative data for 5-50 SOIs, with the sensitivity and dynamic range of PCR-based methods. Such a methodology would be ideally suited for medium-scale modeling of gene regulation networks.

1.2 Objectives

For functional genomics studies, there is a need to first quantify the state of these groups of molecules and then extract from this mass of data functional information about either the correlation between the measured state and some other phenotypic characteristic, or more profoundly, the fundamental interactions, control loops, and kinetics at work. Thus, gene expression analyses comprise two major efforts: the collection of useful information from the transcriptome, and the processing of data in some meaningful way. This thesis introduces a new method for collecting data in Chapter 3, and two new methods for processing data in Chapter 4. One of the key components of the new method of data collection is a class of sequence-specific heteropolymers that contain conjugates of DNA oligonucleotides and peptoids. During the course of investigations, I compiled a collection of synthetic tools for engineering the peptoids (N-substituted glycine oligomers), and I present these tools first in Chapter 2.

1.2.1 Collecting Gene Expression Data

The number of transcripts that exist from a particular gene at any time is a measure of how actively that gene is being transcribed and how quickly the transcript is subsequently being degraded. Although there is not a 1:1 correlation between transcript abundance and protein abundance, measurements of mRNA species are considered to be measures of the extent of gene expression. The transcriptome contains a wide variety of mRNA species. The majority of genes in the genome are expressed in very small quantities, Table 1.1. Furthermore, many of the very tightly regulated genes that are of great interest, are in this scarce category. This distribution must be taken into account in the design of any transcript quantification method.

Transcripts are molecules of mRNA, and the goal of transcript quantification is to create an inventory listing transcripts by what protein they code for, and the abundance of each of these species. Some methods inventory only one transcript at a time, some inventory a large, predefined list, and still others inventory all transcripts, even those that code for genes that have not been identified. Quantification methods also vary in their sensitivity, reproducibility, dynamic range, and ease of use. Many of the methods can only quantify a particular transcript relative to some other transcript. Other methods measure the abundance of transcripts irrespective of a comparator nucleic acid, but no method can simultaneously compare nucleic acid concentrations from more than two samples.

The fundamental property used for the identification and quantification of mRNA molecules is each molecule's sequence. The sequence of an mRNA molecule is

complementary to that of the gene it was transcribed from, so in this manner it can be identified with a particular gene. Also, either in its native form or reverse transcribed into cDNA, the sequence of the molecule can be put to use by probing for it with a labeled complementary nucleic acid, whether it be cDNA (complementary DNA), RNA, or even PNA (peptide nucleic acid). Every assay for mRNA relies on identification by sequence, either through hybridization or through direct sequencing.

In Chapter 3, I present a method for quantifying the absolute abundance of nucleic acid species of a pre-identified sequence using mass spectrometry. I denote this methodology **m**ass-spectrometric **a**nalysis of **g**ene **e**xpression as MAGE. Because nucleic sequences cannot be discriminated by their mass, I introduce a system of hybridizable oligodeoxynucleotide probes conjugated reversibly to peptoid labels serving as mass tags. I engineer the peptoid tags to be optimally suited for mass spectrometric quantification. MAGE is designed to minimize potential sources of error, and rely on controllable physical processes, and be parallelizable up to about 50 sequences of interest. Such a method would be useful for medium-scale studies of groups of related transcripts, especially for quantitative modeling.

1.2.2 Analyzing Gene Expression Data

For the past ten years, investigators have been employing new methods of gene expression data collection to simultaneously measure the abundance of thousands of different species of mRNA. A typical experiment might measure the abundance of transcripts corresponding to 8,000 genes in 20 different biological samples. Data on this

scale are not amenable to conventional statistical tests of significance. Furthermore, many of the hypotheses tested by gene expression experiments are complex or unconventional. At the far extreme of this are experiments designed to *suggest* hypotheses for further testing, when the analytical question becomes "what patterns are present in the data that are worth investigating?" Two major challenges of analyzing highly multivariate gene expression data such as these are first to present, or visualize, the data in an informative manner, and second to robustly identify patterns, especially clusters, in the data. In Chapter 4 of this thesis, I present one solution to each of these problems.

Visualization of data becomes a problem when each element of data, or vector, grows beyond 2 dimensions; the central issue is mapping vectors in multidimensional space onto 2 dimensions. My solution to this problem is to first apply a method of data reduction known as principle components analysis (PCA), and follow that by mapping the data onto an orthogonal set of basis functions known as Andrews curves. This serves to convert each high-dimensional vector to a two dimension wavy line. Two vectors that are close to one another in high-dimension space will become two lines with a similar wave pattern.

Clustering is a task that derives from several basic hypotheses that are frequently tested in large-scale gene expression experiments. One is that genes that have similar expression patterns over a range of samples or conditions are likely to be biologically related. Another is that biological samples that have similar patterns of gene expression are likely share in some other phenotypes. Either hypothesis leads to the analytical task of clustering, which seeks to identify vectors in high dimensional space that are close to

one another. Our solution to this problem is to apply the expectation-maximization algorithm to optimize the parameters of a mixture model. A mixture model is a hypothesis that the data were generated by a linear combination of probability distribution functions. The majority of such models that have been previously studied are based on normal distributions, but I present results that indicate that mixtures of Lorentzian distributions may have significant utility for gene expression analysis.

1.3 References

1. Celis, J. E. et al. Gene expression profiling: monitoring transcription and translation production using DNA microarrays and proteomics. *FEBS Letters* **480**, 2-16 (2000).
2. Hatzimanikatis, V., Choe, L. H. & Lee, K. H. Proteomics: theoretical and experimental considerations. *Biotechnol. Prog.* **15**, 312-318 (1999).
3. Hatzimanikatis, V., Kolstad, J. & Lee, K. H. Mathematical framework for correlating mRNA and protein expression profiles. *BIOT* **#296** (2000).
4. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100**, 8418-8423 (2003).
5. Lapointe, J. et al. Gene expression profiling indentifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA* **101**, 811-816 (2004).
6. Bullinger, L. et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England Journal of Medicine* **350**, 1605-1616 (2004).
7. Barden, C. B. et al. Classification of follicular thyroid tumors by molecular signature: results of gene profiling. *Clinical Cancer Research* **9**, 1792-1800 (2003).
8. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).

9. Murray, J. I. et al. Diverse and specific gene expression responses to stresses in cultured human cells. *Molecular and Cellular Biology* **15**, 2361-2374 (2004).

10. Wang, J. & Kim, S. K. Global analysis of dauer gene expression in *Caenorhabditis elegans*. *Development* **130**, 1621-1634 (2003).

11. Mjolsness, E., Mann, T., Castano, R. & Wold, B. From coexpression to coreuglation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data. *Advances in Neural Information Processing Systems* **12**, 28-934 (2000).

12. Liu, T. et al. Gene expression networks underlying retinoic acid-induced differentiation of acute promyelocytic leukemia cells. *Blood* **96**, 1496-1504 (2000).

13. Ronen, M., Rosenberg, R., Shraiman, B. I. & Alon, U. Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl. Acad. Sci. USA* **99**, 10555-10560 (2002).

14. Goutsias, J. & Kim, S. A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. *Biophysical Journal* **86**, 1922-1945 (2004).

15. Diehn, M. et al. Genomics expression programs and the integration of the CD28 costimulatory signal in T cell activation. *Proc. Natl. Acad. Sci. USA* **99**, 11796-11801 (2002).

16. Ideker, T. & Lauffenburger, D. A. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends in Biotechnology* **21**, 255-262 (2003).

17. Southern, E. Detection of specific sequences among DNA fragments separated by gel-electrophoresis. *Journal of Molecular Biology* **98**, 503 (1975).

18. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA* **96**, 6745-6750 (1977).

19. Ginzinger, D. G. et al. Measurement of DNA copy number at microsatellite loci using quantitative PCR analysis. *Cancer Research* **60**, 5405-5409 (2000).

20. Fink, L. et al. Real-time quantitative RT-PCR after laser-assisted cell picking. *Nature Medicine* **4**, 1329-1333 (1998).

21. Gibson, U. E. M., Heid, C. A. & Williams, P. M. A novel method for real time quantitative RT PCR. *Genome Research* **6**, 995-1001 (1996).

22. Becker-Andre, M. & Hahlbrock, K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Research* **17**, 9437-9446 (1989).

23. Ding, C. & Cantor, C. R. A high-throughput gene epression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proc. Natl. Acad. Sci. USA* **100**, 3059-3064 (2003).

24. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. Serial analysis of gene expression. *Science* **270**, 484-487 (1995).

25. Liang, P. & Pardee, A. B. Differential display of eukaryotic messenger RNA by means of polymerase chain reaction. *Science* **257**, 967-971 (1992).

26. Stein, J. & Liang, P. Differential display technology: a general guide. *Cellular and Molecular Life Sciences* **59**, 1235-1240 (2002).

27. Alberts, B. et al. *Molecular Biology of the Cell* (Garland Publishing, New York, 1994).

28. Chee, M. et al. Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-614 (1996).

29. Lipshutz, R., Fodor, S., Gingeras, T. & Lockhart, D. High density synthetic oligonucleotide arrays. *Nature Genetics* **21**, 20-24 (1999).

30. Hughes, T. R. et al. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology* **19**, 342-342 (2001).

31. Call, D. R., Chandler, D. P. & Brockman, F. Fabrication of DNA microarrays using unmodified oligonucleotide probes. *BioTechniques* **30**, 368-379 (2001).

32. Boncheva, M., Scheibler, L., Lincoln, P., Vogel, H. & Akerman, B. Design of oligonucleotide arrays at interfaces. *Langmuir* **15**, 4317-4320 (1999).

33. Schena, M., Shalon, D., Davis, R. w. & Brown, P. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467-470 (1995).

34. Duggan, D., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. Expression profiling using cDNA microarrays. *Nature Genetics* **21**, 10-14 (1999).

35. Rouse, R. J. D., Espinoza, C. R., Niedner, R. H. & Hardiman, G. Development of a microarray assay that measures hybridization stoichiometry in moles. *BioTechniques* **36**, 464-470 (2004).

36. Lockhart, D. & Winzeler, E. A. Genomics, gene expression, and DNA arrays. *Nature* **405**, 827-836 (2000).

37. Wang, E., Miller, L. D., Ohnmacht, G. A., Liu, E. T. & Marincola, F. M. High-fidelity mRNA amplification for gene profiling. *Nature Biotechnology* **18**, 457-459 (2000).

38. Eberwine, J. Amplification of mRNA populations using aRNA generated from immobilized oligo(dT)-T7 primed cDNA. *BioTechniques* **20**, 584- (1996).

39. Mazzanti, C. et al. Using gene expression profiling to differentiate benign versus malignant thyroid tumors. *Cancer Research* **64**, 2898-2903 (2004).

40. Perou, C. M. et al. Molecular portraits of human breast tumors. *Nature* **406**, 747-752 (2000).

41. Rajeevan, M., Vernon, S. D., Taysavang, N. & Unger, E. R. Validation of Array-Based Gene Expression Profiles by Real-Time (Kinetic) RT-PCR. *Journal of Molecular Diagnostics* **3**, 26-31 (2001).

42. Churchill, G. A. Fundamentals of experimental designs for cDNA microarrays. *Nature Genetics Supplement* **32**, 490-495 (2002).

43. Tseng, G. C., Oh, M.-K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29**, 2549-2557 (2001).

44. Quackenbush, J. Microarray data normalization and transformation. *Nature Genetics Supplement* **32**, 496-501 (2002).

45. Kerr, M. K. & Churchill, G. A. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* **98**, 8961-8965 (2001).

46. Chuaqui, R. F. et al. Post-analysis follow-up and validation of microarray experiments. *Nature Genetics Supplement* **32**, 509-514 (2002).

47. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31-36 (2001).

48. Shchepinov, M., Case-Green, S. C. & Southern, E. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic Acids Research* **25**, 1155-1161 (1997).

49. Southern, E., Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genetics* **21**, 5-9 (1999).

50. Ketomaki, K., Hakala, H., Kuronen, O. & Lonnberg, H. Hybridization properties of support-bound oligonucleotides: the effect of the site immobilization on the stability and selectivity of duplex formation. *Bioconjugate Chemistry* **10.1021/bc0340058** (2003).

51. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415-418 (2004).

52. Blake, W. J., Kaern, M., Cantor, C. R. & Collins, J. J. Noise in eukaryotic gene expression. *Nature* **422**, 633-637 (2003).

53. McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *TIG* **15**, 65-69 (1999).

54. Brown, J. S., Kuhn, D., Wisser, R., Power, E. & Schnell, R. Quantification of sources of variation and accuracy of sequence discrimination in a replicated microarray experiment. *BioTechniques* **36**, 324-332 (2004).

55. Ramakrishnan, R. et al. An assessment of Motorola CodeLink microarray performance for gene expression profiling applications. *Nucleic Acids Research* **30**, e30 (2002).

56. Wilson, C. L., Pepper, S. D., Hey, Y. & Miller, C. J. Amplification protocols introduce systematic but reproducible errors into gene expression studies. *BioTechniques* **36**, 498-506 (2004).

57. Etienne, W., Meyer, M. H., Peppers, J. & Meyer Jr., R. A. Comparison of mRNA gene expression by RT-PCR and DNA microarray. *BioTechniques* **36**, 618-626 (2004).

CHAPTER 2
EXPANDING THE TOOLKIT FOR PEPTOID SYNTHESIS

2.1 Introduction....22
2.2 Experimental....26
2.2.1 General Peptoid Synthesis....26
2.2.2 Specialized Syntheses....28
2.2.3 Analytical Procedures....31
2.3 Results and Discussion....32
2.3.1 Branched Peptoids....32
2.3.2 Peptoid Capping....35
2.3.3 N-terminal Modifications....37
2.3.4 Macrocyclic Peptoids....41
2.3.5 Oligo-Adamantane Peptoids....46
2.3.6 Oligodeoxynucleotide-Peptoid Conjugates....49
2.4 Summary....53
2.5 References....54

2.1 Introduction

Peptoids, Figure 2.1, are N-substituted glycine oligomers, and have become a significant class of peptidomimetics[1].

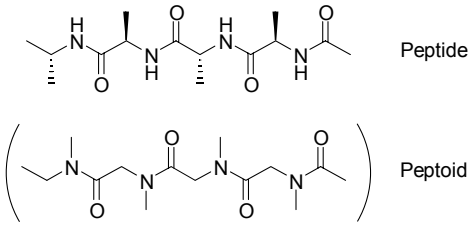


Figure 2.1: Schematic illustrations of peptoids and peptides. Compared to peptides such as oligo-alanine, upper, peptoids, such as oligo-sarcosine, lower, lack stereochemistry and sites for hydrogen bonding, but retain a similar spacing and overall structure.

Originally proposed by Zuckermann *et al.*[2] as a strategy for synthesizing diverse libraries of lead compounds for drug discovery, peptoids have since been designed to form stable secondary structures[3], including a family of α-chiral-side-chain substituted peptoids that form stable helices[4-7]. Several studies have shown that peptoids[8] or peptoid-peptide hybrids[9,10] can be designed to act as protein ligands in the nanomolar to micromolar affinity range. Kodadek *et al*. are currently developing methodologies for the synthesis of large peptoid libraries and subsequent screening and isolation of peptoid ligands from those libraries[11].

A number of studies have demonstrated the use of peptoids in other biomimetic and biotechnological roles. Peptoid nucleic acids extend the biomimetic role of peptoids beyond that of peptide mimicry[12,13]. Peptoids have been designed to serve as cell penetrators[14,15] and gene delivery vehicles[16]. Several studies have demonstrated peptoids with antimicrobial properties[17,18], including helical mimics of magainin-2 amide[19]. Helical peptoids have also been designed to mimic lung surfactant protein C[20], and a series of trialkylglycine peptoids have been shown to have analgesic effect by blocking VR1 channels[21]. Peptoids were also applied as uncharged, water-solubilizing caps for use in membrane-interactive peptides[22].

One of the most attractive features of peptoids is the ease with which diverse, relative pure oligomers can be synthesized. Although several methods for peptoid synthesis have been proposed[2,23-25], the solid-phase, submonomer method of Zuckermann *et al.*[26,27] is the most widely replicated. The central advantage of the submonomer method (Fig. 2.2) is that the growing chain of the peptoid is extended by repeated applications of a single linking chemistry, and that the submonomer providing diversity is a primary amine (Fig. 2.3). Perhaps the greatest disadvantage of the submonomer method, the requirement of on average 3 hours of reaction time per monomer added to the growing chain, has recently been eliminated by Olivos *et al*.[28], who demonstrated microwave-assisted synthesis of peptoids of approximately 1 minute of reaction time per monomer added.

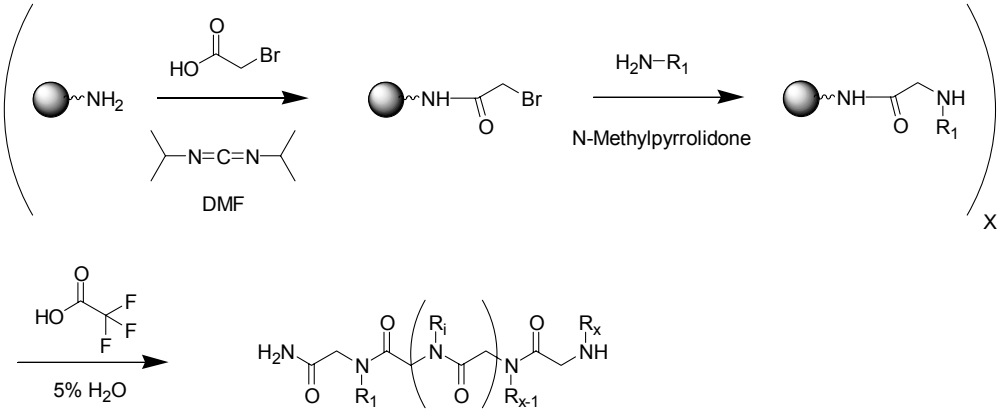


Figure 2.2: The solid-phase submonomer method of peptoid synthesis. This method is executed by repeated, alternating rounds of bromoacetylation and primary amine substitution. Commonly, the finished chain is cleaved by trifluoroacetic acid to form a C-terminal amide cap.

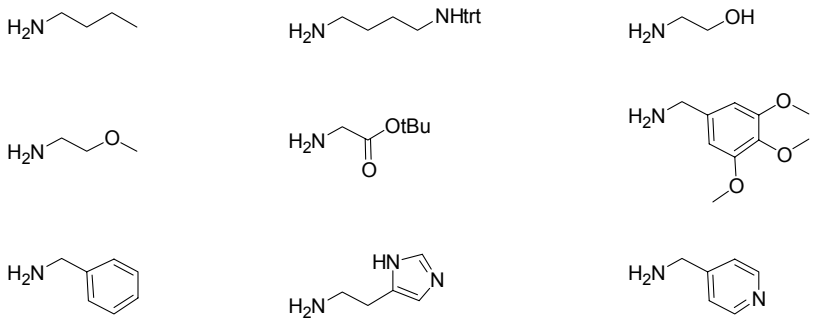


Figure 2.3: Primary amines used for peptoid synthesis. Peptoid diversity is generated through the choice of primary amines (R-NH_2 in Figure 4.2). Some commonly used classes include aromatic, aliphatic, heterocyclic, cationic, anionic, bulky, small, hydrophobic and hydrophilic.

The submonomer synthesis has been extended to allow for a variety of primary amine submonomers[29], including unprotected heterocycles[30]. Depending on the linker chosen, peptoids can be produced with either C-terminal acids[31] or amides[32]. A variety of chemoselective functionalities[33] can be incorporated into peptoids either along the backbone or at the N-terminus, such as aminooxyacetamides, N-(carbamoylmethyl)acetohydrazides, mercaptoacetamides, 2-

pyridinesulfenylmercaptoacetamides, maleimides[34], and aldehydes. Analysis of peptoids is commonly accomplished by RP-HPLC, mass spectrometry[35], and capillary electrophoresis[36,37]. Once synthesized, peptoids can be sequenced by Edman degredation[38].

Here, I seek to further expand the toolkit for peptoid synthesis while staying within the submonomer synthetic methodology. First, I demonstrate a simple method for introducing 2:1 branch into the growing peptoid chain. This would allow the rational synthesis of branched or multiply branched structures, and it could also be incorporated into a parallel synthesis to generate libraries of branched structures. Second, I demonstrate a simple method for capping the growing peptoid chain. Capping is a standard step in the protected-monomer synthesis of DNA, and caps have found application in peptide synthesis as purification tags[39-43]. If a particular primary amine submonomer had a low rate of substitution during peptoid synthesis, that step could be immediately following by the addition of a high-substituting cap. By doing this, that chain would never grow any longer and would not result in a potentially confounding single-deletion sequence. Third, I demonstrate several useful N-terminal modifications, including two types of haloacetamides, two types of carboxylic acids, and two types of primary amines. Fourth, I demonstrate a straightforward technique for peptoid macrocyclization that is chemically orthogonal with many potential peptoid sequences. Because many macrocyclic peptides have potent biological activity[44-47], a general method for peptoid macrocyclization would expand the potential applications of peptoids. Fifth, I demonstrate the synthesis of two different peptoids incorporating four adamantine moieties. Adamantyl species have not yet been reported as peptoid submonomers

because when attached proximal to the growing chain, it can severely hinder chain extension. In my syntheses, adamantane moieties are attached via two different spacers. Finally, I demonstrate a new method for conjugating peptoids and DNA oligodeoxynucleotides (ODNs). Peptoid-ODN conjugates have been applied to demonstrate a drag-tag methodology for electrophoresis[34], but they have not yet been applied to biological studies as many peptide-ODN conjugates have.

2.2 Experimental

2.2.1 General Peptoid Synthesis

Peptoids were synthesized manually using the method of Figliozzi *et al.*[27] The synthesis is described here using 100 mg of resin, but up to 250 mg have been successfully used in the same size synthesis vessel by scaling all other reagents linearly. 100 mg of rink amide MBHA resin (Novabiochem, La Jolla, CA) was loaded into a 10 mL peptide synthesis vessel that had been modified to improve agitation by adding a small pocket on the wall of the reaction chamber. The resin was first washed several times with N,N-dimethylformamide (DMF, Aldrich Chemical Co., Milwaukee, WI). All solvents were purchased anhydrous and kept as dry as possible by careful handling and storage with molecular sieves (3A, EM Industries, Gibbstown, NJ). The resin was agitated by an upward directed flow of argon. A wash step refers to adding 1-2 mL of solvent, agitating for 30 seconds, then draining the vessel under aspirator vacuum.

After the initial washing to swell the resin, the DMF was drained and the Fmoc groups protecting the resin free amines were removed by adding 2 mL of 20% piperidine in DMF, agitating for one minute, draining, and adding another 2 mL of 20% piperidine. The second solution was agitated for 15 minutes and then drained. The resin was washed with DMF six times before peptoid synthesis.

From here, repeated rounds of a single linking chemistry were used (Table 2.1 and Fig. 2.2). First, the free amines were acetylated by adding 850 μL of 0.6 M bromoacetic acid (BAA, Aldrich Chemical Co., Milwaukee, WI) and 200 μL of 3.2 M diisopropylcarbodiimide (DIC, Aldrich Chemical Co., Milwaukee, WI). The slurry was agitated for 30 minutes, drained, and an identical solution was added for a further 30 minutes of agitation. Following this, the mixture was drained, washed twice with DMF, and once with N-methylpyrrolidone (NMP, Aldrich Chemical Co., Milwaukee, WI).

The nucleophilic substitution of a primary amine is the second half of a round of synthesis. The primary amine of choice was dissolved at around 1.5 M in NMP and 1mL of this solution was added to the vessel. The mixture was agitated for two hours, drained, and the resin is subsequently washed twice with NMP and once with DMF. Amines in acid salt form were first neutralized by the addition of 95% the stoichiometric amount of KOH (Aldrich Chemical Co., Milwaukee, WI). The organic layer was separated and used for the synthesis.

Once the peptoid was completed (generally by adding the final primary amine) the resin was washed thoroughly with DMF (four to six washes) and then dichloromethane (four to six washes), and allowed to dry for about an hour in the reaction vessel. Following this, the peptoid was cleaved from the resin by placing the resin into a

glass vial containing 5-10 mL of 95% trifluoroacetic acid in water (TFA, Aldrich Chemical Co., Milwaukee, WI). This mixture was stirred for 30 minutes, filtered, and washed with further TFA and water. The resulting solution was diluted with water and dried using a rotary evaporator or with a stream of dry nitrogen, subsequently frozen and lyophilized. The dried material was resuspended in one of water, dimethylsulfoxide, or a mixture of water and acetonitrile and placed into a tared cryovial for final lyophilization and storage.

Step		Reagent	Reaction Time	Volume (μL)	Repetitions
1	BAA Addition	0.6 M bromoacetic acid in DMF	-	850	-
2	Activation	3.2 M diisopropylcarbodiimide in DMF	-	200	-
3	Acetylation		30 min		2
4	Wash	DMF NMP	30 s 30 s	2000 2000	2 1
5	Displacement	1.5 M primary amine in NMP	2 h	1000	1
6	Wash	NMP DMF	30 s 30 s	2000 2000	2 1

Table 2.1: Peptoid synthesis single linking chemistry scheme of Figliozzi *et al.*, for 100 mg of resin. These steps are repeated for each monomer addition.

2.2.2 Specialized Syntheses

I introduced branches by incorporating 1,4-diaminobutane (Aldrich Chemical Co., Milwaukee, WI) during the primary amine substitution step. Two secondary amine capping agents were tested, piperidine (Aldrich Chemical Co., Milwaukee, WI) and *N,N*-

diethylamine (Aldrich Chemical Co., Milwaukee, WI). These are also introduced during the normal primary amine substitution step.

The N-terminal modifiers were incorporated by making the final step of peptoid synthesis an acetylation, and the modifiers included bromoacetic acid (Fluka, via Aldrich Chemical Co., Milwaukee, WI), iodoacetic acid (Fluka, via Aldrich Chemical Co., Milwaukee, WI), diglycolic acid (Aldrich Chemical Co., Milwaukee, WI), succinic acid (Aldrich Chemical Co., Milwaukee, WI), FMOC-γ-aminobutyric acid (Novabiochem, La Jolla, CA), and FMOC-alanine (Novabiochem, La Jolla, CA).

Macrocyclization was accomplished by creating a free amine near the C-terminus of the peptoid using mono-trityl 1,4-diaminobutane acetic acid salt (Novabiochem, La Jolla, CA). The peptoid N-terminus was modified by diglycolic acid, and the post-cleavage macrocyclization is effected by 3 equivalents of PyBOP (Novabiochem, La Jolla, CA), 10 equivalents of DIPEA (Aldrich Chemical Co., Milwaukee, WI) in 97:3 DCM:DMF (both solvents from Aldrich Chemical Co., Milwaukee, WI) at room temperature for 12 hours.

Adamantane-containing peptoids were synthesized in one of two ways. First, by incorporating mono-trityl 1,4-diaminobutane, deprotecting it with 95:4:1 water:triisopropylsilane:trifluoracetic acid (TIS, Aldrich Chemical Co., Milwaukee, WI), and coupling 1-adamantanecarboxylic acid (Aldrich Chemical Co., Milwaukee, WI) with 4 equivalents of DIC. Second, by coupling 1-adamantanecarboxylic acid with an excess of 2,2′-(ethylenedioxy)diethylamine (Aldrich Chemical Co., Milwaukee, WI) using 1-Ethyl-3-(3-dimethyllaminopropyl)carbodiimide hydrochloride (EDC, Novabiochem, La

Jolla, CA) to form **1** (Fig. 2.4). This adamantane-spacer-amine was then substituted into a peptoid synthesis in the normal manner.

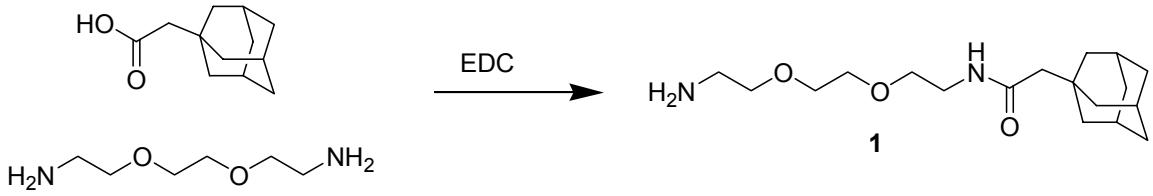


Figure 2.4: Schematic of how adamantane is incorporated into peptoids by first synthesizing (1).

Peptoid-DNA oligodeoxynucleotide (ODN) conjugates were synthesized first preparing N-iodoacetyl peptoids, as described in this work. ODNs with 5′ C6 disulfide modifiers were prepared commercially (formerly, Beckman Institute Biopolymer Synthesis Facility, Pasadena CA, presently, Qiagen, Valencia, CA). These were resuspended at 50 μM concentration in pH 7.2 100 mM NH_4HCO_3 (Aldrich Chemical Co., Milwaukee, WI) in DNase-free water (Gibco, via Invitrogen, Carlsbad, CA). To this, a 20-fold excess of tris-carboxyethylphosphine (TCEP, Pierce Biotechnology, Rockford, IL) was added, and immediately followed by a 40-fold excess of N-iodoacetyl peptoid. The mixture was placed under argon and gently mixed for 72 hours. The product was purified using reverse phase HPLC over a Zorbax 300Extend-C18 column in an Agilent 1100 system. The peaks were eluted with a linear gradient of 1-70% B in A over 50 minutes at 0.3 mL/min (solvent A=100 mM TEAA (Fluka, via Aldrich Chemical Co., Milwaukee, WI) in 100% water, solvent B=100 mM TEAA in 90% acetonitrile, 10% water).

2.2.3 Analytical Procedures

Analysis of peptoids was accomplished by reverse phase HPLC over a Zorbax 300Extend-C18 column in an Agilent 1100 system. The peaks were eluted with a linear gradient of 0-75% B in A over 50 minutes at 0.3 mL/min (solvent A=0.1% TFA in 100% water, solvent B=0.1% TFA in 100% acetonitrile). The column was held at 30 °C, and detection was accomplished by means of a diode array detector at 220 nm.

Analysis of ODNs and peptoid-ODN conjugates was accomplished by reverse phase HPLC over a Zorbax 300Extend-C18 column in an Agilent 1100 system. The peaks were eluted with a linear gradient of 1-70% B in A over 50 minutes at 0.3 mL/min (solvent A=100 mM TEAA in 100% water, solvent B=100mM TEAA in 90% acetonitrile, 10% water). The column was held at 30 °C, and detection was accomplished by means of a diode array detector at 220 nm and 260 nm.

Mass spectrometry of crude samples and HPLC fractions was accomplished by matrix-assisted laser desorption spectrometry with time-of-flight analysis (MALDI-TOF) on an Applied Biosystems Voyager-DE PRO BioSpectrometry Workstation firing a 337 nm nitrogen laser. Peptoids were generally analyzed with a matrix of α-cyano-4-hydroxycinnaminic acid (Aldrich Chemical Co., Milwaukee, WI), formulated at 10 mg/mL with 0.1% TFA in 50:50 water:acetonitrile. ODNs and peptoid-ODN conjugates were generally analyzed with a matrix of 3-hydroxypicolinic acid (Aldrich Chemical Co., Milwaukee, WI), formulated at 5 mg/mL with 0.05 mg/mL dibasic ammonium carbonate (Aldrich Chemical Co., Milwaukee, WI) in 90:10 water:acetonitrile.

2.3 Results and Discussion

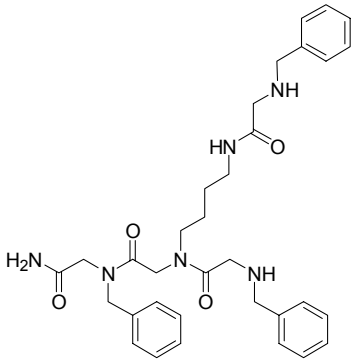
2.3.1 Branched Peptoids

The branched structures BR1, BR2, and BR3 (Fig. 2.5) were synthesized using the submonomer method with the primary amine sequence shown in Table 2.2.

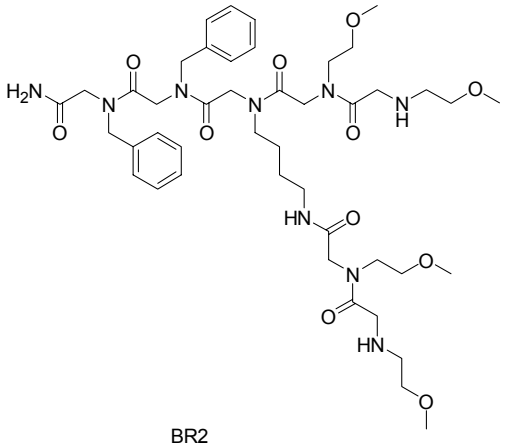
BR1	BR2	BR3
Benzylamine	Benzylamine	Benzylamine
Diaminobutane	Benzylamine	Benzylamine
Benzylamine	**Diaminobutane**	**Diaminobutane**
	Methoxyethylamine	Methoxyethylamine
	Methoxyethylamine	Propylamine
		Diaminobutane
		Benzylamine
		Methoxyethylamine

Table 2.2: Primary amine submonomers used for syntheses of branched peptoids BR1, BR2, and BR3. Branches are introduced by diaminobutane incorporation.

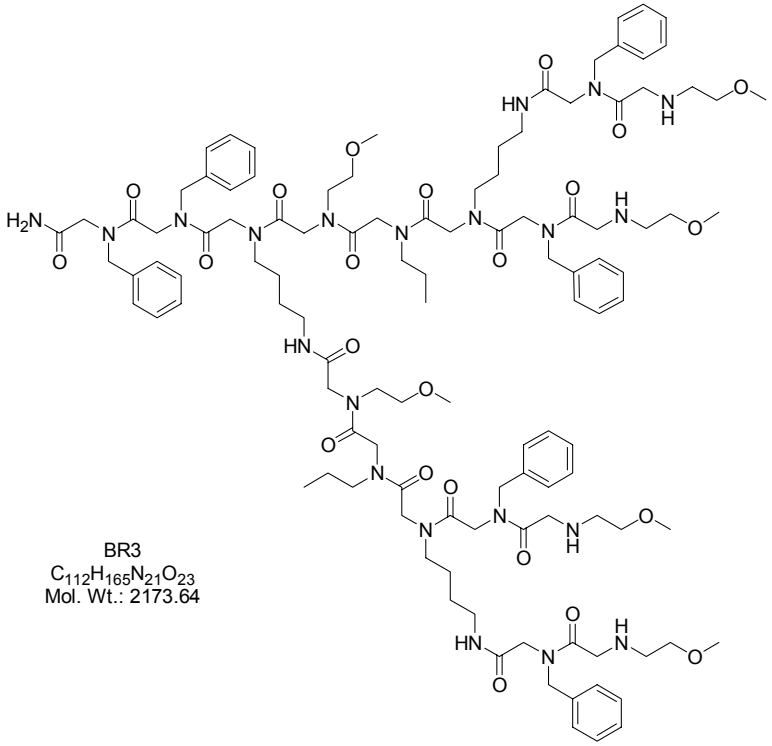
Following the incorporation of diaminobutane, the next bromoacetylation step creates a peptide bond that may not be desirable depending on the application. Successful synthesis of these example sequences is indicated by MALDI-TOF, (Fig. 2.6). RP-HPLC analysis indicated an average of 90% yield, as exemplified by the data shown in Figure 2.7.



BR1
$C_{33}H_{42}N_6O_4$
Mol. Wt.: 586.72



BR2
$C_{44}H_{69}N_9O_{11}$
Mol. Wt.: 900.07



BR3
$C_{112}H_{165}N_{21}O_{23}$
Mol. Wt.: 2173.64

Figure 2.5: Schematic illustrations of three branched structures of increasing size.
Synthesis methods are detailed in Table 2.2.

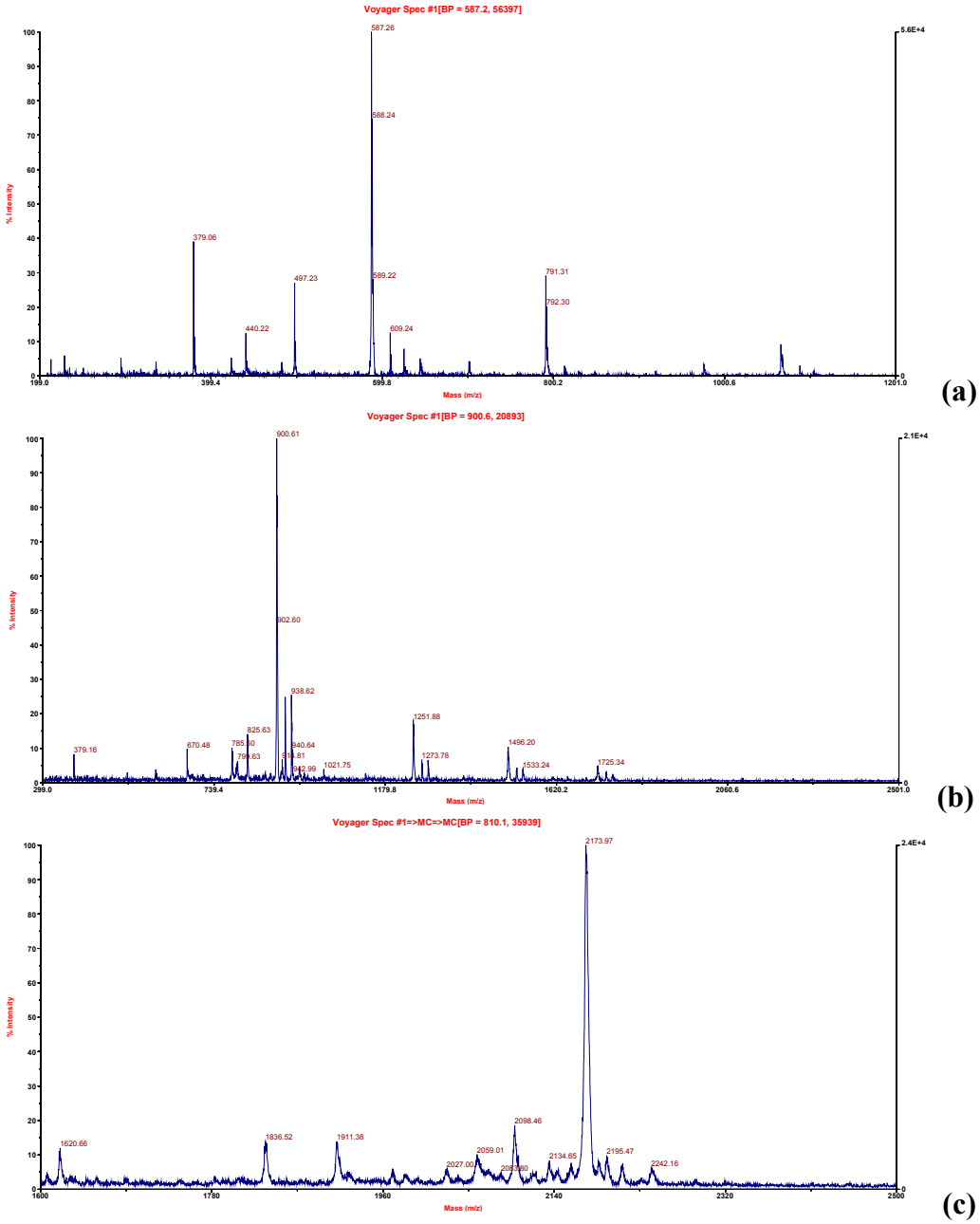


Figure 2.6: MALDI-TOF spectra of three branched structures of increasing size, BR1 (a), BR2 (b), and BR3 (c). The masses of the three peptoids are visible as H^+ adducts.

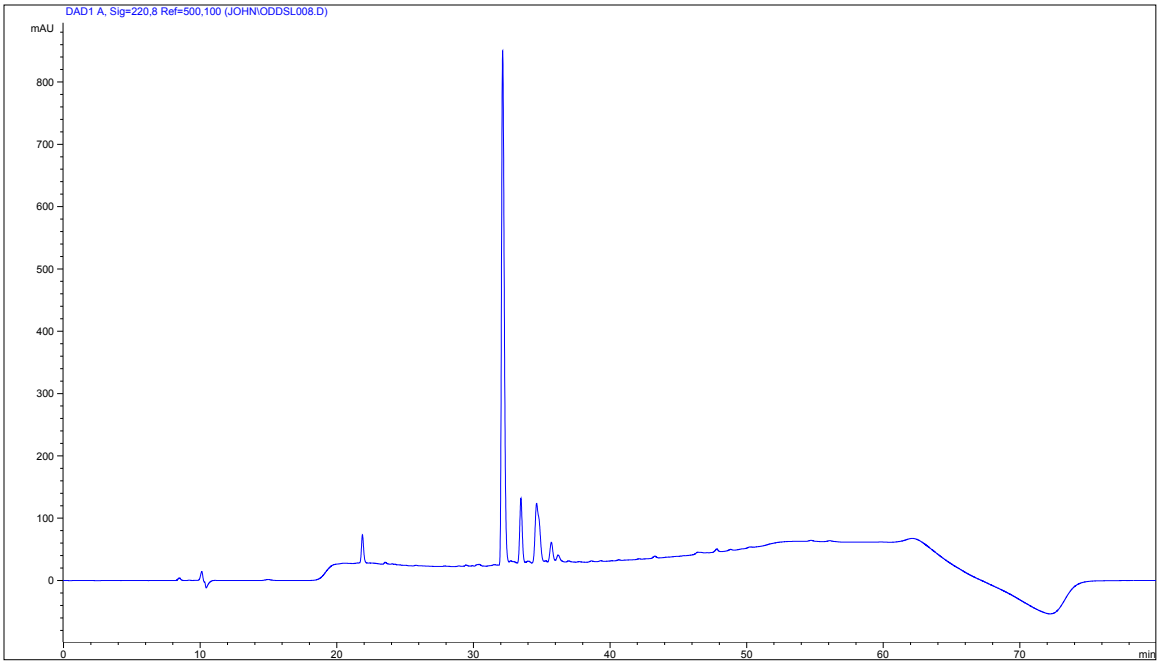


Figure 2.7: RP-HPLC separation and detection at 220 nm of as-made branched peptoid BR2. Analysis indicated approximately 90% yield of the desired product.

2.3.2 Peptoid Capping

Peptoid capping was demonstrated by synthesizing two peptoids (Fig. 2.8) using the submonomer method with the primary amine sequence shown in Table 2.3.

CP1	CP2
Methoxyethylamine	Methoxyethylamine
Methoxyethylamine	Methoxyethylamine
Methoxyethylamine	Methoxyethylamine
N,N-Diethylamine	Piperidine
Methoxyethylamine	Methoxyethylamine
Methoxyethylamine	Methoxyethylamine
Methoxyethylamine	Methoxyethylamine

Table 2.3: Primary amine submonomers used for syntheses of capped peptoids CP1 and CP2.

Three complete rounds of synthesis were executed after the addition of the capping secondary amine. In the case of CP1, MALDI-TOF (Fig. 2.9) indicates that the

N,N-diethylamine largely blocks extension of the peptoid chain, where as in CP2, piperidine fails to block extension of the peptoid chain.

CP1
$C_{21}H_{41}N_5O_7$
Mol. Wt.: 475.58

CP2
$C_{22}H_{41}N_5O_7$
Mol. Wt.: 487.59

Figure 2.8: Schematic illustrations of two capped peptoids, CP1 and CP2.

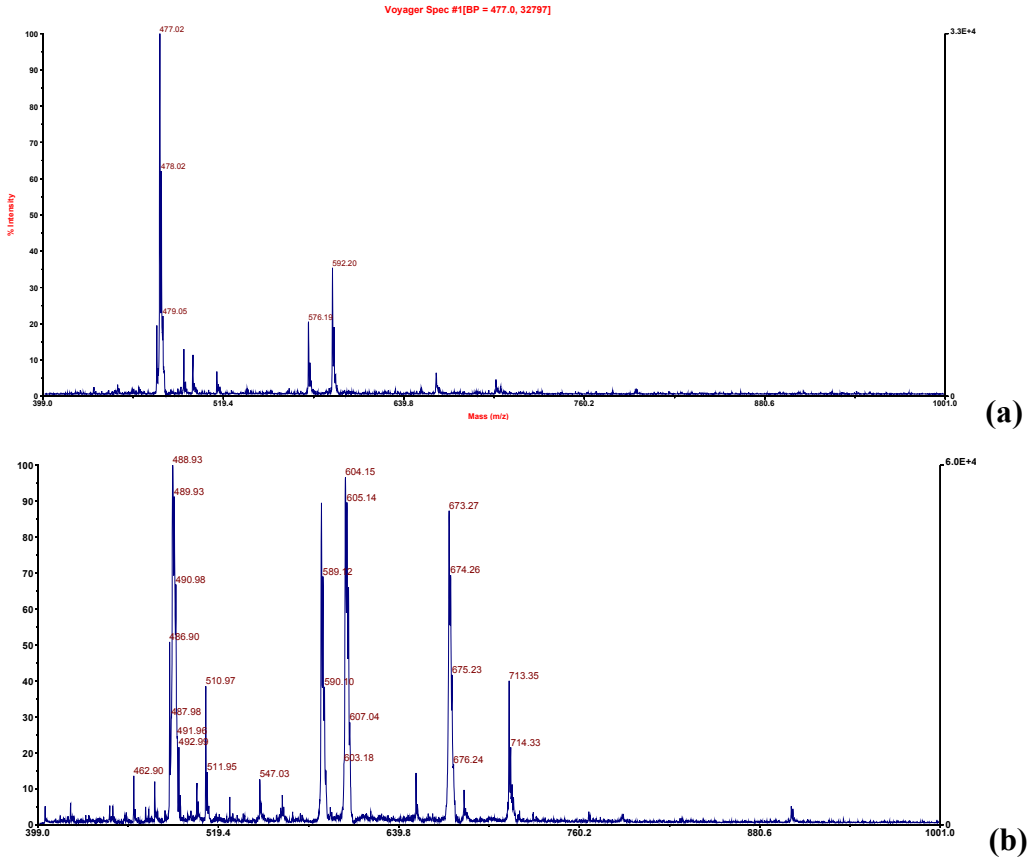


Figure 2.9: MALDI-TOF spectra of product of two attempts to cap the peptoid growing chain. In (a), CP1 largely blocks chain extension, where as in (b), CP2 fails to block chain extension. The masses of the peptoids are visible as H^+ adducts.

