Rice University

Zeolites to Peptides: Statistical Mechanics Methods for Structure Solution and Property Evaluation

by

Ramdas S. Pophale

A ThesisSubmitted
in Partial Fulfillment of the Requirements for the Degree
Doctor of Philosophy

Approved, Thesis Committee:

Michael W. Deem, Chair
John W. Cox Professor, Department of Bioengineering and Department of Physics and Astronomy

Ariel Fernandez
Karl F. Hasselman Professor, Department of Bioengineering

Huey W. Huang
Sam and Helen Worden Chair Professor, Department of Physics and Astronomy

Houston, Texas
March, 2009
UMI Number: 3362385

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3362385
Copyright 2009 by ProQuest LLC
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346
Methods in statistical mechanics are used to study structure and properties of zeolites and peptides. A Monte Carlo method is applied to solve structure of a newly synthesized zeolite. Understanding the structure of a zeolite could lead to optimization of chemical processes it is involved in. Dielectric constant and elastic modulus are calculated using a molecular dynamics method for a pure silica zeolite with and without the structure directing agent used in its synthesis. These properties are of interest due to the potential use of this zeolite as low dielectric constant material in manufacturing integrated circuits. Results of four methods probing energy landscapes in peptides are compared.
for four cyclic peptides. Their ability to equilibrate structural properties and their relative speeds are important in their ability to simulate complex structures. A docking study is carried out to probe interactions between two proteins, Cripto and Snail, and E-cadherin promoter. The study supports experimental evidence that Cripto is involved in expression of E-cadherin through a promoter priming mechanism. Finally, the use of computational models in the design of better vaccines is illustrated through an example of Influenza.
Acknowledgments

First and foremost, I would like to thank my adviser Dr. Michael W. Deem for his guidance, patience, and support through my years at Rice. It's been a privilege to work with him on the various challenging research projects. I would like to thank Dr. Ariel Fernandez, and Dr. Huey Huang for agreeing to be on my thesis committee, and taking the time out to review my work.

In last few years, I have been fortunate to have worked with researchers from various disciplines. Dr. Robin Curtis, Dr. Allen Burton, and Dr. Wataru Shinoda have all made pivotal contributions to my progress in different research projects. I would also like to acknowledge my fellow group members in the Deem group for their useful suggestions, and discussions.

Finally, I would like to dedicate this thesis to my brother, and my parents.
Contents

Abstract ii

Acknowledgments iv

List of Figures xi

List of Tables xv

1 Introduction 1

2 Structure Solution of Zeolite CLS-3 5

2.1 Introduction 5

2.2 ZEFSAII 6

2.2.1 Figure of Merit 8

2.2.2 Intensity Calculations 11

2.2.3 Sampling the figure of merit 15
2.3 ZEFSAI results .................................................. 19
  2.3.1 Data collection and Indexing ............................ 19
  2.3.2 Preparation of the input files ................. 22

2.4 Structure Refinement .......................................... 31
  2.4.1 Rietveld Method .......................................... 31

2.5 Rietveld refinement results .................................. 35
  2.5.1 Method of soft restraints ......................... 36
  2.5.2 Organic molecule as a rigid body ............... 41

2.6 Discussion and Conclusions .................................. 42

3 Determination of dielectric constant and elastic modulus for single crystal ZSM-5 52
  3.1 Low dielectric constant (low-k) zeolite films .... 52
3.1.1 Framework structure of ZSM-5 and location of the tetrapropylammonium ion ............................................ 56

3.2 General Utility Lattice Program (GULP) .................................................. 60

3.3 Calculation of the dielectric constant ....................................................... 64

3.3.1 Results ................................................................................................ 65

3.4 Calculation of the elastic constants ............................................................. 70

3.4.1 MPDyn ................................................................................................ 73

3.4.2 Results ................................................................................................ 76

3.4.3 Discussion and Conclusion ................................................................. 81

4 Comparison of Monte Carlo methods for simulation of cyclic peptides 83

4.1 Introduction ............................................................................................. 83

4.2 Simulation Methods .................................................................................. 87

4.2.1 Configurational bias Monte Carlo method ........................................... 87
5 Probing the role of Cripto in the regulation of E-cadherin with docking methods

5.1 Introduction

5.1.1 Docking

5.1.2 Hex

5.1.3 Methods and results

5.1.4 Discussion and conclusions
6 Computer-assisted vaccine design

6.1 Introduction ......................................................... 143

6.2 Methods ................................................................. 146

6.3 Results ................................................................. 149

6.4 Discussion ............................................................... 156
   6.4.1 Suggestions to improve predictability of vaccine efficacy .... 161

6.5 Vaccination strategies and risk management of flu pandemic .... 164
   6.5.1 Multiple-strain introduction transmission model .......... 165
   6.5.2 Hierarchical scale free network ............................ 166
   6.5.3 Virus transmission and evolution ............................. 167
   6.5.4 Vaccination Strategies ......................................... 168
   6.5.5 WHO FluNet data analysis ..................................... 169
   6.5.6 Multiple-component vaccine .................................. 172
   6.5.7 Population at risk ............................................. 173
List of Figures

2.1 X-Ray diffraction pattern for CLS-3. .......................... 20

2.2 Structure directing agent for CLS-3 synthesis. .......................... 25

2.3 Energy histograms from parallel tempering run. .......................... 27

2.4 Structure and coordinate sequence for solution 1. .......................... 29

2.5 Structure and coordinate sequence for solution 2. .......................... 30

2.6 Framework and template structure after refinement: soft restraints method. 39

2.7 Framework and template structure after refinement: rigid body method. 43

2.8 Experimental, calculated and difference plots for diffraction data for the soft restraints method. .......................... 47

2.9 Experimental, calculated and difference plots for diffraction data for the rigid body method. .......................... 48
3.1 ZSM – 5 – TPA1 framework structure, looking down the straight channel

(b-axis) ................................................................. 68

3.2 Energy conservation during MD simulation of ZSM-5. ......................... 77

3.3 Elastic modulus (ZSM-5). ............................................ 79

3.4 Elastic modulus (ZSM-5-TPA). ..................................... 80

4.1 Analytical Rebridging ................................................. 91

4.2 Energy histograms in parallel tempering simulation. ............................ 99

4.3 Anderson thermostat .................................................. 101

4.4 Movement of the lowest temperature replica through the temperature ladder for dWdDPdVL and dPFSAFF .................................................. 104

4.5 Energy histograms at 300 K for four peptides with four simulation methods. 106

4.6 Histograms in radius of gyration histograms at 300 K for four peptides

with four simulation methods. ........................................... 107
4.7 Proline-dihedral-4-4 ................................................................. 108

4.8 $R_g$ histograms at 1500 K for dPFASFF ........................................ 109

4.9 Histograms in torsion angle at 1500 K for dPFASFF ...................... 110

4.10 Proline-hist-diff-temp .............................................................. 111

4.11 Auto-correlation-functions in $R_g$ at 300 K ............................... 114

4.12 Auto-correlations in $R_g$ at 1500 K for dPFSAFF ...................... 116

4.13 dPAAdPAA configurations ....................................................... 117

4.14 Alanine dipeptide with and without GB .................................... 119

5.1 Lowest energy orientation for 1MEY-P4-IDVA system .................. 140

6.1 Correlation between vaccine efficacy and $p_{epitope}$ .................. 151

6.2 Correlation between vaccine efficacy and $p_{sequence}$ .................. 153

6.3 Correlation between vaccine efficacy and ferret antisera derived antigenic distance ................................................ 155
6.4 Average isolate cases: WHO FluNet and simulation data. 170

6.5 Correlation between vaccine efficacy and $p_{\text{epitope}}$. 171

6.6 Vaccine efficacy for single and multi-component vaccine. 174

6.7 Effect of administration delay on the vaccine efficacy for single and multi-component vaccine. 175

6.8 Population at Risk (PaR). 177

6.9 Sensitivity test for network structure. 183

6.10 Sensitivity test for transmission parameters. 184

6.11 Sensitivity test for immune killing. 185

6.12 Sensitivity test for immune history capacity. 186

6.13 Sensitivity test for mutation and diversity parameters. 187

6.14 Sensitivity test for initially infected group. 188

6.15 Sensitivity test for fixation rate. 189
List of Tables

2.1 Weights for terms in the figure of merit. .......................... 16

2.2 Cell size from indexing of CLS-3. ................................. 21

2.3 Intensity dependent weights in ZEFSAlII. ......................... 23

2.4 Temperatures used in parallel tempering. ......................... 26

2.5 Soft restraints. ................................................. 37

2.6 Average bond lengths for the method of soft restraints. ........... 38

2.7 Rietveld refinement results for soft restraints. ................... 40

2.8 Rietveld refinement results for rigid body. ....................... 44

3.1 Fractional coordinates of the framework atoms in ZSM-5 [1]. .... 58

3.2 Fractional coordinates for one of the two orientations of the template atoms [1]. .................................................. 59