2.3.3 N-terminal Modifications

Normally, the final addition to the peptoid growing chain in the submonomer synthesis is a primary amine, leaving a secondary amine N-terminus. In these six examples, Figure 4.10, the final step was an acetylation using one of the acids in Table 2.4. The calculated masses of these peptoids are evident in MALDI-TOF spectra (Fig. 2.11). The yields of these peptoids varied from at least 85% up to greater than 97% as measured by RP-HPLC, with a typical analysis shown in Figure 2.12.

Structure Code	Final acetylation step submonomer
NT1	Bromoacetic acid
NT2	Iodoacetic acid
NT3	Diglycolic acid
NT4	Succinic acid
NT5	Alanine
NT6	γ-Aminobutyric acid

Table 2.4: Acetylating submonomers used to modify the N-terminus of peptoids NT1-NT6.

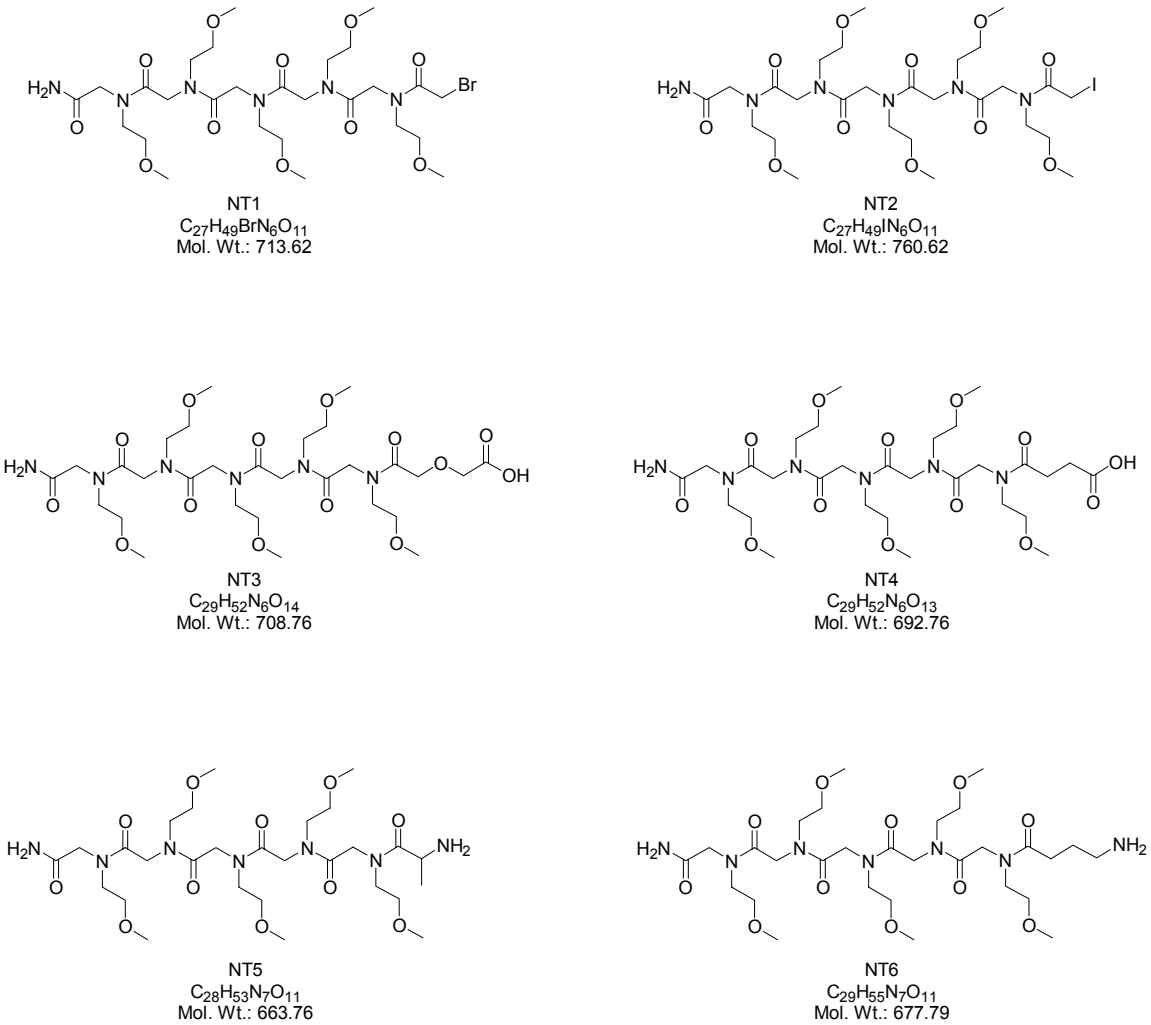
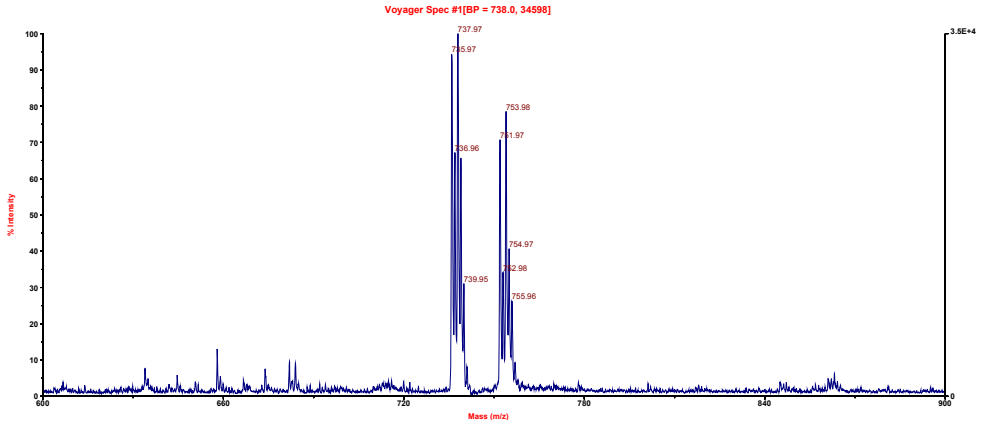
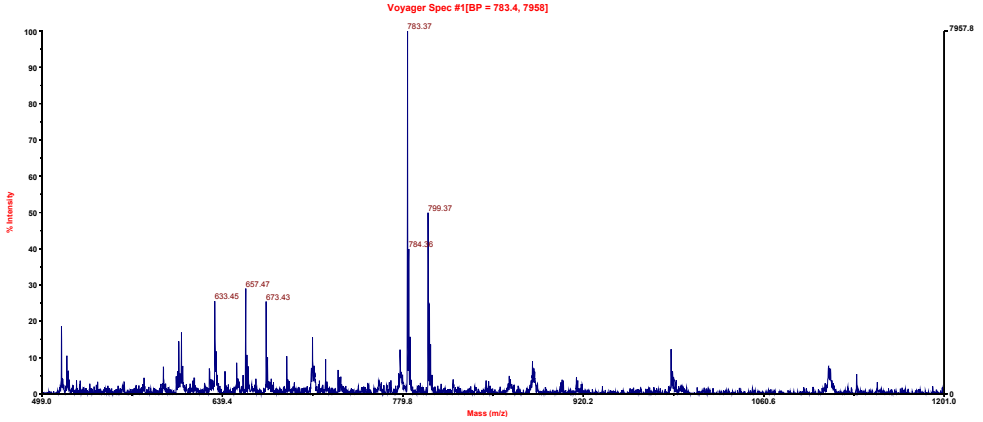


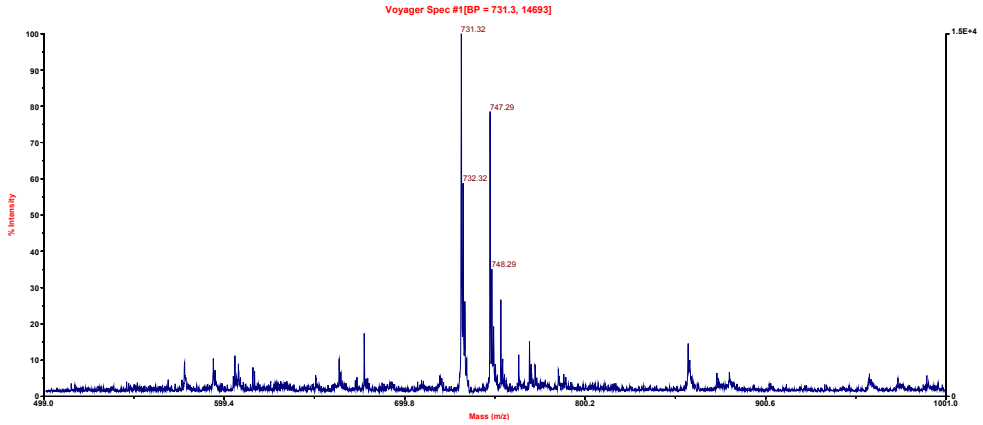
Figure 2.10: Schematic illustration of six peptoids with alternative N-termini. These were formed by terminating the growing peptoid chain with an acetylation step instead of the usual substitution step.


Voyager Spec #1[BP = 738.0, 34598]
737.97
735.97
753.98
751.97
736.96
754.97
752.98
739.95
755.96
% Intensity
Mass (m/z)
3.5E+4

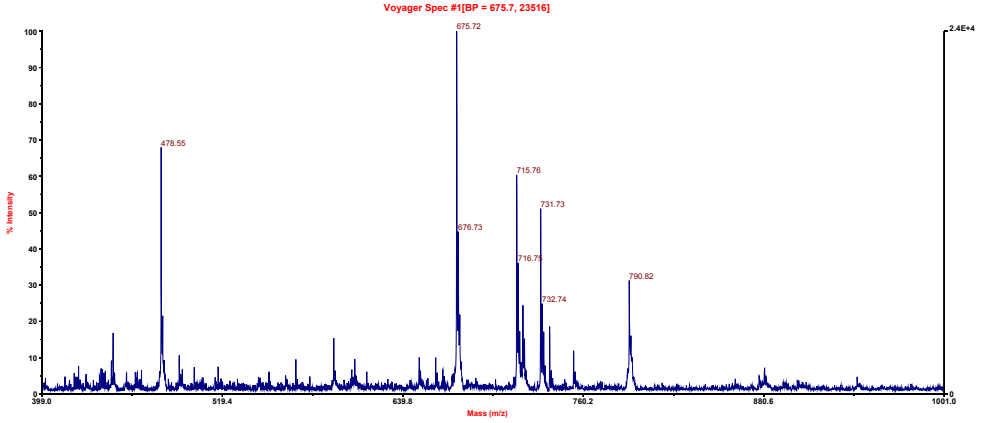
(a)


Voyager Spec #1[BP = 783.4, 7958]
783.37
799.37
784.36
657.47
633.45
673.43
% Intensity
Mass (m/z)
7957.8

(b)


Voyager Spec #1[BP = 731.3, 14693]
731.32
747.29
732.32
748.29
% Intensity
Mass (m/z)
1.5E+4

(c)


Voyager Spec #1[BP = 675.7, 23516]
675.72
478.55
715.76
731.73
676.73
716.75
790.82
732.74
% Intensity
Mass (m/z)
2.4E+4

(d)

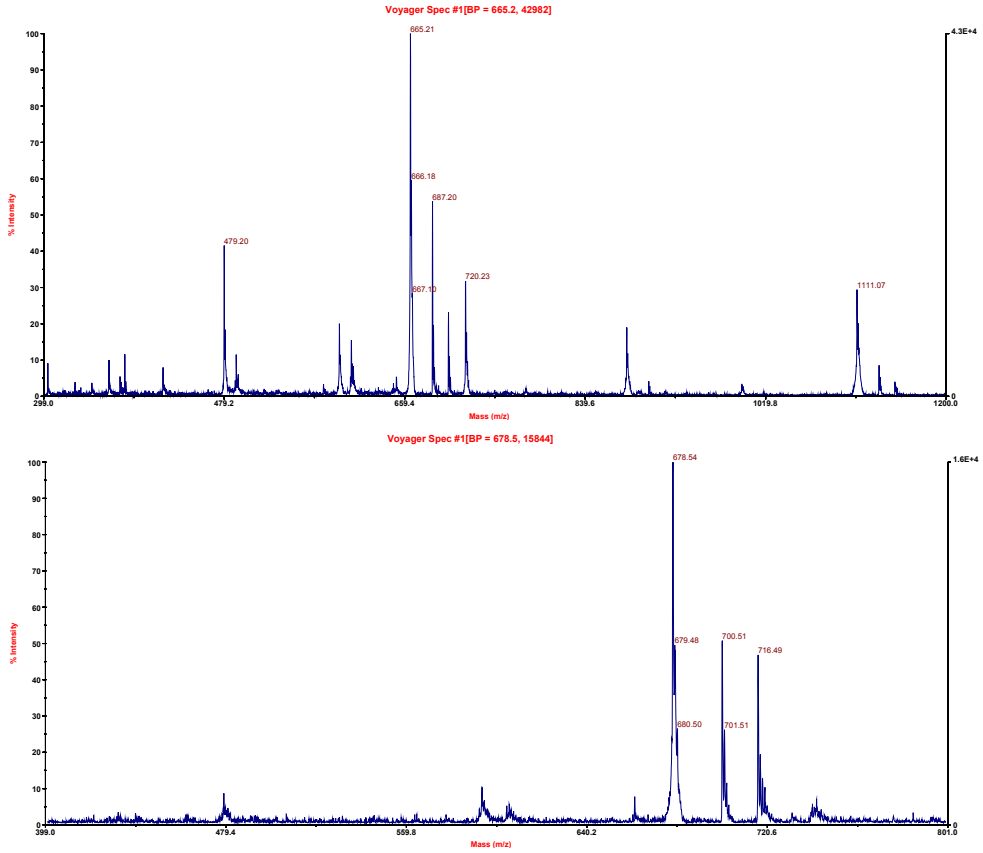


Figure 2.11: MALDI-TOF spectra of six N-terminal-modified peptoids, NT1-NT6. For NT1, (a), two groups of peaks represent the Na^+ and K^+ adducts, and incorporation of bromine accounts for the peak splitting within each of the two groups. For NT2, (b), NT3, (c), NT4, (d), and NT5 (e), the Na^+ and K^+ adducts are evident. For NT6, (f), the H^+ adduct is visible as well as the salt adducts.

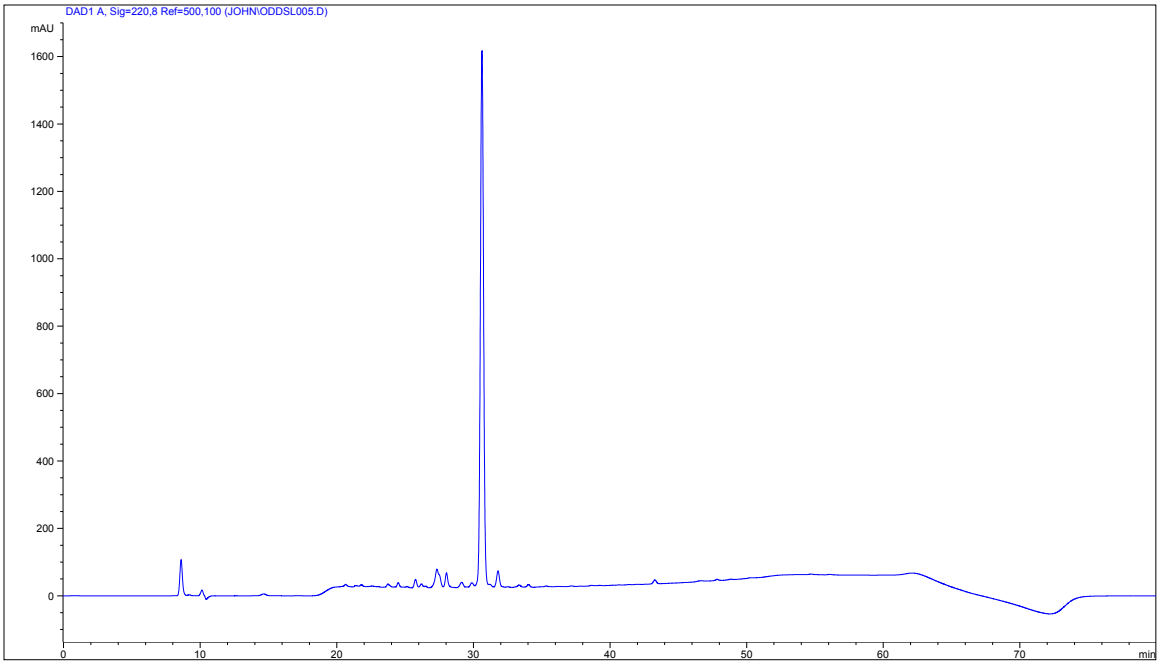


Figure 2.12: **RP-HPLC separation with detection at 220 nm of N-iodoacetyl peptoid NT2 indicating over 95% yield**.

2.3.4 Macrocyclic Peptoid

Peptoid macrocyclization was demonstrated by first synthesizing a difunctionalized free-acid free-amine peptoid CY1 (Fig. 2.13) using the submonomer method with the primary amine sequence shown in Table 2.5. The peptoid was terminated with diglycolic acid. Following this, the peptoid was cyclized (Fig. 2.14), to form CY2. The loss of water during the cyclization is evident in the MALDI-TOF, Figure 2.15. The two species are difficult to separate using RP-HPLC, with retention times differing by less than 30 seconds in our methods. The pre- and post-cyclization RP-HPLC analyses are shown in Figure 2.16.

CY1
Methoxyethylamine
Methoxyethylamine
Mono-trityl 1,4-diaminobutane
Methoxyethylamine
Methoxyethylamine
Methoxyethylamine
Methoxyethylamine
Methoxyethylamine
Methoxyethylamine
Methoxyethylamine
Methoxyethylamine

Table 2.5: Primary amine submonomers used for syntheses of macrocyclic precursor peptoid CY1.

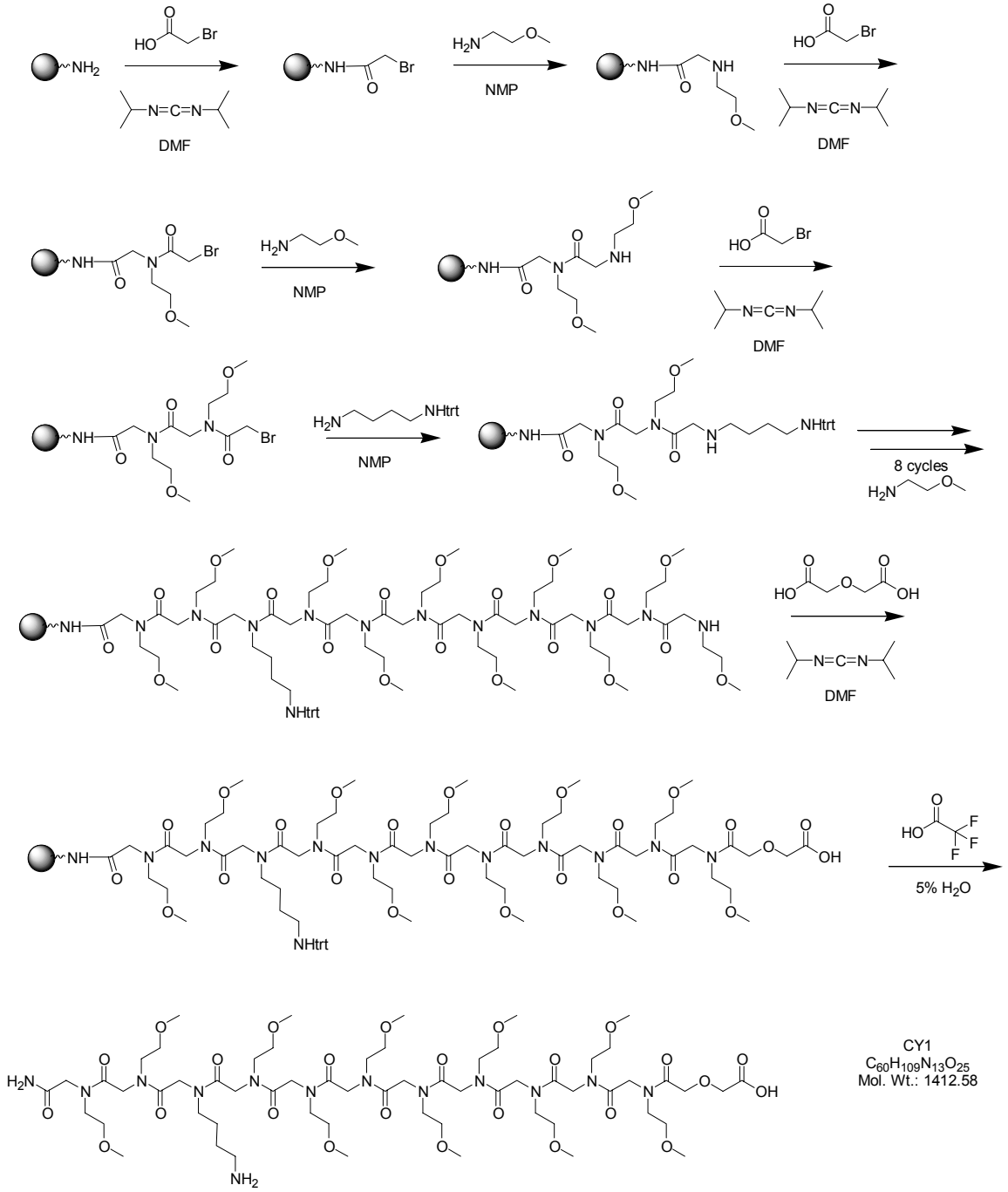


Figure 2.13: Schematic illustrations of the synthesis of CY1 free-acid free-amine peptoid for cyclization.

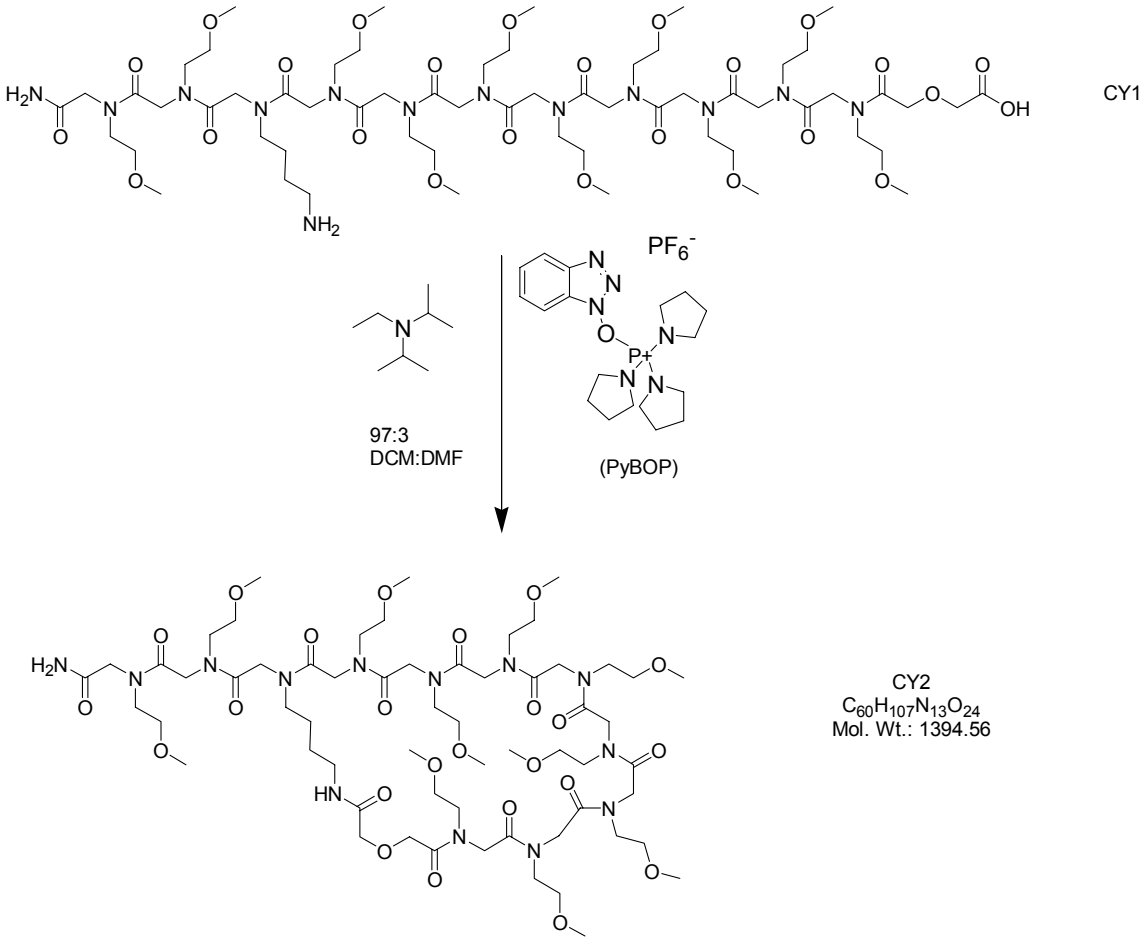


Figure 2.14: Schematic illustration of cyclization of CY1 into CY2 using peptide coupling reagents PyBOP and DIPEA.

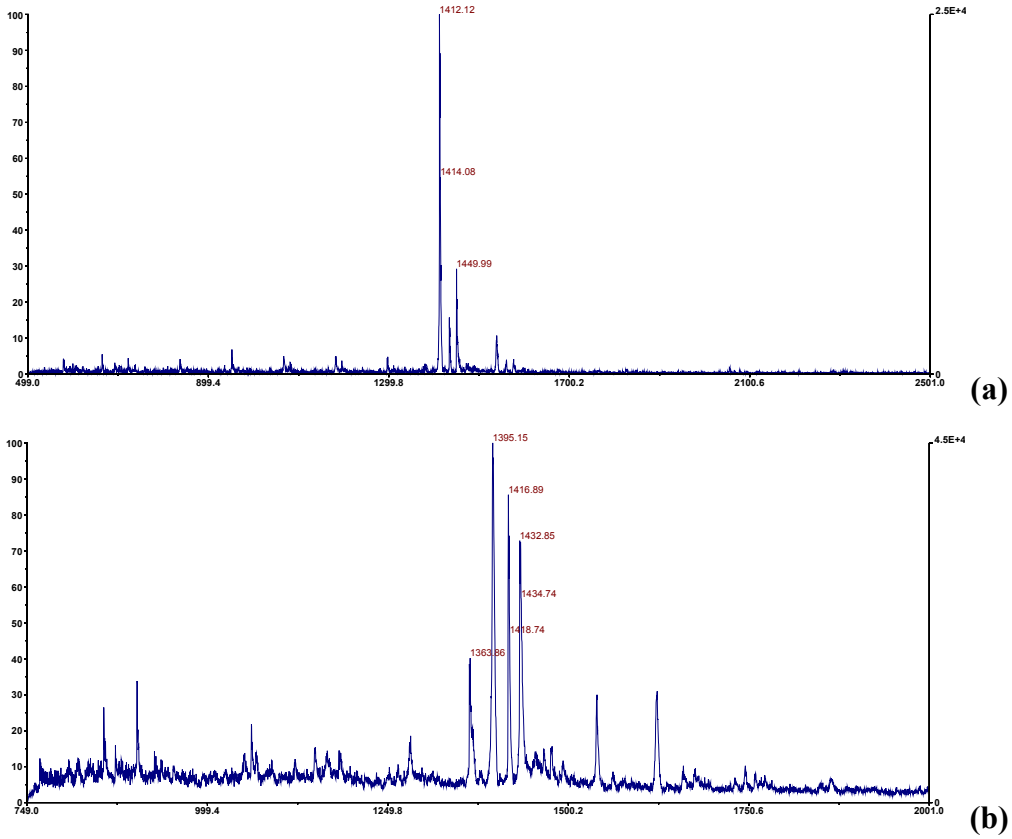


Figure 2.15: MALDI-TOF spectra of free-acid free-amine peptoid, CY1, (a), and the cyclized product, CY2, (b). In (a), the H^+ adduct is evident, while in (b), the H^+, Na^+ and K^+ adducts are evident.

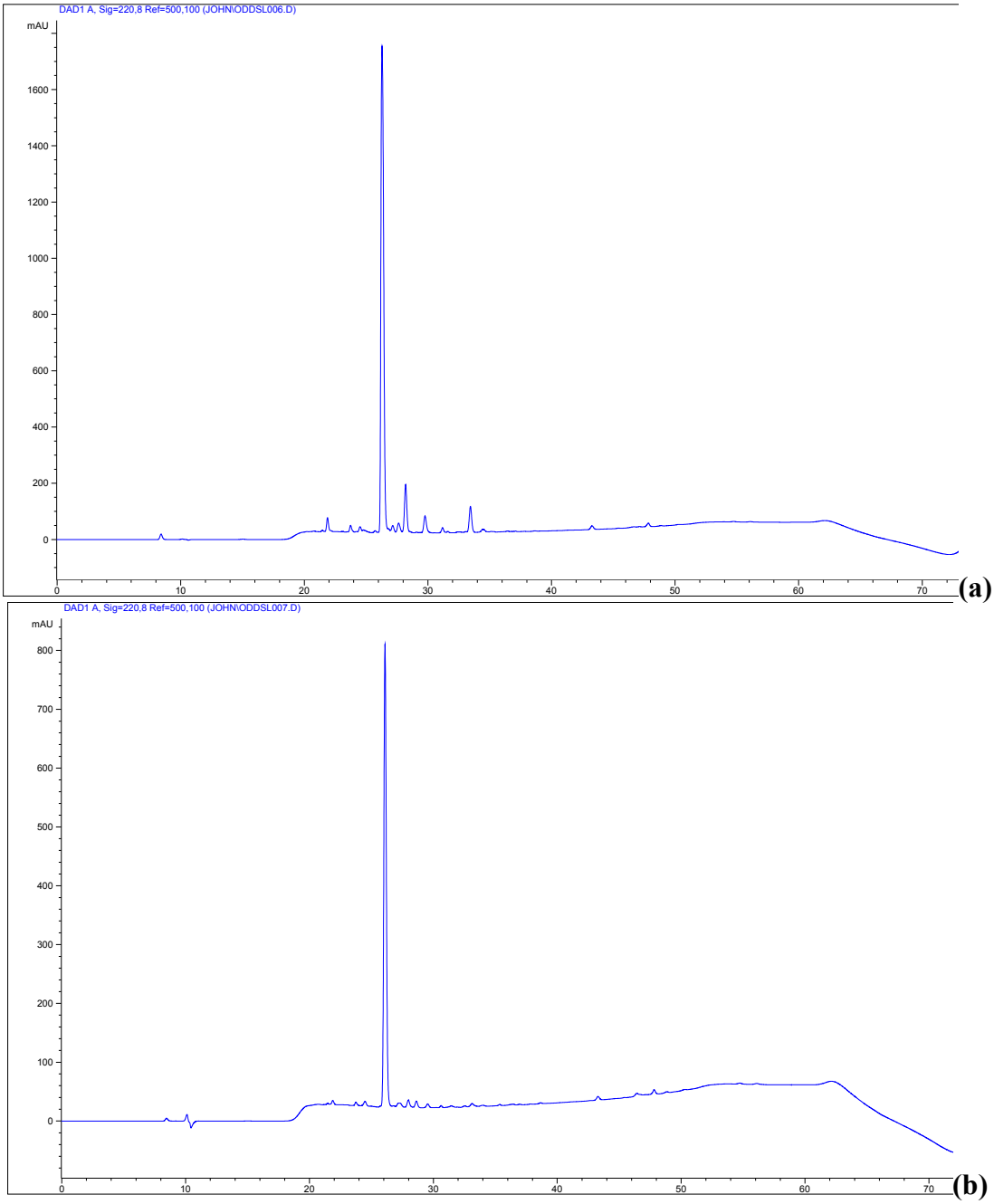


Figure 2.16: RP-HPLC separation with detection at 220 nm of pre- and post-cyclization peptoids CY1 (a) and CY2 (b).

2.3.5 Oligo-Adamantane Peptoids

Two tetra-adamantyl peptoids (Fig. 2.17) were synthesized using the submonomer method with the primary amine sequence shown in Table 2.6. AD2 required no

additional modifications after chain synthesis using **1**, Figure 2.4, but AD1 required a deprotection step following by a peptide coupling step, Figure 2.18. The calculated mass of AD2 is evident in MALDI-TOF spectra, Figure 2.19, but it may be that the synthesis of AD1 resulted in an unexpected outcome.

AD1 (precursor)	AD2
1	Mono-trityl 1,4-diaminobutane
Propylamine	Propylamine
1	Mono-trityl 1,4-diaminobutane
Propylamine	Propylamine
Propylamine	Mono-trityl 1,4-diaminobutane
1	Propylamine
Propylamine	Mono-trityl 1,4-diaminobutane
1	

Table 2.6: Primary amine submonomers used for syntheses of adamantyl-peptoids AD1 and AD2.

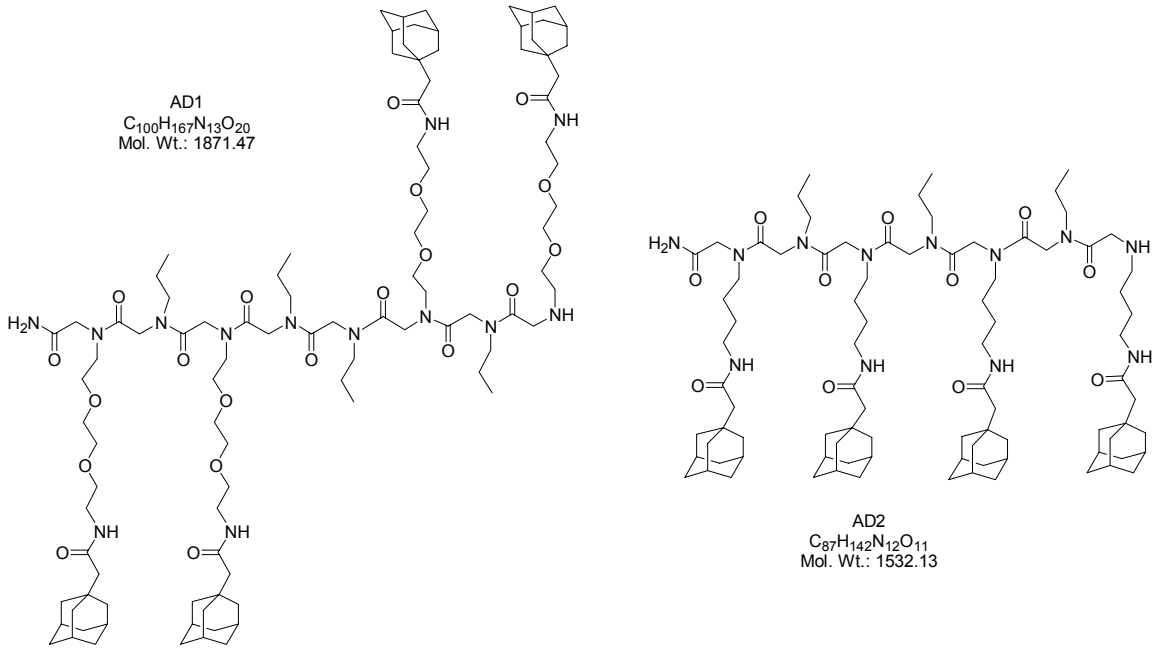


Figure 2.17: Schematic illustrations of AD1 and AD2 tetra-adamantyl peptoids.

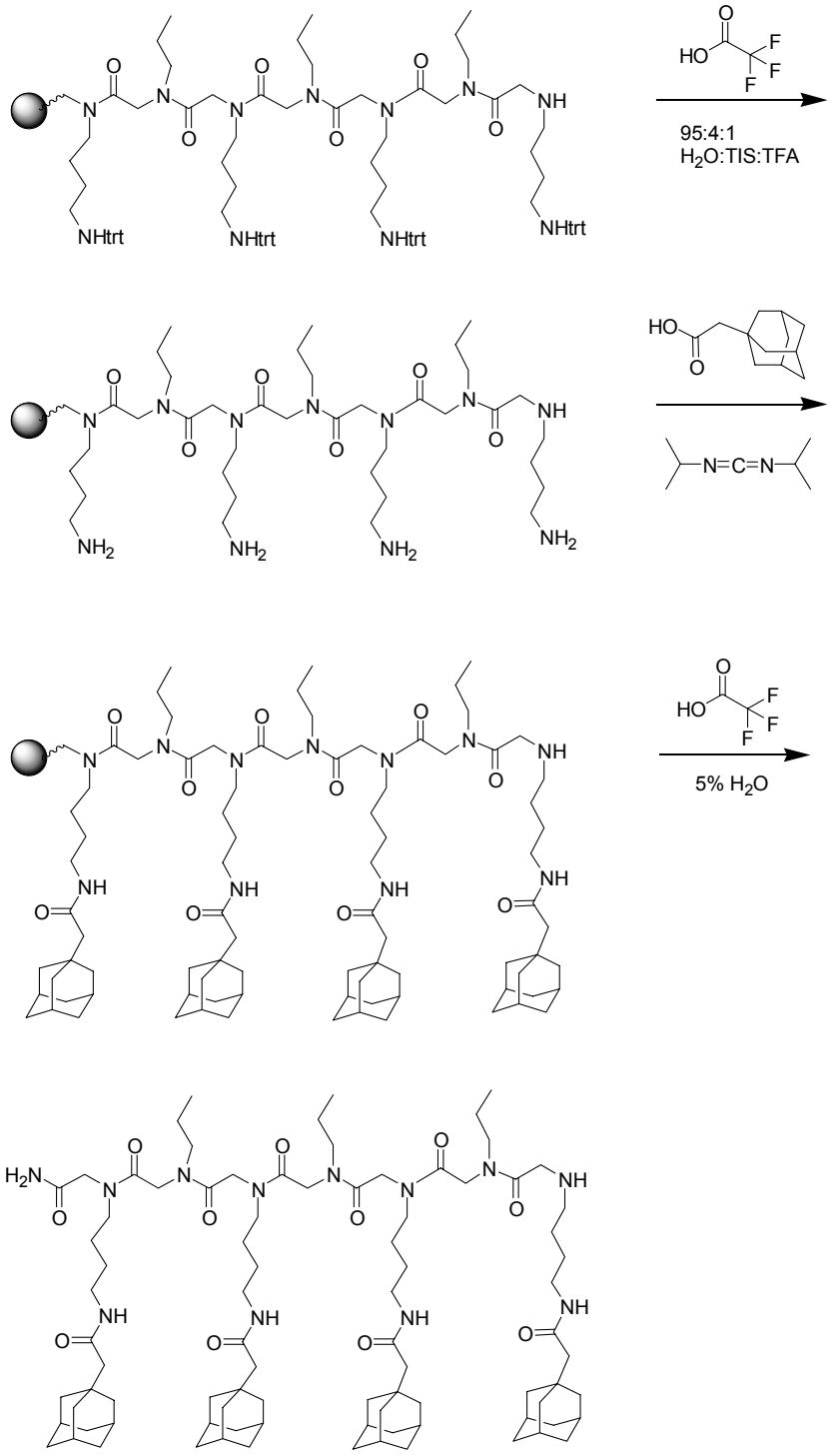


Figure 2.18: Schematic illustration of the process for deprotecting pendant amines and conjugating adamantane carboxylic acid to AD2.

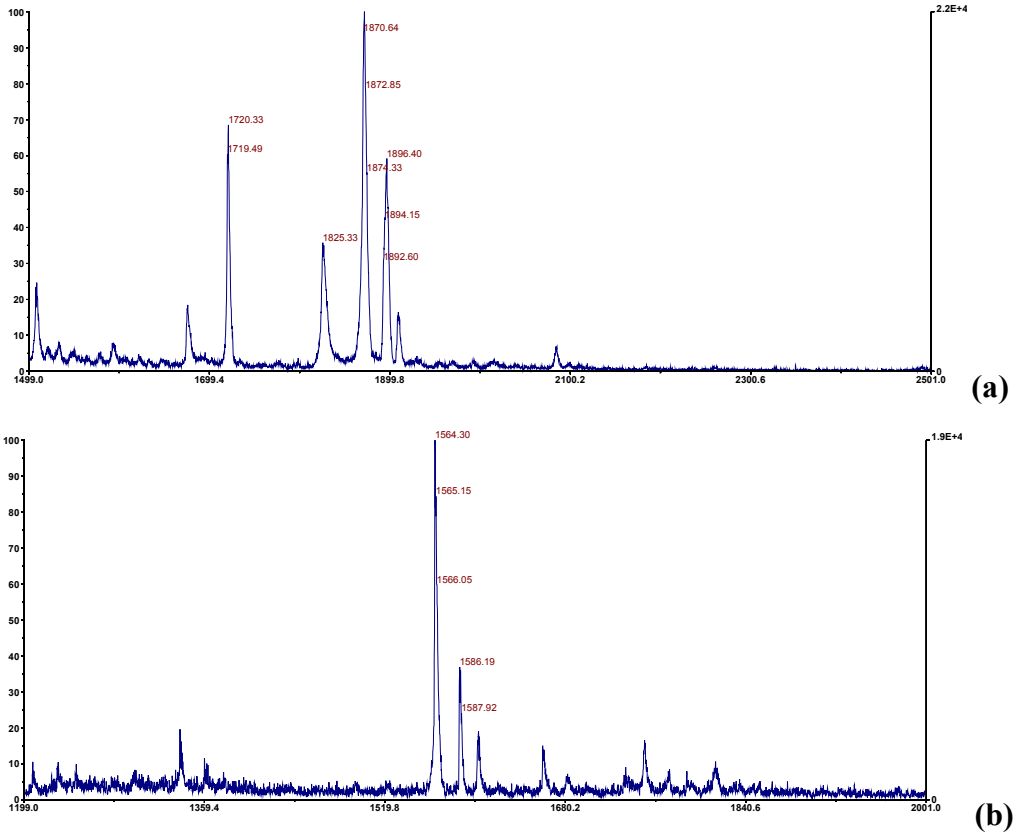


Figure 2.19: MALDI-TOF spectra of tetra-adamantyl peptoids. AD1, (a), is evidence by the H^+ adduct. In (b), the spectrum indicates H^+ and Na^+ adducts that are somewhat different from the theoretical mass of AD2.

2.3.6 Oligodeoxynucleotide-Peptoid Conjugates

The 5′-disulfide ODN IC1 was combined with an excess of TCEP, immediately followed by an excess of N-iodoacetyl peptoid in ammonium bicarbonate buffer, (Fig. 2.20. The three steps of the reaction are tracked with RP-HPLC (Fig. 2.21) and the predicted masses are confirmed by MALDI-TOF (Fig. 2.22). The reaction generally proceeds to greater than 98% yield by ODN.

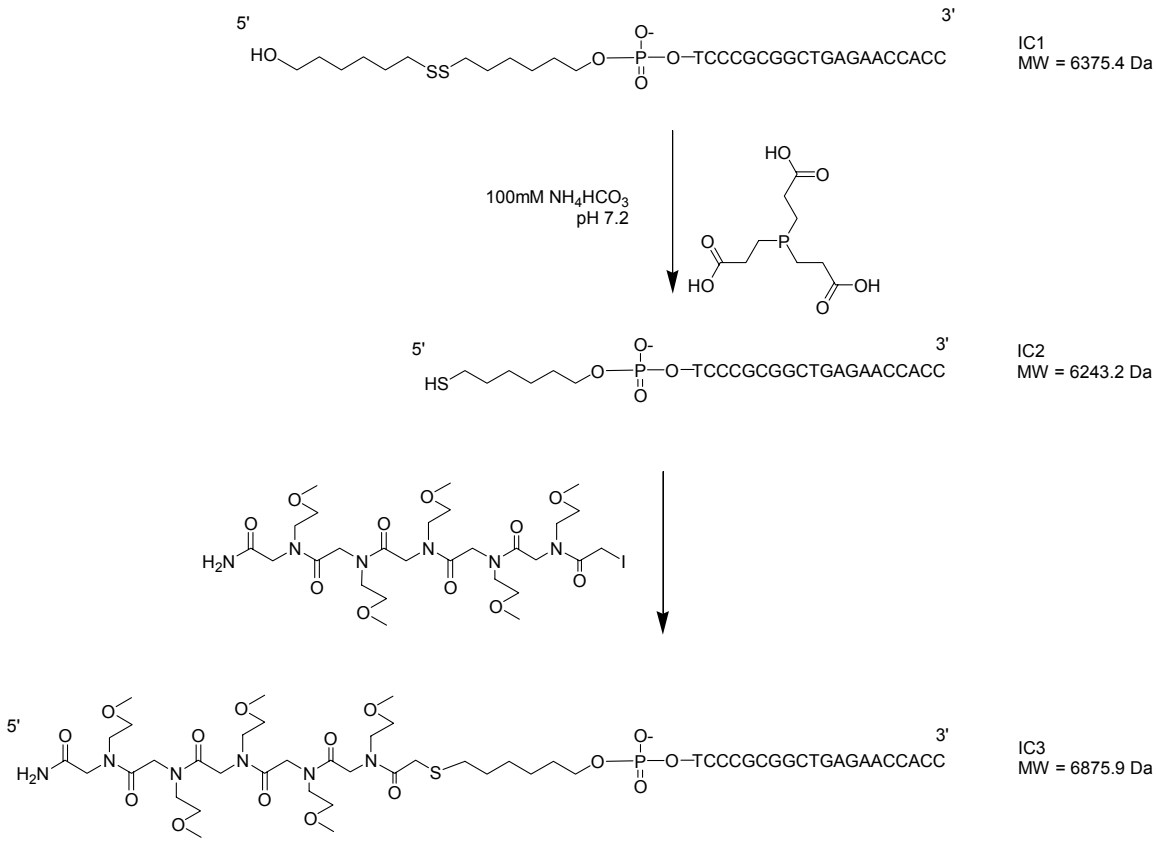


Figure 2.20: Schematic illustration of process for conjugating N-iodoacetyl peptoids to 5′-thiol ODNs.