3.3 ZSM-5: Dielectric tensor constants ............................... 67
3.4 ZSM – 5 – TPA1: Dielectric tensor constants ............................... 67

3.5 ZSM – 5 – TPA2: Dielectric tensor constants ............................... 69

3.6 Averaged cell length and its statistical error under zero external pressure. 78

4.1 Acceptance probability for swapping moves in a parallel tempering simula-

tion. ........................................................................................................ 100

4.2 Comparison of simulation results of dWdDPdVL in terms of relative CPU
time. ........................................................................................................ 105

4.3 Peptide configurations clustered according to RMSD (0.75 Å) ............. 115

5.1 Energies and ligand interacting DNA base pairs in the top 5 orientations

for the systems of P4-1MEY-1DVA and P1-1DVA. ......................... 139
6.1 Summary of results. The table includes the identities and accession numbers of the vaccine and circulating strains for each of the years since 1971 that the H3N2 virus has been the predominant influenza virus and for which vaccine efficacy data are available. Vaccine efficacy is taken from the literature [?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?] and defined by eq. ??.

Where more than one study exists for the same influenza season, the efficacy results are averaged. The dominant epitope is predicted by our theory for all seasons where the vaccine and circulating strains are not a match.

The $p_{\text{Epitope}}$ and $p_{\text{Sequence}}$ values are calculated using eq. ?? and eq. ??, respectively. Two measures of antigenic distance from ferret antisera assays, $d_1$ [?] and $d_2$ [?], are determined from the literature [?, ?, ?, ?, ?, ?, ?, ?].

Where more than one antisera assay has been performed, the calculated distances are averaged. Error bars are calculated assuming binomial statistics for each data set: $\varepsilon^2 = [\sigma_v^2/u^2/N_v + (v/u^2)^2\sigma_u^2/N_u]$, where $\sigma_v^2 = v(1-v)$ and $\sigma_u^2 = u(1-u)$. If two sets of data are averaged in one year, then $\varepsilon^2 = \varepsilon_1^2/4 + \varepsilon_2^2/4$. 

\[ \varepsilon^2 = \varepsilon_1^2/4 + \varepsilon_2^2/4. \]
Chapter 1

Introduction

The main goal of the research topics included in this Doctoral Thesis is the structure determination of biological and synthetic systems using methods based in statistical mechanics. The thesis is organized in five different chapters. Each chapter documents findings from a different research project, and is self-contained.

The first chapter describes efforts in solving the three dimensional structure of a newly synthesized zeolite from powder diffraction data. Monte Carlo techniques, such as configurational bias, and parallel tempering were used to solve for the structure. Solution thus obtained was refined using a non-linear, least squares fit method called the Rietveld method. Two different approaches were tried to model the non-framework atoms from an organic template molecule. The work was able to generate geometrically feasible structures for the zeolite in question. Useful insights were obtained regarding handling
of the impurity within the powder sample in order to facilitate further efforts towards finding the correct solution.

The second chapter documents determination of dielectric constant and elastic modulus for a single crystal zeolite. Thin film zeolites have emerged as attractive substrates for applications in the semiconductor industry, where manufacturers of integrated circuits are on the look out for low dielectric constant materials. An increase in porosity is often cited as a way to reduce the dielectric constant, however, with increasing porosity, a decrease in mechanical strength becomes an issue. A molecular dynamics approach was used to study the calculate these properties for ZSM-5. Also observed, was the effect of a structure directing organic molecule used during zeolite synthesis on these properties.

Chapter three compares four different combinations of simulation techniques for a set of four cyclic peptides. Cyclic peptides tend to have rough free energy profiles with local energy minima are often separated by high energy barriers. Simple Monte Carlo or molecular dynamics methods by themselves fail to sample the relevant configurations
from a distribution of interest. A combination of these methods with advanced techniques such as rebridging and parallel tempering stands a better chance at equilibrating these systems. The rebridging move improves the equilibration in all the cases. A generalized Born model is implemented to account for the solvent effects and is tested for a simple dialanine peptide.

Chapter four describes results of a docking study. Homologues of protein Cripto and protein Snail, and the E-cadherin promoter were docked using a docking program. It was found that Cripto and Snail favourably bind E-cadherin near the so called E-box. The result supports experimental observations, which show Cripto as a regulating agent for the expression of E-cadherin through a promoter (Snail) priming mechanism.

The final chapter talks about design and improvement in efficacy of Influenza vaccine. A new parameter shows a better correlation between the dominant virus strain and the vaccine than the traditional measures of antigenic distance. A study that combines multiple-strain avian influenza model and the new parameter is able to reproduce the
observed viral fixation rates and with isolates data from World Health Organization FluNet Global Influenza Surveillance Network.
Chapter 2
Structure Solution of Zeolite CLS-3

2.1 Introduction

The derivation of an atomic-scale model of the framework crystal structure of a newly synthesized zeolite is a nontrivial task [2]. Most zeolite samples are poly-crystalline in nature. This, along with typical crystallite sizes (below 5μm) makes structure solution a difficult process. Many known zeolite structures are a result of physical model building efforts. Conventional crystallographic methods have yielded solution in only a limited number of cases [3]. This approach requires a high quality scattering data set for a successful solution. Such data set are not usually obtainable from zeolite samples, explaining the limited applicability of this approach. Of late, electron diffraction methods [4, 5] have been used to solve a few zeolite structures, however, this approach is laborious. Charge-flipping methods have shown substantial promise in recent years [6].

In this chapter, I document my efforts to solve the structure of a newly synthesized
zeolite, CLS-3, from X-ray powder diffraction data collected at a synchrotron source.

This chapter is divided into two sections. In section 1, I explain the process of structure determination of the zeolite using ZEFSAlI (zeolite framework solution), our in-house method. In section 2, I describe the refinement process used to validate the structure obtained from ZEFSAlI.

2.2 ZEFSAlI

ZEFSAlI is a direct, real-space method [2] that applies powerful Monte Carlo techniques to the problem of zeolite structure solution via powder diffraction data. ZEFSAlI stands for zeolite framework solution. It requires easily available powder diffraction data and little preconceived bias on behalf of the researcher. At its heart lies the key step of defining a cost function or a figure of merit that is the function of the atomic positions within the crystalline unit cell and that is minimized by the structure corresponding to the experimental material. The figure of merit is defined for a given zeolite sample with
known unit cell size, cell parameters, symmetry, and density. By definition, the global minimum of the figure of merit should correspond to the structure of the zeolite sample under investigation. Locating this global minimum is a fundamentally challenging problem, owing to the presence of multitude of local minima and large barriers in the cost function. Two versions of the method are available, namely, ZEFSAl and ZEFSAlI. The major difference between the two methods lies in the manner in which they go about minimizing the zeolite figure of merit. ZEFSAl combines a simple Metropolis perturbation step with a simulated annealing procedure. It has a nearly 90% success rate on zeolites with 6 or fewer crystallographically-distinct T atoms. A number of different groups have used this method to solve new zeolite structures [7, 8, 9, 10, 11]. However, in case of more complex structures, the method ends up generating multiple hypothetical structures, often failing to find the correct one. ZEFSAlII, on the other hand, combines biased Monte Carlo moves and simulated annealing in order to overcome the barriers. In the most difficult cases (crystallographically-distinct T atoms > 10), the method of parallel
tempering is used to overcome the barriers in the cost function, and proves superior to
simulated annealing. ZEFSII can solve all of the known zeolite structures, and it was
employed in this research. In the next section I will describe composition of the figure of
merit and the origin of the terms therein in some detail.

2.2.1 Figure of Merit

In ZEFSII there are \( n_{\text{unique}} \) atoms placed in the unit cell. Each of the \( n_{\text{symm}} \) operators
when applied to an unique T-atom generates an atom position, so there can be at most
\( n_{\text{unique}} \times n_{\text{symm}} \) positions of atoms in the unit cell. The cell parameters, the space-group
symmetry, and the number of unique T-atoms are kept constant. This leaves positions
of the T-atoms as the only variables. The figure of merit is defined as follows:

\[
H = \alpha_{T-T} H_{T-T} + \alpha_{T-T-T} H_{T-T-T} + \alpha_{<T-T-T>} H_{<T-T-T>}
\]

\[
+ \alpha_M H_M + \alpha_{uc} H_{uc} + \alpha_{NB} H_{NB} + \alpha_{PXD} H_{PXD} + \alpha_{PND} H_{PND}
\]  
(2.1)

The lower the value of \( H \), the more the particular arrangement of T-atoms resembles
a zeolite of the given cell size, symmetry, and density. Each term in $H$ represents a particular contribution, and the $\alpha_i$ are the relative weights. These weights are optimized according to the rate of success of the method on several trial structures (over a decade ago) and are kept fixed. One begins with a random configuration of T-atoms and seeks the minimum of $H$ by moving the T-atoms suitably. Since the space-group symmetry is enforced, all symmetry-related T-atoms move collectively. Also, the crystalline order is enforced so the variables are limited to asymmetric unit of a single unit cell.

Three different types of terms contribute to Eq. 1. These are, the geometric terms, the density terms, and the diffraction terms. The first three terms of $H$ are geometric contributions that are obtained from 32 known high silica zeolites by histogramming the T-T distances, and the T-T-T angles [12, 13]. As expected for tetrahedrally coordinated species, the T-T distances are distributed around 3.1 Å and the T-T-T angles are distributed around 109.5°. The term $<T-T-T>$ represents the average over all angles around a particular T-atom.
The $H_{NB}$ and $H_{uc}$ terms account for the 4-connectedness of silicates. These terms are defined to be nonzero and positive whenever a T-atom is found to have greater than or fewer than 4 first neighbours, respectively. A neighbour is defined as an atom that is within 4 Å. If there are fewer than 4 neighbours, a progressively higher weight is assigned to the atom. In the case where an atom is found to have greater than 4 neighbours, a list of all $N$ neighbours is prepared. From this list, 4 of the atoms are selected as connected neighbours. For the rest, a repulsive potential energy term is included. The list of $N$ neighbours is searched exhaustively for the combination of 4 connected and $N - 4$ unconnected atoms so as to get a minimum energy associated with the central T-atom.

The $H_M$ term favours merging. Whenever a particular atom sits on a special position, meaning a position that is invariant under symmetry operations other than the identity, merging is favoured. As ZEFSaII assigns positions to atoms in a random manner, the probability of finding an atom exactly on a special position is very low. Therefore, a merging range is defined with $r_M = 0.8$ Å. When two or more symmetry related atoms
fall within this range, they are merged, as in replaced by a single atom at the position
of their center of mass. $H_M$ allocates a negative favourable energy to the merged atoms
that is linearly proportional to the distance between original and the merged position
known as the merging distance. Merging is necessary whenever the number of T-atoms
derived from the experimental density, $n_0$ is less than the number created by symmetry,
$n_{\text{unique}} \times n_{\text{symm}}$. Merging is not allowed whenever $n_0 = n_{\text{unique}} \times n_{\text{symm}}$. In order that
atoms do not collapse on a highly symmetrical position, thus, lowering the density below
the observed one, a term $H_D = (n_T - n_0)^2$ is included to enforce the observed density.
Here, $n_T$ is the actual number of atoms in the unit cell after merging.

2.2.2 Intensity Calculations

The experimental information available about the zeolite is incorporated into the
figure of merit through the diffraction terms, $H_{\text{PXD}}$ and $H_{\text{PND}}$. A typical diffraction
powder X-ray diffraction (PXD) pattern is shown in Fig. 1. To our disposal we have
a list of Bragg reflections with Miller indices \((hkl)\) and relative intensities. Using the standard formulas [14], one can calculate the relative intensities for a given arrangement of atoms in a unit cell and for the same list of reflections. In arbitrary units, the intensity of a reflection is given by the following equation as

\[
I(hkl) = p(\theta)|F_{hkl}|^2,
\]

(2.2)

where \(2\theta\) is the angle of the Bragg reflection, and the term \(p(\theta)\) is the polarization term given by \(p(\theta) = [1 + \cos^2(2\theta)]/[2\sin(\theta) \sin(2\theta)]\) for X-rays and \(p(\theta) = 1/[2\sin(\theta) \sin(2\theta)]\) for neutrons. The scattering amplitude, \(F_{hkl}\), is given by the following equation as

\[
F_{hkl} = \sum_j f_j(k) o_j \exp(-B_j k^2/4) \exp(2\pi i k \cdot x_j),
\]

(2.3)

where

\[
k = h\mathbf{b}_1 + k\mathbf{b}_2 + l\mathbf{b}_3
\]

(2.4)

\[
x_j = m_j^{(1)} a_1 + m_j^{(2)} a_2 + m_j^{(3)} a_3
\]

(2.5)
Here, $a_i$ represent the crystal axes, the $m_j(i)$ represent the crystallographic coordinates, and the $b_i$ represent the reciprocal lattice vectors. The $f_j(k)$ are the form factors for the given atomic species [15]; the $o_j$ are the occupancy numbers, which account for cell positions not always filled with an atom or filled with atoms of different types with different probabilities; and the $B_j$ are the Debye-Waller factors that account for thermal vibrations in the lattice. Since the crystal structure is not known, the information about the occupancies or the Debye-Waller factors is unavailable. Therefore, $o_j$ and $B_j$ are set to 1 and 0.5.

The presence of multiple reflections at angles closer than the resolution is one of the major challenges to the use of powder data. In order to compare calculated intensities with the observed ones, a composite peak is put at the average angle with all the intensity of the multiple reflections into the composite peak.

How well a particular configuration of T-atoms matches against the experimental
powder pattern is measured via the following quantity.

\[
H_{PXD} = \frac{1}{N} \min_{s} \left[ \frac{\sum_{i} (I_{i}^{\text{obs}} - sI_{i}^{\text{calc}})^{2}/\omega_{i}}{\sum_{i} 1/\omega_{i}} \right]
\]  

(2.6)

Here, \(i\) runs over all the \(N\) peaks, composite or otherwise. The \(\omega_{i}\) are the weights, and \(s\) is a global scaling factor. The intensities are relative, and the experimental intensities are scaled so that the largest one is 1000. The expression for the global scaling factor,

\[
s_{\min} = \frac{\sum_{i} (I_{i}^{\text{obs}}/I_{i}^{\text{calc}})/\omega_{i}}{\sum_{i}(I_{i}^{\text{calc}})^{2}/\omega_{i}}
\]

(2.7)

The weights \(\omega_{i}\) account for uncertainty in the experimental as well as calculated data both of which are proportional to the intensity itself.

It is important to choose weights in Equation 1 in a judicious manner. The geometric potentials in equation 1 are smooth, so are the density terms. However, the diffraction terms have a rough energy profile in that they are quite sensitive to the positions of the atoms. ZEFSAILII mixes the geometric terms with the diffraction terms in a manner so
as to smooth out the roughness in the diffraction terms. Table 2.1 lists the values of various weights in the figure of merit. The rate of success in finding structure solution is independent of small changes in the parameters [2].

Finally, the merging term $H_M$ is kept fixed at zero at $r_M = 0.8\text{Å}$ and -300 at $r_M = 0\text{Å}$.

2.2.3 Sampling the figure of merit

In this section I will describe the Monte Carlo algorithm that was used to sample the figure of merit. The traditional Metropolis method does not use information about the energy landscape around the current configuration while proposing trial moves. This is an inefficient strategy for sampling the phase space according to the Boltzmann distribution, as most of the moves try to bring the system to high energy regions of the configuration space, and end up being rejected. The biased Monte Carlo scheme has been shown to improve the sampling efficiency [16, 17, 18]. This method probes the configurations around the current one and proposes moves that are more likely to be accepted. In
Table 2.1  Weights for terms in the figure of merit.