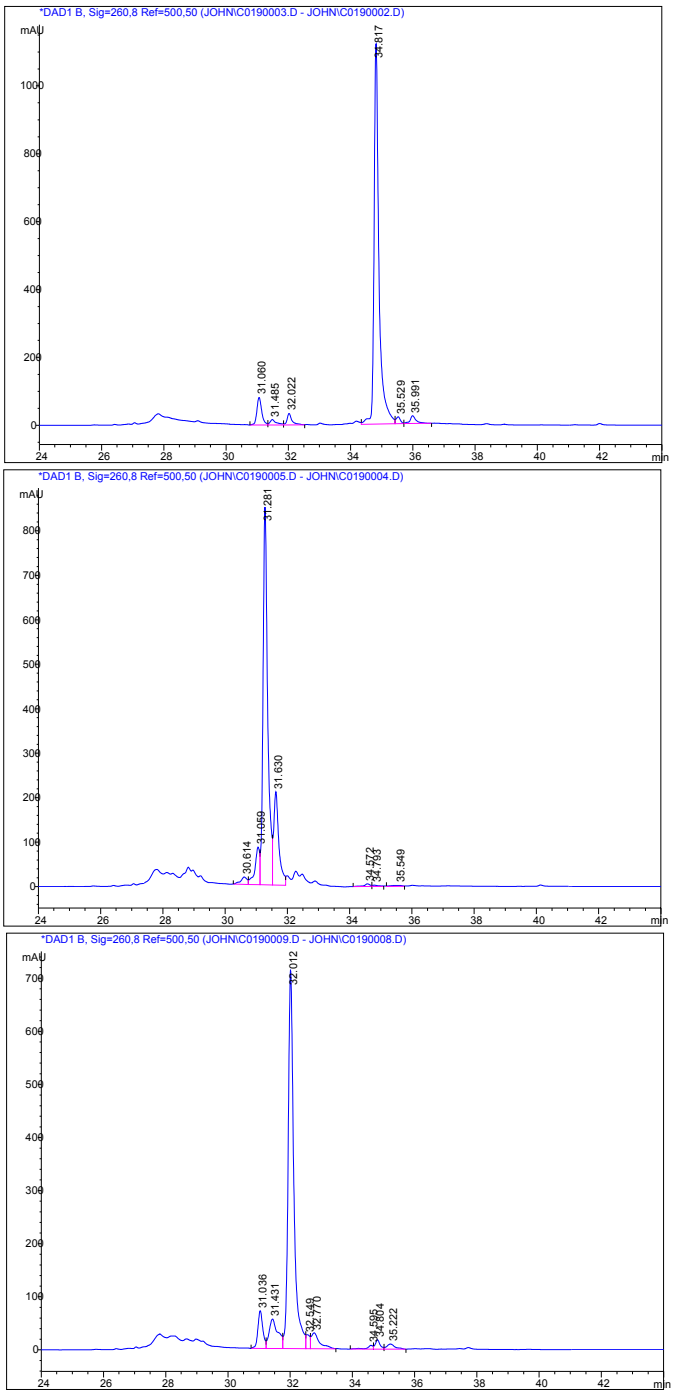


Figure 2.21: RP-HPLC separation and detection of commercially prepared 5′ disulfide ODN IC1, (a), TCEP-reduced ODN IC2, (b), and peptoid-ODN conjugate IC3, (c).

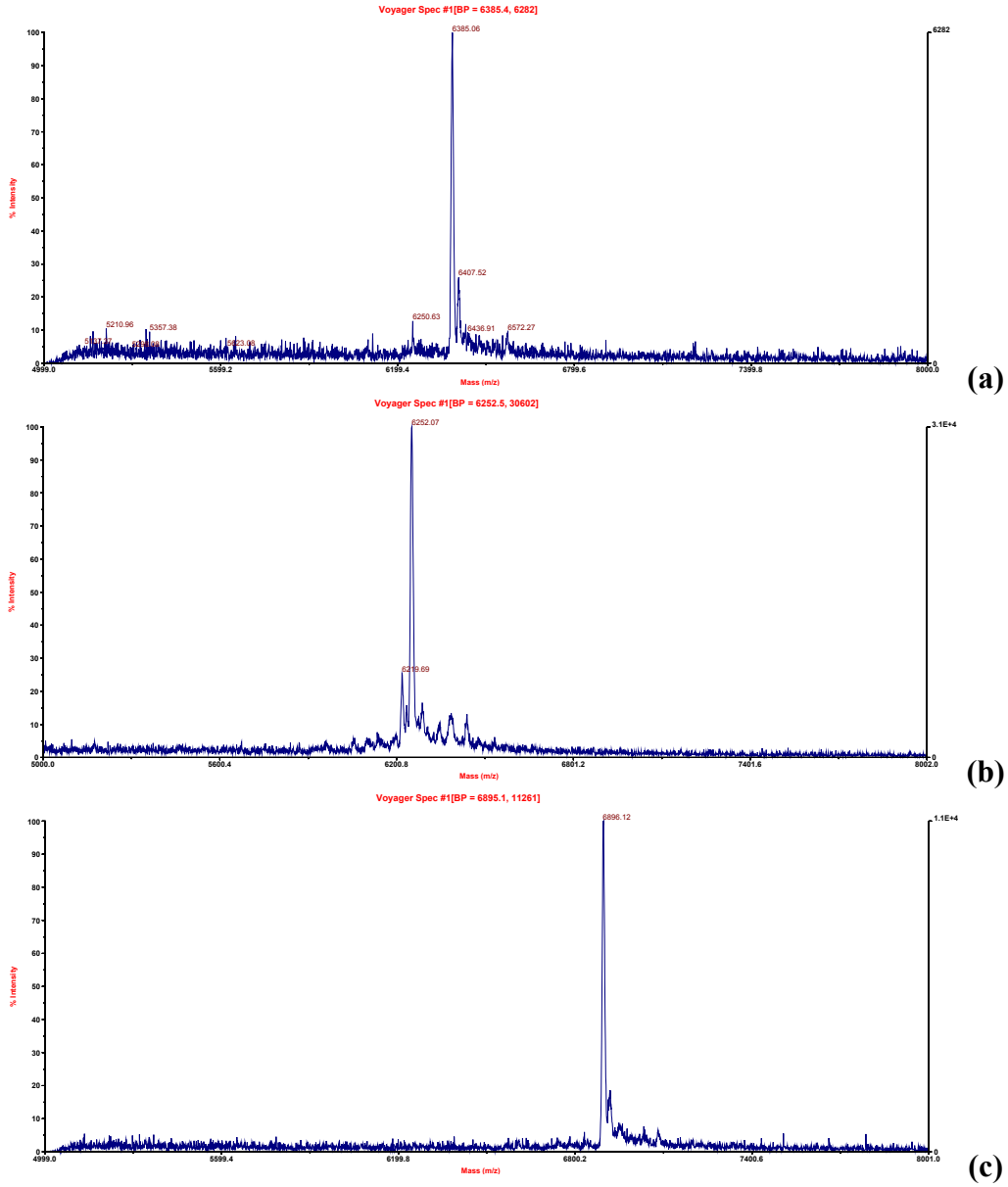


Figure 2.22: MALDI-TOF spectra of commercially prepared 5′ disulfide ODN IC1, (a), TCEP-reduced ODN IC2, (b), and peptoid-ODN conjugate IC3, (c).

2.4 Summary

This work has demonstrated several new methodologies that can expand the range of applications for peptoids. I show that (i) branched peptoids can be synthesized by incorporating an unprotected diamine submonomer, (ii) that N,N-diethylamine can effectively cap the peptoid growing chain, (iii) that six N-terminal modifications for peptoids, including two haloacetamides, two acids, and two amines can be prepared, (iv) that peptoids can be macrocyclized by incorporating a C-terminus-proximal pendant amine, terminating the growing chain with an acid, and using peptide coupling reagents to complete the cyclization, (v) that adamantane can be incorporated into peptoids by including short spacers between the main peptoid chain and the pendant adamantyl moiety, and (vi) that peptoid-oligodeoxynucleotide conjugates can be produced in high yield by combining N-terminal iodoacetyl peptoids and 5′ thiol ODNs.

2.5 References

1. Kirshenbaum, K., Zuckermann, R. & Dill, K. Designing polymers that mimic biomolecules. *Current Opinion in Biotechnology* **9**, 530-535 (1999).
2. Simon, R. J. et al. Peptoids: A modular approach to drug discovery. *Proc. Natl. Acad. Sci. USA* **89**, 9367-9371 (1992).
3. Kirshenbaum, K. et al. Sequence-specific polypeptoids: A diverse family of heteropolymers with stable secondary structure. *Proc. Natl. Acad. Sci. USA* **95**, 4303-4308 (1998).
4. Armand, P. et al. Chiral N-substituted glycines can form stable helical conformations. *Folding and Design* **30**, 369-375 (1997).
5. Armand, P. et al. NMR determination of the major solution conformation of a peptoid pentamer with chiral side chains. *Proc. Natl. Acad. Sci. USA* **95**, 4309-4314 (1998).
6. Wu, C. W., Sanborn, T. J., Zuckermann, R. & Barron, A. Peptoid oligomers with alpha-Chrial, aromatic side chains: effects of chain length on secondary structure. *J. Am. Chem. Soc.* **123**, 2958-2963 (2001).
7. Wu, C. W. et al. Structural and spectroscopic studies of peptoid oligomers with alpha-chiral aliphatic side chains. *Journal of the American Chemical Society* **125**, 13525-13530 (2003).

8. Zuckermann, R. et al. Discovery of nanomolar ligands for 7-transmembrane G-protein-coupled receptors from a diverse N-(substituted)glycine peptoid library. *J. Med. Chem.* **37**, 2678-2685 (1994).

9. Ruijtenbeek, R. et al. Peptoid-peptide hybrids that bind Syk SH2 domains involved in signal transduction. *Chembiochem.* **2**, 171-179 (2001).

10. Nguyen, J. T. et al. Improving SH3 domain ligand selectivity using a non-natural scaffold. *Chemistry & Biology* **7**, 463-473 (2000).

11. Alluri, P. G., Reddy, M. M., Bachhawat-Sikder, K., Olivos, H. J. & Kodadek, T. Isolation of protein ligands from large peptoid libraries. *Journal of the American Chemical Society* **125**, 13995-14004 (2003).

12. Wu, Y., Xu, J.-C., Liu, J. & Jin, Y.-X. Synthesis of N-Boc and N-Fmoc dipeptoids with nucleobase residues as peptoid nucleic acid monomers. *Tetrahedron* **57**, 3373-3381 (2001).

13. Wu, Y. & Xu, J. C. Synthesis of peptoid nucleic acid with thymine as nucleic base. *Chinese Chemical Letters* **11**, 771-774 (2000).

14. Wender, P. A. et al. The design, synthesis, and evaluation of molecules that enable or enhance cellular uptake: Peptoid molecular transporters. *Proc. Natl. Acad. Sci. USA* **97**, 13003-13008 (2000).

15. Peretto, I. et al. Cell penetrable peptoid carrier vehicles: synthesis and evaluation. *Chem. Commun.*, 2312-2313 (2003).

16. Murphy, J. E. et al. A combinatorial approach to the discovery of efficient cationic peptoid reagents for gene delivery. *Proc. Natl. Acad. Sci. USA* **95**, 1517-1522 (1998).

17. Goodson, B. et al. Characterization of novel antimicrobial peptoids. *Antimicrobial Agents and Chemotherapy* **43**, 1429-1434 (1999).

18. Ng, S. et al. Combinatorial discovery process yields antimicrobial peptoids. *Bioorganic and Medicinal Chemistry* **7**, 1781-1785 (1999).

19. Patch, J. A. & Barron, A. Helical peptoid mimics of magainin-2 amide. *Journal of the American Chemical Society* **125**, 12092-12093 (2003).

20. Wu, C. W., Seurynck, S. L., Lee, K. Y. C. & Barron, A. Helical peptoid mimics of lung surfactant protein C. *Chemistry & Biology* **10**, 1057-1063 (2003).

21. Garcia-martinez, C. et al. Attenuation of thermal nociception and hyperalgesia by VR1 blockers. *Proc. Natl. Acad. Sci. USA* **99**, 2374-2379 (2002).

22. Tang, Y.-C. & Deber, C. M. Aqueous solubility and membrane interactions of hydrophobic peptides with peptoid tags. *Biopolymers* **76**, 110-118 (2004).

23. Shin, I. & Park, K. Solution-phase synthesis of aminooxy peptoids in the C to N and N to C directions. *Organic Letters* **4**, 869-872 (2002).

24. Li, S. et al. Photolithographic synthesis of peptoids. *Journal of the American Chemical Society* **126**, 4088-4089 (2004).

25. Ast, T., Heine, N., Germeroth, L., Schneider-Mergener, J. & Wenschuh, H. Efficient assembly of peptomers on continuous surfaces. *Tetrahedron Letters* **40**, 4317-4318 (1999).

26. Zuckermann, R., Kerr, J. M., Kent, S. B. H. & Moos, W. Efficient method for the preparation of peptoids [oligo(N-substituted glycines)] by submonomer solid-phase synthesis. *Journal of the American Chemical Society* **114**, 10646-10647 (1992).

27. Figliozzi, G., Goldsmith, R., Ng, S., Banville, S. & Zuckermann, R. Synthesis of N-substituted glycine peptoid libraries. *Methods in Enzymology* **267**, 437-447 (1996).

28. Olivos, H. J., Alluri, P. G., Reddy, M. M., Salony, D. & Kodadek, T. Microwave-assisted solid-phase synthesis of peptoids. *Organic Letters* **4**, 4057-4059 (2002).

29. Uno, T., Beausoleil, E., Goldsmith, R., Levine, B. & Zuckermann, R. New submonomers for poly N-substituted glycines (peptoids). *Tetrahedron Letters* **40**, 1475-1478 (1999).

30. Burkoth, T. S., Fafarman, A. T., Charych, D. H., Connolly, M. D. & Zuchermann, R. N. Incorporation of unprotected heterocyclic side chains into peptoid oligomers via solid-phase submonomer synthesis. *Journal of the American Chemical Society* **125**, 8841-8845 (2003).

31. Anne, C., Fournie-Zaluski, M., Roques, B. & Cornille, F. Solid phase Synthesis of peptoid derivatives containing a free C-terminal carboxylate. *Tetrahedron Letters* **39**, 8973-8974 (1998).

32. Brown, D. S., Revill, J. M. & Shute, R. E. Merrifield, alpha-methoxyphenyl (MAMP) resin; a new versatile solid support for the synthesis of secondary amides. *Tetrahedron Letters* **39**, 8533-8536 (1998).

33. Horn, T., Lee, B.-C., Dill, K. & Zuckermann, R. Incorporation of chemoselective functionalities into peptoids via solid-phase submonomer synthesis. *Bioconjugate Chemistry* **15**, 428-435 (2004).

34. Barron, A. & Vreeland, W. Free-solution capillary electrophoresis of polypeptoid-oligonucleotide conjugates. *Polymer Preprints* **41**, 1018-1019 (2000).

35. Heerma, W. et al. Comparing mass spectrometric characteristics of peptides and peptoids--2. *Journal of Mass Spectrometry* **32**, 697-704 (1997).

36. Robinson, G., Manica, D., Taylor, E., Smyth, M. & Lunte, C. Development of a capillary electrophoretic separation of an N-(substituted)-glycine-peptoid combinatorial mixture. *Journal of Chromatography B* **707**, 247-255 (1998).

37. Robinson, G., Taylor, E., Smyth, M. & Lunte, C. Application of capillary electrophoresis to the separation of structurally diverse N-(substituted)-glycine-peptoid combinatorial mixtures. *Journal of Chromatography B* **705**, 341-350 (1998).

38. Boeijen, A. & Liskamp, R. M. J. Sequencing of peptoid peptidomimetics by Edman degradation. *Tetrahedron Letters* **39**, 3589-3592 (1998).

39. Ball, H. L. & Mascagni, P. Purification of synthetic peptides using reversible chromatographic probes based on the Fmoc molecule. *Int. J. Peptide Protein Res.* **40**, 370-379 (1992).

40. Ball, H. L., Bertolini, G., Levi, S. & Mascagni, P. Purification of synthetic peptides with the aid of reversible chromatographic probes. *Journal of Chromatography A* **686**, 73-83 (1994).

41. Ball, H. L. & Mascagni, P. N-(2-chlorobenzyloxycarbonyloxy)-succinimide as a terminating agent for solid-phase peptide synthesis: Application to a one-step purification procedure. *Letters in Peptide Science* **2**, 49-57 (1995).

42. Ball, H. L. & Mascagni, P. Chemical synthesis and purification of proteins: a methodology. *Int. J. Peptide Protein Res.* **48**, 31-47 (1996).

43. Mascagni, P., Ball, H. L. & Bertolini, G. Selective purification of synthetic proteins by the use of FMOC- and biotin-based reversible chromatographics probes. *Analytica Chimica Acta* **352**, 375-385 (19997).

44. Li, P., Roller, P. P. & Xu, J. Current synthetic approaches to peptide and peptidomimetic cyclization. *Current Organic Chemistry* **6**, 411-440 (2002).

45. Freidinger, R. M. Design and synthesis of novel bioactive peptides and peptidomimetics. *Journal of Medicinal Chemistry* **46**, 5553-5566 (2003).

46. Forns, P. et al. Constrainted derivatives of stylostatin 1. 1. synthesis and biological evaluation as potential anticancer agents. *J. Med. Chem.* **46**, 5825-5833 (2003).

47. Bray, B. L. Large-scale manufacture of peptide therapeutics by chemical synthesis. *Nature Reviews: Drug Discovery* **2**, 587-593 (2003).

CHAPTER 3
MAGE: MASS-SPECTROMETRIC ANALYSIS OF GENE EXPRESSION

3.1 Introduction....61
3.2 Experimental....67
3.2.1 Oligodeoxynucleotide Probes....67
3.2.2 MAGE Methodology....68
3.2.3 Analytical Procedures....69
3.3 Results and Discussion....70
3.3.1 Oligodeoxynucleotide Probes....70
3.3.2 Mass-Spectrometric Quantification....76
3.3.3 MAGE Methodology....79
3.4 Summary....88
3.5 References....89

3.1 Introduction

The **m**ass-spectrometric **a**nalysis of **g**ene **e**xpression (MAGE) methodology is designed to unambiguously quantify one or more nucleic acids simultaneously in multiple samples, using mass spectrometry for the means of ultimate detection. An assay that makes use of mass spectrometry is desirable because in recent years mass spectrometer sensitivity has reached the zeptomolar region and below, and dynamic ranges in excess of six orders of magnitude; these capabilities could relax the dynamic range restrictions and decrease nucleic acid material requirements inherent in current methods for studying gene expression. MALDI-TOF has been demonstrated to detect as little as 2 attomoles of peptides near 1000 Da in the standard format, and as little as 42 zeptomoles in a microspot format [1]. It is theorized that MALDI-TOF systems and MALDI-TOF/TOF systems are capable of even greater sensitivity, but the field is currently limited by a poor understanding of the origin of noise present in all MALDI-TOF spectra [2]. Isotopic dilution has been demonstrated to function over 4 orders of magnitude in MALDI-TOF systems [3,4]. Further, MAGE relies on solution-phase hybridization for target recognition, a practice thought to contribute to the greater reliability of many single-transcript methods (e.g., real-time PCR) as compared to high-throughput methods (e.g., cDNA microarrays).

Mass spectrometry alone cannot be used to quantify specific sequences in a mixture of nucleic acids. This is not only because many different sequences can share the exact same mass, but also because many constitutively different nucleic acids can have nearly identical masses. Furthermore, the detection of nucleic acids becomes increasingly difficult as their masses increase[5], and lastly, even if a particular sequence could be identified in a mixture, it would not be straightforward to determine its molar abundance. MAGE establishes a one-to-one relationship between nucleic acid sequences of interest and small, inert tag molecules that can be distinguished and quantified using mass spectrometry. The MAGE methodology is outlined in Figure 3.1, and involves the following five steps: (1) Add probe molecules to unknown cDNA mixture and allow them to hybridize; (2) Ligate hybridized ODNs; (3) Separate Biotin-ODNs from mixture; (4) Cleave peptoid tags from ODNs and recover peptoids; (5) Add "heavy" peptoids as internal standards and perform mass spectrometry.

The MAGE methodology can be viewed as a version of a ligase detection reaction[6,7] where the ligation step serves to covalently attach two entities, biotin and a peptoid, when the target SOI has been successfully detected. By making use of a thermostable ligase, the detection phase of MAGE (steps 1 and 2) can be executed close to the melting temperature of the probes and target, thus minimizing spurious detection events.

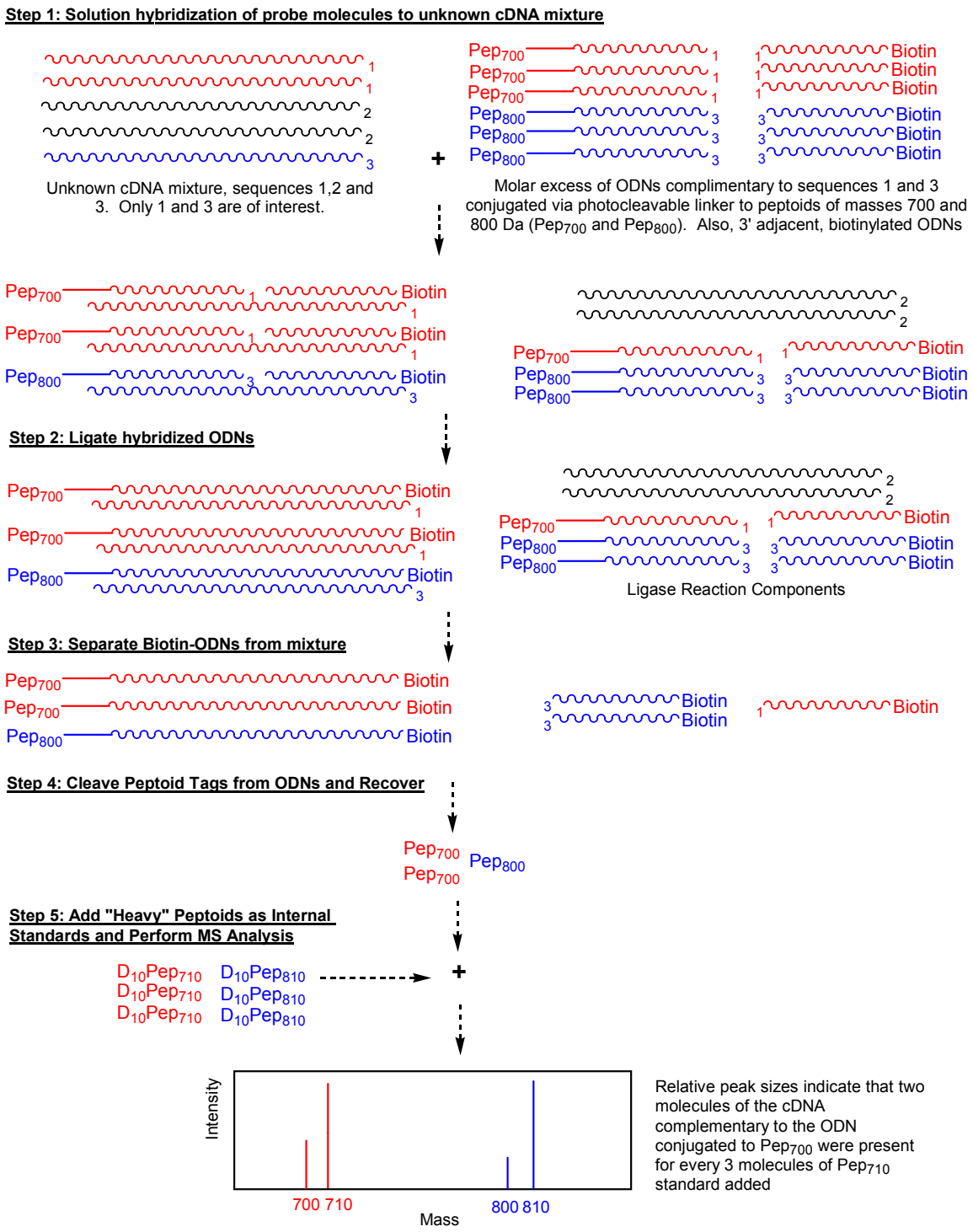


Figure 3.1: The MAGE methodology. The method uses a ligation step to create molecules with both peptoid mass tag and biotin moieties in one-to-one proportion with sequences of interest. The peptoids are subsequently cleaved, and quantified using isotopic dilution mass spectrometry.

The engineering of the ODN-mass tag conjugates is central to MAGE. The inert tags must be chemically compatible with all of the assay steps, they must facilitate quantification, and they must be available in a wide variety of masses. I chose N-substituted glycine oligomers, or peptoids, to serve as mass tags. Peptoids are synthesized by solid-phase methods with a simple submonomer chemistry, where alternating submonomers are primary amines[8]. Because there are hundreds of suitable, commercially available amines, a large variety of oligomers can be synthesized. They can be produced on an individual basis either manually or in a modified peptide synthesizer, or they can be produced in parallel using a mix-and-split method. Peptoids are stable under a variety of thermal, chemical, and biological stresses; for instance, they are not subject to proteolytic degradation. Furthermore, since they are nonnatural, it will be easier to distinguish them from any contaminating cellular components that might be in the target mixture of nucleic acids. The peptoids are chosen to represent a suitable distribution of masses, and they can be designed to offer other properties, such as a chromatographic property set useful for the separation stages of MAGE. Generally, submonomer amines are selected that yield water-soluble, non-cationic peptoids, such as methoxyethylamine and glycine. The peptoids, which are produced typically in greater than 95% yield and purity, can be purified further using HPLC methods similar to those used to purify peptides. Further details on peptoid synthesis can be found in Chapter 2 of this thesis.

Once the peptoid tags are constructed, they are conjugated to DNA oligonucleotides of length 20 to 50. A synthetic method was developed that allows functionalization of the N-terminus of the peptoid that is stable to cleavage from the

peptoid synthesis resin. Using this method, peptoids with N-terminal iodoacetyl groups are produced. The oligonucleotide is commercially produced with a 5′ disulfide modifier. Once reduced, the 5′ sulfhydryl oligonucleotide reacts with the N-iodoacetyl peptoid to form a thioether bond.

Though thioether bonds are not reversible within the conditions used for MAGE, I used a commercially available phosphoramidite that contains a photolabile orthonitrobenzyl group. This bond is stable to acid and base, but cleaves quantitatively when exposed to long-wave UV light for 5 minutes[9-13]. This phosphoramidite is added to the 5′ end of the complete oligonucleotide, and is followed by the 5′ disulfide. Then, when the conjugate is formed, the photocleavable bond is later be used to separate the DNA portion of the conjugate probe from the mass tag (step 4).

The last major challenge is quantifying a mixture of dilute oligomeric species of unique mass using mass spectrometry (step 5). For this, I made use of isotopic internal standards. Isotopic dilution is not only convenient for MAGE, but it has also been shown theoretically to be the most accurate and sensitive method of mass spectrometric calibration[4]. It has also been demonstrated for use in proteomics studies[14-17]. For each peptoid mass tag in our detector library, we make a chemically identical species that has an isotopic shift by using D_3 bromoacetic acid as a replacement submonomer during peptoid synthesis. Each time the deuterated species is substituted, 2 Da are added to the molecular weight of the peptoid (the third deuterium is lost). This can be repeated several times to separate the "heavy" peptoid from its isotopic standard in the mass dimension. Because the tags are chemically identical, they will ionize to almost exactly the same extent, and they will emerge from a chromatographic pre-separation at nearly

the same time. The ratio of the peak sizes of the two species is used to determine their relative quantity. This could be used to determine exact concentrations, or, by hybridizing the two libraries of probes to two different samples, the relative amounts of the target species in the two samples could be determined.

To summarize, the method requires that at least one pair of ODN-cleavable linker-peptoid conjugates be created for each sequence of interest. These probes, along with 3′ adjacent biotin-labeled ODNs of equal length, are used to interrogate a target mixture of cDNA (step 1). Following hybridization, the two adjacent probes are ligated to enhance the specificity of the identification (step 2), and to enable the use of a biotin-affinity column for removal of non-hybridized peptoid tags (step 3). The resulting mixture is exposed to longwave ultraviolet light to release the peptoid tags (step 4), which are quantified using MALDI-TOF mass spectrometry using the isotopically labeled peptoids as internal standards (step 5).

3.2 Experimental

3.2.1 Oligodeoxynucleotide Probes

Oligodeoxynucleotides (ODNs) with 3′ BiotinTEG modifiers and 5′ chemical phosphorylation were prepared commercially (formerly, Beckman Institute Biopolymer Synthesis Facility, Pasadena CA, presently, Qiagen, Valencia, CA). Peptoids were synthesized using the method of Figliozzi *et al*.[8], as detailed in Chapter 2 of this thesis. For MAGE, peptoid 5-mers synthesized uniformly of methoxyethylamine submonomers (Aldrich Chemical Co., Milwaukee WI) were terminated by a final acetylation with N-iodoacetic acid (Fluka, via Aldrich Chemical Co., Milwaukee WI). Isotopically modified peptoids were produced by repeated incorporation of D_3-bromoacetic acid (Cambridge Isotopes, Andover MA).

ODNs with three consecutive 5′ modifiers were prepared commercially (formerly, Beckman Institute Biopolymer Synthesis Facility, Pasadena CA, presently, Qiagen, Valencia, CA). From 3′ to 5′, the modifiers (Glen Research Corp., Sterling VA) were Dabcyl-dT, PC-spacer, and C6-Disulfide. These were resuspended at 50 μM concentration in pH 7.2 100 mM NH_4HCO_3 (Aldrich Chemical Co., Milwaukee, WI) in DNase-free water (Gibco, via Invitrogen, Carlsbad, CA). To this, a 20-fold excess of tris-carboxyethylphosphine (TCEP, Pierce Biotechnology, Rockford, IL) was added, and immediately followed by a 40-fold excess of N-iodoacetyl peptoid. The mixture was

placed under argon and gently mixed for 72 hours. The product was purified using reverse phase HPLC over a Zorbax 300Extend-C18 column in an Agilent 1100 system. The peaks were eluted with a linear gradient of 1-70% B in A over 50 minutes at 0.3 mL/min (solvent A=100 mM TEAA (Fluka, via Aldrich Chemical Co., Milwaukee, WI) in 100% water, solvent B=100 mM TEAA in 90% acetonitrile, 10% water).

3.2.2 MAGE Methodology

Target DNA was suspended in 1x ligation buffer for T4 DNA ligase, 66 mM Tris-HCl, 5 mM $MgCl_2$, 1 mM dithioerythritol, 1 mM ATP, pH 7.5 (Roche Applied Science, Indianapolis, IN). To this, an-approximately 2-fold excess of 5′-peptoid and 3′-biotin probes were added and vortexed briefly. The mixtures were then annealed by heating to 94°C and gradually cooling to room temperature, at which time T4 DNA ligase was added at approximately 1 unit of ligase per 7 pmol of final product. The mixture was ligated at 16°C for at least 12 hours. Following this, Neutravidin (Pierce Biotechnology, Rockford IL) resin was added in approximately 4-fold excess to biotin probes and gently mixed for at least 3 hours. The resulting suspensions were filtered using 0.22 μm Ultrafree-MC centrifugal filters (Millipore, Billerica, MA) and the resin washed several times each with phosphate buffered saline (Gibco, via Invitrogen, Carlsbad, CA), followed by NH_4HCO_3, 100 mM pH 8.2 (Aldrich Chemical Co., Milwaukee WI), followed by DNase-free water (Gibco, via Invitrogen, Carlsbad, CA). The resin is then resuspended in DNase-free water and exposed while stirring to longwave UV light from a B-100AP lamp (UVP Inc., Upland CA) for 20 minutes. Following this, the solution is

removed from the resin using 0.22 μm Ultrafree-MC centrifugal filters, and the resin is washed twice with DNase-free water. The filtrates are pooled and concentrated by lyophilization, then analyzed by MALDI-TOF mass spectrometry.

3.2.3 Analytical Procedures

Analysis of peptoids was accomplished by reverse phase HPLC over a Zorbax 300Extend-C18 column in an Agilent 1100 system. The peaks were eluted with a linear gradient of 0-75% B in A over 50 minutes at 0.3 mL/min (solvent A=0.1% TFA in 100% water, solvent B=0.1% TFA in 100% acetonitrile). The column was held at 30° C, and detection was accomplished by means of a diode array detector at 220 nm.

Analysis of ODNs and peptoid-ODN conjugates was achieved by reverse phase HPLC over a Zorbax 300Extend-C18 column in an Agilent 1100 system. The peaks were eluted with a linear gradient of 1-70% B in A over 50 minutes at 0.3 mL/min (solvent A=100mM TEAA in 100% water, solvent B=100mM TEAA in 90% acetonitrile, 10% water). The column was held at 30° C, and detection was accomplished by means of a diode array detector at 220 nm, 260 nm and 450 nm.

Mass spectrometry of crude samples and HPLC fractions was accomplished by matrix-assisted laser desorption spectrometry with time-of-flight analysis (MALDI-TOF) on an Applied Biosystems Voyager-DE PRO BioSpectrometry Workstation firing a 337 nm nitrogen laser. Peptoids were generally analyzed with a matrix of α-cyano-4-hydroxycinnaminic acid (Aldrich Chemical Co., Milwaukee, WI), formulated at 10 mg/mL with 0.1% TFA in 50:50 water:acetonitrile. ODNs and peptoid-ODN conjugates

were generally analyzed with a matrix of 3-hydroxypicolinic acid (Aldrich Chemical Co., Milwaukee, WI), formulated at 5 mg/mL with 0.05 mg/mL dibasic ammonium carbonate (Aldrich Chemical Co., Milwaukee, WI) in 90:10 water:acetonitrile.

3.3 Results and Discussion

3.3.1 Oligodeoxynucleotide Probes

In order to demonstrate the synthesis and subsequent cleavage of a peptoid-ODN conjugate, the conjugate PC3 was synthesized, Figure 3.2.

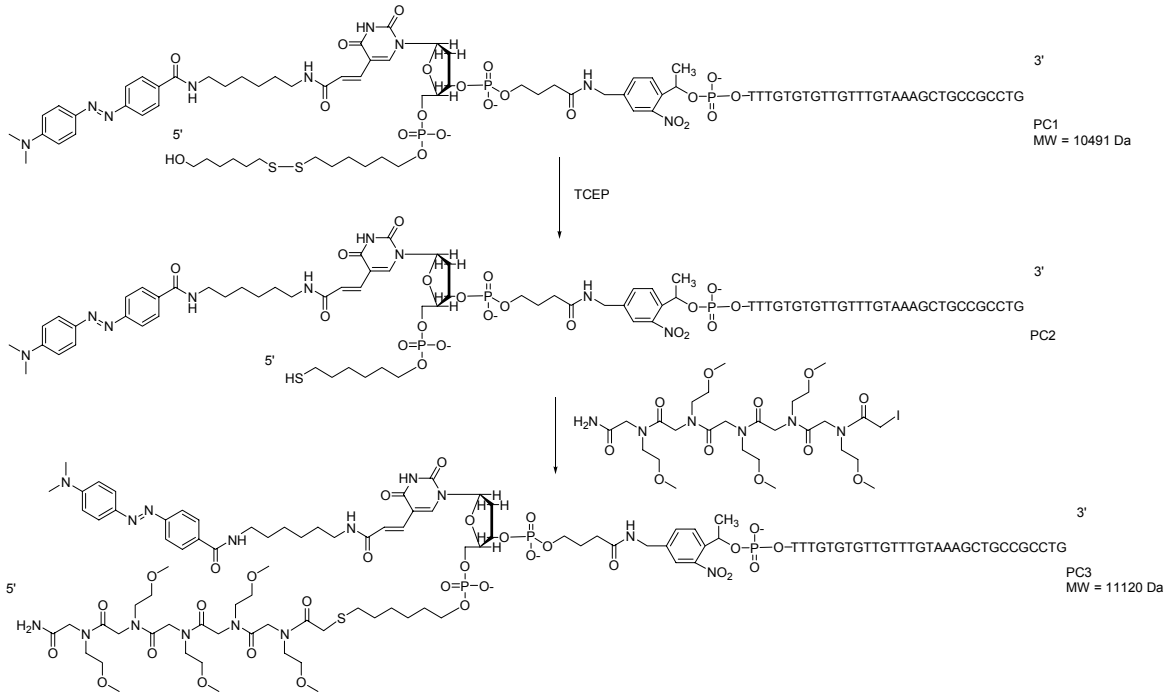


Figure 3.2: Schematic illustrated of the preparation of conjugate PC3. The peptoid-ODN conjugate PC3 is synthesized by first reducing the 5′ disulfide of PC1 and then conjugating an N-iodoacetyl peptoid to PC2.

The dabcyl label immediately 5′ of the main sequence of PC1 is a strongly absorbing chromophore at 450 nm, and it serves in this system to facilitate the tracking and

purification of the peptoid fragments using RP-HPLC. The conjugation reaction proceeded to approximately 50% yield as indicated by RP-HPLC (Fig. 3.3) which was unusually low relative to tests without the dabcyl label.

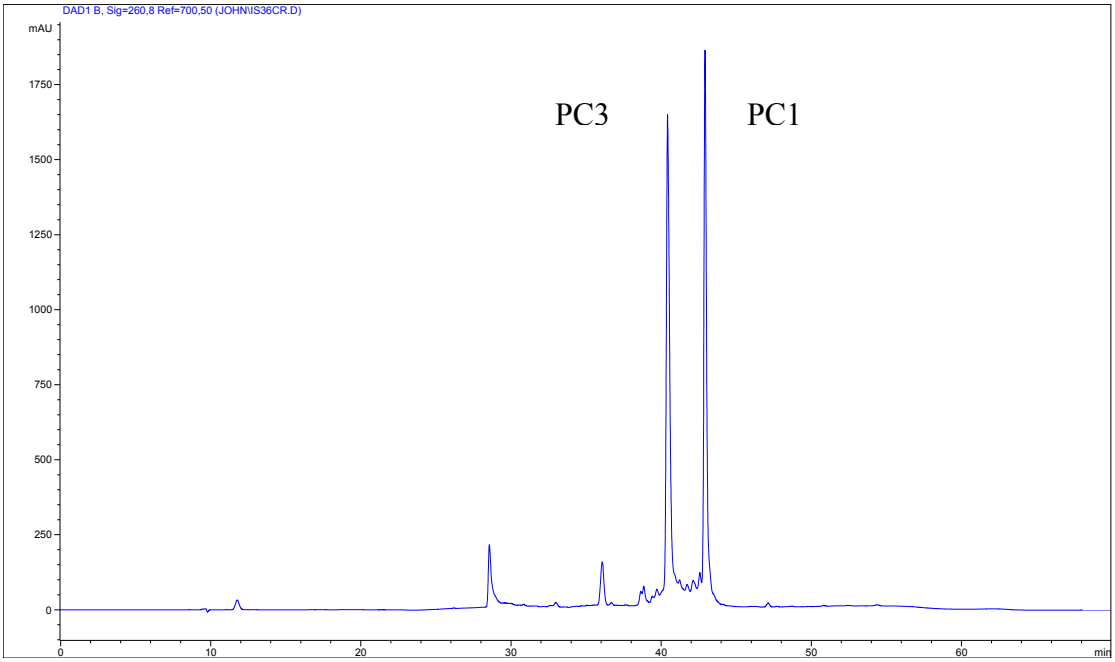


Figure 3.3: RP-HPLC chromatogram of crude PC3 product. RP-HPLC separation with detection at 260 nm of the crude PC3 product indicates approximately 50% yield. The conjugate PC3 elutes earlier than the starting material PC1.

RP-HPLC was used to separate the desired product from the starting material, and MALDI-TOF mass spectrometry confirmed the expected masses of the two species, (Fig. 3.4). Because the MALDI-TOF fires a 337 nm laser, the two species were visible in MALDI-TOF spectra in both their full-length and fragmented states (Fig. 3.5 and Fig. 3.6).

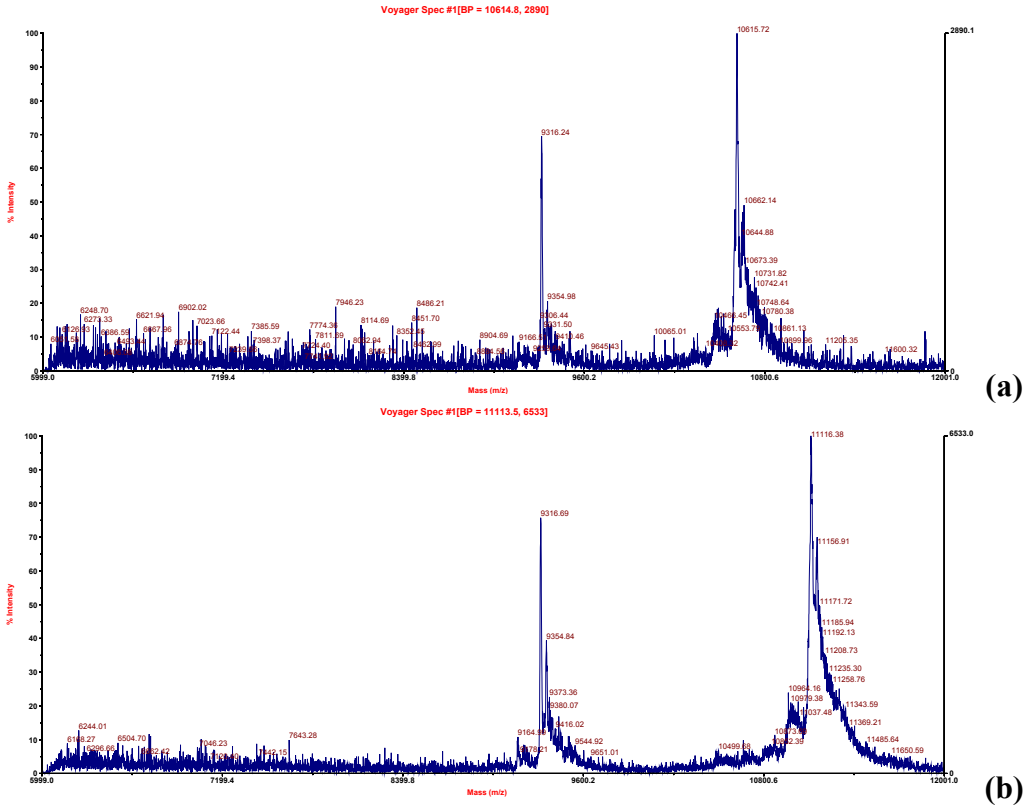


Figure 3.4: MALDI-TOF analysis of the two main peaks of the RP-HPLC separation of crude PC3. For each fraction, two MALDI-TOF peaks were detected because the photocleavable bond is fragmented by the MALDI laser. The masses of the two main peaks agreed with the expected products of (a), Figure 3.4, and (b), Figure 3.5. The smaller peaks in (a) and (b) are identical because only the 5′ side of the PC1 material was modified in the reaction.