<table>
<thead>
<tr>
<th>weight</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_{T-T}$</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha_{T-T-T}$</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha_{T-T-T}$</td>
<td>2</td>
</tr>
<tr>
<td>$\alpha_D$</td>
<td>30</td>
</tr>
<tr>
<td>$\alpha_M$</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha_{uc}$</td>
<td>1</td>
</tr>
<tr>
<td>$\alpha_{NB}$</td>
<td>1.5</td>
</tr>
<tr>
<td>$\alpha_{PXD}$</td>
<td>1-2</td>
</tr>
<tr>
<td>$\alpha_{PND}$</td>
<td>1-2</td>
</tr>
</tbody>
</table>
ZEFSAII, the biased move proceeds as follows.

Let the current configuration be $A_1$. From the following Gaussian distribution, $k$ random displacements, $\Delta x_i$, are extracted that propose $k$ new configurations, $B_i$.

$$p_i^{\text{int}} = \frac{\exp[-\Delta x_i^2/(2\sigma^2)]}{[2\pi\sigma^2]^{3/2}}$$

(2.8)

A Rosenbluth weight $W$ is constructed as

$$W(n) = \sum_{i=1}^{k} \exp[-\beta H(B_i)],$$

(2.9)

and a normalized probability is assigned to each configuration $B_i$ as

$$p_i^{\text{ext}} = \exp[-\beta H(B_i)]/W(n).$$

(2.10)

One of these configurations, $B_n$, is selected randomly according to its probability. The configuration $B_n$ then becomes the proposed move. The lower the energy is, the more likely it is for the corresponding configuration to be selected. In order to satisfy the detailed balance, the acceptance probability of the proposed move is modified. One needs to calculate the probability of the reverse move $B_n \rightarrow A_1$. The super detailed
balance condition that ensures detailed balance, is satisfied by defining a set of \( k - 1 \) new trial moves \( A_j \), from the proposed configuration \( B_n \) [19]. The set \( \{A_1, A_j\} \) defines the reverse Rosenbluth weight as

\[
W(o) = \exp[-\beta H(A_1)] + \sum_{j=2}^{k} \exp[-\beta H(A_j)].
\] (2.11)

The normalized probability of selecting the reverse move is

\[
p_o^{ext} = \frac{\exp[-\beta H(A_1)]}{W(o)}.
\] (2.12)

The super detailed balance condition is written as

\[
\pi(A_1)T(A_1 \rightarrow B_n)acc(A_1 \rightarrow B_n) = \pi(B_n)T(B_n \rightarrow A_1)acc(B_n \rightarrow A_1),
\] (2.13)

where \( \pi(A) \propto \exp[-\beta H(A)] \) is the limiting distribution that one wants to sample.

The probability of accepting the proposed move is \( acc(A_1 \rightarrow B_n) \). The forward transition probability is \( T(A \rightarrow B_n) \) is just the probability of selecting the configuration, \( p_n^{int} p_o^{ext} \), and the reverse transition probability is \( p_o^{int} p_o^{ext} \). The super detailed balance condition is given by
\[
\frac{\text{acc}(A_1 \rightarrow B_n)}{\text{acc}(B_n \rightarrow A_1)} = \frac{\pi(B_n)p_o^{\text{int}}p_o^{\text{ext}}}{\pi(A_1)p_n^{\text{int}}p_n^{\text{ext}}} = \frac{W(n)}{W(o)}. \tag{2.14}
\]

The acceptance probability is chosen to be

\[
\text{acc}(A_1 \rightarrow B_n) = \min\left(1, \frac{W(n)}{W(o)}\right). \tag{2.15}
\]

A value of \(k = 5\) is used for the biased move. However, at low temperatures, mere biased Monte Carlo moves are unable to sample the rough figure of merit efficiently. For difficult structures (\(n \geq 9\)), sampling is improved dramatically by employing the method of parallel tempering. A temperature ladder is chosen that allows for a reasonable overlapping of energy histograms between the neighbouring temperatures.

### 2.3 ZEFSAII results

#### 2.3.1 Data collection and Indexing

Synchrotron X-ray powder diffraction data (Fig. 2.1) for as-made CLS-3 were collected on the X16C beamline at Brookheaven National Laboratory in Debye-Scherrer
Figure 2.1  X-Ray diffraction pattern for CLS-3.
mode. The data sample were collected for the 'as made' material at ambient tempera-
ture using a step size of 0.005° from 3.5° to 50° 2θ with a wavelength of $\lambda = 1.19958\text{Å}$.

The diffraction pattern was indexed in a triclinic lattice with $\overline{P}_1$ symmetry using TREOR [20]. The cell parameters are given in Table 2.2.

**Table 2.2** Cell size from indexing of CLS-3.

<table>
<thead>
<tr>
<th>cell side</th>
<th>Å</th>
<th>cell angles</th>
<th>°</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>7.508</td>
<td>$\alpha$</td>
<td>94.309</td>
</tr>
<tr>
<td>b</td>
<td>9.509</td>
<td>$\beta$</td>
<td>109.699</td>
</tr>
<tr>
<td>c</td>
<td>15.652</td>
<td>$\gamma$</td>
<td>64.941</td>
</tr>
</tbody>
</table>

The NMR studies of the sample showed 5 unique T atoms with two of the atoms having 3 unique T atoms (instead of the usual 4) in their first coordinate shells. There was a clear indication that the zeolite in question had a layered structure.
2.3.2 Preparation of the input files

The data from diffraction experiment was modified into a suitable format before it was used as an input for the simulations. For given cell dimensions, symmetry operators, and X-ray wavelength ZEFSAlI outputs all of the reflections along with their multiplicities. It assigns random intensities to these peaks. Some of the peaks that are too close to be resolved experimentally are grouped together into a single composite peak. The random intensities were replaced by the experimental intensities. In a composite peak, all other peaks except the last one were assigned a negative intensity. The sum of all the intensities in that group was then assigned to the last reflection in the group. Once all the composite peaks were identified, the intensities were normalized to 1000. Weights were assigned to the peaks according to their intensities as explained in Table 2.3. Larger peak indicates more uncertainty in intensity calculation. These peaks were assigned larger weights so as to count them less while comparing calculated and experimental patterns.
Table 2.3  Intensity dependent weights in ZEFSAII.

<table>
<thead>
<tr>
<th>Scaled intensity range</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-100</td>
<td>1</td>
</tr>
<tr>
<td>100-200</td>
<td>2</td>
</tr>
<tr>
<td>200-300</td>
<td>3</td>
</tr>
<tr>
<td>300-500</td>
<td>4</td>
</tr>
<tr>
<td>≥500</td>
<td>5</td>
</tr>
</tbody>
</table>
The initial fractional coordinates for the Si and O atoms were generated randomly. A density of 10 T atoms/1000 Å³ was considered as suggested by chemical analysis of the ample. The presence of structure directing agent (SDA) and the low symmetry triclinic setting of the unit cell meant one had to account for the scattering density of the SDA. Chemical analysis of the material pointed towards 1 template molecule per unit cell. Therefore, one SDA molecule was included within the unit cell. The positional and rotational degrees of freedom of the SDA molecule entered the figure of merit only through the diffraction terms; energetic interactions between framework atoms and SDA were not considered. Fig. 2.2 shows the SDA molecule. Information about the detailed procedure for running ZEFSAI2 simulations could be found on Deem group's web page [21].

In order to determine temperatures for the parallel tempering, several short simulations with different temperatures were run. Resulting energy histograms were visually inspected. The final temperatures were decided; based on the extent of the overlapping
Figure 2.2  Structure directing agent for CLS-3 synthesis.
between adjacent histograms. Table 2.4 lists the selected temperatures for the cor-responding replicas in parallel tampering simulations.

Table 2.4 Temperatures used in parallel tempering.

<table>
<thead>
<tr>
<th>Replica</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>130</td>
</tr>
<tr>
<td>6</td>
<td>480</td>
</tr>
</tbody>
</table>

In all, 25 runs were carried out for a given set of input files. Using different random numbers, the figure of merit was sampled in independent ensembles. The structures from the runs were scanned in order to collect those structures that had reasonably low
energies ($\leq 1000$ Kcal/mol), and had in their structure at the most two T atoms that were 3 coordinated. Fig. 2.3 shows the energy histogram for a representative run. The histograms overlap reasonably well indicating that the replicas were able to move freely across the temperature ladder.

**Figure 2.3** Energy histograms from a representative parallel tempering run at different temperatures ($k_B = 1$).
Scanning the results of all the simulations resulted in two structures being found in 18 of the 25 runs. Out of the 18 runs, in 11 cases one of the two structures had the lowest energies for that simulation. The structures were closely related to each other. Solution 1 had two T atoms that were 3-coordinated, whereas, solution 2 had all 5 T atoms that were 4-coordinated. The structures are composed of cages with 2 four-rings, and 4 five-rings. These cages were connected together to form chains in the two directions within the layers. Fig. 2.4 and Fig. 2.5 represent the coordination sequence and structure for the two cases. Structure 2 can be obtained by adding two bonds along the a-axis in structure 1.
minimum occupancy: 3
maximum occupancy: 4
number of T-atoms: 10

<table>
<thead>
<tr>
<th>IZA Atlas coordination sequence</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 9 15 19 30 35 33 39 51 254</td>
<td></td>
</tr>
<tr>
<td>4 9 13 20 23 26 31 40 41 43 250</td>
<td></td>
</tr>
<tr>
<td>3 7 12 19 24 27 31 37 43 45 248</td>
<td></td>
</tr>
<tr>
<td>4 10 14 16 25 28 31 36 42 46 252</td>
<td></td>
</tr>
<tr>
<td>3 7 12 21 24 23 35 38 39 46 248</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1252</td>
</tr>
</tbody>
</table>

Figure 2.4 Structure and coordinate sequence for solution 1.
minimum occupancy: 4
maximum occupancy: 4
number of T-atoms: 10

<table>
<thead>
<tr>
<th>IZA Atlas coordination sequence</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 10 16 21 27 36 42 47 54 64</td>
<td>321</td>
</tr>
<tr>
<td>4 10 15 24 27 34 40 54 52 62</td>
<td>322</td>
</tr>
<tr>
<td>4 10 16 25 31 35 41 51 56 59</td>
<td>328</td>
</tr>
<tr>
<td>4 10 16 21 27 36 42 47 54 64</td>
<td>321</td>
</tr>
<tr>
<td>4 10 16 25 31 35 41 51 56 59</td>
<td>328</td>
</tr>
</tbody>
</table>

Figure 2.5 Structure and coordinate sequence for solution 2.
2.4 Structure Refinement

Solution 1 obtained from ZEFSAlII was further refined using the Rietveld refinement method [22]. The software GSAS/EXPGUI was used to carry out the refinement process. GSAS (General Structure Analysis System) [23] is a comprehensive set of tools to refine the structural models to both X-ray and neutron diffraction data. EXPGUI [24] is a graphical user interface (GUI) editor for GSAS experiment files and shell which allows all other GSAS programs to be executed with a GUI. In the following section, I explain the Rietveld method for structure refinement in brief.

2.4.1 Rietveld Method

A powder diffraction pattern of a crystalline material comprises of individual reflection profiles. Each profile has a peak position, peak height, breadth, tails that decay gradually from the peak position, and an integrated area proportional to the Bragg intensity, $I_{hkl}$, where $hkl$ represent the miller indices. Analysis of the measurements of these peaks,
can reveal several aspects of the material’s structure. In order to obtain such structural information from diffraction experiments, Rietveld proposed a method that uses a least squares approach to refine a theoretical line profile till it matches the experimentally measured profile. The method was first applied to neutron data, but was soon extended to X-ray diffraction data [25, 26]. Rietveld refinement is now commonly applied to both laboratory and synchrotron X-ray data [27].

For powder data obtained from a constant wavelength source, numerical intensity value, \( y_i \), is measured at each increment, \( i \), in scattering angle \( 2\theta \). With a good starting point at hand, the objective is to obtain the best least-squares fit to all of the thousands of \( y_i \)’s simultaneously. The refinement minimizes the residual, \( S_y \).

\[
S_y = \sum_i w_i (y_i - y_{Ci})^2
\]  

(2.16)

where \( w_i = 1/y_i \), statistical weight,

\( y_i = \) observed intensity at the \( i \)th step,
\[ y_{ci} = \text{calculated intensity at the } i\text{th step}, \]

and the sum is over all data points.

In most cases, not all the peak profiles are resolved, but rather partially overlap one another. Normally, many Bragg reflections contribute to the intensity, \( y_i \), observed at any point, \( i \), in the pattern. In order to determine the calculated intensities, \( y_{ci} \), contributions from neighbouring Bragg reflections are taken into account along with the background:

\[
y_{ci} = s \sum_{hkl} L_{hkl} |F_{hkl}|^2 \phi(2\theta_i - 2\theta_{hkl}) P_{hkl} A + y_{bi}\tag{2.17}
\]

where \( s \) is the scale factor,

\( L_{hkl} \) comprises of the Lorentz, polarization, and multiplicity factors,

\( \phi \) is the reflection profile function,

\( P_{hkl} \) is the preferred orientation function,

\( 2\theta_i \) is the peak position

\( 2\theta_{hkl} \) is the center of the peak
A is an absorption factor,

\( F_{hkl} \) is the structure factor for the \( hkl \) Bragg reflection, and

\( y_{bi} \) is the background intensity at the \( i \)th step.

Minimization procedures via Newton-Raphson with least squares lead to a set of normal equations that involve derivatives of all the calculated intensities with respect to each adjustable parameter. These equations are solved by inverting the normal matrix given by the following equation.

\[
M_{jk} = - \sum_i 2w_i[(y_i - y_{ci})\frac{\partial^2 y_{ci}}{\partial x_j \partial x_k}]
\]  \( \text{(2.18)} \)

Here the parameters \( x_i, x_j \) represent the adjustable parameters. As the residual function is non-linear, an iterative procedure is required to find the solution. The shifts, \( \Delta x_k \), are given as follows:

\[
\Delta x_k = \sum M_{jk}^{-1} \frac{\partial S_y}{\partial X_k}
\]  \( \text{(2.19)} \)

An improved model is obtained by applying the shifts to the original parameters.
The procedure is then repeated until convergence to a minimum. Due to the non-linear relationship between the intensities and the adjustable variables, the starting point needs to be close to the correct model. Otherwise, the procedure will either diverge or lead to a false minimum. It is important to note that the Rietveld method is a structure refining technique, not a structure solution method, and needs a good starting point to achieve convergence.

2.5 Rietveld refinement results

The Rietveld refinement of the as-made CLS-3 from synchrotron X-ray diffraction data was performed using the space group $P1$. The intense 001 peak was excluded from the refinement because of severe peak shape problems at low angles. During indexing of the powder data, three peaks were unaccounted for. These peaks were at $2\theta$ values of 17.3°, 21.3°, and 23.0° respectively. These peaks were thought to be arising from a small zeolite beta impurity, and were consequently dropped from the refinement.
The selected structure from ZEFSAI II runs was used as a starting point. Cerius package was used to insert oxygens and to sever the long Si-O-Si bonds between the layers and to terminate them with the silanol-siloxy pairs for the refinement.

The pseudo-Voigt function of Thompson et. al. [28] and the asymmetry correction described by Finger et. al. [29] were used to model the peak profiles. The 'Chebyshev polynomial of the first kind' was used to model the background contribution. Atoms of the same type were grouped together and were constrained to have the same isotropic thermal displacement factors. Two different methods were employed for the treatment of the organic molecule during the refinement stage.

2.5.1 Method of soft restraints

Method of soft restraints treats the SDA as a flexible molecule. GSAS provides a set of stereo chemical restraints with an appropriate standard deviation assigned to each. A weight factor determines the extent of effect a restraint would have on the minimization
process. In the case of CLS-3, bond-length restraints were applied. Table 2.5 lists the restraints placed on different types of bonds involved, and their respective standard deviations. The O-Si-O angles were restrained indirectly by restraining the Si-O and O-O distances.

<table>
<thead>
<tr>
<th>Bond type</th>
<th>Restraint (Å)</th>
<th>Standard deviation (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si-O</td>
<td>1.60</td>
<td>0.03</td>
</tr>
<tr>
<td>O-O</td>
<td>2.61</td>
<td>0.05</td>
</tr>
<tr>
<td>C-C</td>
<td>1.54</td>
<td>0.03</td>
</tr>
<tr>
<td>C=C</td>
<td>1.40</td>
<td>0.03</td>
</tr>
<tr>
<td>C-N</td>
<td>1.47</td>
<td>0.05</td>
</tr>
</tbody>
</table>

At the beginning of refinement the weights were set to a large value (250) to avoid diverging refinement. Table 6 provides the refined atomic position parameters for the
zeolite as well as for the organic template. Table 2.6 provides the average bond distances and the maximum and minimum values of the bond lengths for the different types of bonds.