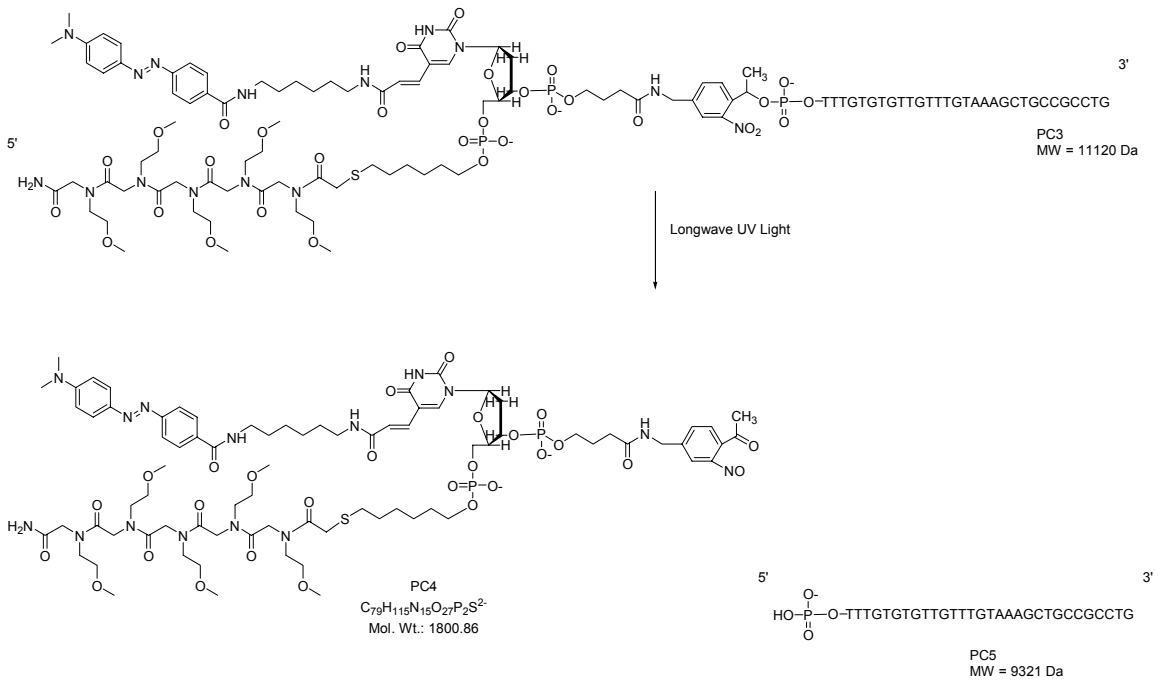


Figure 3.5: Schematic illustration of how exposure to longwave UV light by a lamp or MALDI laser resulted in the creation of two fragments PC4 and PC5 from the PC3 material.

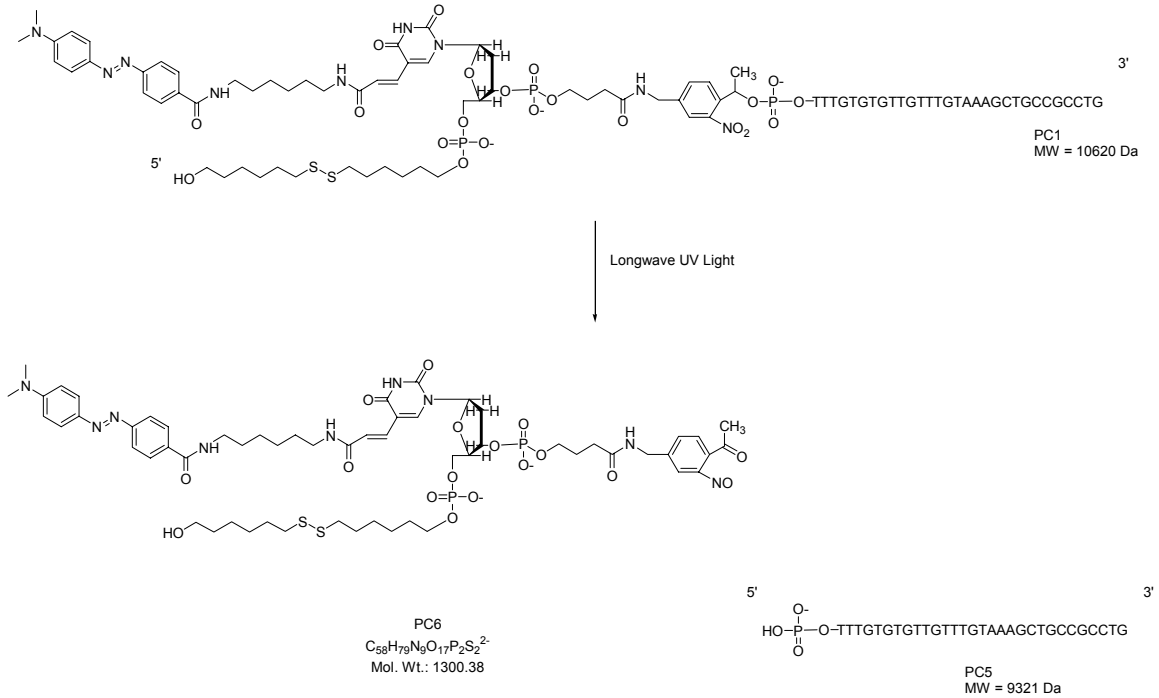
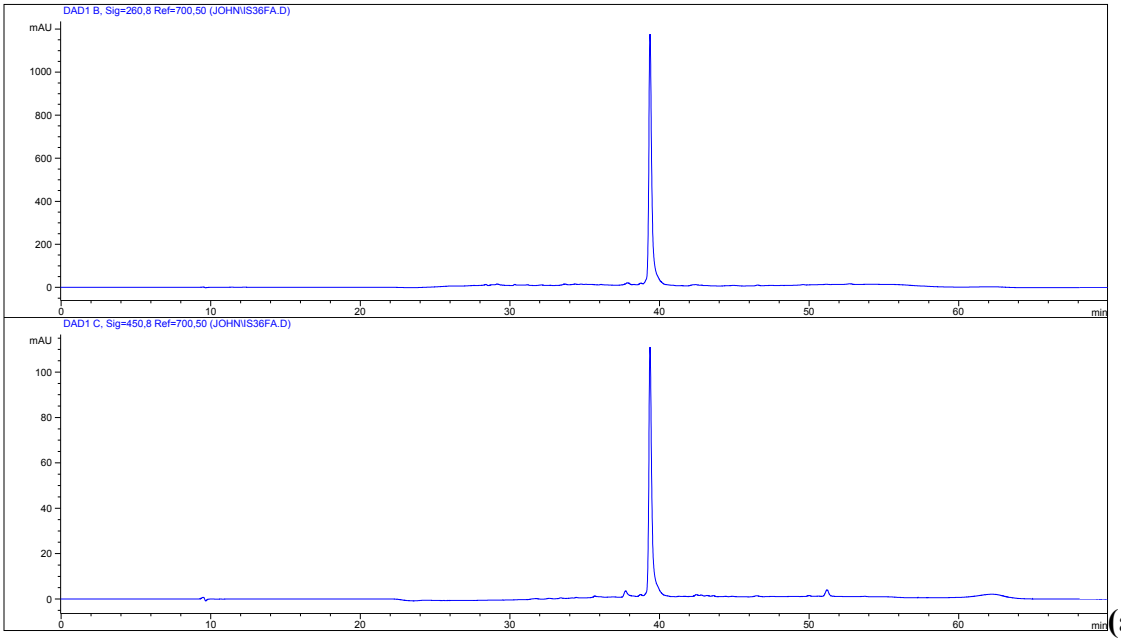
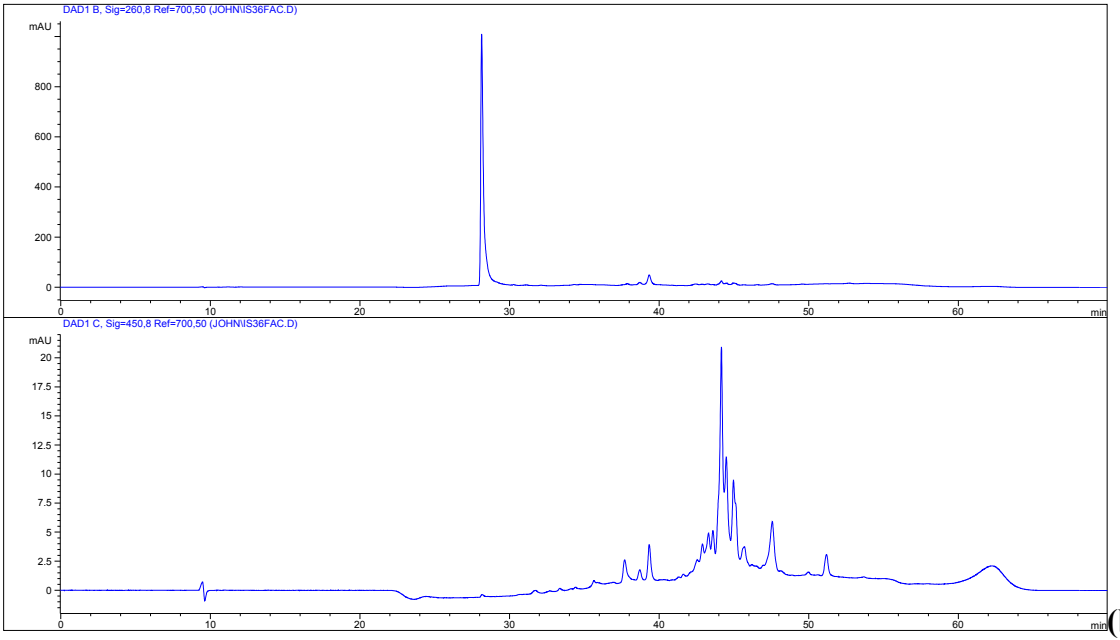


Figure 3.6: Schematic illustration of how exposure to longwave UV light by a lamp or MALDI laser resulted in the creation of two fragments PC6 and PC5 from the PC1 material.

Following purification (Fig. 3.7(a)) the PC3 conjugate was intentionally cleaved by 10 minutes of exposure to longwave UV light, and the resulting mixture was analyzed by RP-HPLC (Fig. 3.7(b)). The large DNA fragment, PC5, was detected at 260 nm, while the smaller peptoid fragment, PC4, was detected at 450 nm. The peptoid fragment was visible in a number of smaller peaks, possibly due to rearrangement during the photocleavage. The calculated area under the several peaks shown in Figure 3.7(b, bottom) is within 95% of the area under the single peak shown in Figure 3.7(a, bottom). The expected masses of PC4 and PC5 are observed when they are collected by RP-HPLC and analyzed by MALDI-TOF (Fig. 3.8).



(a)



(b)

Figure 3.7: RP-HPLC chromatogram of cleavage process. The purified PC3 peptoid-ODN conjugate was analyzed by RP-HPLC with detection at 260 nm (a, top) and 450 nm (a, bottom). Exposure to longwave UV light resulted in cleavage of PC3 into PC4 and PC5. The DNA fragment PC5 post cleavage is visible at 260 nm (b, top), while the peptoid fragment PC4 is visible at 450 nm (b, bottom) because of the dabcyl label.

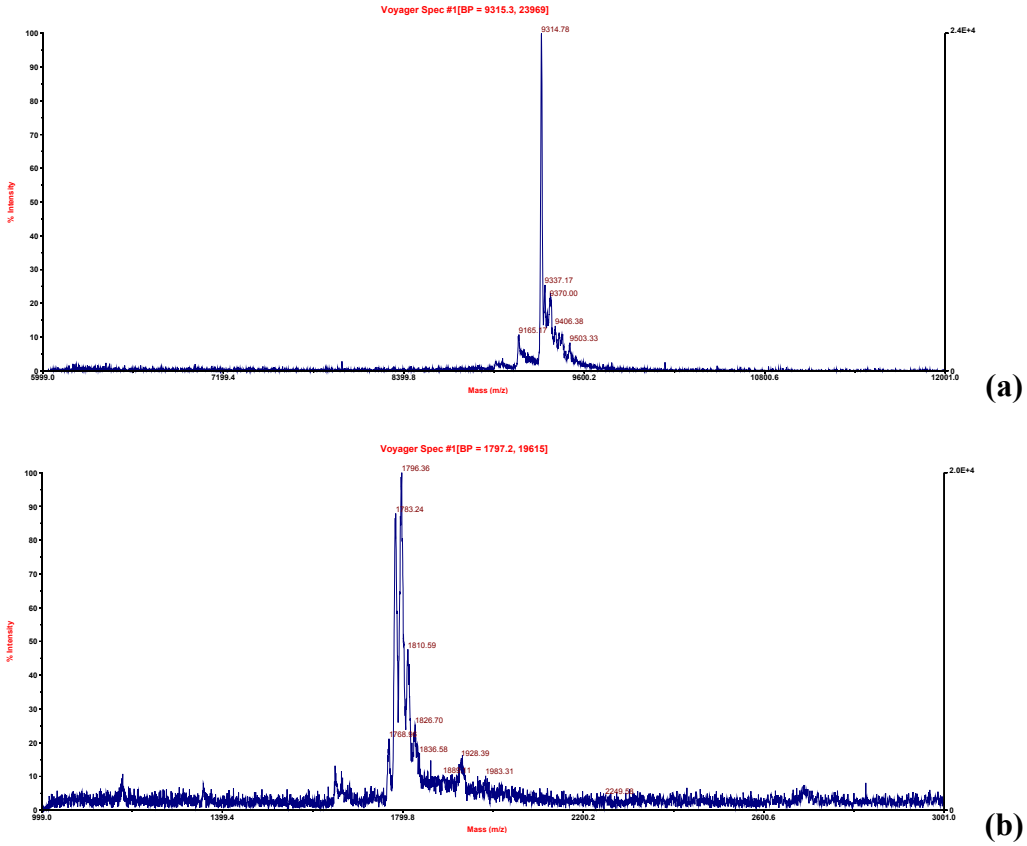


Figure 3.8: MALDI-TOF analyses of cleavage process. After exposure to longwave UV light, the cleavage products of PC3 were collected and analyzed by MALDI-TOF. The fragment visible in 260 nm (a) had a mass that agreed with PC5, and the fragment visible in 450 nm (b) had a mass that agreed with PC4.

3.3.2 Mass-Spectrometric Quantification

Two 5-mer uniform methoxyethylamine peptoids were synthesized to demonstrate quantification using isotopic dilution MALDI-TOF (Fig. 3.9). Following post-synthetic lyophilization, the two peptoids were dissolved in water at equal molar concentrations, and combined at specific volumetric fractions. The resulting mixtures were analyzed by MALDI-TOF, and three spectra were recorded for each mixture. Because of naturally occurring ^{13}C isotopes, as well as hydrogen incorporation into ID2 due to impurities in D_3-bromoacetic acid, it was necessary to calculate the area under a

curve spanning several Dalton (Fig. 3.10). The ratios of the areas corresponding to ID1 and ID2 were calculated and compared to the volumetrically measured mole fractions (Fig. 3.11 and Table 3.1). An example spectra for a 1:1 mixture of the two peptoids ID1 and ID2 (Fig. 3.9). The method was tested over 5 orders of magnitude. It is observed that the method functions optimally within 2. This is due to a combination of noise from impurities in the peptoid solutions as well as noise inherent to the MALDI process. A useful feature of isotopic dilution is that the amount of standard can be adjusted so that the comparison measurement is made within an optimal range.

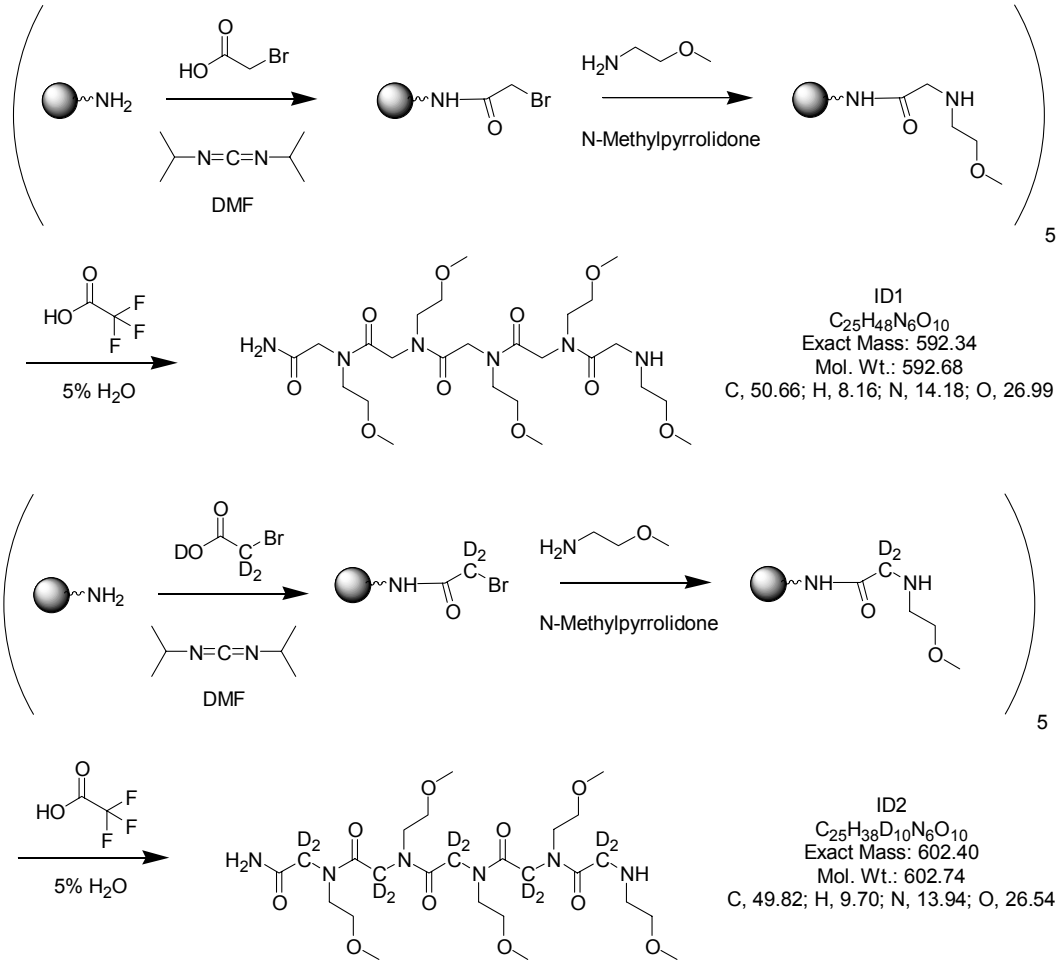


Figure 3.9: Schematic illustrations of ID1 and ID2 syntheses. Isotopically-shifted "heavy" peptoids such as ID2 are synthesized by incorporating D_3-bromoacetic acid.

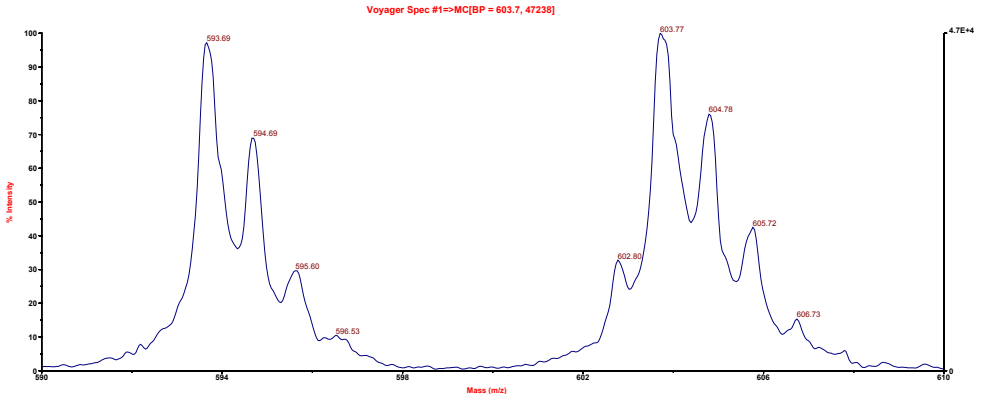


Figure 3.10: MALDI-TOF spectrum of 1:1 mixture of ID1 and ID2.

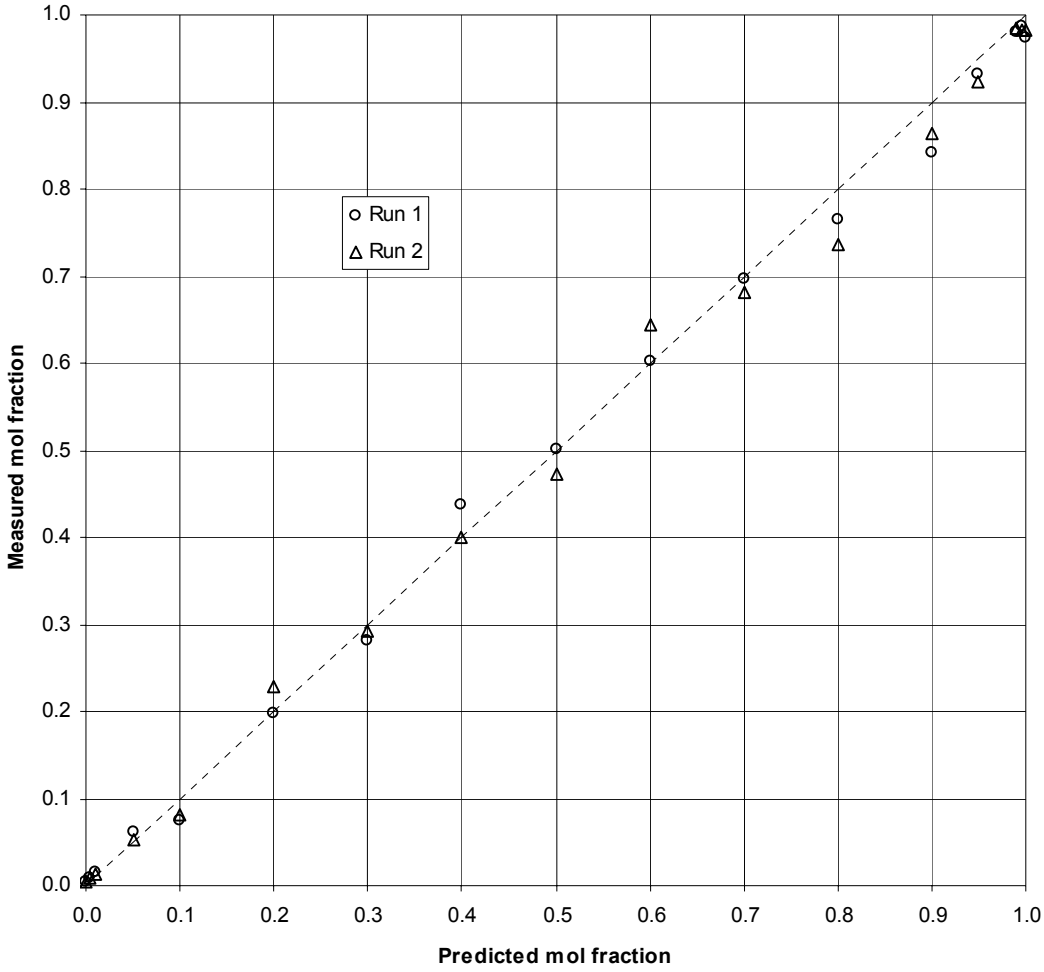


Figure 3.11: Measured mole fractions compared to predicted mole fractions. Peptoid ID1 and 10-fold deuterated peptoid ID2 were mixed in various proportions, and analyzed by MALDI-TOF. The areas under the curves respective to each peptoid were used to estimate the relative mole fractions and compare those estimates to the volumetrically measured mole fractions.

Volumetric		MALDI-TOF Prediction		Volumetric		MALDI-TOF Prediction	
ID1	ID2	ID1	ID2	ID1	ID2	ID1	ID2
0.001	0.999	0.005	0.995	0.500	0.500	0.505	0.495
0.001	0.999	0.004	0.996	0.500	0.500	0.501	0.499
0.005	0.995	0.009	0.991	0.600	0.400	0.603	0.397
0.005	0.995	0.008	0.992	0.600	0.400	0.644	0.356
0.010	0.990	0.016	0.984	0.700	0.300	0.696	0.304
0.010	0.990	0.014	0.986	0.700	0.300	0.681	0.319
0.010	0.990	0.050	0.950	0.750	0.250	0.665	0.335
0.010	0.990	0.060	0.940	0.750	0.250	0.639	0.361
0.010	0.990	0.026	0.974	0.800	0.200	0.764	0.236
0.050	0.950	0.062	0.938	0.800	0.200	0.737	0.263
0.050	0.950	0.053	0.947	0.900	0.100	0.841	0.159
0.100	0.900	0.074	0.926	0.900	0.100	0.863	0.137
0.100	0.900	0.082	0.918	0.900	0.100	0.874	0.126
0.100	0.900	0.178	0.822	0.900	0.100	0.785	0.215
0.100	0.900	0.115	0.885	0.900	0.100	0.876	0.124
0.100	0.900	0.134	0.866	0.950	0.050	0.933	0.067
0.200	0.800	0.197	0.803	0.950	0.050	0.922	0.078
0.200	0.800	0.229	0.771	0.990	0.010	0.980	0.020
0.250	0.750	0.340	0.660	0.990	0.010	0.984	0.016
0.250	0.750	0.223	0.777	0.990	0.010	0.978	0.022
0.300	0.700	0.281	0.719	0.990	0.010	0.946	0.054
0.300	0.700	0.293	0.707	0.990	0.010	0.984	0.016
0.400	0.600	0.438	0.562	0.995	0.005	0.986	0.014
0.400	0.600	0.400	0.600	0.995	0.005	0.982	0.018
0.500	0.500	0.501	0.499	0.999	0.001	0.973	0.027
0.500	0.500	0.472	0.528	0.999	0.001	0.982	0.018

Table 3.1: Measured mole fractions compared to predicted mole fractions. Peptoid ID1 and 10-fold deuterated peptoid ID2 were mixed in various proportions, and analyzed by MALDI-TOF. The areas under the curves respective to each peptoid were used to estimate the relative mole fractions and compare those estimates to the volumetrically measured mole fractions.

3.3.3 MAGE Methodology

In order to test that the MAGE methodology can discriminate targeted sequences from incorrect sequences, four target mixtures were analyzed for a 60mer sequence from mouse muscle cells:

(A) 0.5 nmol 60-mer ODN of antisense myogenin

(B) 1.0 pmol APETALA cDNA

(C) 1.0 pmol each APETALA and Myogenin "full length" cDNA

(D) 1.0 pmol Myogenin "full length" cDNA

Here, cDNAs were random-primed first-strand syntheses of clones from mouse C2C12 tissue culture (Myogenin) or *Arabadopsis thaliana* (APETALA). Target A represents a high-concentration positive control, target D represents a negative control, and targets B and C are intermediate points. The sequence of target A is: 5′ AAA CAC ACA ACA AAC ATT TCG ACG GCG GAC TGG TTC CAG AGG ACA CGA CTA CTA TGG CCC 3′.

To interrogate these targets, the probes M1 and M2 were prepared (Fig. 3.12). M2 is a 30-mer ODN that is 5′ phosphorylated and immediately 3′ of the 30-mer ODN M1. The combined 60-mer sequence is exactly complementary to the ODN target A, and is also complementary to sequences found within the majority of the random-primed myogenin cDNA targets.

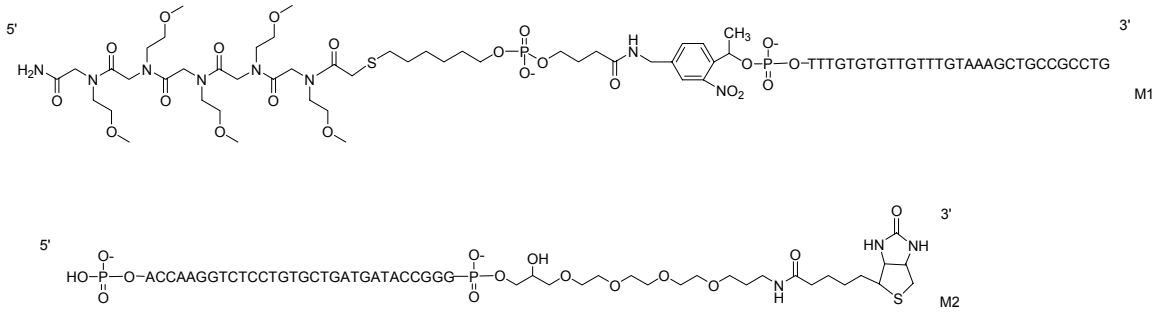


Figure 3.12: Probes M1 and M2. Probe M1 is a 30-mer ODN 5′ reversibly conjugated to a peptoid. Probe M2 is a 30-mer ODN that is 5′ phosphorylated and 3′ biotinylated. When arranged 5′ M1 M2 3′, they form a 60-mer probe for the mouse gene myogenin.

The MAGE assay was executed on each sample. After the photocleavage stage, those target mixtures where M1 and M2 were ligated (indicating a successful detection event) should contain the peptoid fragment M1P (Fig. 3.13). At the quantification stage, no isotopic standard peptoid was added; only the presence of the correct probe was sought.

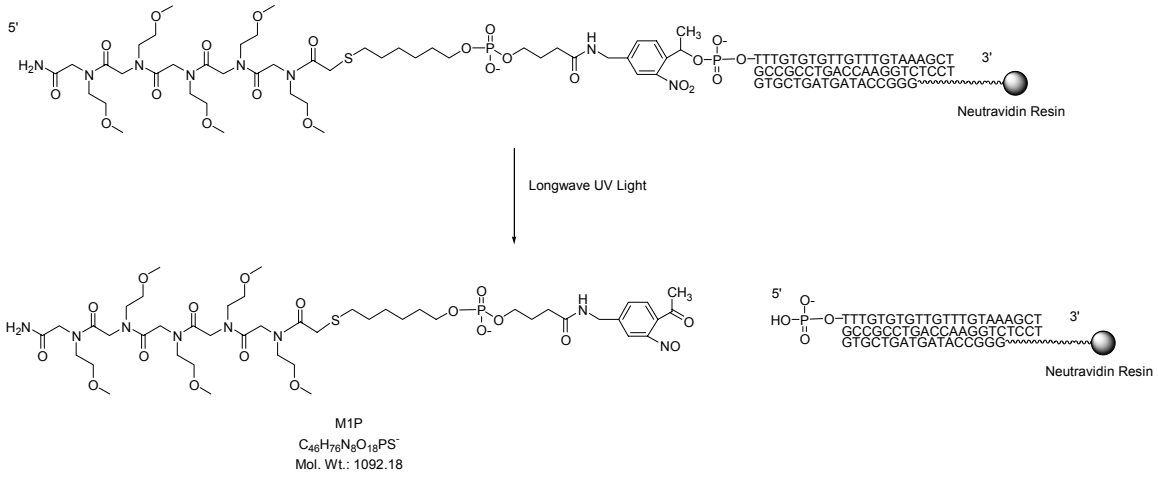


Figure 3.13: Release of M1P tag. During MAGE, the successfully ligated M1-M2 probes are captured by Neutravidin resin. The peptoid fragments M1P are freed by exposure to longwave UV light.

MALDI-TOF mass spectrometry showed the presence of the expected peptoid fragment MP1 only in target mixture A, the higher-concentration ODN target (Fig. 3.14(a)). In the other mixtures, no signal was detected (Fig. 3.14(b)).

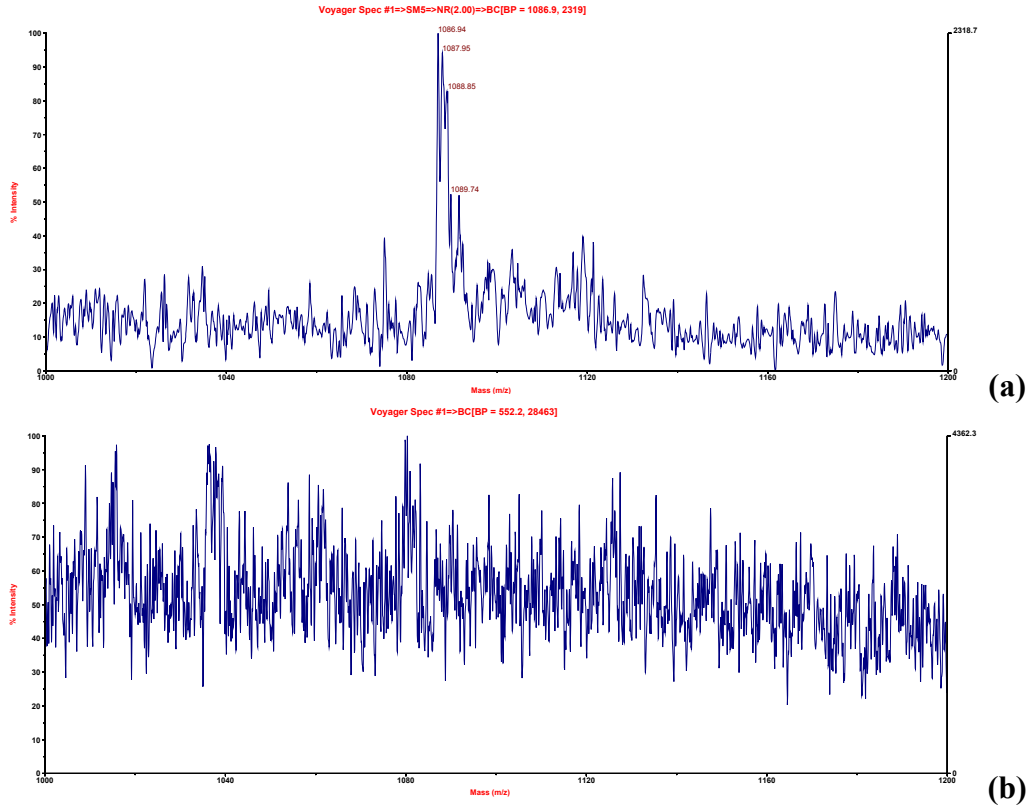


Figure 3.14: MALDI-TOF analysis of M1P tag. The MAGE methodology was employed to detect a 60-nucleotide segment of the myogenin gene in four mixtures. Mixture A contained 500 pmols of a synthetic ODN myogenin target, Mixtures B and C contained 1 pmol of myogenin cDNA, and mixture D contained only cDNA from the APETELA gene. MAGE detected the target in mixture A, (a), but did not in mixtures B, C or D (representative spectrum, b).

A second test of the MAGE methodology employed six target mixtures of two synthetic 60-mer anti-sense ODNs, T1 and T2; T1 is from mouse myogenin, T2 is from mouse paraoxonase. The six mixtures combine the two ODNs in various proportions as shown in Table 3.2.

	pmols T1	pmols T2
A	250	0
B	125	125
C	25	225
D	2.5	247.5
E	0.25	249.75
F	0	250

Table 3.2: Relative amounts of T1 and T2 in targets. The targets for experiment 2 are mixtures of anti-sense oligonucleotides representing mouse myogenin and mouse paraoxonase in various proportions.

The sequence of the T1 target is: 5′ AAA CAC ACA ACA AAC ATT TCG ACG GCG GAC TGG TTC CAG AGG ACA CGA CTA CTA TGG CCC 3′, while the sequence of the T2 target is: 5′ CCG TGA CAC AAG GTG TTT CGA GAA ATG ACA CTA GAC ACT GTT CGG TCG ACG TGC GTG CAG 3′.

To interrogate these targets, the probes M3 and M4 were prepared (Fig. 3.15). M3 is a 30-mer ODN that is 5′ phosphorylated and immediately 3′ of the 30-mer ODN M4. The combined 60-mer sequence is exactly complementary to the ODN target T1, and is also complementary to sequences found within the majority of the random-primed cDNA targets.

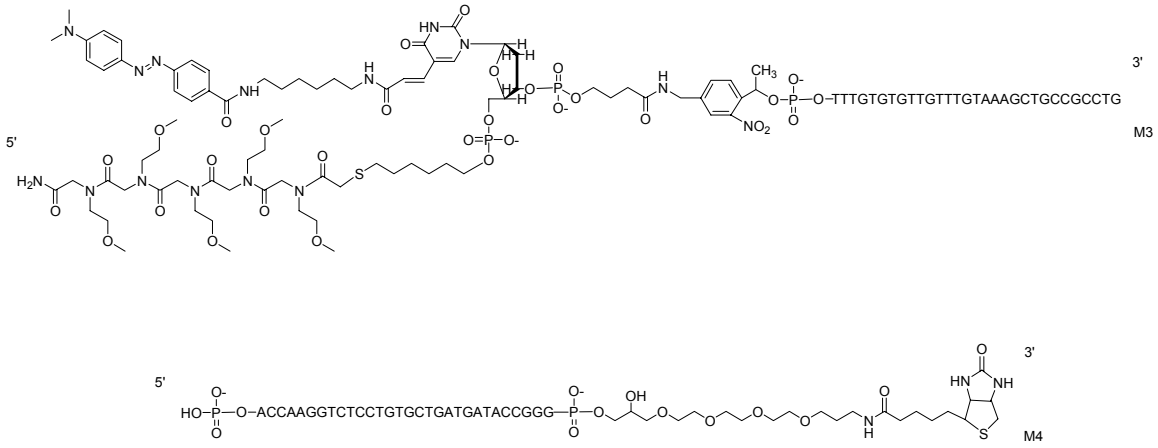


Figure 3.15: Probes M3 and M4. Probe M3 is a 30-mer ODN 5′ reversibly conjugated to a peptoid. Probe M4 is a 30-mer ODN that is 5′ phosphorylated and 3′ biotinylated. When arranged 5′ M3 M4 3′, they form a 60-mer probe for the mouse gene Myogenin.

The MAGE assay was executed on each sample. After the photocleavage stage, those target mixtures where M3 and M4 were ligated (indicating a successful detection event) should contain the peptoid fragment M3P (Fig. 3.16). At the quantification stage, no isotopic standard peptoid was added; only the presence of the correct probe was sought.

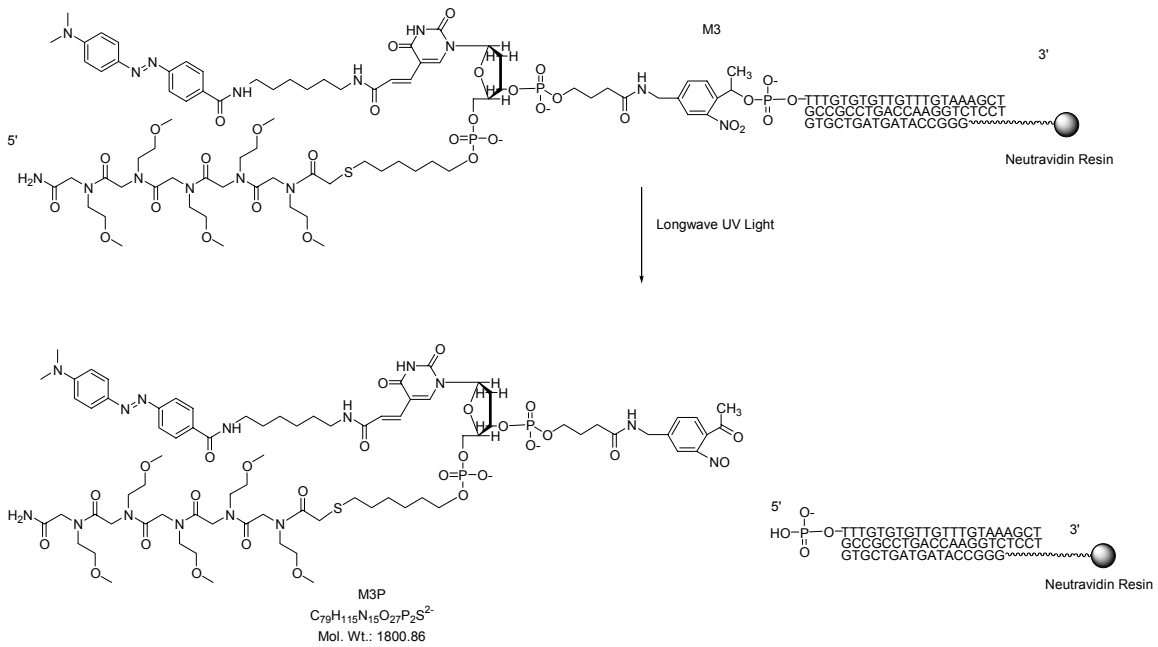
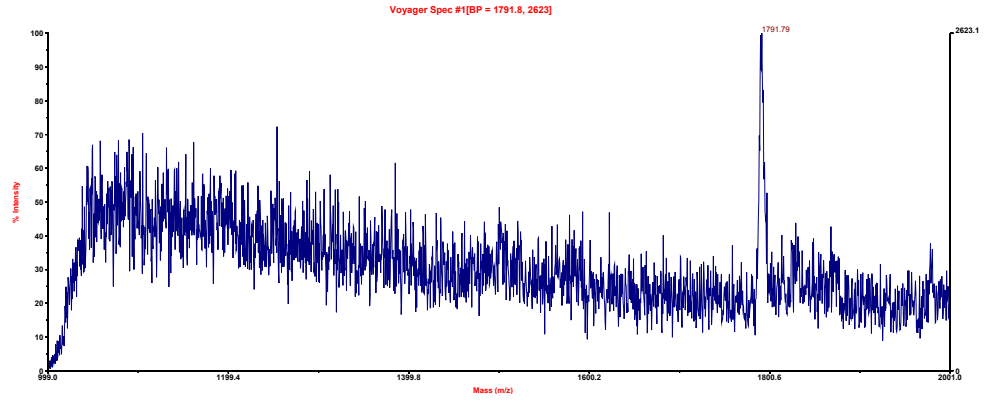
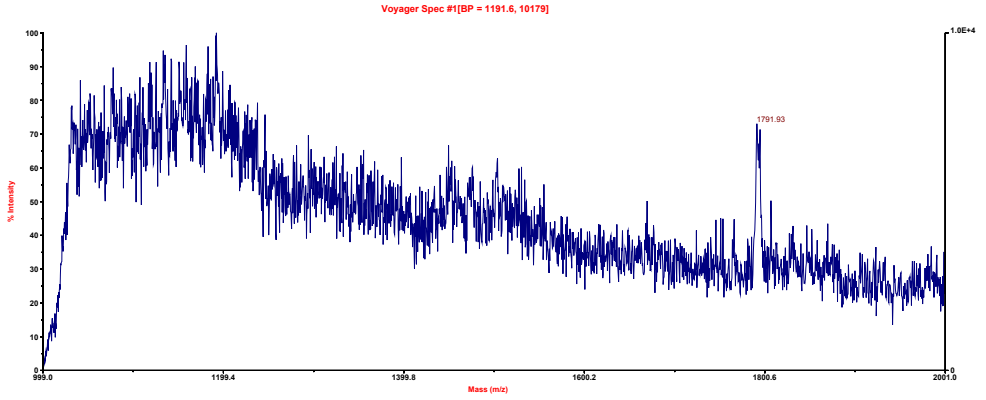


Figure 3.16: Release of M3P tag. During MAGE, the successfully ligated M1-M2 probes are captured by Neutravidin resin. The peptoid fragments M1P are freed by exposure to longwave UV light.

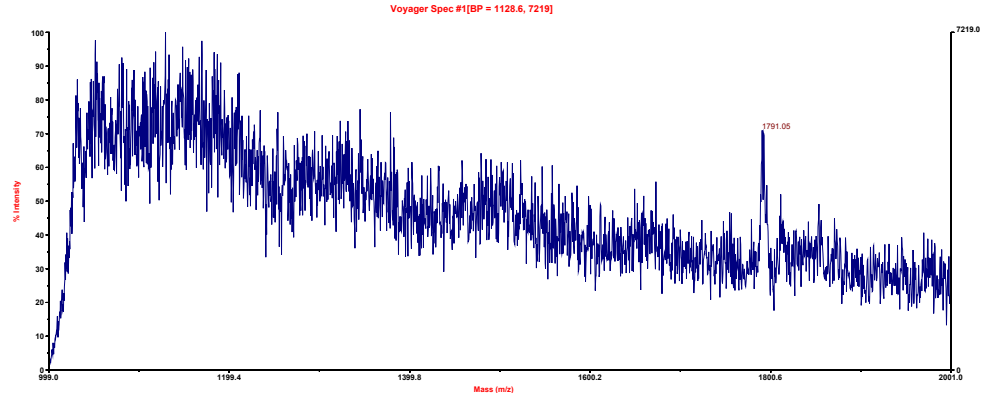
MALDI-TOF mass spectrometry showed the presence of the expected peptoid fragment MP3 in mixtures A, B, C, D, and E (Fig. 3.14(a-e)). In the negative control, mixture F, the peptoid MP3 was not detected, (Fig. 3.14(f)). During the course of the MALDI-TOF analysis, the settings of the spectrometer were altered to obtain the best possible signal for each sample. Thus, the relative peak sizes in Figures 3.14(a-e) are not indicative of the abundances of T1. Furthermore, since the kinetics of hybridization between the probes and targets used in this experiment were not studied independently, and the hybridization conditions were fixed across mixtures A-F, it is not possible to deconvolute the effect of slower hybridization from the effect of lower target concentration.


Voyager Spec #1[BP = 1791.8, 2623]
1791.79
2623.1
% Intensity
Mass (m/z)
999.0
1199.4
1399.8
1600.2
1800.6
2001.0

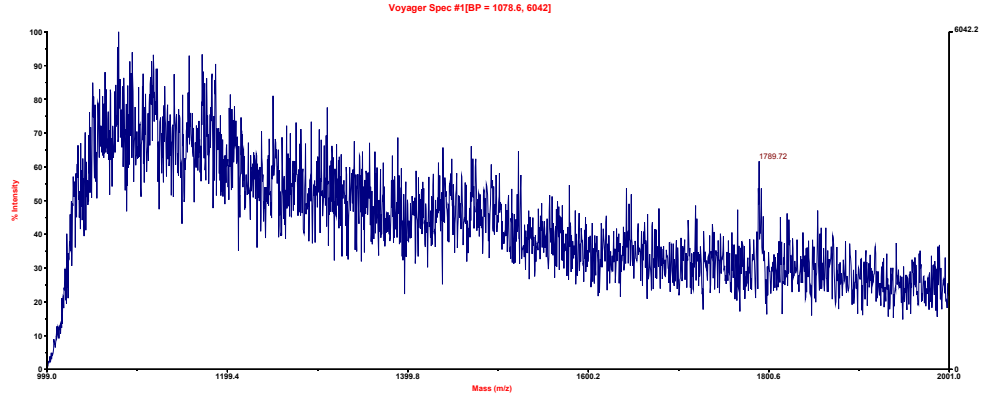
(a)


Voyager Spec #1[BP = 1191.6, 10179]
1791.93
1.0E+4
% Intensity
Mass (m/z)
999.0
1199.4
1399.8
1600.2
1800.6
2001.0

(b)


Voyager Spec #1[BP = 1128.6, 7219]
1791.05
7219.0
% Intensity
Mass (m/z)
999.0
1199.4
1399.8
1600.2
1800.6
2001.0

(c)


Voyager Spec #1[BP = 1078.6, 6042]
1789.72
6042.2
% Intensity
Mass (m/z)
999.0
1199.4
1399.8
1600.2
1800.6
2001.0

(d)

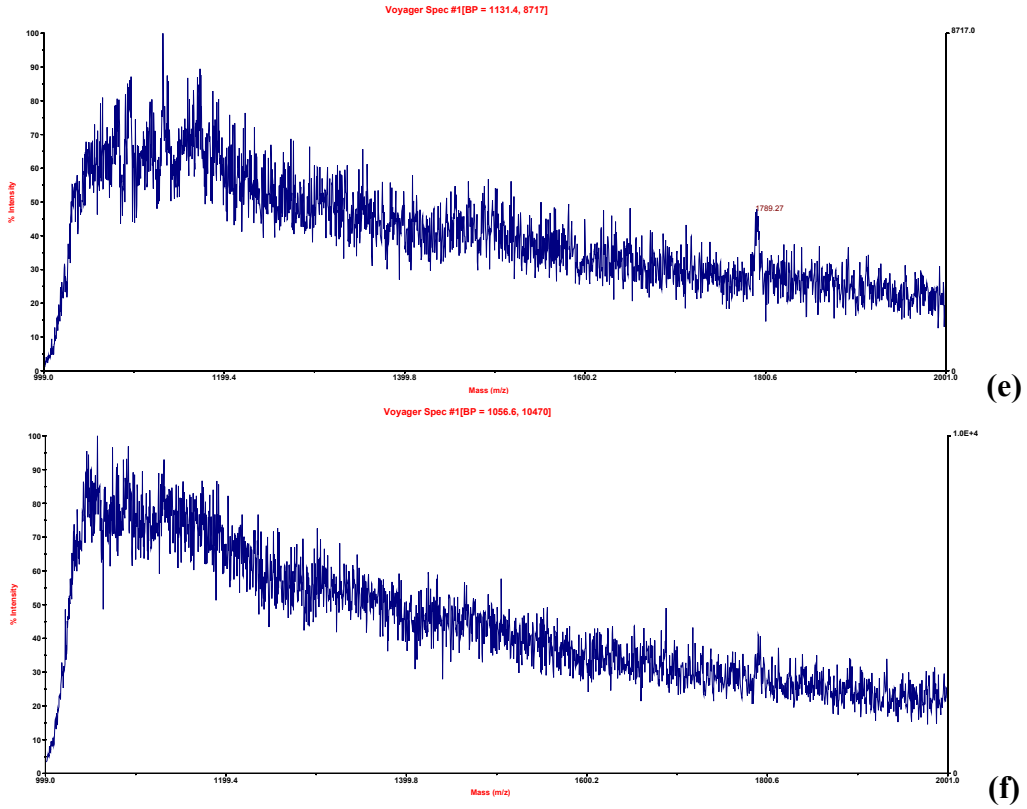


Figure 3.17: MALDI-TOF Detection of M3P tags. The MAGE methodology was employed to detect a 60-nucleotide anti-sense ODN of the myogenin gene in six mixtures. Each mixture contains 250 pmols of ODN. Mixture A contains entirely T1, mixture F contains entirely T2, and B-E are intermediate amounts listed in Table 3.3. The correct expected mass of the peptoid fragment MP3 was detected in samples A-E (a-e), but not clearly in F (f).

3.4 Summary

A methodology for measuring the absolute abundances of specific nucleic acid sequences by means of mass spectrometry has been developed. The method involves 5 steps: (1) add probe molecules to unknown cDNA mixture and allow them to hybridize; (2) ligate hybridized ODNs; (3) separate Biotin-ODNs from mixture; (4) cleave peptoid tags from ODNs and recover peptoids; (5) add "heavy" peptoids as internal standards and perform mass spectrometry.

The chemistry of the reversible peptoid mass tag-ODN probes has been demonstrated in detail using a dabcyl label to facilitate RP-HPLC purification of the photocleavage fragments. The effectiveness of isotopic dilution for quantitative MALDI-TOF has been demonstrated over several orders of magnitude. Finally, the complete MAGE methodology has been executed on two sets of target mixtures. The results show that MAGE may be capable of discriminating correct sequences from incorrect sequences, but further study is necessary to determine is MAGE can function quantitatively.

3.5 References

1. Keller, B. O. & Liang, L. Detection of 25,000 molecules of substance P by MALDI-TOF mass spectrometry and investigations into the fundamental limits of detection in MALDI. *J. Am. Soc. Mass Spectrom.* **12**, 1055-1063 (2001).
2. Krutchinsky, A. N. & Chait, B. T. On the nature of the chemical noise in MALDI mass spectra. *J. Am. Soc. Mass Spectrom.* **13**, 129-134 (2002).
3. Berggren, W. T. et al. Multiplexed gene expression analysis using the Invader RNA assay with MALDI-TOF mass spectrometry detection. *Anal. Chem.* **74**, 1745-1750 (2002).
4. Yu, L. L., Fassett, J. D. & Guthrie, W. F. Detection limit of isotope dilution mass spectrometry. *Anal. Chem.* **74**, 3887-3891 (2002).
5. Chen, X., Westphall, M. S. & Smith, L. M. Mass spectrometric analysis of DNA mixtures: instrumental effects responsible for decreased sensitivity with increasing mass. *Anal. Chem.* **75**, 5944-5952 (2003).
6. Landegren, U., Kaiser, R., Sanders, J. K. M. & Hood, L. A ligase-mediated gene detection technique. *Science* **241**, 1077-1080 (1988).

7. Cao, W. Recent developments in ligase-mediated amplification and detection. *Trends in Biotechnology* **22**, 38-44 (2004).

8. Figliozzi, G., Goldsmith, R., Ng, S., Banville, S. & Zuckermann, R. Synthesis of N-substituted glycine peptoid libraries. *Methods in Enzymology* **267**, 437-447 (1996).

9. Hahner, S. et al. Matrix-assisted laser desorption/ionization mass spectrometry of DNA using photocleavable biotin. *Biomolecular Engineering* **16**, 127-133 (1999).

10. Olejnik, J., Sonar, S., Krzymanska-Olejnik, E. & Rothschild, K. Photocleavable biotin derivatives: A versatile approach for the isolation of biomolecules. *Proc. Natl. Acad. Sci. USA* **92**, 7590-7594 (1995).

11. Olejnik, J., Krzymanska-Olejnik, E. & Rothschild, K. Photocleavable biotin phosphoramidite for 5′-end-labeling, affinity purification and phosphorylation of synthetic oligonucleotides. *Nucleic Acids Research* **24**, 361-366 (1996).

12. Olejnik, J., Krzymanska-Olejnik, E. & Rothschild, K. Photocleavable aminotag phosphoramidites or 5′-termini DNA/RNA labeling. *Nucleic Acids Research* **26**, 3572-3576 (1998).

13. Olejnik, J. et al. Photocleavable peptide-DNA conjugates: synthesis and applications to DNA analysis using MALDI-MS. *Nucleic Acids Research* **27**, 4626-4631 (1999).

14. Gygi, S. et al. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* **17**, 994-999 (1999).

15. Zhou, H., Ranish, J. A., Watts, J. D. & Aebersold, R. Quantitative proteome analysis by solid-phase isotope tagging and mass spectrometry. *Nature Biotechnology* **19**, 512-515 (2002).

16. Smolka, M. B., Zhou, H., Purkayastha, S. & Aebersold, R. Optimization of the isotope-coded affinity tag-labeling procedure for quantitative proteome analysis. *Analytical Biochemistry* **297**, 25-31 (2001).

17. Griffin, T. J., Gygi, S. P., Rist, B. & Aebersold, R. Quantitative Proteomic Analysis Using a MALDI Quadrupole Time-of-Flight Mass Spectrometer. *Anal. Chem.* **73**, 978-986 (2001).

CHAPTER 4
VISUALIZATION AND ANALYSIS OF GENE EXPRESSION DATA WITH ANDREWS CURVES AND MODEL-BASED CLUSTERING

4.1 Introduction....93
4.2 Experimental....103
4.2.1 Theory....103
4.2.2 Computational Methods....106
4.2.1 Sources of Data....106
4.3 Results and Discussion....107
4.3.1 Andrews Curves....107
4.3.2 EM Clustering of Synthetic Data....118
4.3.3 EM Clustering of Gene Expression Data....129
4.4 Summary....136
4.5 References....138

4.1 Introduction

In their most general form, gene expression data are one or more measurements of the extent to which one or more genes of interest are being transcribed in a sample of tissue or cells. In simpler examples of gene expression analysis, such as Northern blots[1], one or several genes are studied at one or several conditions. In these cases, sufficient analysis comprises simple estimations of confidence intervals[2]. Recently, techniques such as hybridization-based microarrays have been developed that can simultaneously measure the expression levels of thousands of genes of interest[3,4]. Data such as these are not easily visualized in a helpful way, and significant analysis presents a further challenge. Once the data have been cursorily screened for genes of exceptional interest or outlying points, investigators must turn to more sophisticated techniques for locating statistically significant information within their data.

The unit of gene expression data presented for analysis is a matrix described by several features. Each element of the matrix is the value assigned the expression of a particular gene in a particular condition. The matrix can be viewed as a set of column vectors, with each vector carrying either the data for one gene at all conditions of interest, or the data for all genes of interest at one condition. In the first case, the gene vectors exist in a potentially reducible condition space of a number of dimensions equal to the number of conditions of interest. In the second case, the condition vectors exist in the

potentially reducible gene space where the number of dimensions is equal to the number of genes of interest. The actual value stored at a given location in the matrix could describe the gene expression in a variety of ways. Most commonly, the value is a positive function of the number of mRNA transcripts of the gene of interest found in a sample of tissue or cells at the condition of interest[5]. The nature of this function is a combination of the choice of gene expression data-collection method and the preprocessing steps that are applied to the raw data. Often, the contribution to this function from the controlling physics of the method of choice is unknown[6].

Several pioneering studies of large-scale gene expression analysis generated data describing the expression of several thousands of genes over 5 to 100 conditions[7-9]. In some of these studies, the investigators examined the relative locations of the gene vectors in condition space. An example of this was the study by Chu *et al.* of the transcriptional program of sporulation in budding yeast[7], which sought to identify related or coexpressed genes, and hypothesized that such genes would be represented by gene vectors somehow close to one another in condition-dimensional space. It remains a subject of great interest how best to define close, and how best to determine statistically significant groups of close genes. In other studies, the investigators examined the relative locations of the condition vectors in gene space. An example of this was the study by Sorlie *et al.* of human breast tumors[9,10], which hoped to correlate some reduction of the gene expression data with the outcome of the diseased patient, and hypothesized that the condition vectors associated with tumors from patients with similar outcomes would be similarly close to one another in gene-dimensional space.

A significant and currently unavoidable challenge of large-scale gene expression analysis is that because thousands of genes are simultaneously analyzed, and each experiment represents a significant cost, the matrix of data presented for analysis will be highly rectangular; the genes dimension will be much larger than the conditions dimension[11]. In the Chu *et al.* example, the space for analysis is extremely rich, and highly unlikely to be degenerate. The Sorlie *et al.* example, and those like it[12,13], on the other hand, presents a very sparsely populated space. It is far more difficult to analyze the Sorlie space using robust techniques of linear algebra and multivariate statistical inference, not only because the vectors fail to fully define the space, but also because so many analytical techniques suffer from what is known as the curse of dimensionality[14]. The curse is a catch-all term for problems arising from the fact that the volume, and computational complexity, of a space grows exponentially with the number of dimensions.

Two important goals of an analysis like those employed in the canonical studies discussed above are to produce intelligible visual representations of the high-dimensional data, and to develop and apply algorithms capable of identifying significant correlations in the data that investigators could not easily notice without this aid. These algorithms might be designed to search the data for correlations that support a predetermined hypothesis, or they might be designed to find unspecified correlations that might aid the investigator in creating testable hypotheses. The most common type of analysis applied to gene expression data is some form of clustering algorithm designed to identify subsets of data that are similar.

Visualization Strategies

By far the most common visualization strategy for gene expression data is to simply depict the relevant portion of the data matrix with each element colored to represent the value of the element[7,9,15-17]. This method has the main advantages of being easily implemented and spatially compact; it is a simple task to locate data of interest. However it does not offer what many modern multivariate data visualization techniques offer, which is either some mathematical consistency, such as projection onto orthogonal basis functions[18], or a representation that is especially suited for analysis by the human brain, such as Chernoff faces[19,20].

Andrews proposed to plot multivariate data in two dimensions by mapping the data vectors onto a simple trigonometric polynomial basis function[21]. If each vector is represented as $\tilde{v} = (v_1,...,v_n)$, where n is the number of dimensions, then the Andrews plot of the vector is generated by the function

$$F_v(t) = \frac{v_1}{\sqrt{2}} + v_2 \sin(t) + v_3 \cos(t) + v_4 \sin(2t) + v_5 \cos(2t) + ... \qquad \text{(Eq. 4.1)}$$

over the domain $(-\pi < t < \pi)$.

Andrews curves allow multiple points of multivariate data to be plotted in a single two dimensional space simultaneously, and allow clusters to be visually distinguished. The reason for this clustering behavior is that Andrews curves, like all orthogonal basis functions, preserve the means, distances, and variances of untransformed data. In particular, the Euclidian distance between any two vectors is proportional to a straightforward view of the distance between two corresponding transformed functions.

$$\left\|\tilde{v} - \tilde{u}\right\|^2 \propto \int_{-\pi}^{\pi} \left(F_v(t) - F_u(t)\right)^2 dt \qquad \text{(Eq. 4.2)}$$

Andrews curves have been applied to data of biological interest such as pharmaceutical formulation data[22] and psychiatric data[23], and they have been generalized with wavelet theory[18]. Here, I present the application of Andrews curves to the Chu *et al.* data set and the Sorlie *et al.* data set.

Data Reduction Strategies: Principal Components Analysis

Even though Andrews curves can map vectors of any length to two-dimensional space, they generally lose meaning as the number of dimensions increase; the curves become confusing rather than elucidating[21]. Further, since the first, low frequency, terms in the Fourier series have the most influence on the visual appearance of the plot, some preprocessing of the data is helpful. Andrews and others[24] suggest that the data be subjected to a deterministic data reduction strategy such as principal components analysis (PCA) in order to both reduce the number of dimensions and to sort the dimensions in order of importance. In algorithmic terms, PCA identifies the direction of greatest variance in a set of vectors, ranks this as the first eigenvector, and then proceeds to iteratively identify the direction of next-greatest variance that is orthogonal to all those eigenvectors previously identified. The algorithm can be terminated by design, if only a small number of eigenvectors account for a satisfactory fraction of the variance, or it is terminated by necessity when the number of eigenvectors reaches whichever is smaller, either the number of vectors, or the length of the vectors.

PCA has been successfully applied to gene expression data to differentiate between eigenvectors associated with artifacts, noise, and biological processes[25-27]. PCA is simple to execute computationally using singular value decomposition (SVD)[28], but it is limited by its linear nature. This is illustrated by two examples of two-dimensional

data, Figure 4.1a where PCA reduces the data to a single dimension successfully, and Figure 4.1b where PCA finds the variance to be equally accounted for by any choice of two eigenvectors; the simple nonlinear pattern is not extracted. Attempts have been made to design non-linear feature extraction algorithms[29], which often rely on transforming the data into a set of distances between the input vectors, and subjecting those distances to a linear analysis. In this work, I apply PCA to reduce unwieldy data sets for visualization with Andrews curves.

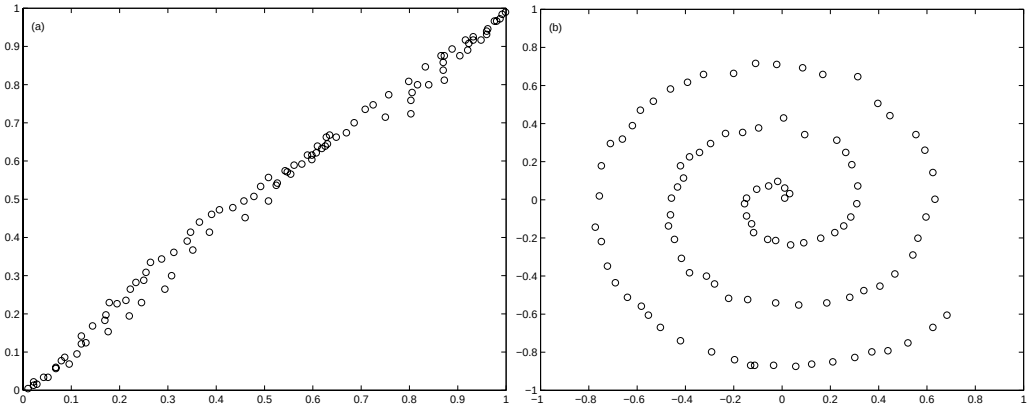


Figure 4.1: Two example data sets illustrating effective use of PCA. PCA is capable of extracting a linear pattern such as in (a), where the variance in the data exists almost entirely along one eigenvector, but it is not capable of extracting a nonlinear pattern such as in (b), where PCA would report a roughly equal amount of variance along any two eigenvectors.

Data Reduction Strategies: Model-Based Clustering

Another challenge of gene expression analysis is performing computational tasks in a manner that is robust to noise, especially noise arising from the chosen method of gene expression data collection. Cellular circuitry is subject to noise of biological origin, and sometimes relies on it for critical operations[30]. It is a daunting challenge to separate the sources of noise in a complex process such as a gene regulation network, and although a number of sophisticated statistical tests have been proposed to help separate

machine noise from biological noise in gene expression studies[31-35], a central goal of investigators is to develop methods for analysis that penetrate all types of noise to infer persistent correlations from the data. The two general categories of noise that confound analytical techniques are strong outliers, which are small numbers of points that are relatively distant from associated points, and dense noise, which describes data where persistent corollaries are obscured by closely associate data that is easily confused for the feature of interest.

The primary analytical method employed in the studies of Chu *et al.*, Sorlie *et al.*, and many other studies is a clustering algorithm that recursively binds the two vectors with the highest Pearson correlation coefficient into nodes until a single node is reached, resulting in a binary tree, or dendrogram[15]. Such dendrograms can also be generated by top-down recursive bisection[36]. Another common clustering algorithm that does not rely on multivariate statistics is k-means[37]. K-means, after an initial set of cluster centroids is provided, alternates between assigning vectors to the nearest centroid and recalculating the values of the centroids from the vectors assigned to it, until convergence. All of these heuristic methods are computationally efficient, parallelizable, and avoid pitfalls such as over-fitting and the curse of dimensionality. However, these methods are not statistically robust.

A variety of more sophisticated alternatives have been proposed that employ model-based clustering[38-46], where the vectors are assumed to have been generated by some combination of probability distribution functions, and an algorithm is applied to compute one or more such combinations that aptly describe the data. Although these methods can be very powerful, they challenge the investigator with new problems. For

example, whereas the method of Eisen *et al.* is deterministic, most model-based methods are probabilistic searches for local extrema. Furthermore, these methods require care to avoid overfitting and require significant preprocessing to help reduce the computational time for high-dimensional data.

Mixture modeling is not widely applied to gene expression data. In general, the technique is more suited to classifying genes than it is to classifying conditions, such as in the Sorlie *et al*. study. The reason for this is that when classifying conditions, there are a small number of vectors (representing, e.g., $10\text{-}10^2$ tissues) in a high-dimensional space ($10^3\text{-}10^4$ genes), and the large nondiagonal covariance matrices used to describe the clusters in this space will frequently become singular during the EM estimation. Thus, without further modification, mixture modeling is best suited to cluster a large number of low-dimensional vectors, such as those from the Chu *et al*. data.

A common, widely applied strategy for model-based clustering is to assume that the data arise from a linear combination of multivariate Gaussian distributions[38,42,47], and employ an algorithm such as expectation-maximization[48] to calculate the parameters of such a combination that maximize the probability that the model generated the input data. This method of clustering allows the investigator to use knowledge of inherent physics or experimental experience to select a model that will produce the most informative results for the data of interest. Furthermore, because these methods are based on well-studied statistical models, other analytical problems such as selecting the number of clusters can also be approached from a fundamental statistical perspective. Because Gaussian distributions will model data that lacks inherent Gaussian behavior poorly, investigators

have applied a number of heuristic methods for accounting for non-Gaussian behavior such as strong outliers[32,40].

McLachlan *et al.* has proposed making use of mixtures of T-distributions (MoT), with multivariate form

$$T(x;\mu,\Sigma,\upsilon) = \frac{\Gamma(\frac{\nu+d}{2})\left|\Sigma\right|^{-1/2}}{(\pi\nu)^{\frac{1}{2}d}\Gamma(\frac{\nu}{2})\left\{1+\frac{\delta(x,\mu,\Sigma)}{\nu}\right\}^{\frac{1}{2}(\nu+d)}}$$

where (Eq. 4.3)

$$\delta(x,\mu,\Sigma) = (x-\mu)^T\Sigma^{-1}(x-\mu)$$

in order to model data that are significantly more noisy than those which are likely to be created by Gaussian distributions[39,49,50]. In a T-distribution, the parameter ν is known as the degrees of freedom. As this approaches infinity, the T-distribution tends to the Gaussian distribution (Fig. 4.2d), whereas in the limit of ν=1, the T-distribution tends to the Lorentzian, or Cauchy distribution (Fig. 4.2a).

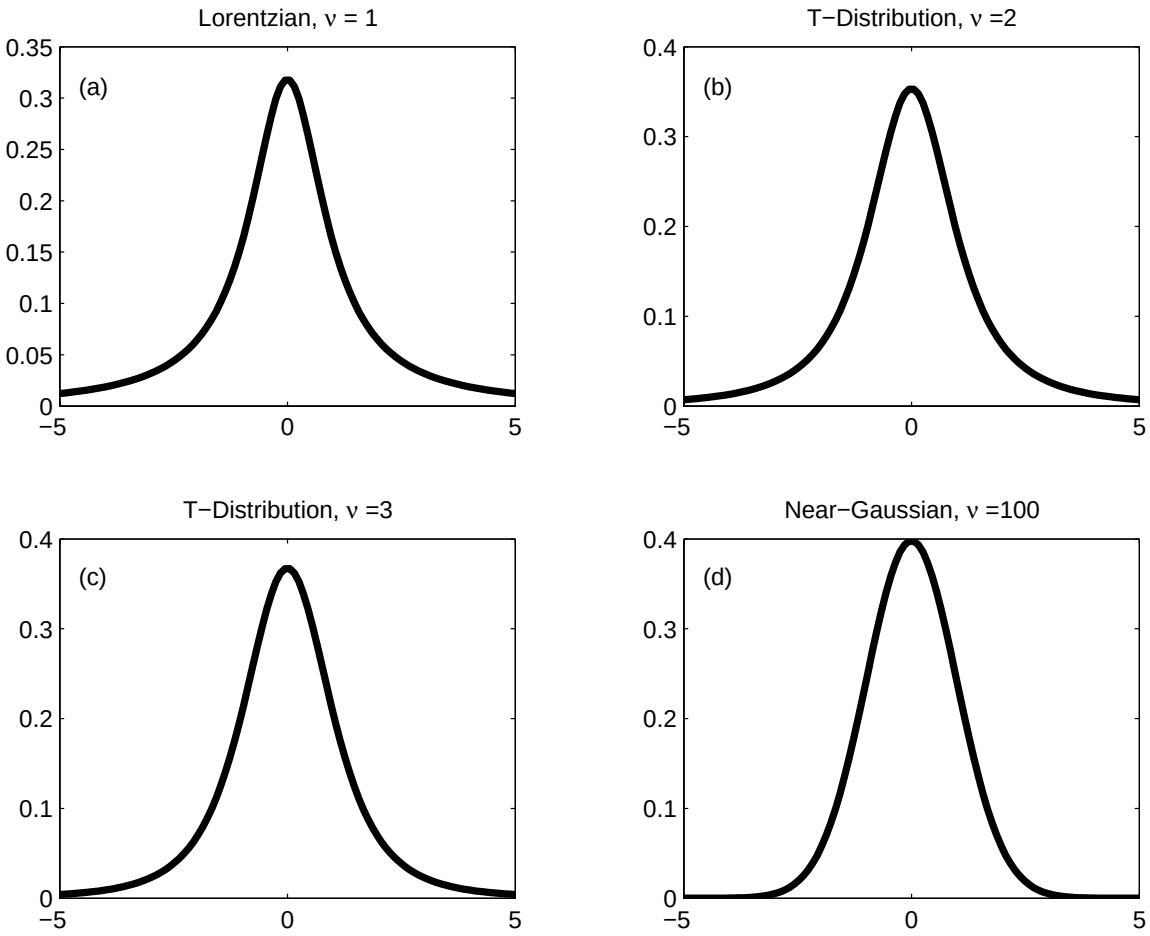


Figure 4.2: Univariate T-distributions at several degrees of freedom. This illustrates that as the number of degrees of freedom increase, the distributions become more Gaussian, and less permissive to outliers.

Studies suggest that gene expression data generated by cDNA microarrays have a generally Gaussian distribution[51]. The standard method of preprocessing popularized by Brown actually yields Lorentzian distributions because the final data is what is termed "ratio of medians," and involves dividing one Gaussian-distributed data set by another (see Appendix A). Despite this, it remains unknown what models serve to extract the most useful biological clusters from gene expression data. Even if machine noise were entirely eliminated, investigators would still choose different models depending on the information they hope to gain. Tolerance to outliers and dense noise remains a critical

requirement of any gene expression data clustering technique. This work describes making use of Mixtures of Lorentzian distributions (MoL) to consistently identify persistent clusters despite the presence of both types of noise.

4.2 Experimental

4.2.1 Algorithm for Clustering by Expectation-Maximization

Clustering algorithms that make use of expectation-maximization (EM) rely on Baye's rule,

$$p(x,d) = p(x|d)p(d) = p(d|x)p(x)$$
$$p(x|d) = \frac{p(d|x)p(x)}{p(d)} \qquad \text{(Eq. 4.4)}$$
$$p(d) = \int p(d|x)p(x)dx$$

as their fundamental basis. Heuristically, Baye's rule updates a prior hypothesis with posterior experimental knowledge. The EM algorithm is designed to continuously improve the likelihood of the data over a set of model parameters by alternating between two steps. The E-step calculates the log-likelihood of the full data set over the proposed parameters, and the M-step calculates a new set of parameters that maximize the log-likelihood of the E-step. This can be applied to optimize the parameters of a finite mixture of k Gaussian distributions (MoG) of weight π (MoG),

$$p(x|d) = \sum_{i=1}^{k} \pi_i G_x(\mu_i, \Sigma_i), \qquad \text{(Eq. 4.5)}$$

where μ_i is the mean and Σ_i is the covariance of the ith D-dimensional Gaussian distribution,

$$G_x(\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma (x-\mu)}. \quad \text{(Eq. 4.6)}$$

The E-step, which is a form of Eq. 4.4, requires calculating the posterior probability τ representing the responsibility for each point of each of the k Gaussian distributions,

$$\tau_{in} = \frac{G_{xn}(\mu_i \Sigma_i)\pi_i}{\sum_{j=1}^{k} G_{xn}(\mu_j \Sigma_j)\pi_j} \quad \text{(Eq. 4.7)}$$

The M-step,

$$\mu_i = \frac{\sum_{n=1}^{N} \tau_{in} x_n}{\sum_{n=1}^{N} \tau_{in}} \quad \text{(Eq. 4.8)}$$

$$\Sigma_i = \frac{\sum_{n=1}^{N} \tau_{in} (x_n - \mu_i)(x_n - \mu_i)^T}{\sum_{n=1}^{N} \tau_{in}} \quad \text{(Eq. 4.9)}$$

$$\mu_i = \frac{1}{N} \sum_{n=1}^{N} \tau_{in} \quad \text{(Eq. 4.10)}$$

updates the parameters over all data N, which are the means and covariances of the k Gaussian distributions, as well as the weighting factors, π. Convergence is measured by the fractional change in the log likelihood of the complete model over the data; it tends to zero as the algorithm reaches a maximum log likelihood. Initial values for the model parameters are provided either randomly or strategically by the investigator, and the

choice determines whether the EM algorithm will eventually reach a local or global maximum log likelihood.

The algorithm for Mixture of Lorentzians (MoL), adapted from McLachlan *et al.,*[39] differs slightly from that of MoG (Eqs. 4.7-4.10). The E-step requires calculating the posterior probability, τ, and a second weighting factor, u, which is a function of δ, the Mahalanobis squared distance between x and μ.

$$\tau_{in}^{(k)} = \frac{\pi_i^{(k)} L_{xn}(\mu_i^{(k)}, \Sigma_i^{(k)})}{\sum_{m=1}^{K} \pi_m^{(k)} L_{xn}(\mu_m^{(k)}, \Sigma_m^{(k)})} \qquad \text{(Eq. 4.11)}$$

$$u_{in}^{(k)} = \frac{1+d}{1+\delta(x_n, \mu_i^{(k)}, \Sigma_i^{(k)})} \qquad \text{(Eq. 4.12)}$$

Once these are computed for the k^{th} iteration, the M-step is performed to update the mean and scatter. The weighting vector π is updated in the same manner as MoG.

$$\mu_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} x_j}{\sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)}} \qquad \text{(Eq. 4.13)}$$

$$\Sigma_i^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{ij}^{(k)} u_{ij}^{(k)} (x_j - \mu_i^{(k+1)})(x_j - \mu_i^{(k+1)})^T}{\sum_{j=1}^{n} \tau_{ij}^{(k)}} \qquad \text{(Eq. 4.14)}$$

The method of McLachlan *et al.* was further adapted by adding a heuristic cluster deletion algorithm to handle a problem of EM clustering of multivariate data. In both MoG and MoL, the covariance matrix must be inverted, and if the covariance of a distribution has become close to singular, the calculation will fail. This problem most commonly arises when the observations assigned to one cluster form a lower-dimensional

linear subspace. In order to overcome this, in our algorithm, when a cluster is nearly singular, it is deleted and the weighting factors for the remaining clusters are recomputed. This has the consequence of allowing the log likelihood to decrease between steps of the EM algorithm. I refer to these algorithms as adaptive mixture of Gaussians or mixture of Lorentzians.

4.2.2 Computational Methods

All experiments were conducting using MATLAB Release 13 (The Mathworks, Inc.), on one of several machines: Macintosh G4 running System X, IBM Pentium M running Windows XP Pro, Dual Pentium III running Redhat Linux, or Sun Ultra 60 running Solaris 8. Complete code for the Adaptive Mixture of Lorentzians, Adaptive Mixture of Gaussians, and utility software are in appendix B.

4.2.1 Sources of Data

All the data analyzed in this study were obtained from the Stanford Microarray Database (http://genome-www5.stanford.edu/). In particular, the yeast data of Chu *et al.*[7] served as an example of a richly populated, low-dimensional space, and the extensive study of human breast cancer first described by Perou *et al.*[52] served as an example of a sparsely populated, high-dimensional space. Specifically, the breast cancer data were taken from Sorlie *et al.*[10], supplemental table 6, and the yeast data was taken from the

entire data set of Chu *et al*. No preprocessing beyond that embodied by particular algorithms was applied (for example, PCA requires centering of data).

4.3 Results and Discussion

4.3.1 Andrews Curves

The study of Chu *et al*. uses cDNA microarrays to measure the expression of 6118 genes of *Sacchraromyces cerevisiae* at seven time points during sporulation: 0, 0.5, 2, 5, 7, 9, and 11 hours. The genes that showed the greatest induction or repression during the experiment were classified into seven groups: Metabolic (52 genes), Early I (62), Early II (47), Early-Mid (95), Middle (158), Mid-Late (61), and Late (5). For each of these classes, a subgroup of representative genes was used to create average expression patterns for the class (Table 4.1).

Metabolic	Early I	Early II	Early-Mid	Middle	Mid-Late	Late
ACS1	ZIP1	KGD2	YBL078C	YSW1	CDC27	SPS100
PYC1	YDR374C	AGA2	QRI1	SPR28	DIT2	YKL050C
SIP4	DMC1	YPT32	PDS1	SPS2	DIT1	YMR322C
CAT2	HOP1	MDR1	APC4	YLR227C		YOR391C
YOR100C	IME2	SPO16	KNR4	ORC3		
CAR1		NAB4	STU2	YLL005C		
		YPR192W	YNL013C	YLL012W		

Table 4.1: Genes used by Chu *et al.* to create average expression patterns for each of seven classifications.

I process the data in their entirety using principal components analysis (PCA) implemented by singular value decomposition (SVD). In this case, the data are represented by 6118 vectors in 7-dimensional space. The PCA analysis does not reduce

the length of the vectors. The resulting 7 orthogonal subspaces are ranked in order of the fraction of the variance they account for in Figure 4.3.

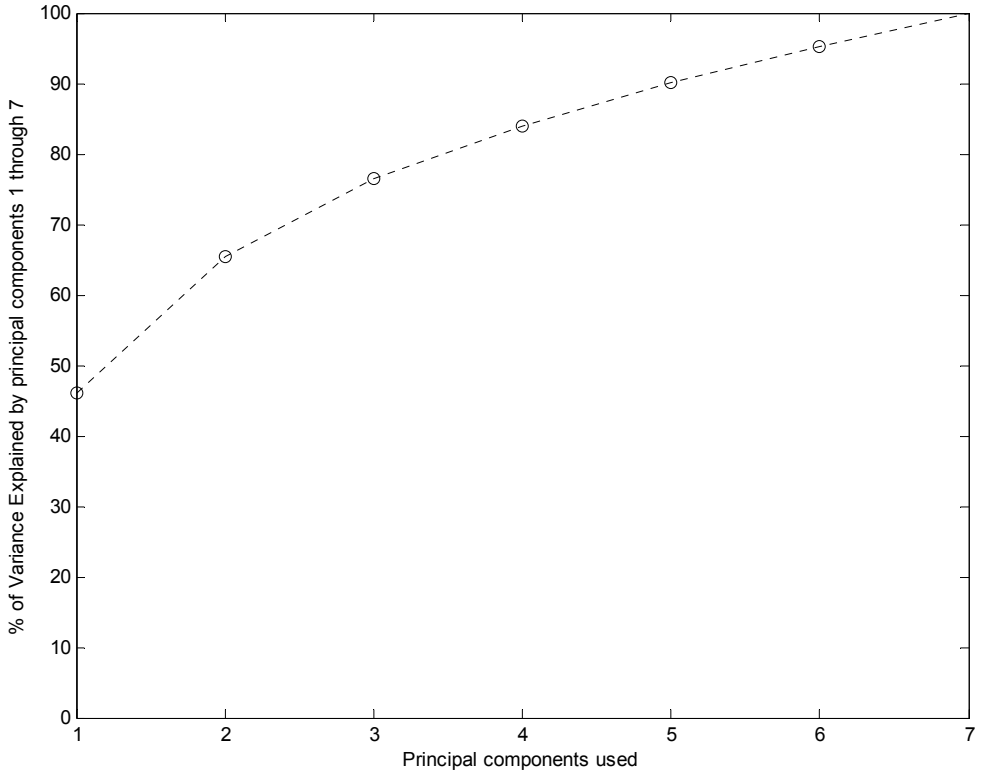


Figure 4.3: The full data of Chu *et al.* is processed by PCA implemented with SVD. The resulting eigenvalues are used to show the fraction of the total variance explained as more principal components (eigenvectors) are included.

Chu *et al.* present the seven-dimensional data using the method of Eisen *et al.* described earlier, by assigning the expression level a color on a map from red (induced) to green (repressed) with black indicated an unchanged expression level. Our alternative to this using Andrews curves maps the PCA-arranged vectors for the genes of interest onto an orthogonal Fourier subspace that preserves the mathematical relationships between the original vectors and allows viewers to properly infer the distance between vectors as the distance between lines on the Andrews plot.

The first visualization task is to distinguish clustered vectors from unclustered vectors. This is demonstrated in Figure 4.4, where six clustered metabolic genes (ACS1, PYC1, SIP4, CAT2, ORF YOR100C, CAR1) yield proximal Andrews curves, and three randomly selected, unclustered genes (ORFs YAR052C, YAR053W, YAR060C) do not. This visualization can be accomplished with as few as three principal components (Fig. 4.5).

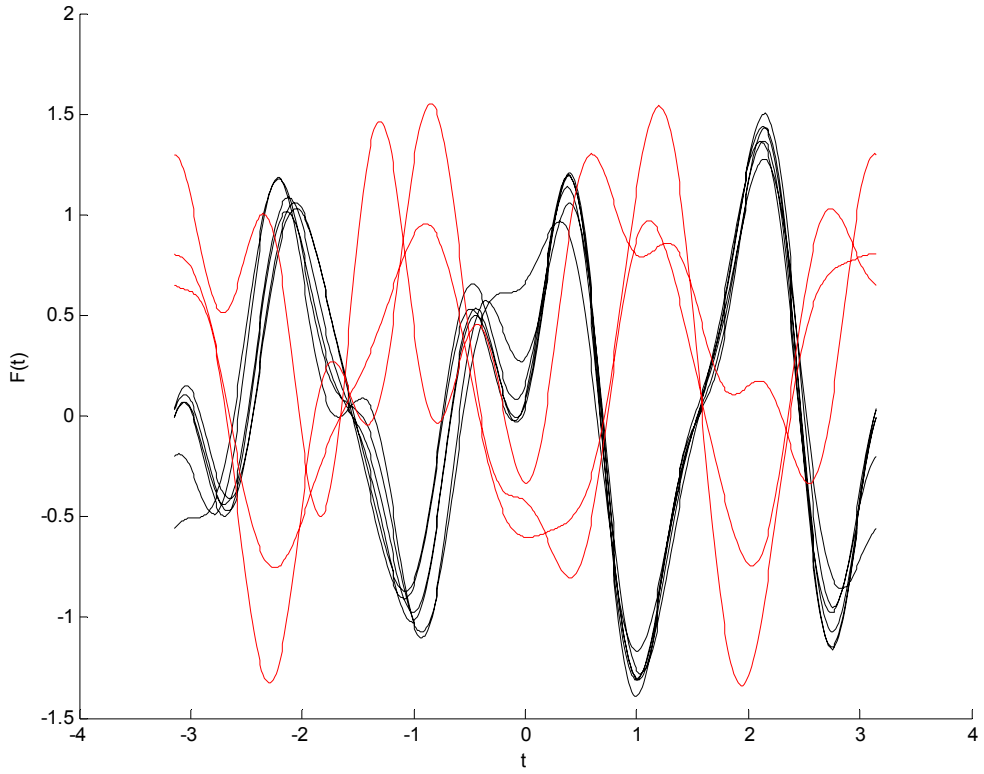


Figure 4.4: After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted. In black, six genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4, CAT2, ORF YOR100C, and CAR1), and in red, three random genes (ORFs YAR052C, YAR053W, and YAR060C). Here, all seven principal components are used for the Andrews plot.

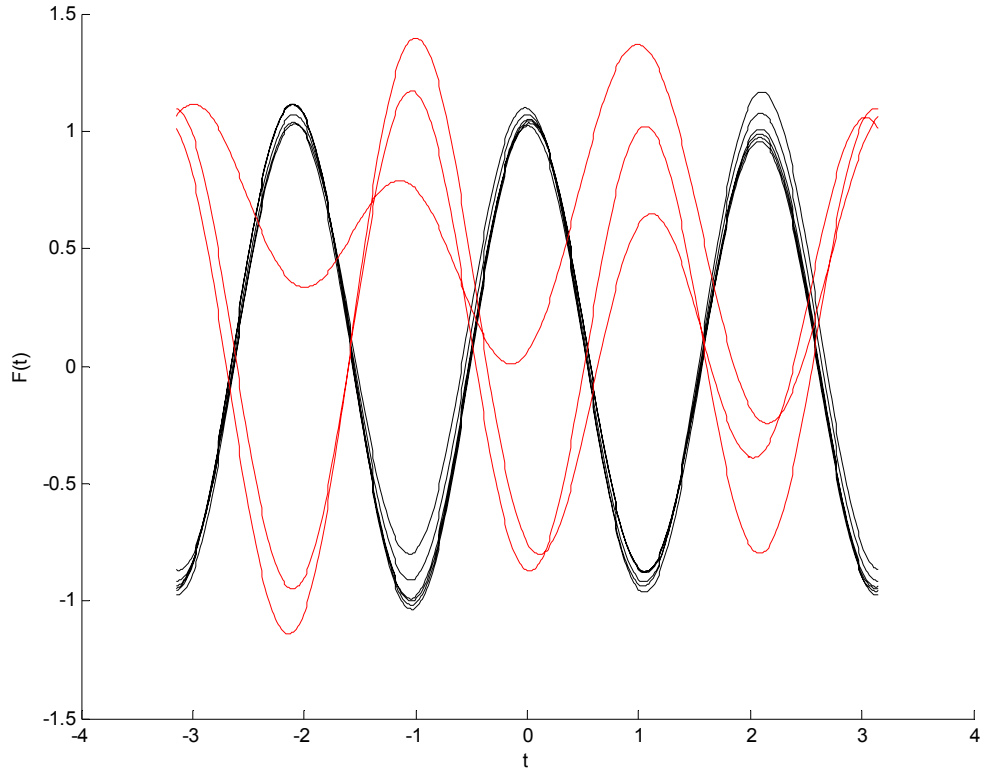


Figure 4.5: After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted. In black, six genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4, CAT2, ORF YOR100C, and CAR1), and in red, three random genes (ORFs YAR052C, YAR053W, and YAR060C). Here, only **three** principal components are used for the Andrews plot.

The second visualization task is to distinguish one group of clustered vectors from another similarly clustered group. In Figure 4.6, Andrews curves are used to distinguish 3 metabolic genes (ACS1, PYC1, SIP4) from 3 middle genes (YSW1, SPR28, SPS2).

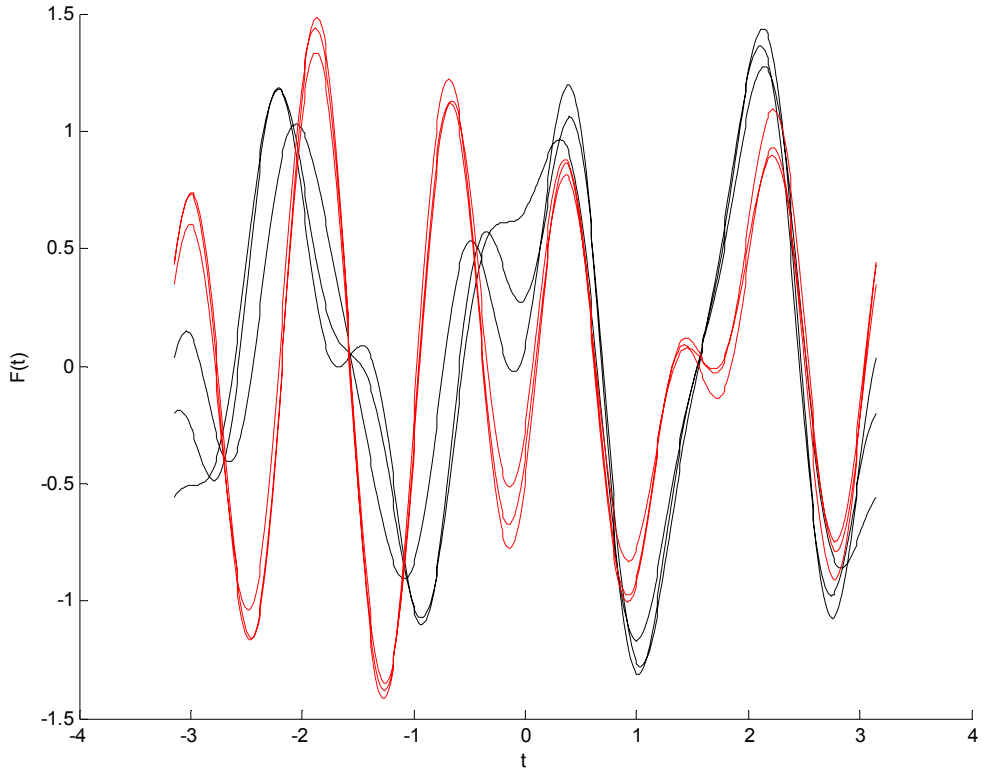


Figure 4.6: After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted. In black, three genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4), and in red, three genes designated as belonging to the middle class (YSW1, SPR28, SPS2). Here, all seven principal components are used for the Andrews plot.

The effect of altering the number of principal components used for Andrews plot is illustrated in Figure 4.7, where a different, but usable perspective is presented using between 4 and 7 principal components. It is especially helpful to examine a range of principal components when distinguishing between groups of proximal vectors. For example, if three genes from the middle class (ORC3, ORF YLL005C, ORF YLL012W) are compared to three genes from the mid-late class (CDC27, DIT2, DIT1), the resulting Andrews plots vary widely in their usefulness across the range of 4 to 7 principal components (Fig. 4.8).