**Table 2.6** Average bond lengths for the method of soft restraints.

<table>
<thead>
<tr>
<th>Bond type</th>
<th>Average bond distance (Å)</th>
<th>Max (Å)</th>
<th>Min (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si-O</td>
<td>1.61</td>
<td>1.65</td>
<td>1.56</td>
</tr>
<tr>
<td>O-O</td>
<td>2.61</td>
<td>2.69</td>
<td>2.51</td>
</tr>
<tr>
<td>C-C</td>
<td>1.56</td>
<td>1.61</td>
<td>1.54</td>
</tr>
<tr>
<td>C=C</td>
<td>1.42</td>
<td>1.44</td>
<td>1.41</td>
</tr>
<tr>
<td>C-N</td>
<td>1.52</td>
<td>1.56</td>
<td>1.51</td>
</tr>
</tbody>
</table>

After nearly 700 cycles, the refinement converged to a $R_p$ value of 22.14% and a $R_{wp}$ value of 27.03%. The refinement results are summarized in Table 2.7. The structure at the end of refinement (with restraint weight = 250) is shown in Figure 2.6.
Figure 2.6  Framework and template structure after refinement for the soft restraints method.
<table>
<thead>
<tr>
<th>space group</th>
<th>$\overline{P}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>7.50Å</td>
</tr>
<tr>
<td>$b$</td>
<td>9.51Å</td>
</tr>
<tr>
<td>$c$</td>
<td>15.66Å</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>94.30°</td>
</tr>
<tr>
<td>$\beta$</td>
<td>109.72°</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>64.97°</td>
</tr>
<tr>
<td>wavelength</td>
<td>1.19958Å</td>
</tr>
<tr>
<td>profile range</td>
<td>7.05°—50.0°</td>
</tr>
<tr>
<td>$R_p$</td>
<td>21.91%</td>
</tr>
<tr>
<td>$R_{wp}$</td>
<td>26.93%</td>
</tr>
<tr>
<td>$d$</td>
<td>0.075</td>
</tr>
</tbody>
</table>
2.5.2 Organic molecule as a rigid body

The second method treats the organic template as a rigid molecule. This is a much more involved process than that of soft restraints. It is especially useful when working with powder data. Treating the template as a rigid body avoids meaningless changes within the molecule. The molecule undergoes only translational or rotational changes. Theoretically, only 6 parameters are needed to handle the rigid body, 3 parameters to define the position \((x, y, z)\) of the center, and 3 rotation parameters. GSAS, however, reads in 3 additional rotation parameters. Such treatment of the template saves computing time, and enables inclusion of light atoms (hydrogens) from the beginning. Overall, this method leads to a more stable refinement where the probability of divergence is reduced, and one is more likely to converge on to the correct structure [30].

In GSAS, the rigid body coordinates are given as Cartesian coordinates. For small molecules, the coordinates for rigid bodies are calculated from the knowledge of bond
lengths and bond angles \([30, 31]\). Bigger molecules are rather difficult, but if one knows an already existing structure, where the atomic coordinates are known, a shortcut could be applied \([32]\). A center is chosen for the rigid body. The center of gravity is a good choice. The respective coordinates of the center of gravity are subtracted from the respective coordinates of all other atoms \((x\text{ coordinate of the center from } x\text{ coordinate of an atom, and so on})\). These new coordinates are inserted under the option of rigid body in GSAS.

The results of the refinement following inclusion of the SDA as a rigid body are tabulated in Table 2.8. After nearly 850 cycles, the refinement converged to a \(R_p\) value of 28.31\% and a \(R_{wp}\) value of 32.99\%. The structure of the zeolite and the SDA template are shown in Figure 2.7.

2.6 Discussion and Conclusions

The Rietveld refinement tries to minimize the residual in equation 20 through adjustments to the refinable parameters until the 'best fit' is obtained between the calculated
Figure 2.7  Framework and template structure after refinement for the rigid body method.
Table 2.8  Rietveld refinement results for rigid body.

<table>
<thead>
<tr>
<th>space group</th>
<th>$\overline{P}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>7.50Å</td>
</tr>
<tr>
<td>$b$</td>
<td>9.51Å</td>
</tr>
<tr>
<td>$c$</td>
<td>15.67Å</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>94.30°</td>
</tr>
<tr>
<td>$\beta$</td>
<td>109.72°</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>64.99°</td>
</tr>
<tr>
<td>wavelength</td>
<td>1.19958Å</td>
</tr>
<tr>
<td>profile range</td>
<td>7.05°–50.0°</td>
</tr>
<tr>
<td>$R_p$</td>
<td>28.31%</td>
</tr>
<tr>
<td>$R_{wp}$</td>
<td>32.99%</td>
</tr>
<tr>
<td>$d$</td>
<td>0.059</td>
</tr>
</tbody>
</table>
and the observed pattern. There are several criteria available to check the goodness of the 'best fit' obtained during the Rietveld refinement. A particular 'best fit' depends on the adequacy of the model, and on whether the procedure reached a global or a local (false) minimum.

In literature, people often provide the R-pattern ($R_p$) and R-weighted pattern ($R_{wp}$) values to indicate the goodness of the fit obtained. For a correctly solved structure that is well refined, the R terms range from 5% – 15% [33, 34]. These values are defined as follows:

$$R_p = \frac{\sum_i |y_i(\text{obs}) - y_i(\text{calc})|}{\sum_i y_i(\text{obs})}$$ (2.20)

$$R_{wp} = \left( \frac{\sum_i w_i [y_i(\text{obs}) - y_i(\text{calc})]^2}{\sum_i w_i (y_i(\text{obs}))^2} \right)^{1/2}$$ (2.21)

The R-values have in their numerator, the residual that is being minimized, hence from mathematical point of view, these values are important, and best reflect progress of the refinement. Another statistic that is often quoted is the Durbin-Watson statistic,
It is defined as

\[ d = \frac{\sum_{i=2}^{N} (\Delta y_i - \Delta y_{i-1})^2}{\sum_{i=1}^{N} \Delta y_i^2}, \]  

(2.22)

where \( \Delta y_i = y_i(\text{obs}) - \bar{y} \).

It reveals the serial correlation between the successive \( y_i \) values. The ideal value for this statistic is 2.00 [27]. In case that the calculated and observed profiles do not match well due to reasons of shape and/or area, there is a strong serial correlation of the residuals. The D-W \( d \) in both, soft restraints and rigid body cases are far from 2.00.

Although numerical criteria are important, these values could be misleading at times. The difference plot, along with the observed and calculated peak profiles provides a better indication of the direction in which the refinement is heading. Errors in such parameters as scale factor, lattice parameters, zero offset, and structure are obvious in the plots, but not necessarily in the R values. These plots for both, soft restraints and rigid body treatment of SDA can be seen in Fig. 2.8 and Fig. 2.9 respectively.
Figure 2.8  Experimental, calculated and difference plots for diffraction data for the soft restraints method. Peaks at 2θ values of 17.3°, 21.3°, and 23.0° that could not be indexed were omitted from the refinement.
Figure 2.9  Experimental, calculated and difference plots for diffraction data for the rigid body method. Peaks at 2θ values of 17.3°, 21.3°, and 23.0° that could not be indexed were omitted from the refinement.
Rietveld refinement of the structure obtained from ZEFSAII led to $R_p$ and $R_{wp}$ values between 25 % - 30 % range. The background function looks reasonably well fitted. However, the difference plots show a disagreement between the observed and fitted peak profiles. Few of the observed peaks are missing from the calculated profile, indicating a possible problem with unit cell size or space group. However, since the indexing was able to account for all but three peaks (that were thought to be due to a zeolite beta impurity), it is more likely that the starting point of the refinement is not close enough to the correct solution. The non-linear refinement method is unable to overcome this handicap, and ends up settling into what looks like a false minima. This fact is reflected in the less than ideal D-W $d$ values for both the cases.

It is clear from Fig. 6 that in spite of the use of the restraints with larger weights, the structure of the SDA does not look chemically sensible. Clearly, the atoms in the template are too close to each other. The rigid body treatment fared better in this sense, however, refinement with that procedure led to higher $R_{wp}$ and $R_w$ values, and a more
pronounced difference plot. It is clear that the soft restraints method is able to achieve better statistics through unphysical movement of the atoms involved. Further reduction in the weights led to complete condensation of the template atoms (not shown). Rigid body method on the other hand is better in keeping the geometry intact of larger template molecules.