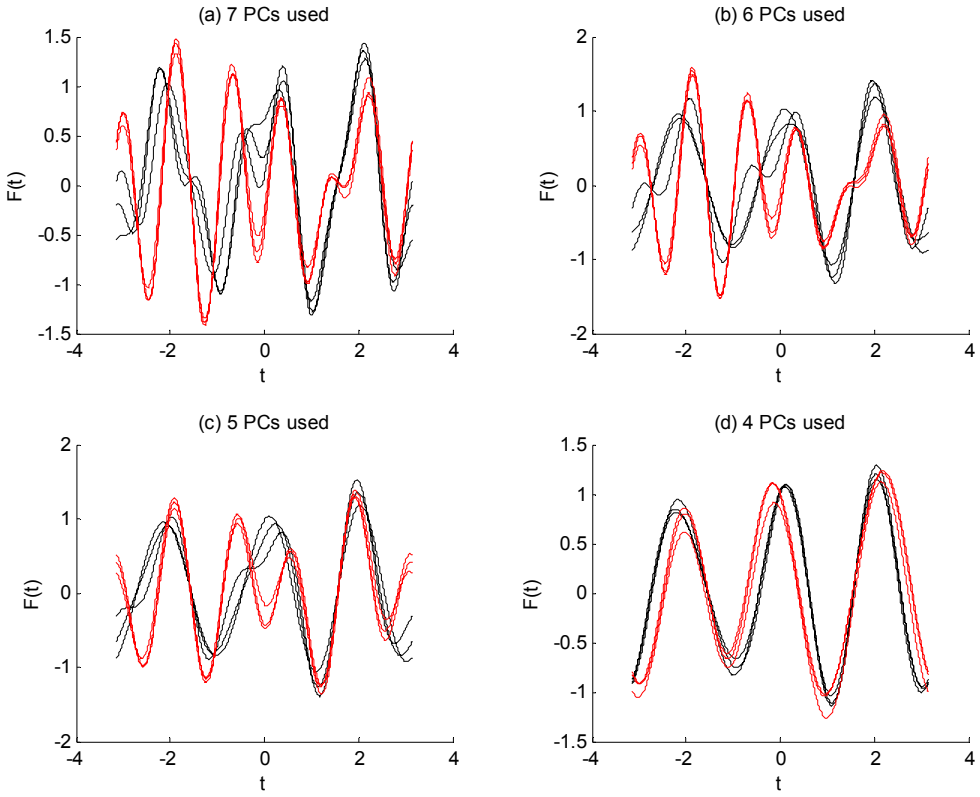


Figure 4.7: After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted. In black, three genes designated as belonging to the metabolic class (ACS1, PYC1, SIP4), and in red, three genes designated as belonging to the middle class (YSW1, SPR28, SPS2). Here, the data are plotted for four choices of number of principal components.

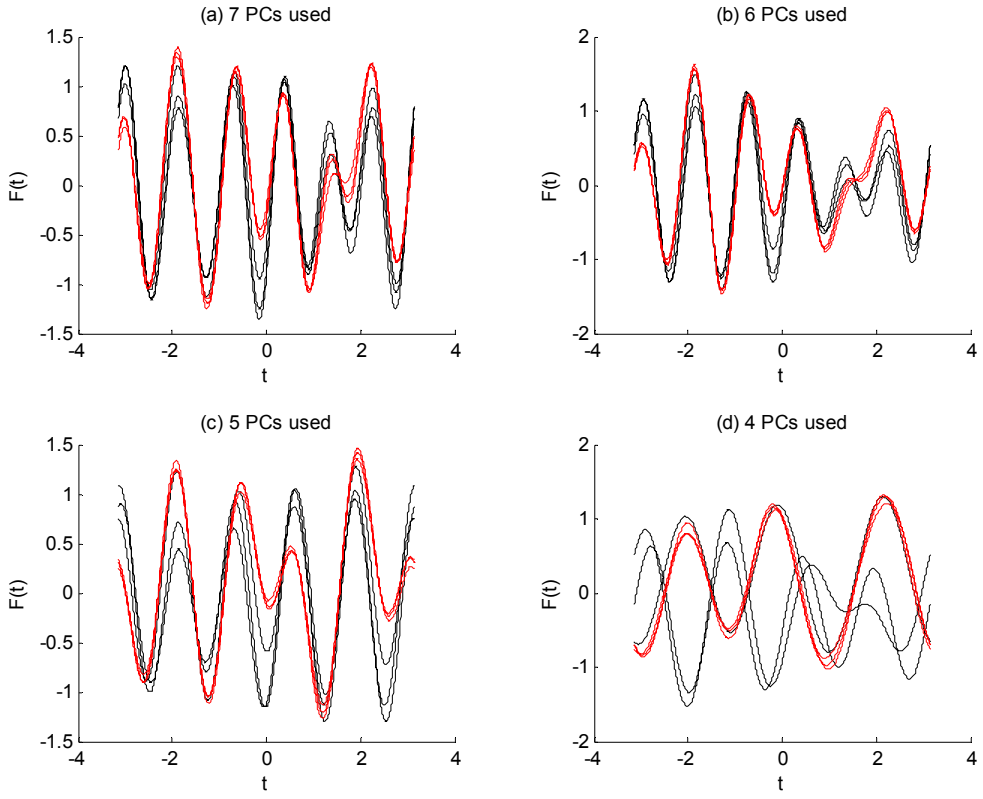


Figure 4.8: After PCA processing, the data of Chu *et al.* is mapped onto the Andrews space and plotted. In black, three genes designated as belonging to the mid-late class (CDC27, DIT2, DIT1), and in red, three genes designated as belonging to the middle class (ORC3, ORF YLL005C, ORF YLL012W). Here, the data are plotted for four choices of number of principal components.

I also apply the method of Andrews to the reduced breast cancer data of Sorlie *et al.*, which are represented as a set of 122 vectors in 552-dimensional space. All of the microarray analyses conducted using the 122 tissue samples measured the expression level of many thousands of genes, but the data were subsequently reduced to a subset of 552 genes that met Sorlie *et al.*'s definition of "intrinsic." Briefly, these genes' expression levels vary greatly among patients, but minimally between pairs of samples drawn from the same patient. The investigators robustly cluster the tumor data into five classes: Normal, Luminal A, Luminal B, Basal, and ERBB2+. These classes strongly

correlate with patient outcome as measured by time to distant metastasis. Outcomes are best for patients whose tumors are classified as Luminal A, and become progressively worse for Luminal B, Basal, and are the worst for ERBB2+.

The 122x552-dimensional space is sparsely populated, and cannot support more than 122 independent linear subspaces. Further, it will be necessary to reduce the number of dimensions to fewer than 10, the point where Andrews curves become intelligible. Thus, for these data, PCA will accomplish significant dimensional reduction, as well as ordering the eigenvectors in order of statistical importance. Compared to the data of Chu *et al.*, the Sorlie *et al.* data is much less easily explained by a small number of principal components (Fig. 4.9).

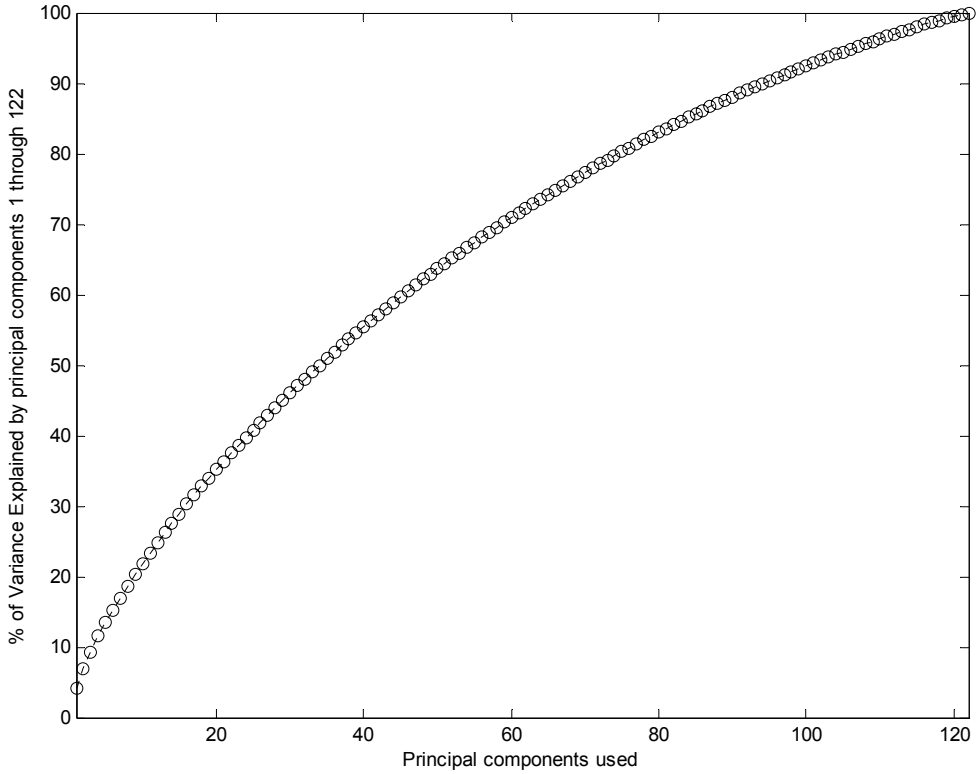


Figure 4.9: Plot of analysis of PCA reduction of data of Sorlie *et al.* implemented with SVD. The resulting eigenvalues are used to show the fraction of the total variance explained as more principal components (eigenvectors) are included.

PCs	Percent of total variance explained
1	4.20
2	6.92
3	9.31
4	11.5
5	13.5
6	15.3
7	17.0
8	18.7
9	20.3
10	21.9

Table 4.2: The reduced data of Sorlie *et al.* is processed by PCA implemented with SVD. The largest number of principal components (PCs) that can realistically be visualized with Andrews curves accounts for approximately 21.9% of the variance in the data.

Despite the limited explaining power of 1-10 principal components for the Sorlie *et al.* data, Andrews curves can still helpfully visualize tumor classes. When applied to three tumors from the Luminal A class (Norway FU15-BE, Norway FU37-BE, Norway FU16-BE) and three tumors from the ERRB2+ class (Northway FU18-BE, Norway FU04-BE, Norway 65-2ndT), the resulting Andrews curves vary in usefulness over the range from 4 to 7 principal components used (Fig. 4.10); notably, 5 principal components (Figure 4.10c) may offer more visual appeal than 4, 6 and 7. Similarly, in Figure 4.11, three tumors from the Normal class (Benign STF 37, Benign STF 20, Benign STF 11 can be distinguished from three from the Basal class (Norway FU12-BE, Norway FU23-BE, Norway FU39-BE), but most easily with 4 principal components (Fig. 4.11d).

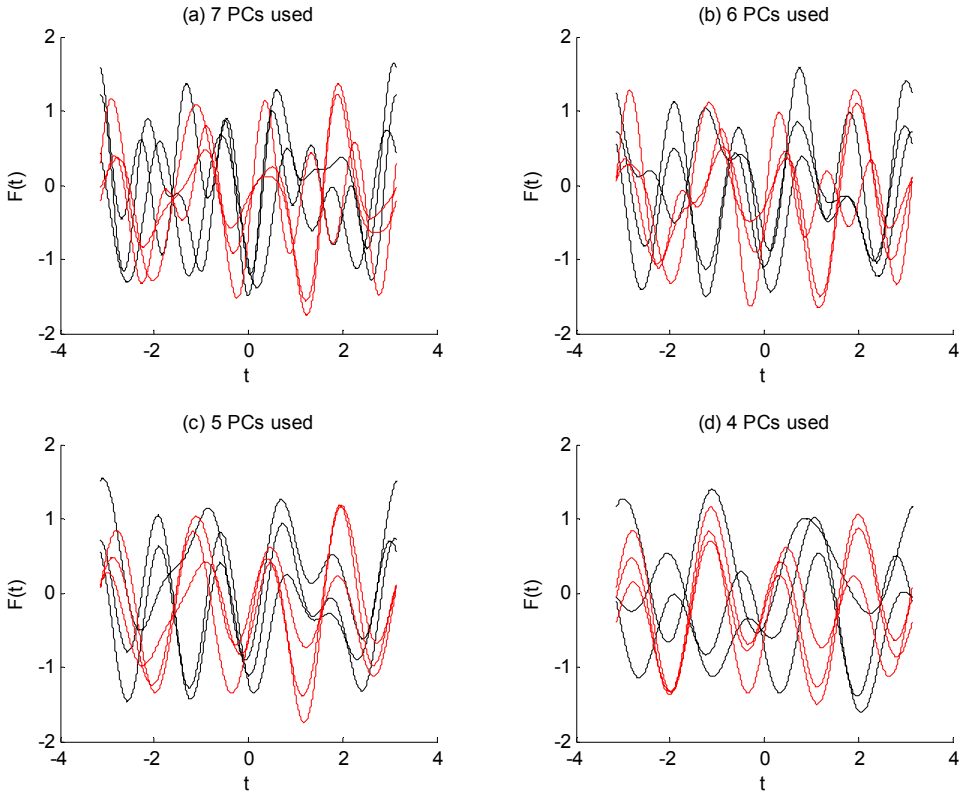


Figure 4.10: After PCA processing, the data of Sorlie *et al.* is mapped onto the Andrews space and plotted. In black, three tumors designated as belonging to the Luminal A class (Norway FU15-BE, Norway FU37-BE, Norway FU16-BE), and in red, three tumors from the ERRB2+ class (Northway FU18-BE, Norway FU04-BE, Norway 65-2ndT). Here, the data are plotted for four choices of number of principal components.

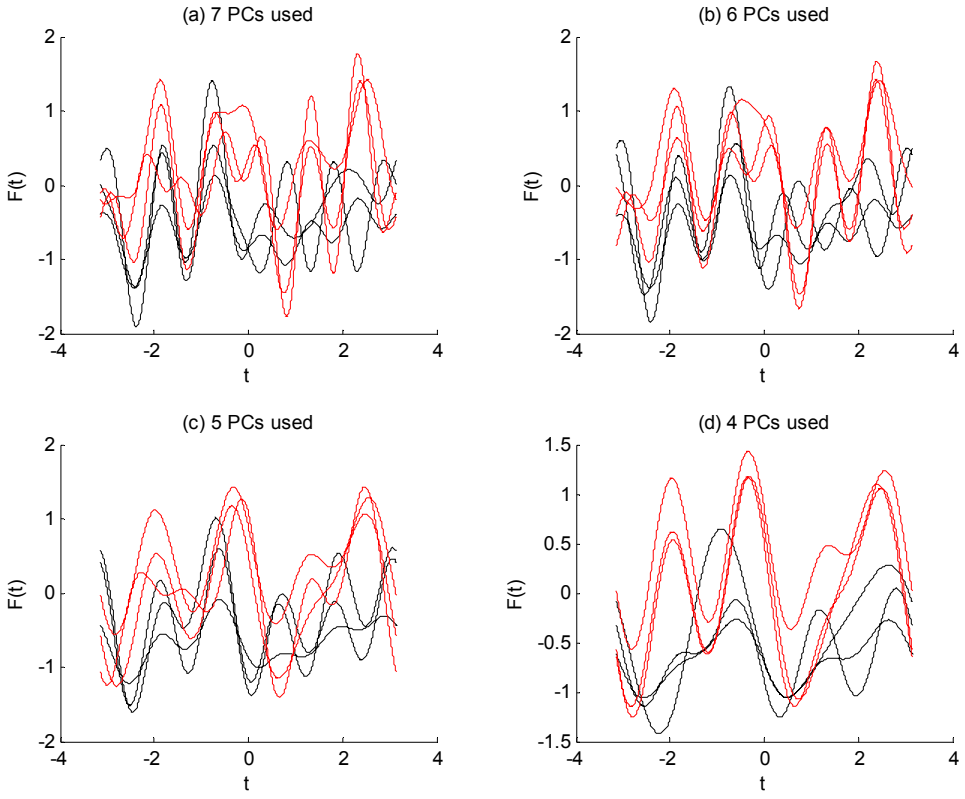


Figure 4.11: After PCA processing, the data of Sorlie *et al.* is mapped onto the Andrews space and plotted. In black, three tumors designated as belonging to the Normal class (Benign STF 37, Benign STF 20, Benign STF 11), and in red, three tumors from the Basal class (Norway FU12-BE, Norway FU23-BE, Norway FU39-BE). Here, the data are plotted for four choices of number of principal components.

4.3.2 EM Clustering of Synthetic Data

In order to illustrate the differences between the MoG and MoL algorithm, experiments were conducted on a series of strategically designed artificial 2-dimensional data sets. The first of these, Artificial Data Set 1 (ADS1), uses the data depicted in Figure 4.12 to illustrate the difficulty MoG has in locating the mean of a persistent cluster of points when noise is present. The MoG result shown in Figure 4.13(a) is a large Gaussian distribution with a mean far from the mean of the main cluster of points. This

happens because the Gaussian distribution pays a large likelihood penalty for distant outliers. On the other hand, the MoL result shown in Figure 4.13(b) correctly identifies the mean of the persistent cluster.

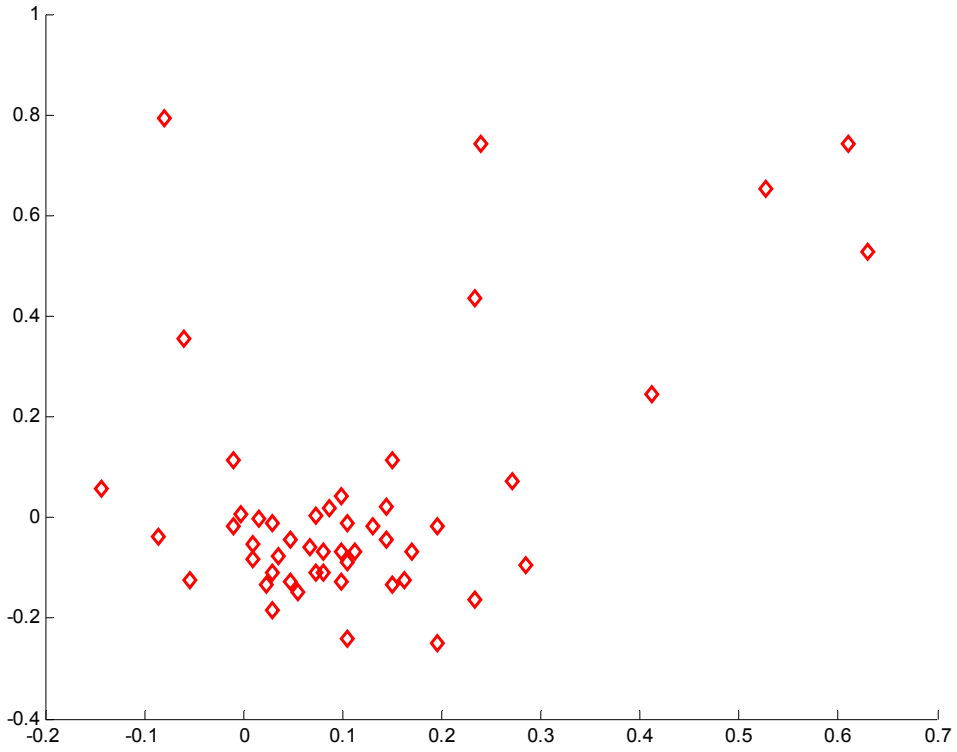
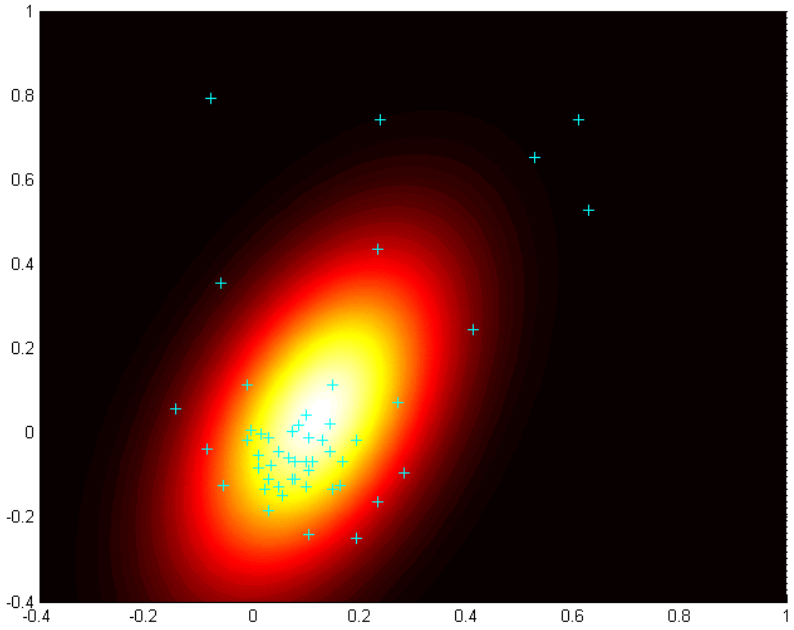
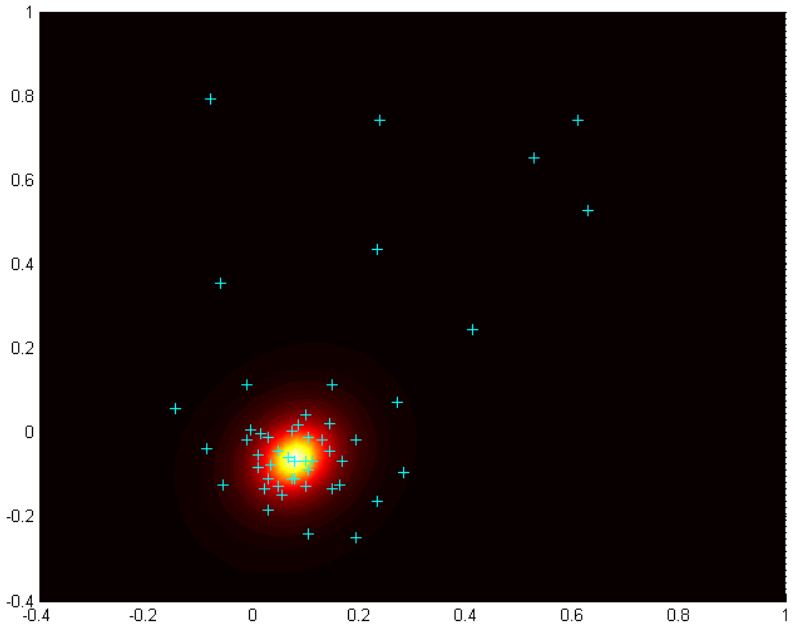


Figure 4.12: Artificial Data Set 1, which contains one persistent cluster and a small number of outliers.



(a)



(b)

Figure 4.13: Artificial Data Set 1, after execution of the EM-MoG (a) or EM-MoL (b) algorithm. In (a), the probability density function for the Gaussian distribution must be quite large in order to explain the distant points, where as the Lorentzian PDF in (b) located the mean of the persistent cluster and ignores the distant points.

The second artificial data set (ADS2, Fig. 4.14) is designed to further illustrate the manner in which the MoG algorithm is confounded by a small number of outliers. In this case, it is clear to the naked eye that there are two clusters. In 100 trials, the MoG algorithm failed to find the true clusters every time, as shown in Figure 4.15(a), whereas the MoL algorithm successfully found the true clusters every time, as shown in Figure 4.15(b). The plots of the PDFs indicate that the MoG algorithm (Fig. 4.16(a)), is forced to include all of the outliers in one extremely diffuse cluster in order to account for them, whereas the MoL algorithm (Fig. 4.16(b)) results in two tight PDFs.

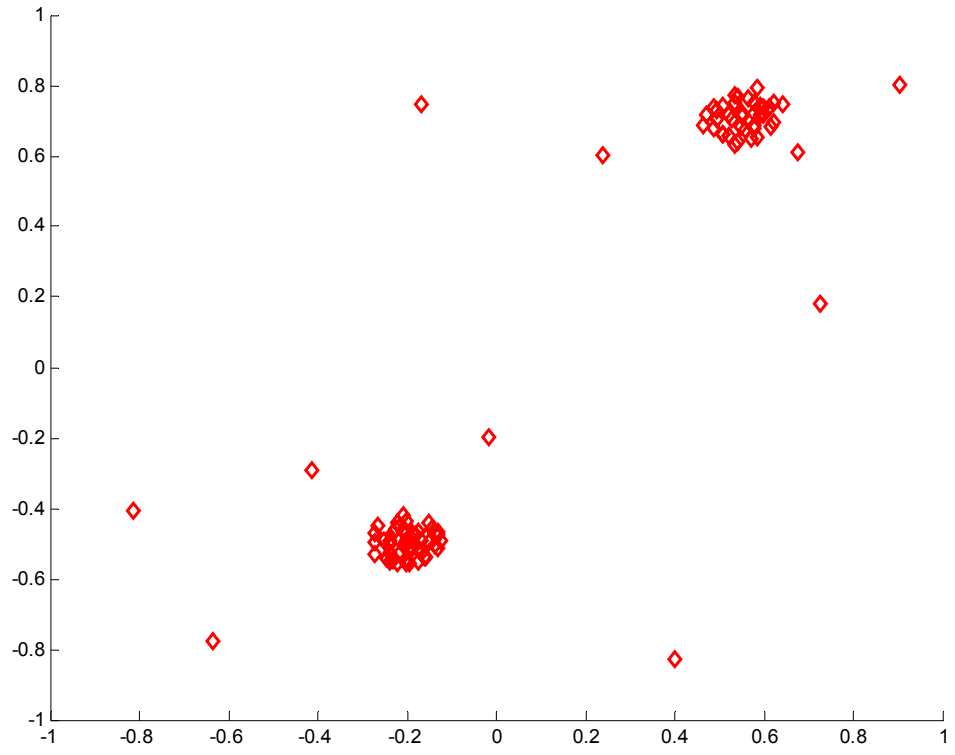
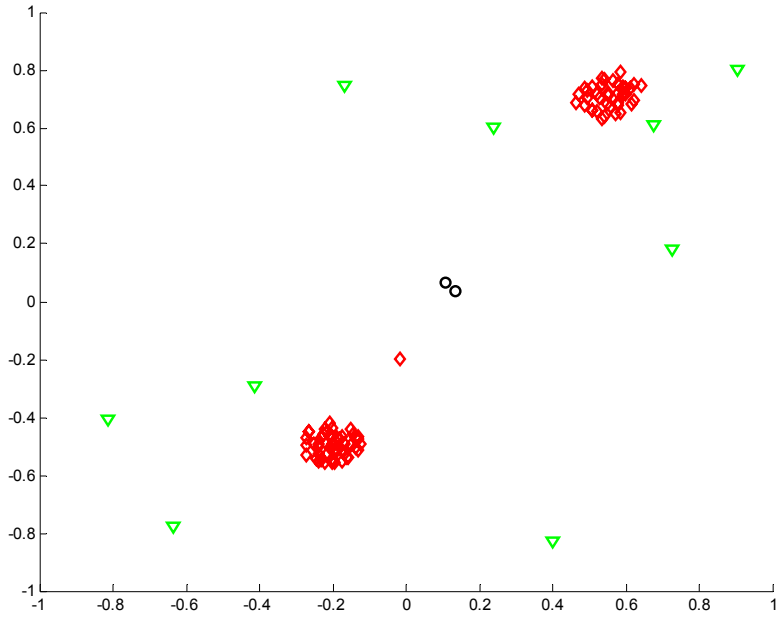
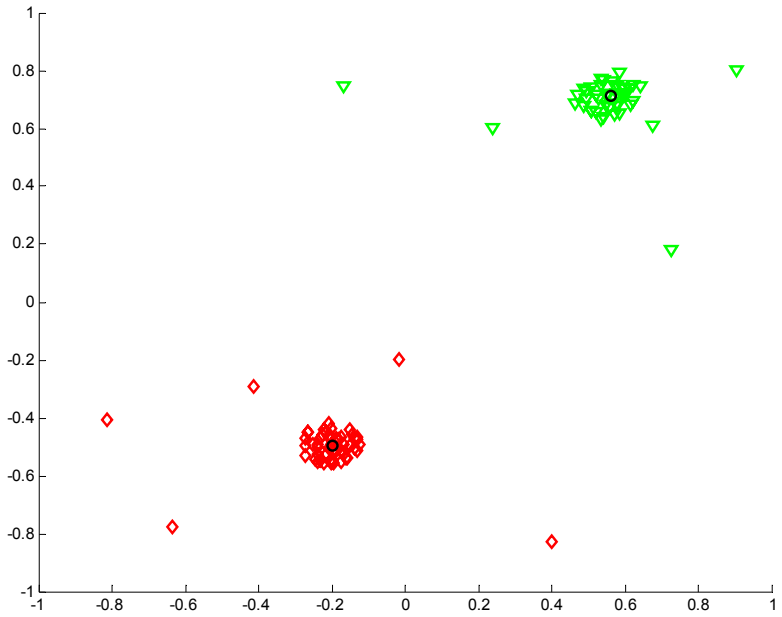


Figure 4.14: Artificial Data Set 2, which contains two tight clusters and a small number of outliers.



(a)



(b)

Figure 4.15: Artificial Data Set 2, after execution of the EM-MoG (a) or EM-MoL (b) algorithm. In (a), the MoG algorithm is confounded by the outliers and reaches a maximum likelihood at two clusters with their means (black circles) between the true clusters. In (b), the MoL algorithm correctly identifies the true clusters.

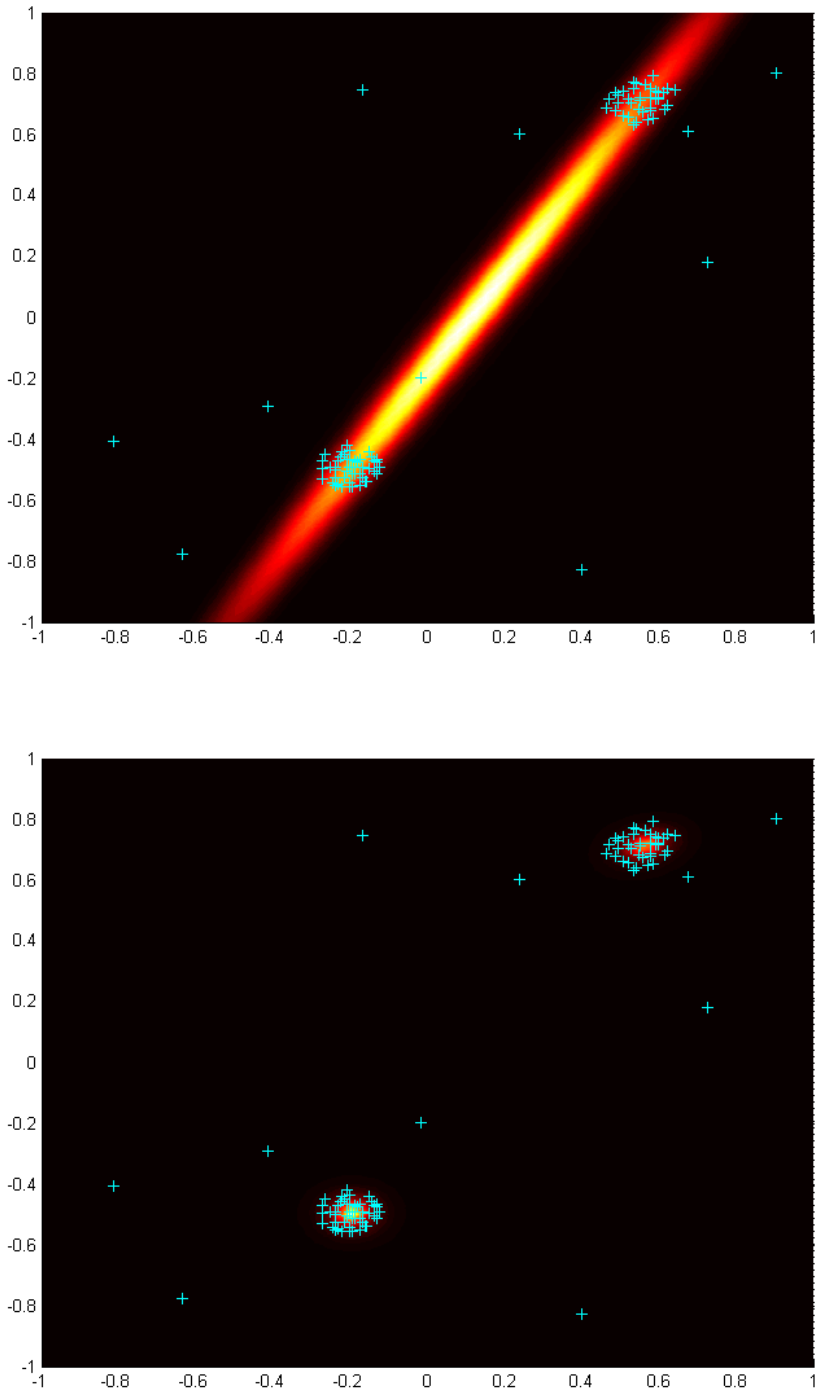


Figure 4.16: Artificial Data Set 2, after execution of the EM-MoG (a) or EM-MoL (b) algorithm. In (a), the PDFs of the two Gaussian clusters overlap along the centerline of the data, whereas in (b) the PDFs are small and centered at the true clusters.

The third artificial data set (ADS3, Fig. 4.17) is designed to illustrate the confounding effect of dense noise on the MoG algorithm. Unlike ADS1 and ADS2, the non-deterministic nature of EM-MoG and EM-MoL is an important factor for ADS3. Depending on the initial choices for the mean and covariance, the EM algorithm will converge to a variety of local maxima. For ADS3, each algorithm was executed 100 times with random initializations. The most probable outcomes (Fig. 4.18) are successful identification of the "faint" clusters by MoL, and complete failure by MoG. The PDFs corresponding to these outcomes (Fig. 4.19) verify that MoL correctly identifies the persistent clusters and MoG is confounded by the dense noise.

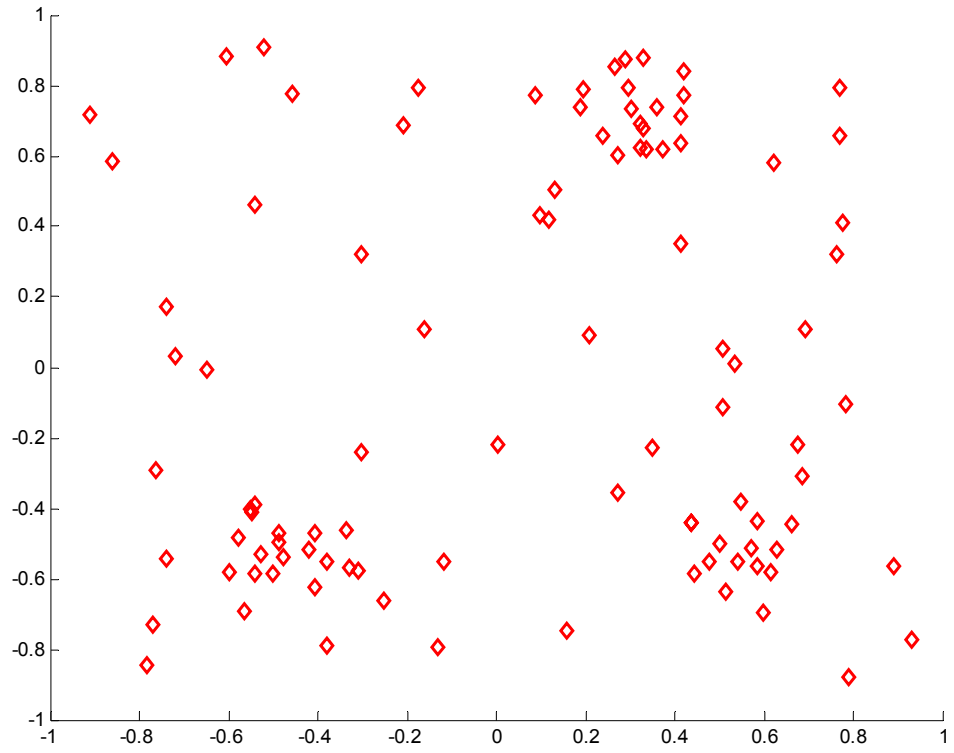
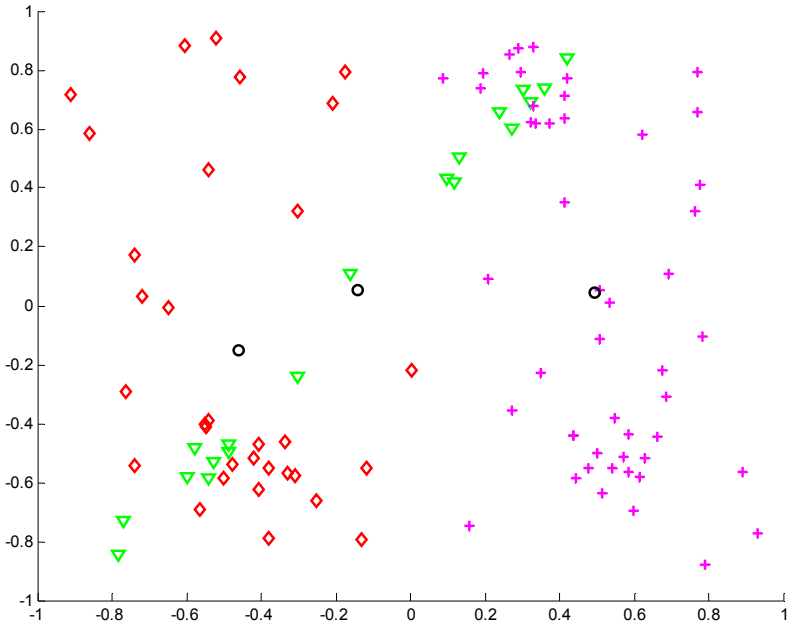
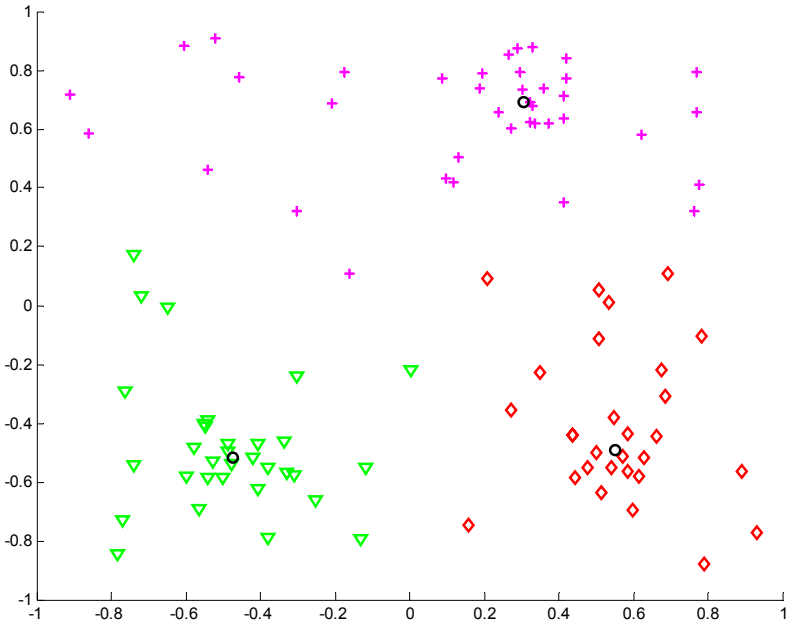


Figure 4.17: Artificial Data Set 3, which contains three faint clusters and a large amount of dense noise.

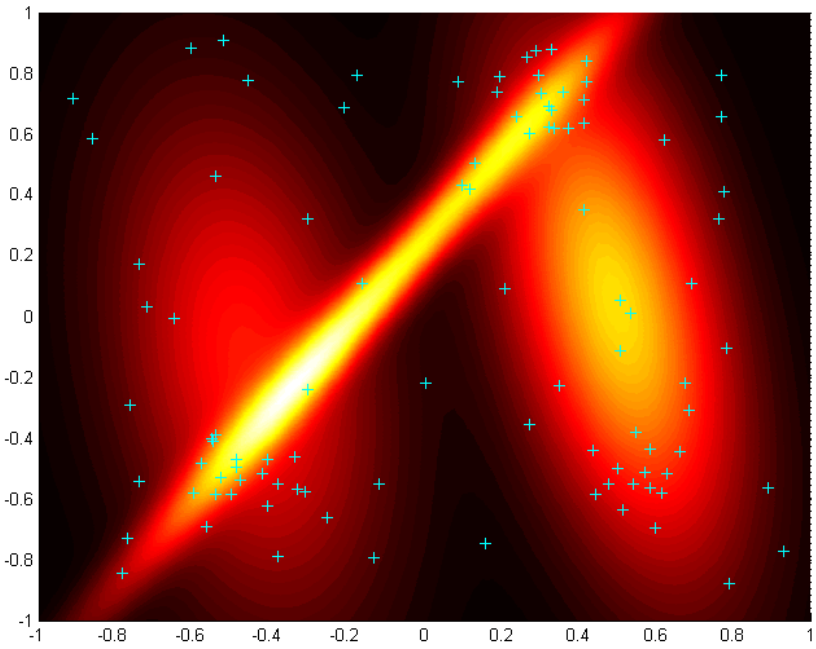


(a)

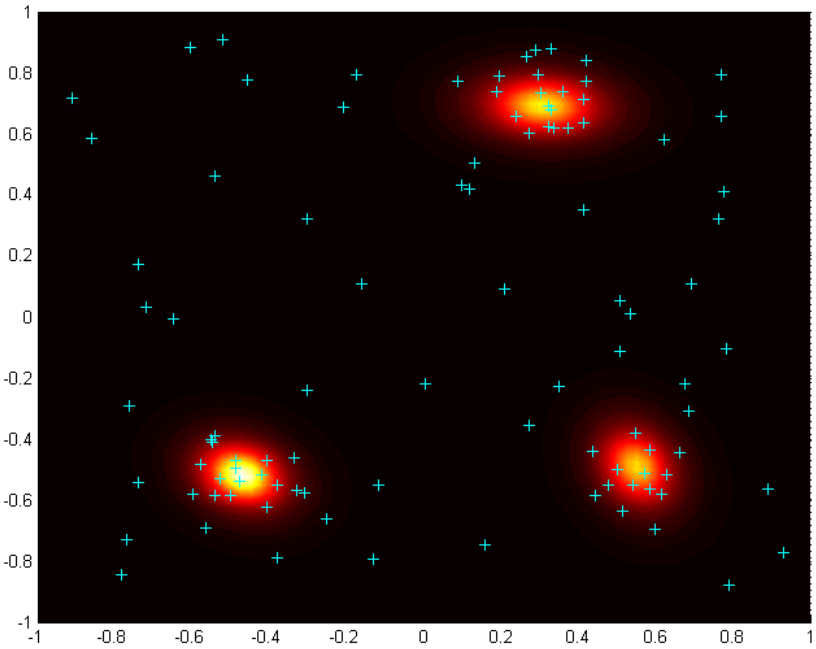


(b)

Figure 4.18: Artificial Data Set 3, after execution of the EM-MoG (a) or EM-MoL (b) algorithm, with the most probable outcome shown. In (a), the MoG algorithm is confounded by the noise and provides three clusters with incorrect means (black circles). In (b), the MoL algorithm correctly identifies the true clusters, although the assignment of points near the borders is heuristically arbitrary.



(a)



(b)

Figure 4.19: Artificial Data Set 3, after execution of the EM-MoG (a) or EM-MoL (b) algorithm. In (a), the PDFs of the three Gaussian clusters spread across the data in an unintuitive manner, whereas in (b) the PDFs are small and centered at the true clusters.

ADS3 also provides an opportunity to test the robustness of the EM-MoL and EM-MoG algorithms to random initial conditions. This may be especially important for gene expression data, when unsupervised clustering is often done in the absence of prior information for the initialization. In 100 trials (Fig. 4.20) the MoL algorithm identifies the correct clusters nearly every time (Fig. 4.21(a)), whereas the MoG algorithm rarely identifies the correct clusters (Fig. 4.21(b)).

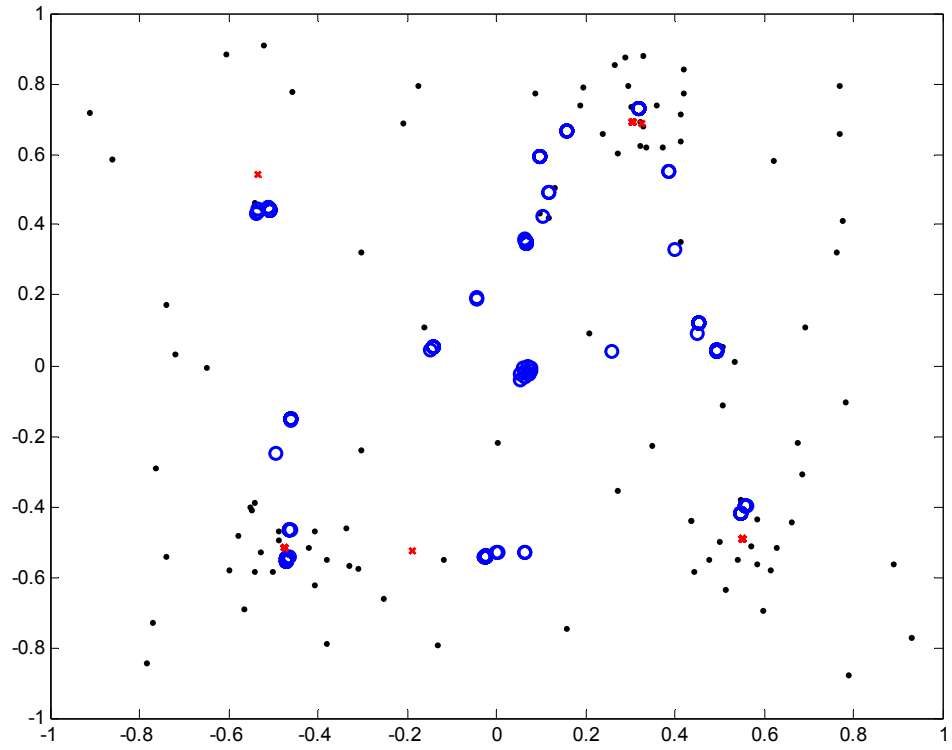
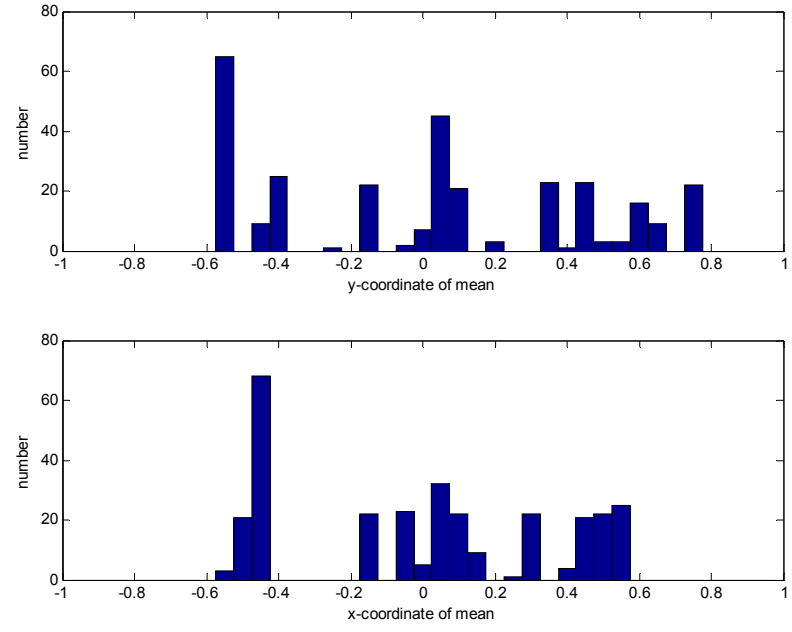
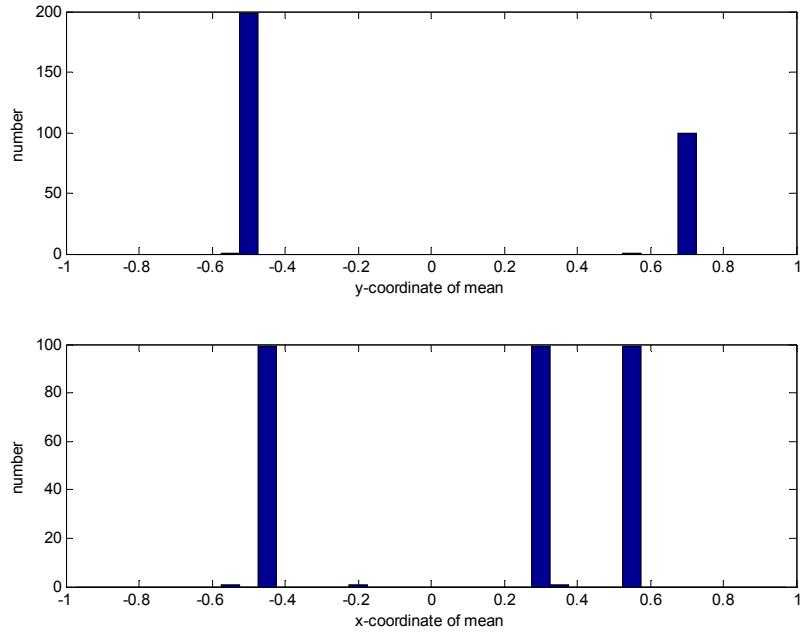


Figure 4.20: Artificial Data Set 3, after execution of the EM-MoG and EM-MoL algorithm over 100 trials. The data points are the black dots, the means proposed by MoG are the blue circles, and the means proposed by MoL are the red crosses.



(a)



(b)

Figure 4.21: Artificial Data Set 3 was processed by 100 trials of the EM-MoG (a) and EM-MoL (b) algorithms. The means of the proposed PDFs were binned into a histogram. Because the y-coordinates of the means of two of the true clusters are very similar, the y-coordinate histogram of (b) has a 200-count bin.

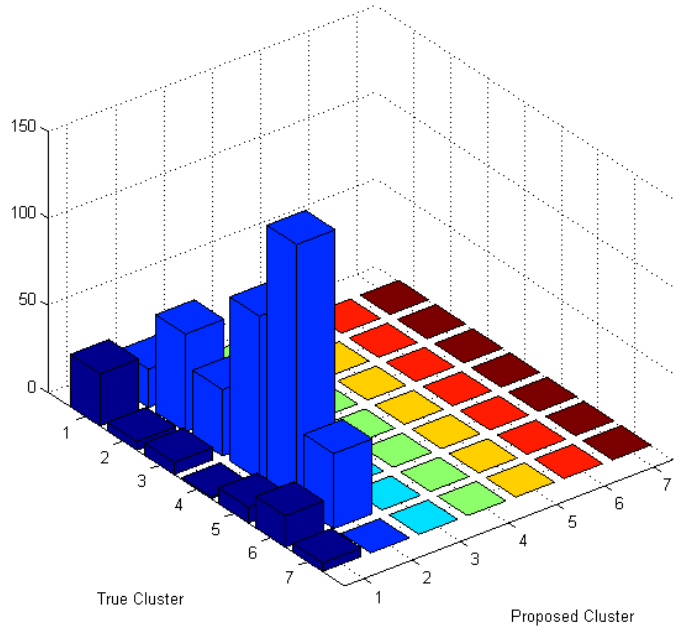
In most artificial data sets, the MoL algorithm not only produces heuristically superior results, but it also produces higher values for the log likelihoods of the proposed models when compared to MoG. This may not be a good way to compare the relative performance of MoG and MoL across all data sets. The Gaussian distribution gives very large likelihood bonuses to points near its mean, since the distribution is quite peaked. Thus, even for clusters with outliers, the MoG likelihood may be higher than the MoL likelihood, even when the means are mislocated.

4.3.3 EM Clustering of Gene Expression Data

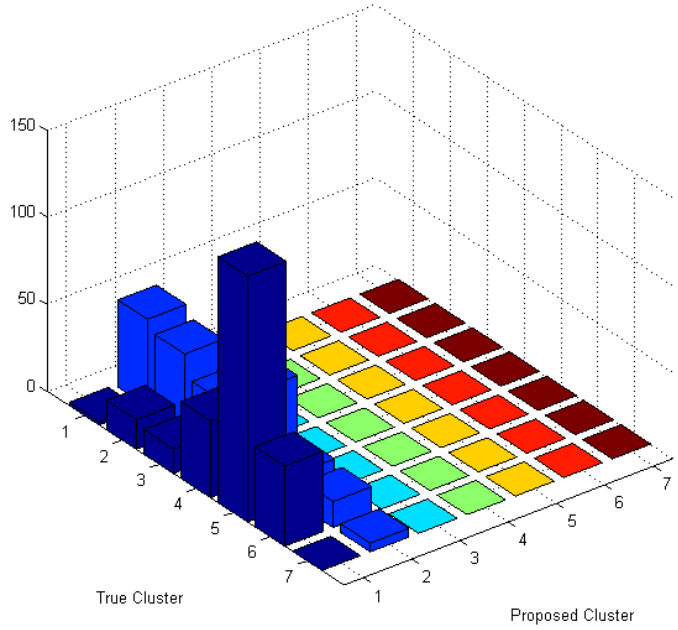
I also test the performance of the EM-MoG and EM-MoL algorithms on a reduced form of the data set from the Chu *et al.* study. The data were limited to 477 genes that were identified by Chu *et al.* as belonging to one of the seven temporal classes defined by a common time of induction. The seven true clusters are numbered from 1 to 7, Metabolic, Early I, Early II, Mid-Early, Middle, Mid-Late, and Late. Mixture models will not be able to mimic this classification perfectly without further preprocessing, because two genes with similar times for induction may have quite different overall temporal profiles, and thus distant vectors in seven-dimensional space.

With this potential limitation in mind, the algorithms were executed with 2 randomly initialized clusters, with the hope that the algorithms would classify the points into two groups, an early-induction group and a late-induction group. In both algorithms, after the criteria for maximum likelihood is reached, the algorithm-assigned labels of 477 genes are plotted in a confusion matrix against the true labels. In the case of seven

clusters, if the data are perfectly classified, then each rank or file of the confusion matrix will have genes in only one file or rank. In the case of two clusters, as in Figure 4.21, a qualitatively correct proposed cluster would contain one or more adjacent true clusters. In Figure 4.22(a), the MoG algorithm yields proposed cluster 1 that contains more than half of the true cluster 1 (Metabolic), but also a significant amount of true cluster 6 (Mid-Late) and all of true cluster 7 (Late). This may be analogous to the effect seen in Figure 4.15(a), where the Gaussian distribution is unable to partition two clusters surrounded by a great deal of noise. By comparison, the MoL algorithm (Fig. 4.22(b)) yields two clusters well relatively well defined boundaries. Proposed cluster 1 contains most of the Middle to Late genes, whereas proposed cluster 2 contains the early genes, with the exception of true cluster 7 (Late), which was improperly grouped with the early genes.



(a)



(b)

Figure 4.22: Confusion matrices for results of 477 pre-classified genes from the data set of Chu *et al.* analyzed by EM-MoG and MoL with 2 randomly initialized clusters. In (a), MoG proposes two relatively diffuse clusters, whereas in (b), MoL proposes two clusters that more precisely partition the true clusters.

When the EM-MoG and EM-MoL algorithms seek seven clusters to match the seven temporal classes, the results are not easily distinguished from one another (data not shown). Neither algorithm reproducibly generates qualitatively accurate confusion matrices.

Another problem of mixture modeling approaches to clustering is selecting he optimal number of clusters. In some cases, such as with Chu *et al.*, there is an established classification scheme. In other cases, the number of clusters must also be hypothesized. Here, I make use of the minimum description length (MDL) term[53] to penalize the likelihood of the mixture model as the number of parameters increases. For this case, the MDL penalty reduces to subtracting half of the number of parameters times the log likelihood from the as-computed likelihood. I then executed the MoG (Fig. 4.23(a)) and MoL (Fig. 4.23(b)) algorithms on the full Chu *et al.* data set, and applied the MDL penalty term. The results are also tabulated in Table 4.3.

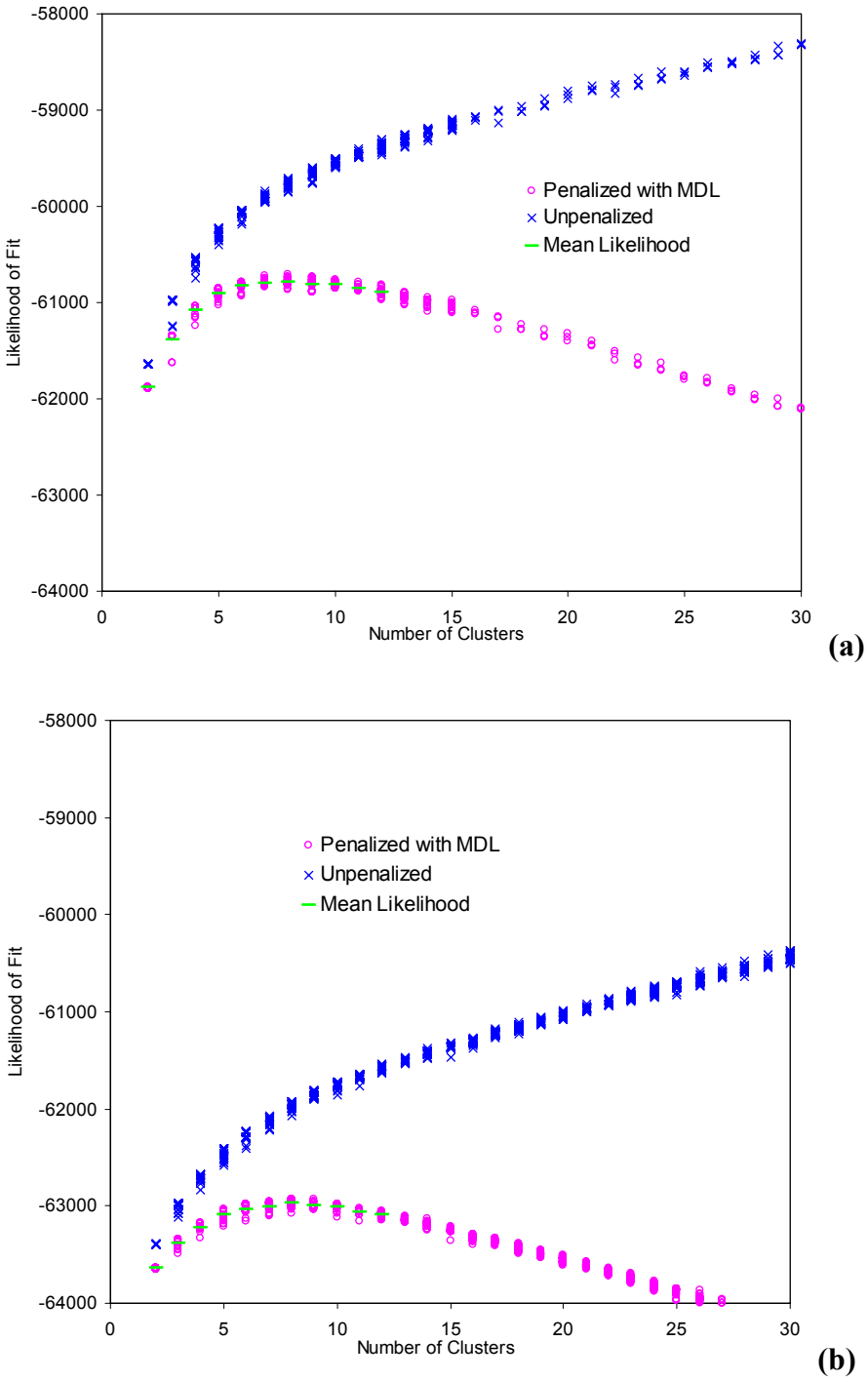
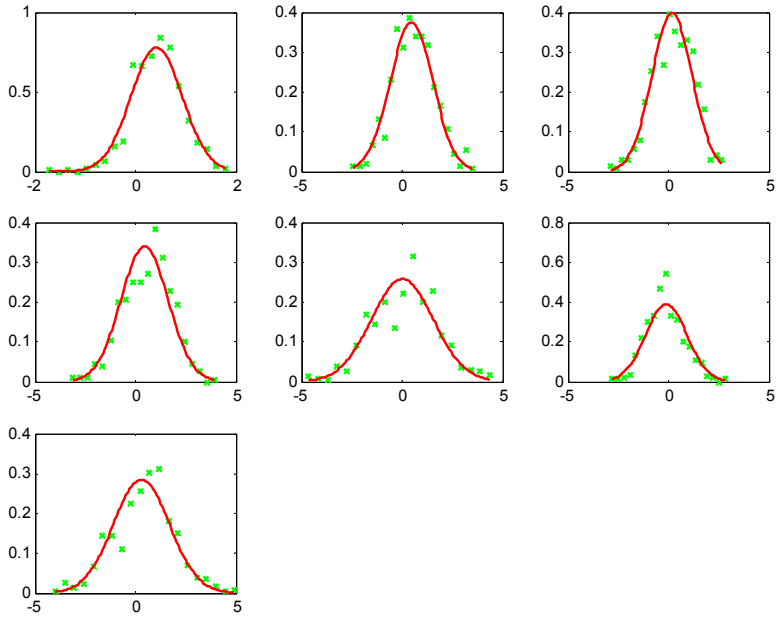


Figure 4.23: Plot of the likelihood of fit of full data set of Chu *et al* analyzed by EM-MoG (a) and EM-MoL (b) with 1 to 30 randomly initialized clusters, with and without the inclusion of the MDL penalty term. With the penalty term, both clustering methods find 7-9 clusters to be optimal, which agrees well with the 7 classes of Chu *et al*.

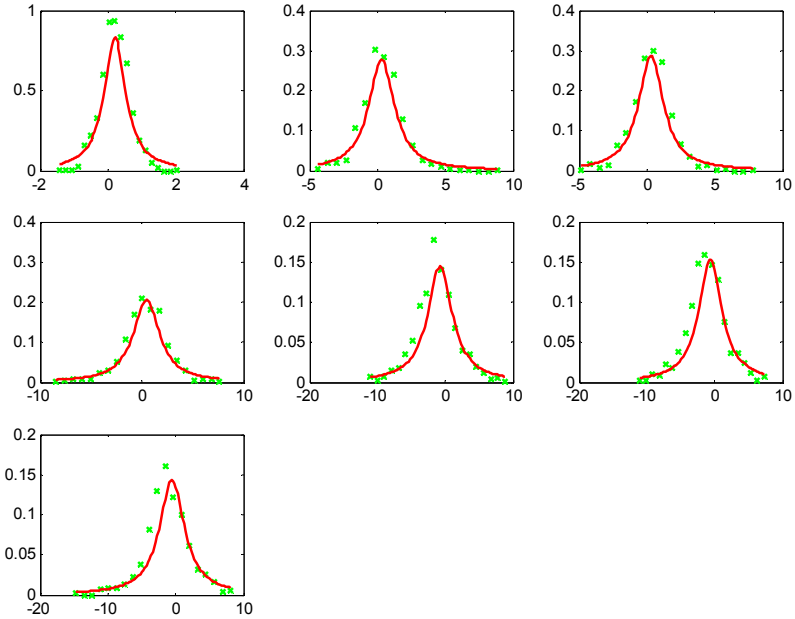
Cluster Size	Mean for MoL	St. Dev	Mean for MoG	St. Dev.
2	-63644	2.9	-61886	0.4
3	-63385	42.8	-61383	87.0
4	-63217	38.3	-61079	55.7
5	-63095	54.9	-60912	52.2
6	-63035	48.5	-60833	45.5
7	-63003	45.3	-60797	32.2
8	**-62973**	**39.0**	**-60787**	**41.2**
9	-62990	29.2	-60807	49.7
10	-63015	35.8	-60807	30.1
11	-63058	28.9	-60851	25.2
12	-63088	27.5	-60892	44.8
Average Standard Deviation		**39.0**		**46.4**

Table 4.3: Likelihoods of fit for the full data set of Chu *et al* analyzed by EM-MoG and EM-MoL with 1 to 12 randomly initialized clusters, with the inclusion of the MDL penalty term. For each choice of cluster size, each algorithm was executed 30 times to generate the statistics shown.

The behavior of the two algorithms can be illustrated by examining the PDF of one of the proposed clusters. In Figure 4.24(a), a representative cluster from the MoG is shown in seven panels where each panel is a histogram of the nth-dimension value of each vector assigned to that cluster. The data in the cluster is distributed in a roughly Gaussian fashion, as shown by the fits. The MoL algorithm similarly finds Lorentzian distributions (Fig. 4.24(b)), but qualitatively generates clusters that are more characteristically Lorentzian.



(a)



(b)

Figure 4.24: Histograms of cluster fits for the full data set of Chu *et al* analyzed by EM-MoG (a) and EM-MoL (b) with 7 randomly initialized clusters. One of the clusters was broken down into its seven dimensions, and each dimension was binned into a histogram and fit by the appropriate PDF.

4.4 Summary

I have demonstrated the combination of principal components analysis and Andrews curves for the visualization of gene expression data. The combination method was applied to visualizing gene vectors in condition space using data from the study of Chu *et al.*, as well as to visualizing condition vectors in gene space using data from the study of Sorlie *et al.* The method is based in well-studied linear algebraic theory, and results in heuristically useful depictions of multivariate gene expression data. Using this method, clustered vectors can be distinguished from both unclustered vectors and differently clustered vectors.

I have implemented the expectation-maximization algorithm to optimize a linear mixture of Lorentzian distributions for clustering. When compared to the equivalent algorithm for Gaussian distributions, the EM-MoL algorithm offers several advantages when applied to artificial two-dimensional data. The EM-MoL algorithm more accurately calculates the means of clusters it identifies, it is more robust to far outliers, and it robustly identifies persistent clusters in a field of dense noise.

I have further compared the EM-MoG and EM-MoL algorithms by clustering the data of Chu *et al.* Neither algorithm consistently replicates the classifications of the original study, however when tested with two clusters, the EM-MoL algorithm reliably partitions the data into earlier-induction and later-induction groups, whereas the EM-MoG algorithm does not. I applied the minimum description length penalty term to show that both algorithms estimate that there are between 7 and 9 clusters in the Chu *et al.*

data. Lastly, I showed that the algorithms generate clusters that are well modeled by their respective probability distribution functions, and that with the Chu *et al.* data, EM-MoG forms more irregular clusters than EM-MoL.

4.5 References

1. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl. Acad. Sci. USA* **96**, 6745-6750 (1977).
2. Taniguchi, M., Miura, K., Iwao, H. & Yamanaka, S. Quantitative assessment of DNA microarrays-comparison with Northern blot analyses. *Genomics* **71**, 34-39 (2001).
3. DeRisi, J., Iyer, V. & Brown, P. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-686 (1997).
4. Chee, M. et al. Accessing genetic information with high-density DNA arrays. *Science* **274**, 610-614 (1996).
5. Quackenbush, J. Microarray data normalization and transformation. *Nature Genetics Supplement* **32**, 496-501 (2002).
6. Churchill, G. A. Fundamentals of experimental designs for cDNA microarrays. *Nature Genetics Supplement* **32**, 490-495 (2002).
7. Chu, S. et al. The transcriptional program of sporulation in budding yeast. *Science* **282**, 699-705 (1998).
8. Perou, C. M. et al. Molecular portraits of human breast tumors. *Nature* **406**, 747-752 (2000).

9. Sorlie, T. et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869-10874 (2001).

10. Sorlie, T. et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. USA* **100**, 8418-8423 (2003).

11. Slonim, D. From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics Supplement* **32**, 502-508 (2002).

12. Lapointe, J. et al. Gene expression profiling indentifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA* **101**, 811-816 (2004).

13. van't Veer, L. J. et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536 (2002).

14. Murtagh, F., Starck, J.-L. & Berry, M. W. Overcoming the curse of dimensionality in clustering by means of the wavelet transform. *The Computer Journal* **43**, 107-120 (1999).

15. Eisen, M., Spellman, P., Brown, P. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863-14868 (1998).

16. Alon, U. et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745-6750 (1999).

17. Bullinger, L. et al. Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *The New England Journal of Medicine* **350**, 1605-1616 (2004).

18. Embrechts, P., Herzberg, A., Kalbfleisch, H., Traves, W. & Whitla, J. An introduction to wavelets with applications to Andrews' plots. *Journal of Computational and Applied Mathematics* **64**, 41-56 (1995).

19. Chernoff, H. Ues of faced to represent points in K-dimensional space graphically. *Journal of the American Statistical Association* **68**, 361-368 (1973).

20. Hamner, C. G., Turner, D. W. & Young, D. M. Comparisons of Several Graphical Methods for Representing Multivariate Data. *Comput. Math. Applic.* **13**, 647-655 (1987).

21. Andrews, D. F. Plots of high-dimensional data. *Biometrics* **28**, 125-136 (1972).

22. Horhota, S. & Aitken, C. Multivariate cluster analysis of pharmaceutical formulation data using andrews plots. *Journal of Parmaceutical Sciences* **80**, 85-90 (1991).

23. Cairns, V. Plotting n-dimensional psychiatric data in 2 dimensions using Andrews' method. *Psychological Medicine* **12**, 169-176 (1982).

24. Chang, W.-C. On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics* **32**, 267-275 (1983).

25. Holter, N. S. et al. Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl. Acad. Sci. USA* **97**, 8409-8414 (2000).

26. Alter, O., Brown, P. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101-10106 (2000).

27. Raychaudhuri, S., Stuart, J. & Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. (1999).

28. Lay, D. C. *Linear algebra and its applications* (Addison Wesley, Reading, MA, 1996).

29. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-2323 (2000).

30. Paulsson, J. Summing up the noise in gene networks. *Nature* **427**, 415-418 (2004).

31. Kerr, M. K. & Churchill, G. A. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* **98**, 8961-8965 (2001).

32. Li, C. & Wong, W. H. Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. USA* **98**, 31-36 (2001).

33. Tseng, G. C., Oh, M.-K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29**, 2549-2557 (2001).

34. Brown, J. S., Kuhn, D., Wisser, R., Power, E. & Schnell, R. Quantification of sources of variation and accuracy of sequence discrimination in a replicated microarray experiment. *BioTechniques* **36**, 324-332 (2004).

35. Lee, M., Kuo, F. C., Whitmore, G. A. & Sklar, J. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Natl. Acad. Sci. USA* **97**, 9834-9839 (2000).

36. Wang, R., Scharenbroich, L., Hart, C., Wold, B. & Mjolsness, E. Clustering analysis of microarray gene expression data by splitting algorithm. *Journal of Parallel and Distributed Computing* **63**, 692-706 (2003).

37. Selim, S. & Ismail, M. K-Means-Type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on pattern analysis and machine intelligence* **PAMI-6**, 81-87 (1984).

38. Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E. & Ruzzo, W. L. Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977-987 (2001).

39. McLachlan, G. & Peel, D. in *Advanced in Pattern Recognition* (eds. Amin, A., Dori, D., Pudil, P. & Freeman, H.) 658-666 (Springer, Sydney, Australia, 1998).

40. Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**, 611-631 (2002).

41. Ben-Dor, A., Shamir, R. & Yakhini, Z. Clustering gene expression patterns. *Journal of Computational Biology* **6**, 281-297 (1999).

42. Ghosh, D. & Chinnaiyan, A. M. Mixture Modelling of gene expression data from microarray experiments. *Bioinformatics* **18**, 275-286 (2002).

43. Allison, D. B. et al. A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics and Data Analysis* **39**, 1-20 (2002).

44. Domany, E. Cluster analysis of gene expression data. *Journal of Statistical Physics* **110**, 1117-1139 (2003).

45. D'haeseleer, P., Liang, S. & Somogyi, R. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* **16**, 707-726 (2000).

46. Shannon, W., Culverhouse, R. & Duncan, J. Analyzing microarray data using cluster analysis. *Pharmacogenomics* **4**, 41-52 (2003).

47. Ouyang, M., Welsh, W. J. & Georgopoulos, P. Gaussian mixture clustering and imputation of microarray data. *Bioinformatics* **20**, 917-923 (2004).

48. Dempster, A. P., Laird, N. M. & Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* **39**, 1-38 (1977).

49. McLachlan, G., Bean, R. W. & Peel, D. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* **18**, 413-422 (2002).

50. Peel, D. & McLachlan, G. Robust mixture modelling using the t distribution. *Statistics and computing* **10**, 339-348 (2000).

51. Brody, J. P., Williams, B. A., Wold, B. J. & Quake, S. R. Significance and statistical errors in the analysis of DNA microarray data. *Proc. Natl. Acad. Sci. USA* **99**, 12975-12978 (2002).

52. Perou, C. M. et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **96**, 9212-9217 (1999).

53. Rissanen, J. Modeling by shortest data description. *Automatica* **14**, 465-471 (1978).

CHAPTER 5
CONCLUSIONS AND RECOMMENDATIONS

5.1 Peptoid Synthesis 145
5.2 MAGE Methodology 148
5.3 Visualization and Model-Based Clustering 152
5.4 Overall.....154
5.5 References.....155

5.1 Peptoid Synthesis

I demonstrated several new synthetic strategies for expanding the versatility of the peptoid platform. Three further areas may be worth investigation.

The first area includes three new modifications for use in our MAGE assay, where I made use of peptoid-DNA conjugates that were formed by modifying the N-terminus of the peptoids with iodoacetic acid. It would be useful to transfer two other features currently embedded at the 5′ end of the oligodeoxynucleotide in MAGE: a pendant visible dye, and a backbone photocleavable linker. By moving these specialized functionalities from the ODN to the peptoid, it would increase the generality and decrease the cost of MAGE. I have demonstrated (data not shown) a system where the dye is added at the N-terminus, and a pendant iodoacetyl group is used for subsequent conjugation to the ODN thiol, but the yield of the synthesis remains poor (Appendix C.1). Attempts to incorporate photocleavable units into the backbone with pendant orthonitrobenzenes and similar moieties have failed (Appendix C.2). It is likely that a successful strategy will incorporate not only this functionality but also a bond that is more subject to photo-initiated degradation than the amide bonds in the peptoid. Finally, because peptoids can incorporate great diversity through the amine submonomers, it may be possible to improve the mass spectrometric efficiency of the peptoids by carefully choosing submonomers. For example, the extent of fragmentation could be reduced, or the ionization efficiency could be increased. I synthesized a number of peptoids

incorporating polar, aromatic primary amine submonomers in an attempt to increase the MALDI ionization efficiency of the peptoid tags. The motivation for this was that many MALDI matrices share structural features with the chosen submonomers. There was no evidence that this tactic was effective (Appendix C.3).

The second area is gene therapy. Cationic peptoids have been studied for gene therapy[1,2], but thus far the only cationic amine submonomers have been primary amines. The literature of gene therapy suggests that other charge centers, such as amidines, have intriguing properties[3]. Because they are sequence-specific heteropolymers, peptoids would be an ideal platform for structure-function studies of oligoamines and oligoamidines. Amidine groups could be incorporated into peptoids by making use of mono-protected diamines, deprotecting those amines at the end of the synthesis, and modifying them with a reagent such as ethyl acetimidate. My work and the work of others[4] has demonstrated a wide variety of chemoselective functionalities that can be incorporated into peptoids, so these oligomers could be conjugated to dyes and other entities of interest to gene therapy.

The third area is extended molecular structures, exemplified by the work of Mirkin *et al.*[5,6] In general, of the field of nanoscale assembly, Mirkin reports that "a major limitation in nanoparticle-based materials chemistry is the lack of suitable assembly methods for preparing extended two- and three-dimensional architectures with synthetically programmable building block and assembly parameters." He also claims of a system of DNA-block copolymer conjugates that "while interesting, these DNA/polymer hybrids are limited with respect to their degree of tailorability, ill-defined compositions, and poor solubilities and dispersities, as well as function." Peptoids could

be a suitable system for these types of investigations. For example, to create extended molecular structures of peptoids and DNA, the branching scheme presented in this thesis could be combined with our method of peptoid-DNA conjugation. By tailoring the structure of the peptoid, the properties of the resulting extended structure could be controlled. It would also be possible to incorporate other nanoscale objects of interest into these scaffolds, such as large biomolecules, or quantum dots. Quantum dots are themselves an interesting potential application of peptoids. Modifying the surface properties of gold and silver and other nanoparticles affects their behavior as light emitters, as well as their stability in solution. Attaching peptoids to nanoparticles (e.g., via a thiol-based linker) gives a tunable way to modify the properties of the particles through the sequence of the peptoid.

5.2 MAGE Methodology

Each component of the MAGE methodology was demonstrated with elementary proof-of-principle experiments. Currently, the major limitation of MAGE is that it can only be as sensitive as the mass spectrometer that is used for the final quantitation. Immediately stemming from the first experiments, I would suggest that (i) the hybridization kinetics of the MAGE probes and likely targets be studied, (ii) the quantitativeness of the ligation step be assessed, (iii) the ultimate sensitivity of peptoid-containing MAGE fragments be studied in a variety of mass spectrometers, and (iv) "reversed" MAGE probes be tested to try to eliminate the multiple photocleavage products. Experiments (i) and (ii) are critical for proving that MAGE establishes a 1:1 or other predictable relationship between mass tags and sequences of interest. In particular, it might be especially important to identify an optimal time and temperature of hybridization for the MAGE probes and targets by tracking duplex formation for a variety of choices of target initial concentration, temperature, and time. These studies might also determine what relative concentration of probes is necessary to ensure pseudo-first-order kinetics. Currently, it is unknown what effect the peptoid and biotin may have on the hybridization kinetics. Experiment (iv) could be conducted with commercially available phosphoramidites from Glen Research, including 5′-biotin phosphoramidite, 3′-thiol-modifier C3 S-S CPG, PC spacer phopsphoramidite, and 3′-phosphate CPG, resulting the new probes M3 and M4 in Figure 5.1. This new configuration would mean that the peptoid fragment would carry the phosphate half of the photocleavage reaction

instead of the linker residue (Fig. 5.2). The phosphate half did not show by RP-HPLC and MALDI-TOF the multiple products that the linker half did (Chapter 3).

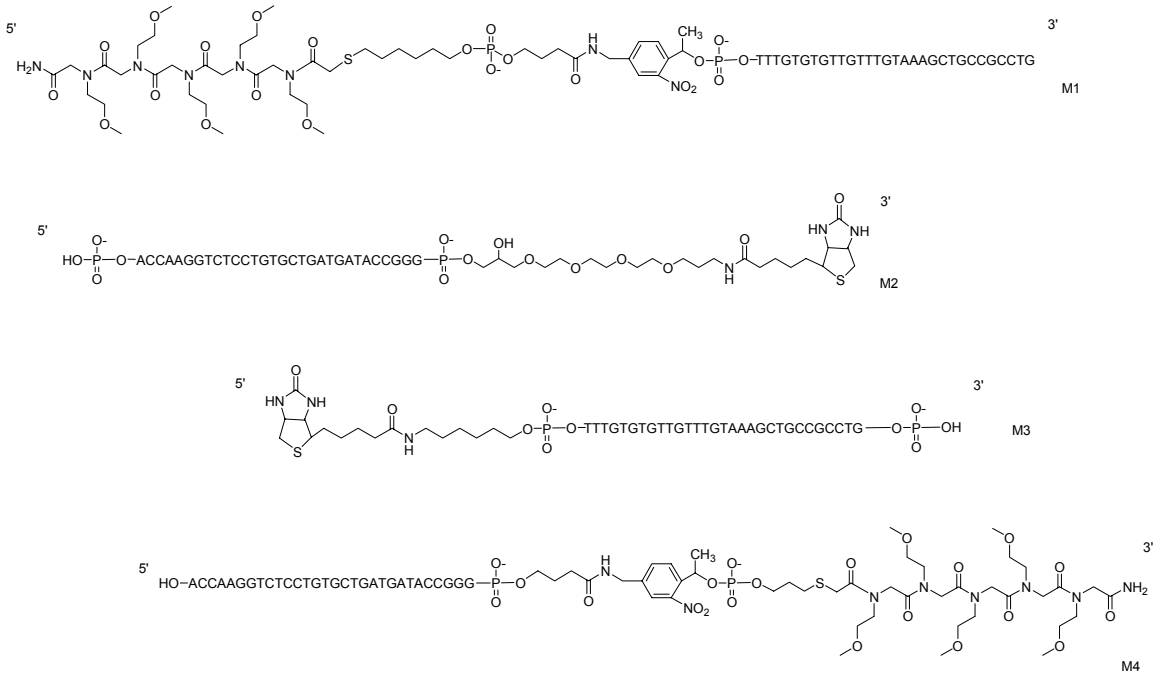


Figure 5.1: Schematic illustrated of MAGE probes in different orientations. Probes M1 and M2 are in the configuration used in this thesis. By reversing the probes to the configuration shown in M3 and M4, the peptoid fragment of photocleavage would be a single species, instead of several as in this thesis.

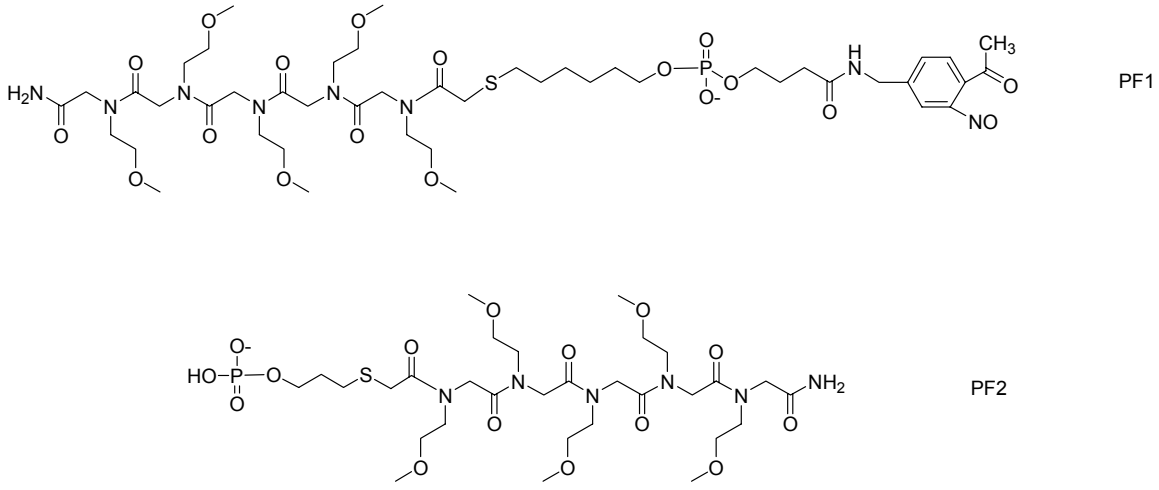


Figure 5.2: Schematic illustrated of alternative MAGE peptoid fragments. By reversing the configuration of the probes, the peptoid fragment of MAGE is changed from PF1, which was present in multiple products, to PF2.

Following these fundamental experiments, I would suggest attempting to (i) multiplex MAGE by synthesizing a small group of peptoid mass tags, (ii) compare MAGE measurements to those of competing technologies, (iii) test the discriminating power of the ligation step by using a thermostable ligase and probing for targets with slightly mismatched competitors present, and (iv) execute the final quantitation on more sensitive mass spectrometers than the one employed for our studies.

Three other potential modifications of MAGE might be of interest. First, the flexibility of the isotopic labeling system could be used to enable "multi-color" MAGE. Here, instead of comparing the sample of interest to a calibrated internal standard, two or more target samples are simultaneously interrogated by probes with peptoid tags of identical sequence but differing amounts of isotopic labeling. For example, four peptoids could be synthesized, each different by six Daltons. One of the four would be used for each of four different target mixtures, and the resulting cleaved tags would be mixed prior to mass spectrometry. This would be especially useful for examining time-course gene expression data, and it would be the equivalent of a four-color microarray.

Second, if the state-of-the-art mass spectrometry equipment does not yield sufficient sensitivity, amplification schemes should be considered. The focus in the field of mass spectrometry has been mass accuracy, not sensitivity. This is in part why two schemes of mass-spectrometric gene expression analysis have relied on amplification[7,8]. The most obvious choices for an amplification scheme for MAGE would be those based on the ligation detection reaction[9-11]. The two common schemes both require adding additional cycles of ligation. In the current implementation of MAGE, after the probes

have been annealed and ligated, the mixture could be melted and reannealed, and ligated again. Assuming that an excess of probes was added, this would result in a linear amplification of the signal. A more drastic step would be to add a second set of probes that are complementary to the first set (and thus, identical to the targets). Here, repeated rounds of ligation results in an exponential amplification, known as ligase chain reaction (LCR). In order to ensure that the signal is not confounded by blunt-end ligation, a gap could be left between the two probes instead of a nick[12,13], known as Gap-LCR.

Third, in order to decrease costs and increase reliability, the MAGE methodology could be implemented in a chip format. All of the technological steps of MAGE have been demonstrated in some combination on chips, starting from the purification of mRNA from cell lysate[14], and ending with mass spectrometry[15].

5.3 Visualization and Model-Based Clustering

Andrews curves are a mathematically rigorous, visually effective method for displaying multivariate data. There are several key limitations that were evident in my work. First, it is not effective to attempt to plot more than about 10 vectors simultaneously, because the plot becomes too densely packed to interpret. Second, it is not effective to attempt to plot more than the first 7-10 terms of the Fourier series, because the curves become exceedingly squiggly. Third, the terms of the Fourier series have a decreasing impact on the appearance of the plot[16]. Because of these limitations, it is critical that some preprocessing step be applied to the high-dimensional gene expression data. Principle components analysis (PCA) is effective, as I demonstrated, but it is not optimal in this application. PCA creates a set of orthogonal linear spaces in decreasing order of their eigenvalues, serving to extract linear features from the data. There is some evidence suggesting that there are significant linear features within gene expression data[17], and my results seem to suggest this as well. Ultimately, because of the highly non-linear nature of massively-fedback networks such as those present in the cell, linear methods will fail to extract all of the significant features. In further pursuit of this visualization method, non-linear preprocessing methods should be considered[18-21].

As evidenced by the proliferation of alternative methods, it is quite difficult to determine the most appropriate method for clustering analysis on gene expression data. The mixture of Lorentzians model I presented offers an extremely permissive set of clusters that avoids heuristic “outlier rejection” methods that are often employed. The

advantages of my method are evident from artificial data, but it remains difficult to conclude that it represents a superior clustering algorithm for gene expression data in general. Part of the reason for this is that it remains largely unknown what the true nature of the biological noise of transcription is, as well as the machine noise inherent to various types of gene expression analysis systems. Currently, the most successful strategies seek to test a number of different clustering schemes in combination with non-parameterized statistical tests such as Monte Carlo cross-validation or bootstrap analysis[22]. Another important feature of a fully-developed clustering scheme is a scheme for developing a hierarchy of clusters. Superior hierarchical schemes are two-way, where each node in the hierarchy can have multiple children as well as multiple parents. Further development of the mixture of Lorentzians algorithm should seek to include it in a hierarchical algorithm that includes a non-parameterized statistical test.

5.4 Overall

Studies of functional genomics draw from a diverse pool of experimental and analytical methods. In recent years, investigators have demanded high quality data with robust statistical analyses to support their increasingly quantitative studies. Here, we have offered a new method for unambiguously measuring the abundance of specific sequences of nucleic acids. Because it is grounded in well-studied physics, and results in absolute abundances, the MAGE methodology is especially suited for these highly quantitative studies. In order to meet the demand for new, statistically robust analytical methods, I have presented several new tactics for large-scale gene expression analysis. Used in combination, my methods allow investigators to reduce their expansive data to a more manageable subset using principle components analysis, visualize that data in a mathematically consistent and useful manner with Andrews curves, and cluster their data using noise-permissive Lorentzian distributions.

5.5 References

1. Huang, C. Y. et al. Lipitoids-novel cationic lipids for cellular delivery of plasmid DNA in vitro. *Chemistry & Biology* **5** (1998).
2. Murphy, J. E. et al. A combinatorial approach to the discovery of efficient cationic peptoid reagents for gene delivery. *Proc. Natl. Acad. Sci. USA* **95**, 1517-1522 (1998).
3. Plank, C., Tang, M. X., Wolfe, A. R. & Szoka Jr., F. C. Branched cationic peptides for gene delivery: role of type and number of cationic residues in formation and in vitro activity of DNA polyplexes. *Human Gene Therapy* **10**, 319-332 (1999).
4. Horn, T., Lee, B.-C., Dill, K. & Zuckermann, R. Incorporation of chemoselective functionalities into peptoids via solid-phase submonomer synthesis. *Bioconjugate Chemistry* **15**, 428-435 (2004).
5. Mirkin, C. A., Letsinger, R. L., Mucic, R. C. & Storhoff, J. J. A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature* **382**, 607-609 (1996).
6. Watson, K. J., Park, S.-J., Im, J.-H., Nguyen, S. T. & Mirkin, C. A. DNA-Block Copolymer Conjugates. *J. Am. Chem. Soc.* **123**, 5592-5593 (2001).
7. Ding, C. & Cantor, C. R. A high-throughput gene epression analysis technique using competitive PCR and matrix-assisted laser desorption ionization time-of-flight MS. *Proc. Natl. Acad. Sci. USA* **100**, 3059-3064 (2003).

8. Berggren, W. T. et al. Multiplexed gene expression analysis using the Invader RNA assay with MALDI-TOF mass spectrometry detection. *Anal. Chem.* **74**, 1745-1750 (2002).

9. Wiedmann, M. et al. Ligase chain reaction (LCR)-overview and applications. *PCR Methods and Applications* **3**, S51-S64 (1994).

10. Lee, H. H. Ligase Chain Reaction. *Biologicals* **24**, 197-199 (1996).

11. Cao, W. Recent developments in ligase-mediated amplification and detection. *Trends in Biotechnology* **22**, 38-44 (2004).

12. Abravaya, K., Carrino, J. J., Muldoon, S. & Lee, H. H. Detection of point mutations with a modified ligase chain reaction (Gap-LCR). *Nucleic Acids Research* **23**, 675-682 (1995).

13. Harden, S. V. et al. Real-time gap ligase chain reation. *Clinical Cancer Research* **10**, 2379-2385 (2004).

14. Hong, J. W., Studer, V., Hang, G., Anderson, W. F. & Quake, S. A nanoliter-scale nucleic acid processor with parallel architecture. *Nature Biotechnology* **22**, 435-439 (2004).

15. Little, D. P. et al. MALDI on a chip: Analysis of Arrays of Low-Femtomole to Subfemtomole Quantities of Synthetic oligonucleotides and DNA Diagnostic Products Dispensed by a Piezoelectric Pipet. *Anal. Chem.* **69**, 4540-4546 (1997).

16. Andrews, D. F. Plots of high-dimensional data. *Biometrics* **28**, 125-136 (1972).

17. Alter, O., Brown, P. & Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97**, 10101-10106 (2000).

18. Roweis, S. T. & Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **290**, 2323-2326 (2000).

19. Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319-2323 (2000).