Failure of the refinement method thus implies that the starting structure is not close enough to the correct one. There are three possible reasons. The CLS-3 has an open framework structure, in that there are 2 atoms that are 3-coordinated. ZEFSAlI favours 4-coordinated T atoms. For 3-coordinated structures, there are many more possible structures. The second factor is the presence of the organic template. In its presence, in the regions that are occupied by the non-framework atoms, the $H_{\text{PXD}}$ term tends to favour nonzero scattering density. This leads to ambiguity in the diffraction term. Due to these two factors the method ends up finding several incorrect structures that could still be feasible from a geometrical point of view. A diffraction pattern obtained
from a calcined zeolite sample, where the template molecule has been removed, would be an ideal starting point for ZEFSAII. However, layered materials are usually unable to remain stable after calcination. The layers are unable to connect with one another to form an intact zeolite structure. The opposing layers do not line up well enough for the layers to neatly condense. The efforts to calcine the CLS-3 sample failed as the material became completely amorphous. The third difficulty is that the sample may be disordered or multiple phase. For example, the peaks at 2θ values of 17.3°, 21.3° and 23.0° in the PXD indicate some potential presence of zeolite beta.

In conclusion, the solution to the layered material CLS-3 has yet proved to be elusive to ZEFSAII. More efforts on both the experimental and computational level could yet lead to the correct solution that could then be verified through the Rietveld refinement. The existence of another phase also needs to be investigated since the indexing fails to account for 3 peaks.
Chapter 3

Determination of dielectric constant and elastic modulus for single crystal ZSM-5

3.1 Low dielectric constant (low-\( k \)) zeolite films

This project was a collaborative effort with Dr. Yushan Yan, University of California at Riverside (UCR), and with Dr. Wataru Shinoda, National Institute of Advanced Industrial Science and Technology, Japan.

A distinct feature of zeolites is the uniform nanometer-scale pore structure (diameter 3 - 10 Å) of these materials. This feature has been exploited in industry with great success. Zeolites have been used as catalysts and separation media for decades. Micro-meter scale zeolite crystals are usually combined with a binder in order to form millimeter-scale granules or pellets in these applications. Recently, zeolite nano-particles in a thin film configuration have been proposed, studied, and shown to hold promise for a plethora of new applications [35]. These applications include low dielectric constant films critical to
the speed of future generation computer chips among others.

Since the inception of the integrated circuit (IC) in 1961, the evolution of the IC industry has taken place at a remarkable speed resulting in the manufacturing of fast and affordable computers. The ability to increase the number of transistors on a single chip through miniaturization has been at the root of this speedy evolution. In order to continually increase the number of transistors per chip, a low-$k$ material is needed for effective insulation of the metal wires connecting the transistors that will minimize the crosstalk noise, power dissipation, and RC delay [36, 37]. The International Semiconductor Technology road map (www.sematech.org) portends the need of a low-$k$ material with a $k$ value of 1.6 by 2010. No manufacturing processes that give such a low value are known at present. A number of candidates have been looked into as potential low-$k$ materials. For dielectric films, nonporous silica ($k = 4$) has been the preferred material since the beginning of the semiconductor industry. It has been suggested to increase the porosity of silica to lower its $k$ value, since $k_{\text{air}} = 1$. Higher porosity, however, weakens
the film and renders it fragile during the chemical and mechanical processing. A balance between low dielectric constant and high mechanical strength is crucial to the use of any material as a thin film substrate for integrated circuits. There are three major classes of porous silica, namely sol-gel silica, surfactant-templated mesoporous silica and pure-silica-zeolite (PSZ) that are being looked into for possible uses as thin dielectric films.

With the use of super critical drying in aero-gel creation, sol-gel silica can be designed to have very high porosity resulting into a very low $k$ value ($\sim 1.2$) [38, 39, 40]. However, there are several practical problems. The high porosity in sol-gel silica results into low mechanical strength and low thermal conductivity. The threshold for elastic modulus as established by the semiconductor industry for the low-$k$ materials is 6 GPa. The drying induced shrinkage is another drawback, so is the wide-pore size distribution which can cause electric breakdown. Finally, since sol-gel silica is hydrophilic, it tends to absorb water. Due to its high dielectric constant ($k=80-90$) even minor adsorption of water can
drastically increase the $k$ value.

Surfactant-templated mesoporous silica has a pore size distribution that is narrow compared to sol-gel silica resulting in a better mechanical strength. However, due to its amorphous nature, mesoporous silica still suffers from low mechanical strength and high hydrophobicity.

Pure silica zeolites (PSZs) on the other hand posses high mechanical strength and heat conductivity because of their dense crystalline structure. They have small pores (< 1 nm) and a very narrow pore size distribution, thus minimizing the electric breakdown problem. Also, PSZs are hydrophobic, resulting in reduced water adsorption. Due to these qualities, researchers at UCR have studied PSZ low-$k$ films for past few years [41]. The efforts have been focused on MFI (ZSM-5). Not only because it is highly hydrophobic and has high porosity ($\sim 33\%$), but also because there are several synthesis procedures available for MFI.

For low-$k$ film applications, two of the most important properties are dielectric con-
stant and elastic modulus. Calculation of these and other mechanical properties for two systems, ZSM-5 and ZSM-5 with an embedded tetrapropylammonium (TPA) ion, is described in this chapter.

3.1.1 Framework structure of ZSM-5 and location of the tetrapropylammonium ion

The initial structure for ZSM-5 and TPA system was taken from Koningsveld et al. [1]. The structure was solved using X-ray diffraction method on a single crystal of Si$_{11.96}$Al$_{0.04}$O$_{24.5}$NC$_{12}$H$_{28}$OH. All twelve SiO$_4$ groups in the solved structure resemble the ideal tetrahedral geometry. The average Si-O distance in each tetrahedron is 1.587 Å, and the average O-Si-O angle is 109.47°. ZSM-5 has ten-membered rings that define a three dimensional channel system. There are two types of channels, straight, and sinusoidal. The straight channels are along [010] and sinusoidal channels are parallel to [100]. The structure was refined in Pnma symmetry (8 symmetry operators). The framework symmetry deviation was found to be too small to refine in a lower-symmetry group.
Pn2$_1$a. The dimensions for the orthorhombic unit cell were reported as $a = 20.022(2)$ Å, $b = 19.899(2)$ Å, and $c = 13.383(1)$ Å.

The TPA$^+$ ion populates the intersection of the straight and sinusoidal channels in two different orientations in a ratio of 3:2. The positional parameters for the framework are given in Table 3.1, and those for one of the two orientations of the template are given in Table 3.2.
Table 3.1  Fractional coordinates of the framework atoms in ZSM-5 [1].