20. Kim et al. General Nonlinear Framework for the Analysis of Gene Interaction via Multivariate Expression Arrays. *Journal of Biomedical Optics* **5**, 411-424 (2000).

21. Goutsias, J. & Kim, S. A nonlinear discrete dynamical model for transcriptional regulation: construction and properties. *Biophysical Journal* **86**, 1922-1945 (2004).

22. Kerr, M. K. & Churchill, G. A. Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci. USA* **98**, 8961-8965 (2001).

APPENDIX A: DERIVATION OF A LORENTZIAN DISTRIBUTION FROM THE RATIO OF TWO INDEPENDENT NORMALLY DISTRIBUTED RANDOM VARIABLES

A.1 Derivation.....159

Let X and Y be two independent normally distributed random variables with mean zero and standard deviation 1,

X=G(0,1); Y=G(0,1)

$$f_x = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x^2)} \qquad f_y = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y^2)}$$

Since X and Y are independent, their joint probability distribution function is the product of their individual probability distribution functions,

$$f_{xy} = f_x f_y = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2 + y^2)}$$

Now we want to examine the probability distribution of their ratio. Let z = x/y and w = y,

$$f_{zw} = \frac{1}{2\pi} e^{-\frac{1}{2}(z^2 w^2 + w^2)} \left\| \begin{matrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial w} \end{matrix} \right\| = \frac{1}{2\pi} e^{-\frac{1}{2} w^2 (z^2 + 1)} \left\| \begin{matrix} y & 0 \\ \frac{-x}{z^2} & 1 \end{matrix} \right\|$$

$$f_{zw} = \frac{1}{2\pi} e^{-\frac{1}{2} w^2 (z^2 + 1)} |y| = \frac{1}{2\pi} |w| e^{-\frac{1}{2} w^2 (z^2 + 1)}$$

Now we can obtain the probability distribution for z = x/y by integrating out the w dependence of f_{zw}

$$f_z = \int_{-\infty}^{\infty} f_{zw}\, dw = \frac{1}{2\pi} \int_{-\infty}^{\infty} |w| e^{-\frac{1}{2} w^2 (z^2 + 1)} dw = 2 \frac{1}{2\pi} \int_0^{\infty} w e^{-\frac{1}{2} w^2 (z^2 + 1)} dw = \frac{1}{2\pi} \int_0^{\infty} 2w e^{-\frac{1}{2} w^2 (z^2 + 1)} dw$$

Now with a substitution of $u=w^2$ and $a = (z^2+1)/2$

$$f_z = \frac{1}{2\pi}\int_0^{\infty} e^{-au}\,du = -\frac{1}{2\pi a}e^{-a\infty} + \frac{1}{2\pi a}e^{-a0} = \frac{1}{2\pi a} = \frac{1}{\pi(z^2+1)}$$

This is the form of the Lorentzian distribution with mean zero and spread 1,

$$f_z = \frac{1}{\pi(z^2+1)}$$

APPENDIX B: MATLAB CODE FOR VISUALIZATION AND CLUSTERING ALGORITHMS

B.1 Master algorithm function for generating Andrews curves 162
B.2 Utility for preprocessing data by PCA 164
B.3 Utility for implementing PCA algorithm with SVD 165
B.4 EM adaptive mixture of Lorentzians.....166
B.5 EM adaptive mixture of Gaussians 169
B.6 Mixture of Lorentzians with plotting for 2D data sets.....170
B.7 Utility for running multiple clustering experiments.....173
B.8 Utility for plotting T-distributions.....174
B.9 Utility for plotting histograms of cluster values.....175
B.10 Utility for plotting cluster means and standard deviations.....176

B.1 Master algorithm function for generating Andrews curves

```
function [] = Demon_Andrews(LabeledData,npcs);
%Andrews Curve Demonstrator
%LabeledData should have the data arrange in columns, so that the
%number of rows is the number of dimensions, and the number of columns
%is the number of samples.  The last row should contain labels, which
are
%integers from 1 up.
%npcs is the number of principle components you wish to use.  This
number
%should be at least 1, and at most either the number of samples or the
number
%of dimensions, whicever is lower.

D=size(LabeledData,1)-1;
N=size(LabeledData,2);
X=LabeledData(1:D,:);

if npcs>D
      disp('Too Many PCs')
      return
end
if npcs>N
      disp('Too Many PCs')
      return
end

%First we preprocess the data to reduce the data to the number of PCs
%desired

[ProcData] = preproc(LabeledData,X,npcs);
size(ProcData)

%We illustrate the effect of this processing

[pc, explained, score, latent, tsquare] = MakePCA(X);

sumexp(1)=explained(1);

for i=2:length(explained)
sumexp(i)=sumexp(i-1)+explained(i);

sumexp
end
figure(1)
clf
plot([1:1:length(explained)],sumexp,'ko:');
%title('Cumulative explaining power of all principal components')
ylabel('% of Variance Explained by principal components 1 through 122')
xlabel('Principal components used')
hold
%line([npcs npcs],[0 100])
axis([1 length(explained) 0 100])
```

```
%Now the data is processed using Andrews' Method
%Andrews' Method is a simple mapping onto an orthogonal basis function
that
%preserves the mathematical integrity of the data

for i=1:N
    val(i)=norm(ProcData(1:npcs,i));
end

ProcData(1:npcs,:)=ProcData(1:npcs,:)./repmat(val,npcs,1);

ps=500;
t=linspace(-pi,pi,ps)';
AndData=zeros(ps,N);
AndData=repmat(ProcData(1,:)/sqrt(2),ps,1);
for i=2:npcs
      if mod(i,2) == 0

      AndData=AndData+(repmat(ProcData(i,:),ps,1).*repmat((sin(i*t)),1,
N));
      else

      AndData=AndData+(repmat(ProcData(i,:),ps,1).*repmat((cos(i*t)),1,
N));
      end
end
AndData;
size(AndData)

%Now we take the Andrews Method processed Data and show it
%Note that the coloring and number of lines is data depedant here

figure(2)
clf
hold
plot(t,AndData(:,112),'k-')
plot(t,AndData(:,113),'k-')
plot(t,AndData(:,114),'k-')
plot(t,AndData(:,21),'r-')
plot(t,AndData(:,22),'r-')
plot(t,AndData(:,23),'r-')
%plot(t,AndData(:,117),'r-')
%plot(t,AndData(:,40),'r-')
%plot(t,AndData(:,54),'r-')
%plot(t,AndData(:,375),'m-')
%plot(t,AndData(:,192),'c-')
%plot(t,AndData(:,2689),'y-')
%title('Andrews Curve of your PCA Reduced Data Set')
ylabel('F(t)')
xlabel('t')
```

B.2 Utility for preprocessing data by PCA

```
function [ProcData] = preproc(Labeled,X,npcs)

assignedata=Labeled;
D=size(X,1);
N=size(X,2);

%PCA
A=MakePCA(X');
PCAdata=A*X;
LPCA=zeros(D+1,N);
for i=1:N
   LPCA([1:D],i)=PCAdata(:,i);
   LPCA(D+1,i)=assignedata(D+1,i);
end
ProcData=LPCA([1:npcs D+1],:);
```

B.3 Utility for implementing PCA algorithm with SVD

```
function [pc, explained, score, latent, tsquare] = MakePCA(q);

[m,n] = size(q);
avg = mean(q);
centerx = (q - avg(ones(m,1),:));

[U,latent,pc] = svd(centerx./sqrt(m-1),0);
score = centerx*pc;
explained=100*(diag(latent))/(sum(diag(latent)));

if nargout < 4, return; end
latent = diag(latent).^2;

if nargout < 5, return; end
tmp = sqrt(diag(1./latent))*score';
tsquare = sum(tmp.*tmp)';
```

B.4 EM adaptive mixture of Lorentzians

```
function [Likelihood, Mixture_coefficients, Means, Covariances] =
MoL(X,K)

% Initialization

D=size(X,1);
N=size(X,2);
Norm=(gamma((D+1)/2))/(pi^(D+1)/2);

for i=1:K

      Mean(:,i)=randMean(D,1);
      test=500;
      while test>=100
            Cov(:,:,i)=randCovariance(D,20);
            test=cond(Cov(:,:,i));
      end
      F(i)=1/K;

end

% E-step

ITS=0;
crit=1;
ClusterFlag=0;

while crit>=1e-5
      ITS=ITS+1;

      for i=1:K
            U=X-repmat(Mean(:,i),1,N);
            V=U.*(inv(Cov(:,:,i))*U);
            M(i,:)=sum(V);
      end

      P1=Norm.*(1+M).^(-(D+1)/2);

      for i=1:K

            P1(i,:)=P1(i,:)*F(i)*(det(Cov(:,:,i)))^(-.5);
      end

      P=P1./repmat(sum(P1,1),K,1);
      U=(1+D).*(1+M).^(-1);
      L(ITS)=sum((log(sum(P1))),2);
      CurrentIteration=ITS
      CurrentLikelihood=L(ITS)
      if ITS>1
            crit=abs((L(ITS)-L(ITS-1))/L(ITS));
      end
      ConvergenceCriteria=crit
```

```
      %M-step
      R=P.*U;
      S=sum(R');
      S1=repmat(S,D,1);

      Mean=(X*(R'))./S1;
      T=sum(P,2);
      F=1/N*(sum(P,2));
      ClusterFlag=0;
      for i=1:K

            M1=repmat(Mean(:,i),1,N);
            Y=X-M1;
          R1=repmat(R(i,:),D,1);
            Cov(:,:,i)=((Y.*R1)*Y')/T(i);

            if cond(Cov(:,:,i))>=1e10
                  disp('Covariance Failure, eliminating')
                  ClusterFlag=i
                  PointsInCluster=N*F(i)
            end

      end
      if ClusterFlag~=0
            g=1;
            for j=1:K
                  if j~=ClusterFlag
                        Mean2(:,g)=Mean(:,j);
                        Cov2(:,:,g)=Cov(:,:,j);
                        F2(g)=F(j);
                        g=g+1;
                  end
            end
            Mean=Mean2;
            Cov=Cov2;
            F=F2/sum(F2);
            K=K-1;
            clear M;clear U;clear V;clear R;clear P1;clear S;
            clear T;clear M1;clear R1;clear Y;
            end
end

Means=Mean;
Covariances=Cov;
Mixture_coefficients=F;
Likelihood=L(ITS);
```

B.5 EM adaptive mixture of Gaussians

```
function [Likelihood, Mixture_coefficients, Means, Covariances] =
MoGF(X,K)

% Initialization

D=size(X,1);
N=size(X,2);
Norm=(2*pi)^(-D/2);

for i=1:K

      Mean(:,i)=randMean(D,1);
      test=500;
      while test>=100
            Cov(:,:,i)=randCovariance(D,20);
            test=cond(Cov(:,:,i));
      end
      F(i)=1/K;

end

% E-step

ITS=0;
crit=1;
ClusterFlag=0;

while crit>=1e-5
      ITS=ITS+1;

      for i=1:K
            U=X-repmat(Mean(:,i),1,N);
            V=U.*(inv(Cov(:,:,i))*U);
            M(i,:)=sum(V);
      end

      P1=Norm.*exp(-M/2);

      for i=1:K

            P1(i,:)=P1(i,:)*F(i)*det(Cov(:,:,i))^(-.5);
      end

      P=P1./repmat(sum(P1,1),K,1);
      L(ITS)=sum((log(sum(P1))),2);
      CurrentIteration=ITS
      CurrentLikelihood=L(ITS)
      if ITS>1
            crit=abs((L(ITS)-L(ITS-1))/L(ITS));
      end
      ConvergenceCriteria=crit
```

```
      %M-step
      S=sum(P');
      S1=repmat(S,D,1);
      Mean=(X*(P'))./S1;
      F=1/N*(sum(P,2));
      ClusterFlag=0;
      for i=1:K


          R1=repmat(P(i,:),D,1);
            Cov(:,:,i)=((X.*R1)*X')/S(i)-Mean(:,i)*Mean(:,i)';

            if cond(Cov(:,:,i))>=1e10
                  disp('Covariance Failure, eliminating')
                  ClusterFlag=i
                  PointsInCluster=N*F(i)
            end

      end
      if ClusterFlag~=0
            g=1;
            for j=1:K
                  if j~=ClusterFlag
                        Mean2(:,g)=Mean(:,j);
                        Cov2(:,:,g)=Cov(:,:,j);
                        F2(g)=F(j);
                        g=g+1;
                  end
            end
            Mean=Mean2;
            Cov=Cov2;
            F=F2/sum(F2);
            K=K-1;
            clear M;clear U;clear V;clear R;clear P1;clear S;
            clear T;clear M1;clear R1;clear Y;
            end
end

Means=Mean;
Covariances=Cov;
Mixture_coefficients=F;
Likelihood=L(ITS);
```

B.6 Mixture of Lorentzians with plotting for 2D data sets

```
function [Likelihood, Mixture_coefficients, Means, Covariances] =
MoL(X,K)

% Initialization

D=size(X,1);
N=size(X,2);
Norm=(gamma((D+1)/2))/(pi^(D+1)/2);

for i=1:K

      Mean(:,i)=randMean(D,1);
      test=500;
      while test>=100
            Cov(:,:,i)=randCovariance(D,20);
            test=cond(Cov(:,:,i));
      end
      F(i)=1/K;

end

% E-step

ITS=0;
crit=1;
ClusterFlag=0;

while crit>=1e-5
      ITS=ITS+1;

      for i=1:K
            U=X-repmat(Mean(:,i),1,N);
            V=U.*(inv(Cov(:,:,i))*U);
            M(i,:)=sum(V);
      end

      P1=Norm.*(1+M).^(-(D+1)/2);

      for i=1:K

            P1(i,:)=P1(i,:)*F(i)*(det(Cov(:,:,i)))^(-.5);
      end

      P=P1./repmat(sum(P1,1),K,1);
      U=(1+D).*(1+M).^(-1);
      L(ITS)=sum((log(sum(P1))),2);
      CurrentIteration=ITS
      CurrentLikelihood=L(ITS)
      if ITS>1
            crit=abs((L(ITS)-L(ITS-1))/L(ITS));
      end
```

```
      ConvergenceCriteria=crit

      %M-step
      R=P.*U;
      S=sum(R');
      S1=repmat(S,D,1);

      Mean=(X*(R'))./S1;
      T=sum(P,2);
      F=1/N*(sum(P,2));
      ClusterFlag=0;
      for i=1:K

            M1=repmat(Mean(:,i),1,N);
            Y=X-M1;
          R1=repmat(R(i,:),D,1);
            Cov(:,:,i)=((Y.*R1)*Y')/T(i);

            if cond(Cov(:,:,i))>=1e10
                  disp('Covariance Failure, eliminating')
                  ClusterFlag=i
                  PointsInCluster=N*F(i)
            end

      end
      if ClusterFlag~=0
            g=1;
            for j=1:K
                  if j~=ClusterFlag
                        Mean2(:,g)=Mean(:,j);
                        Cov2(:,:,g)=Cov(:,:,j);
                        F2(g)=F(j);
                        g=g+1;
                  end
            end
            Mean=Mean2;
            Cov=Cov2;
            F=F2/sum(F2);
            K=K-1;
            clear M;clear U;clear V;clear R;clear P1;clear S;
            clear T;clear M1;clear R1;clear Y;
            end
end

%Now we will assign data to a particular cluster. 3rd row is
clusternumber
assigneddata=zeros(3,length(X));
assigneddata([1,2],:)=X;

for i=1:N
      [Y1,T1]=max(P(:,i));
      assigneddata(3,i)=T1;
end

%This part plots all the data, different colors for different clusters,
up to
%10 clusters... beyond that they'll repeat
```

```
colordef black
clf
figure(1)
hold on
for i=1:N
      switch mod(assigneddata(3,i),10)
            case 0,
                  plot(X(1,i),X(2,i),'c*')
            case 1,
                  plot(X(1,i),X(2,i),'rd')
            case 2,
                  plot(X(1,i),X(2,i),'gv')
            case 3,
                  plot(X(1,i),X(2,i),'m+')
            case 4,
                  plot(X(1,i),X(2,i),'bs')
            case 5,
                  plot(X(1,i),X(2,i),'yo')
            case 6,
                  plot(X(1,i),X(2,i),'wx')
            case 7,
                  plot(X(1,i),X(2,i),'yp')
            case 8,
                  plot(X(1,i),X(2,i),'r<')
            case 9,
                  plot(X(1,i),X(2,i),'g>')
            otherwise,
                  plot(X(1,i),X(2,i),'wh')
      end
end

for i=1:K
      plot(Mean(1,i),Mean(2,i),'yo')
end

Means=Mean;
Covariances=Cov;
Mixture_coefficients=F;
Likelihood=L(ITS);
```

B.7 Utility for running multiple clustering experiments

```
function [] = molexp(reps,m,n,fname)
%molexp(reps,m,n,fname)
%Reps is the number of times to conduct the MDL experiment
%Experiment does MoG, m-n clusters, reps times
%saves results as fname

load chudata;
f=1;

for i=m:n
      for j=1:reps
            currentrepetition=j
            currentcluster=i
            [Likelihood, Mixture_coefficients, Means, Covariances] =
MoLv4(X,i);
            Results(f,1)=Likelihood;
            Results(f,2)=length(Mixture_coefficients);
            Results(f,3)=i;
            Results(f,4)=j;
            f=f+1;
      end
end

Results
save(fname,'Results');
```

B.8 Utility for plotting T-distributions

```
M=0;
C=1;
D=1;
Z=linspace(-5,5);
L=length(Z);


for j=1:L
      Y(j,1)=Tdist(Z(j),M,C,1,D);
end

for j=1:L
      Y(j,2)=Tdist(Z(j),M,C,2,D);
end

for j=1:L
      Y(j,3)=Tdist(Z(j),M,C,3,D);
end

for j=1:L
      Y(j,4)=Tdist(Z(j),M,C,100,D);
end

clf
colordef black
figure(1)
subplot(2,2,1)
plot(Z,Y(:,1),'r-','linewidth',3)
title('Lorentzian, F=1')
subplot(2,2,2)
plot(Z,Y(:,2),'r-','linewidth',3)
title('T-Distribution, F=2')
subplot(2,2,3)
plot(Z,Y(:,3),'r-','linewidth',3)
title('T-Distribution, F=3')
subplot(2,2,4)
plot(Z,Y(:,4),'r-','linewidth',3)
title('Near-Gaussian, F=100')



function y=Tdist(Z,M,C,F,D)

%y=Tdist(Z,M,C,F,D)
%Z=Data Point
%M=Mean of Data
%C=Scatter Matrix
%F=Degrees of Freedom
%D=Dimensionality of Data
y=(gamma((F+D)/2)*(det(C))^(-.5))*(1/((pi*F)^(D/2)*gamma(F/2)))...
*(1+((Z-M)'*(inv(C))*(Z-M))/F)^(-(D+F)/2);
```

B.9 Utility for plotting histograms of cluster values

```
function [N,X,F] = Histosv3(Labeled,Mean,Cov,cn)

g=1;
for i=1:(length(Labeled))
      if Labeled(8,i)==cn
            keep(:,g)=Labeled(:,i);
            g=g+1;
      end
end
figure(1)
clf
colordef black

for i=1:7

subplot(3,3,i)

[N,X]=hist(keep(i,:),20)
plot(X,N/(sum(N)*(X(2)-X(1))),'gx')
hold on
M=Mean(i,cn)
C=Cov(i,i,cn)
D=1;
Z=linspace(X(1),X(20));
L=length(Z);

for j=1:L
      Y(j,1)=Tdist(Z(j),M,C,1,D);
end
plot(Z,Y(:,1),'r-')
end
```

B.10 Utility for plotting cluster means and standard deviations

```
function [] = Meanplot(Means,Covariances)

clf;
hold on;
for i=1:8
Upper(:,i)=Means(:,i)+abs(diag(abs(Covariances(:,:,i))).*Means(:,i));
Lower(:,i)=Means(:,i)-abs(diag(abs(Covariances(:,:,i))).*Means(:,i));
end
A=[1 2 3 4 5 6 7];
Upper
Lower
plot(A,Means(:,1),'y-')
plot(A,Upper(:,1),'y:')
plot(A,Lower(:,1),'y:')
plot(A,Means(:,2),'m-')
plot(A,Upper(:,2),'m:')
plot(A,Lower(:,2),'m:')
plot(A,Means(:,3),'c-')
plot(A,Upper(:,3),'c:')
plot(A,Lower(:,3),'c:')
plot(A,Means(:,4),'r-')
plot(A,Upper(:,4),'r:')
plot(A,Lower(:,4),'r:')
plot(A,Means(:,5),'g-')
plot(A,Upper(:,5),'g:')
plot(A,Lower(:,5),'g:')
plot(A,Means(:,6),'b-')
plot(A,Upper(:,6),'b:')
plot(A,Lower(:,6),'b:')
plot(A,Means(:,7),'w-')
plot(A,Upper(:,7),'w:')
plot(A,Lower(:,7),'w:')
```

APPENDIX C: ADDITIONAL PEPTOID SYNTHESES

C.1 Peptoids incorporating fluorescent label and iodoacetamide.....178
C.2 Peptoids incorporating orthonitrobenzyl moiety.....182
C.3 Peptoids incorporating polar, aromatic side groups.....185

C.1 Peptoids incorporating fluorescent label and iodoacetamide

Two peptoids FI1 (Fig C.1) and FI2 (Fig. C.2) were synthesized by the methods described in Chapter 2 of this thesis. The key steps are to (i) introduce a C-terminus-proximal mono-protected diamine, (ii) terminate the growing chain with 2 successive peptide couplings of a hexyl spacer (Novabiochem, La Jolla CA) and a carboxy-fluorescein (Aldrich Chemical Co., Milwaukee WI), and (iii) reveal the C-proximal primary amine and iodoacetylate it.

MALDI-TOF analyses of the two syntheses (Fig. C.3) indicate that minimal desired product was produced in the synthesis of FI1, while a mixture of the iodoacetylated and uniodoacetylated products were produced in the synthesis of FI2.

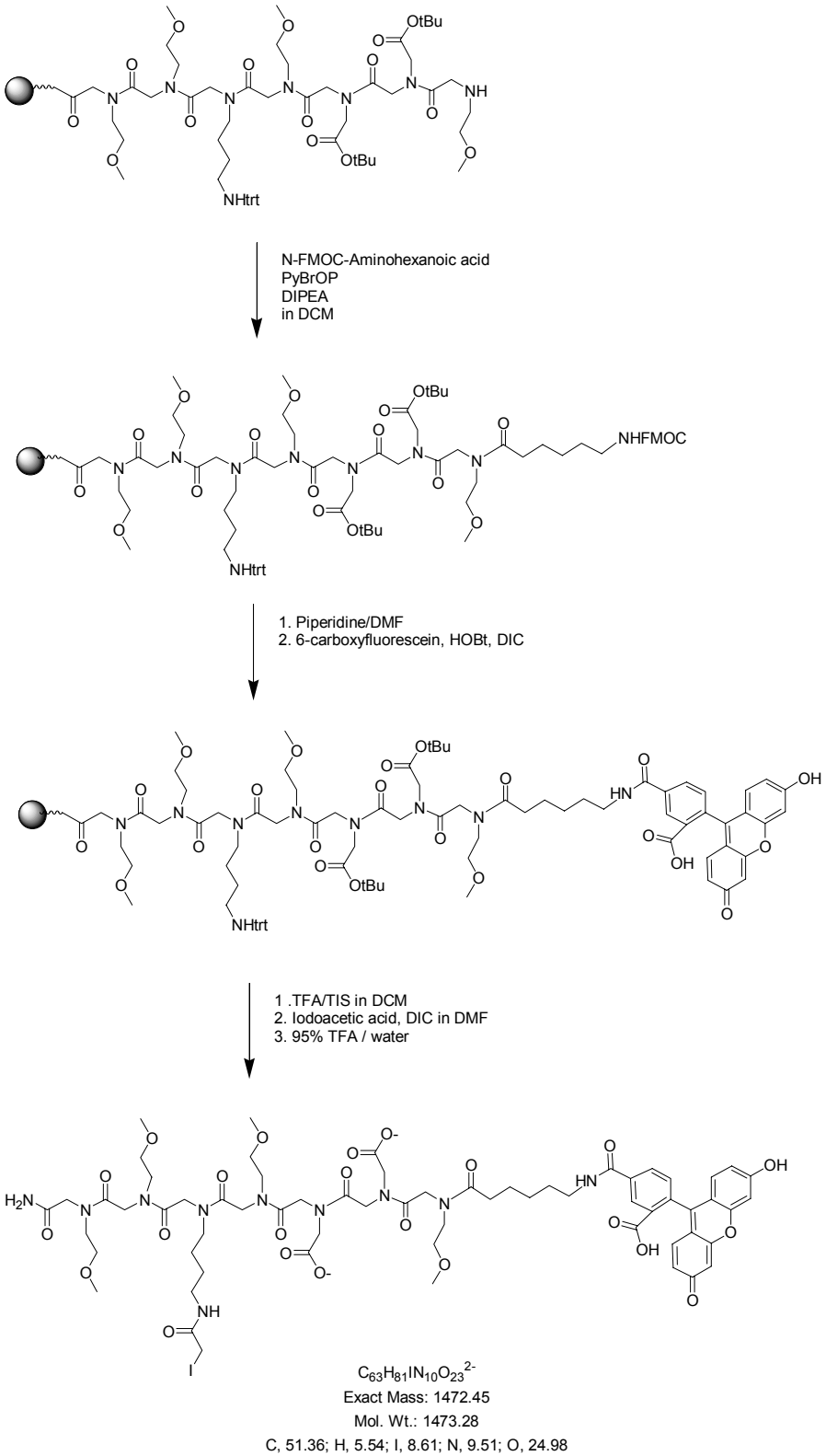


Figure C.1: Schematic illustration of fluoro-iodo peptoid FI1. The iodoacetamide is added near the C-terminus for conjugation to 5′-thiol ODNs.

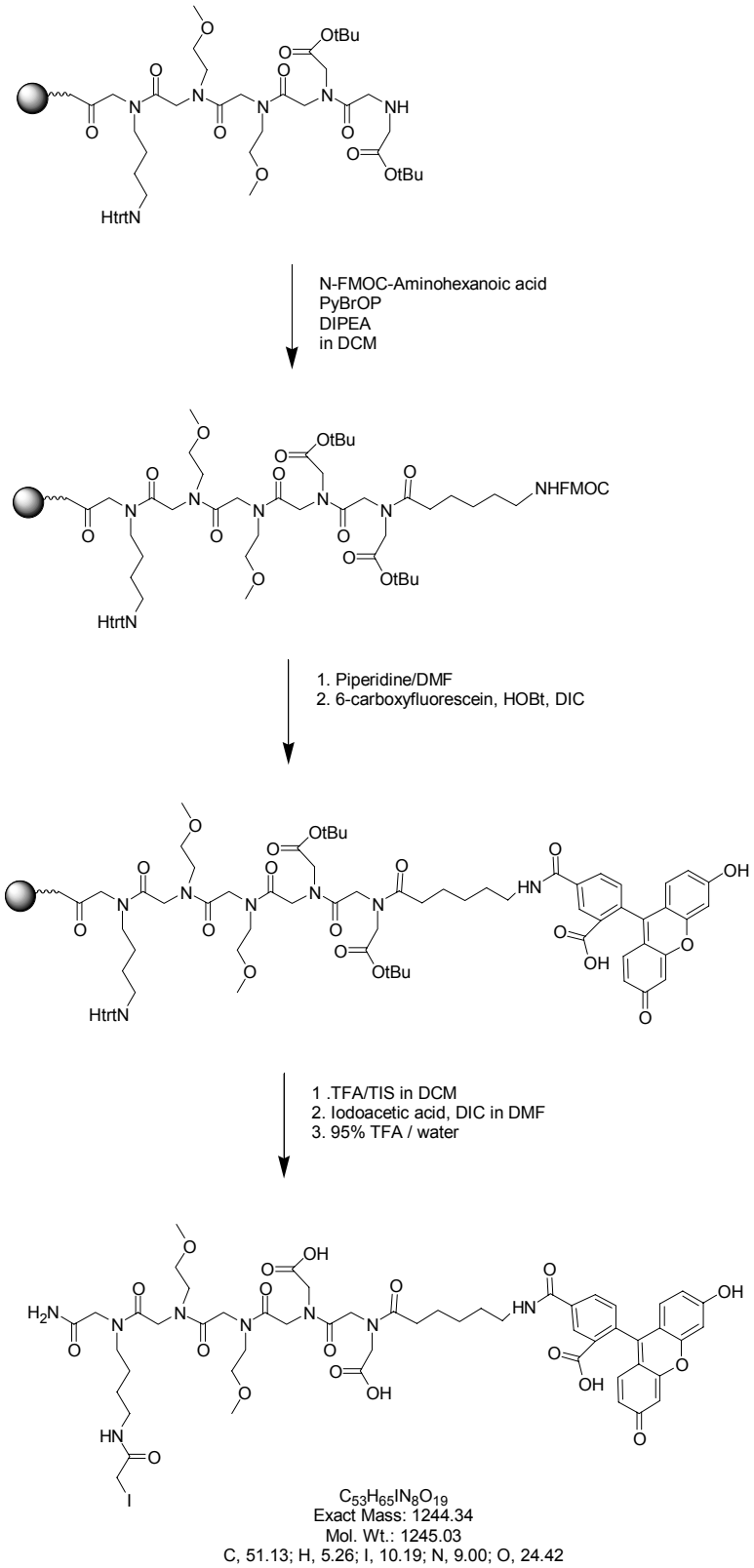


Figure C.2: Schematic illustration of fluoro-iodo peptoid FI2. The iodoacetamide is added near the C-terminus for conjugation to 5′-thiol ODNs.

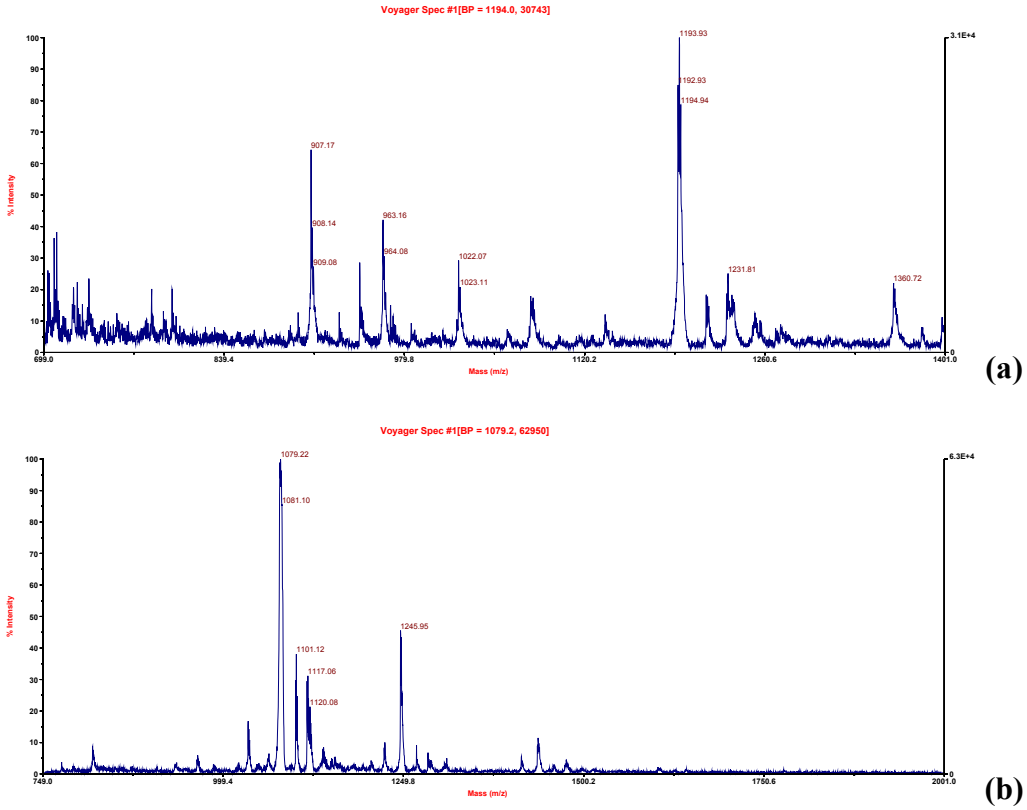
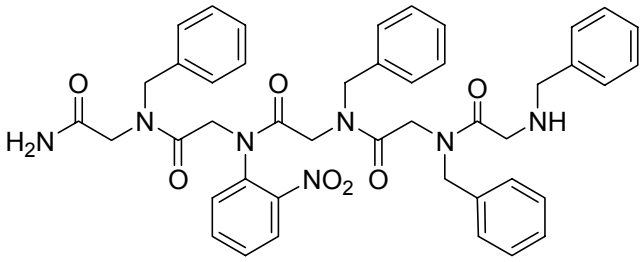


Figure C.3: MALDI-TOF analyses of FI1 (a) and FI2 (b). The iodoacetamide is added near the C-terminus for conjugation to 5′-thiol ODNs.

C.2 Peptoids incorporating orthonitrobenzyl moiety

Two peptoids ON1 and ON2 (Fig. C.4) were produced by the methods described in Chapter 2 of this thesis. They each incorporated the primary amine submonomer ortho-nitroaniline (Aldrich Chemical Co., Milwaukee WI) at the second position in an attempt to introduce a photocleavage site into the peptoids. ESI-Quadrapole analyses of the product mixtures indicate incomplete yield (Fig. C.5 and Fig C.6(a)). Exposure to broad-spectrum UV light for one hour resulted in no appearance of identifiable cleavage products (Fig. C.6).



ON1, ON2
$C_{44}H_{45}N_7O_7$
Exact Mass: 783.34
Mol. Wt.: 783.87
C, 67.42; H, 5.79; N, 12.51; O, 14.29

Figure C.4: Schematic illustration of ortho-nitro aniline-incorporating peptoid ON1 and ON2.

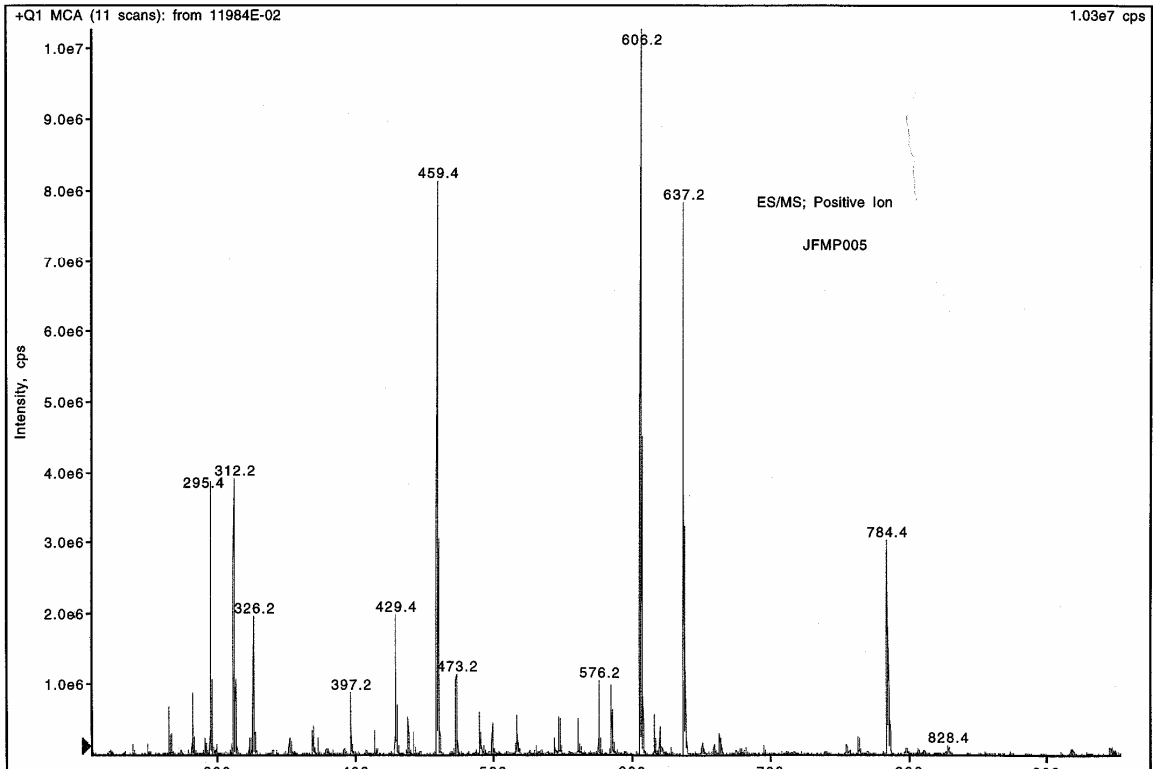


Figure C.5: ESI-Q analysis of ON1 peptoid. The spectrum indicates incomplete yield.

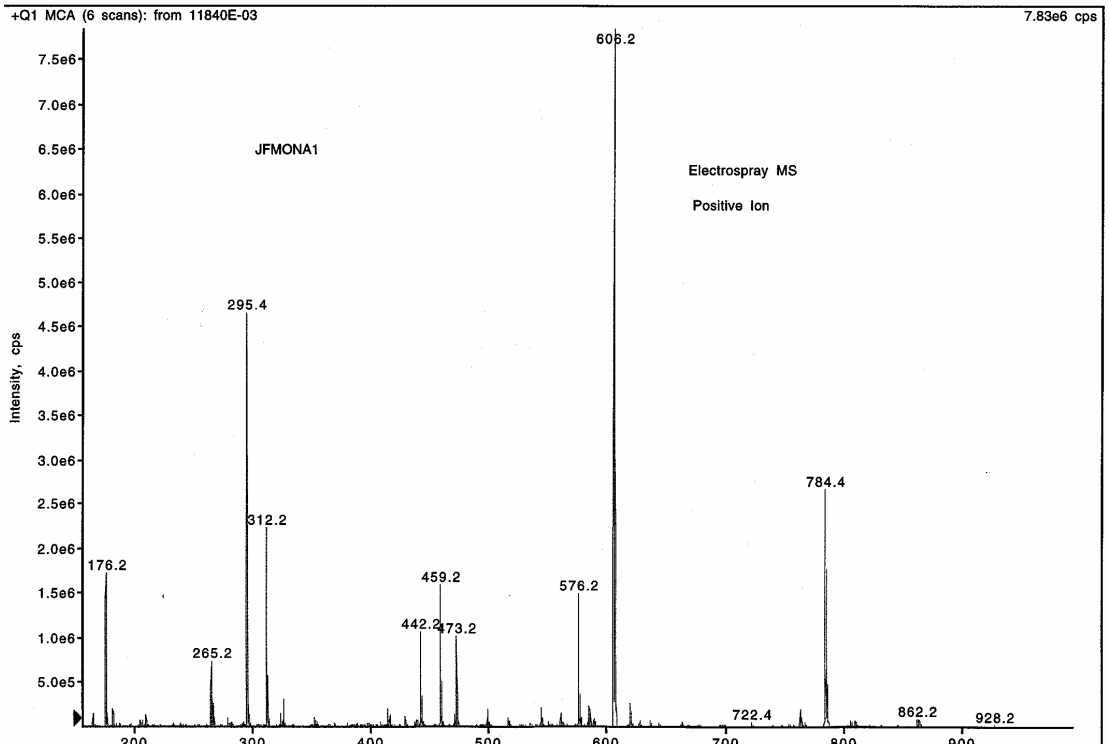


Figure C.6: ESI-Q analyses of ortho-nitro aniline incorporating peptoid ON2 before (a) and after (b) UV radiation.

C.3 Peptoids incorporating polar, aromatic side groups

A series of 11 peptoid 5mers (Fig. C.8) were synthesized by the methods described in Chapter 2 of this thesis. Primary amine submonomers were chosen for these peptoids by comparing them to four common MALDI matrices (Fig C.7).

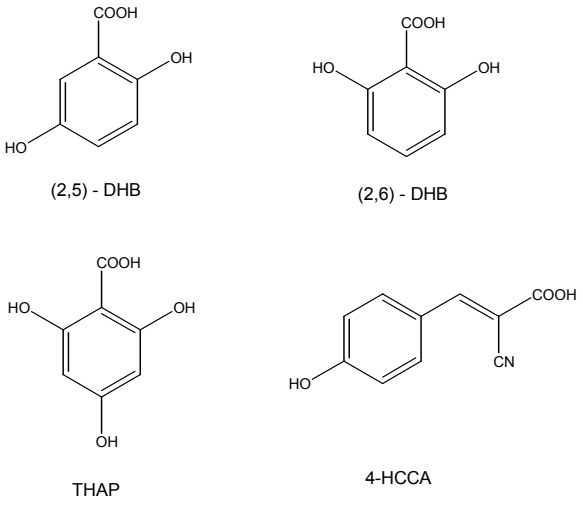
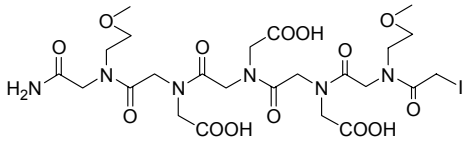
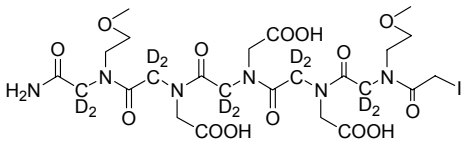


Figure C.7: Schematic illustration of four common MALDI matrices.

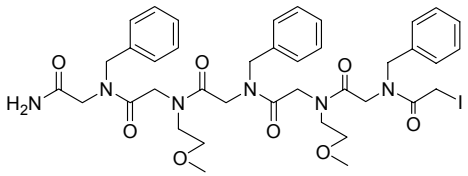
The resulting peptoids were analyzed by MALDI-TOF mass spectrometry using several matrices to determine if they could be detected at lower concentrations than peptoids with non-polar, non-aromatic side groups. The results did not show any sensitivity-enhancing effect of these submonomer choices (data not shown). Two of the peptoids, P057 and P061, were conjugated to 5′-thiol ODNs and photocleaved (see Chapter 2, 3), but the resulting peptoid fragments could not be detected (data not shown).



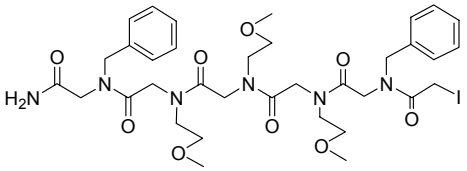
P053: Mol. Wt.: 760.49



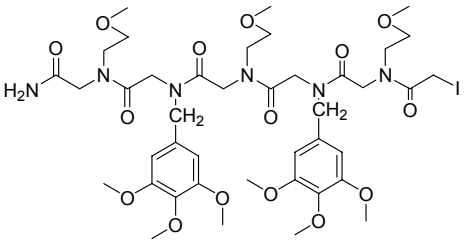
P054: Mol. Wt.: 770.55



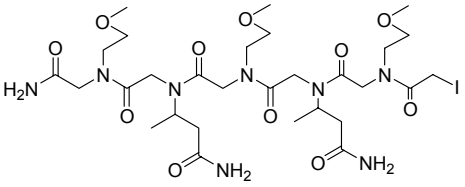
P055: Mol. Wt.: 856.75



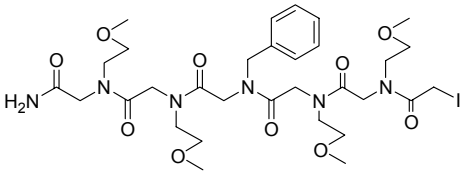
P056: Mol. Wt.: 824.70



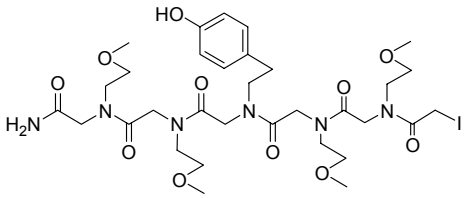
P057: Mol. Wt.: 1004.86



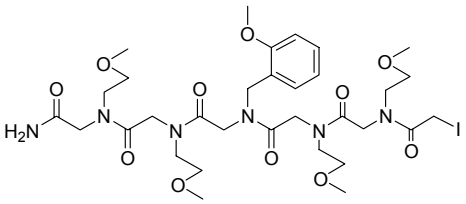
P058: Mol. Wt.: 814.67



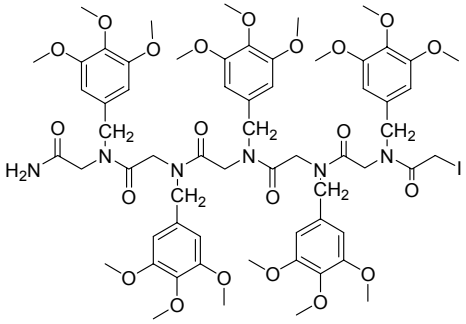
P060: Mol. Wt.: 792.66



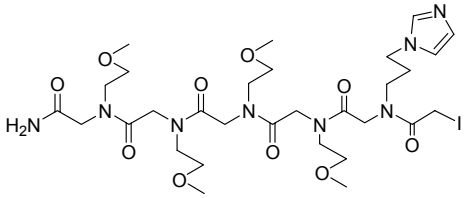
P061: Mol. Wt.: 822.69



P062: Mol. Wt.: 822.69



P064: Mol. Wt.: 1371.22



P063: Mol. Wt.: 810.68

Figure C.8: Schematic illustration of peptoids incorporating polar, aromatic side chains.