<table>
<thead>
<tr>
<th>Atom</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
<th>Atom</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si(1)</td>
<td>0.42238(5)</td>
<td>0.05650(6)</td>
<td>-0.33598(9)</td>
<td>O(8)</td>
<td>0.3085(2)</td>
<td>-0.1552(2)</td>
<td>-0.0728(3)</td>
</tr>
<tr>
<td>Si(2)</td>
<td>0.30716(6)</td>
<td>0.02772(6)</td>
<td>-0.18930(9)</td>
<td>O(9)</td>
<td>0.1980(2)</td>
<td>-0.1554(2)</td>
<td>0.0288(3)</td>
</tr>
<tr>
<td>Si(3)</td>
<td>0.27911(6)</td>
<td>0.02772(6)</td>
<td>-0.18930(9)</td>
<td>O(10)</td>
<td>0.0910(2)</td>
<td>-0.1614(2)</td>
<td>-0.0777(3)</td>
</tr>
<tr>
<td>Si(4)</td>
<td>0.12215(6)</td>
<td>0.06298(6)</td>
<td>0.02670(9)</td>
<td>O(11)</td>
<td>0.1169(2)</td>
<td>-0.1578(2)</td>
<td>-0.2694(3)</td>
</tr>
<tr>
<td>Si(5)</td>
<td>0.07128(6)</td>
<td>0.02722(6)</td>
<td>-0.18551(9)</td>
<td>O(12)</td>
<td>0.2448(2)</td>
<td>0.1594(3)</td>
<td>0.2422(3)</td>
</tr>
<tr>
<td>Si(6)</td>
<td>0.18641(5)</td>
<td>0.05896(6)</td>
<td>-0.32818(8)</td>
<td>O(13)</td>
<td>0.3047(3)</td>
<td>-0.0510(2)</td>
<td>-0.1866(3)</td>
</tr>
<tr>
<td>Si(7)</td>
<td>0.42265(6)</td>
<td>-0.17250(6)</td>
<td>-0.32718(9)</td>
<td>O(14)</td>
<td>0.0768(2)</td>
<td>-0.0519(2)</td>
<td>-0.1769(3)</td>
</tr>
<tr>
<td>Si(8)</td>
<td>0.30778(6)</td>
<td>-0.13016(6)</td>
<td>-0.18548(9)</td>
<td>O(15)</td>
<td>0.4161(2)</td>
<td>0.1276(2)</td>
<td>-0.3896(3)</td>
</tr>
<tr>
<td>Si(9)</td>
<td>0.27554(6)</td>
<td>-0.17279(6)</td>
<td>30.00109(9)</td>
<td>O(16)</td>
<td>0.4086(2)</td>
<td>-0.0017(2)</td>
<td>-0.4136(3)</td>
</tr>
<tr>
<td>Si(10)</td>
<td>0.12058(6)</td>
<td>-0.17310(6)</td>
<td>0.02979(9)</td>
<td>O(17)</td>
<td>0.4020(2)</td>
<td>-0.1314(2)</td>
<td>-0.4239(3)</td>
</tr>
<tr>
<td>Si(11)</td>
<td>0.07044(6)</td>
<td>-0.13037(6)</td>
<td>-0.18200(9)</td>
<td>O(18)</td>
<td>0.1886(2)</td>
<td>0.1298(2)</td>
<td>-0.3836(3)</td>
</tr>
<tr>
<td>Si(12)</td>
<td>0.18706(6)</td>
<td>-0.17327(6)</td>
<td>-0.31933(9)</td>
<td>O(19)</td>
<td>0.1940(2)</td>
<td>0.0007(2)</td>
<td>-0.4082(3)</td>
</tr>
<tr>
<td>O(1)</td>
<td>0.3726(2)</td>
<td>0.0534(3)</td>
<td>-0.2442(3)</td>
<td>O(20)</td>
<td>0.1951(2)</td>
<td>-0.1291(2)</td>
<td>-0.4190(3)</td>
</tr>
<tr>
<td>O(2)</td>
<td>0.3084(2)</td>
<td>0.0587(2)</td>
<td>-0.0789(2)</td>
<td>O(21)</td>
<td>-0.0037(1)</td>
<td>0.0502(2)</td>
<td>-0.2080(2)</td>
</tr>
<tr>
<td>O(3)</td>
<td>0.2007(2)</td>
<td>0.0592(3)</td>
<td>0.0289(4)</td>
<td>O(22)</td>
<td>-0.0040(2)</td>
<td>-0.1528(2)</td>
<td>-0.2078(3)</td>
</tr>
<tr>
<td>O(4)</td>
<td>0.0969(2)</td>
<td>0.0611(2)</td>
<td>-0.0856(3)</td>
<td>O(23)</td>
<td>0.4192(3)</td>
<td>-0.2500</td>
<td>-0.3540(4)</td>
</tr>
<tr>
<td>O(5)</td>
<td>0.1149(2)</td>
<td>0.0541(2)</td>
<td>-0.2763(2)</td>
<td>O(24)</td>
<td>0.1884(3)</td>
<td>-0.2500</td>
<td>-0.3538(4)</td>
</tr>
<tr>
<td>O(6)</td>
<td>0.2435(2)</td>
<td>0.0553(3)</td>
<td>-0.2460(3)</td>
<td>O(25)</td>
<td>0.2883(3)</td>
<td>-0.2500</td>
<td>0.0579(4)</td>
</tr>
<tr>
<td>O(7)</td>
<td>0.3742(2)</td>
<td>-0.1561(2)</td>
<td>-0.2372(3)</td>
<td>O(26)</td>
<td>0.1085(3)</td>
<td>-0.2500</td>
<td>0.0611(4)</td>
</tr>
</tbody>
</table>
Table 3.2  Fractional coordinates for one of the two orientations of the template atoms [1].

<table>
<thead>
<tr>
<th>Atom</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
<th>Atom</th>
<th>$x$</th>
<th>$y$</th>
<th>$z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>N(1)</td>
<td>0.47625</td>
<td>0.250</td>
<td>-0.1095(6)</td>
<td>N'(1)</td>
<td>0.47625</td>
<td>0.250</td>
<td>-0.1095(6)</td>
</tr>
<tr>
<td>C(1)</td>
<td>0.495(1)</td>
<td>0.233(1)</td>
<td>-0.221(2)</td>
<td>C'(1)</td>
<td>0.413(2)</td>
<td>0.229(2)</td>
<td>-0.166(4)</td>
</tr>
<tr>
<td>C(2)</td>
<td>0.568(2)</td>
<td>0.250</td>
<td>-0.241(2)</td>
<td>C'(2)</td>
<td>0.355(1)</td>
<td>0.272(1)</td>
<td>-0.150(2)</td>
</tr>
<tr>
<td>C(3)</td>
<td>0.578(2)</td>
<td>0.250</td>
<td>-0.362(2)</td>
<td>C'(3)</td>
<td>0.285(3)</td>
<td>0.250</td>
<td>-0.195(4)</td>
</tr>
<tr>
<td>C(4)</td>
<td>0.399(1)</td>
<td>0.274(1)</td>
<td>-0.100(2)</td>
<td>C'(4)</td>
<td>0.534(2)</td>
<td>0.279(2)</td>
<td>-0.185(3)</td>
</tr>
<tr>
<td>C(5)</td>
<td>0.355(1)</td>
<td>0.228(1)</td>
<td>-0.150(2)</td>
<td>C'(5)</td>
<td>0.546(2)</td>
<td>0.225(2)</td>
<td>-0.263(4)</td>
</tr>
<tr>
<td>C(6)</td>
<td>0.278(2)</td>
<td>0.250</td>
<td>-0.147(3)</td>
<td>C'(6)</td>
<td>0.601(3)</td>
<td>0.250</td>
<td>-0.332</td>
</tr>
<tr>
<td>C(7)</td>
<td>0.505(2)</td>
<td>0.319(2)</td>
<td>-0.089(2)</td>
<td>C'(7)</td>
<td>0.468(2)</td>
<td>0.315(2)</td>
<td>-0.060(3)</td>
</tr>
<tr>
<td>C(8)</td>
<td>0.473(2)</td>
<td>0.334(2)</td>
<td>0.019(3)</td>
<td>C'(8)</td>
<td>0.516(2)</td>
<td>0.335(2)</td>
<td>0.020(3)</td>
</tr>
<tr>
<td>C(9)</td>
<td>0.496(1)</td>
<td>0.412(1)</td>
<td>0.045(2)</td>
<td>C'(9)</td>
<td>0.496(1)</td>
<td>0.412(1)</td>
<td>0.045(2)</td>
</tr>
<tr>
<td>C(10)</td>
<td>0.508(1)</td>
<td>0.199(1)</td>
<td>-0.032(2)</td>
<td>C'(10)</td>
<td>0.508(1)</td>
<td>0.199(1)</td>
<td>-0.032(2)</td>
</tr>
<tr>
<td>C(11)</td>
<td>0.490(1)</td>
<td>0.124(1)</td>
<td>-0.045(2)</td>
<td>C'(11)</td>
<td>0.490(1)</td>
<td>0.124(1)</td>
<td>-0.045(2)</td>
</tr>
<tr>
<td>C(12)</td>
<td>0.529(1)</td>
<td>0.088(1)</td>
<td>0.033(1)</td>
<td>C'(12)</td>
<td>0.529(1)</td>
<td>0.088(1)</td>
<td>0.033(1)</td>
</tr>
</tbody>
</table>
3.2 General Utility Lattice Program (GULP)

The dielectric constant for ZSM-5 and ZSM-5+TPA systems was calculated using GULP [42]. GULP is designed to perform a variety of simulations for solid systems. In GULP, the internal energy is assumed to be an expansion in terms of increasingly higher order interactions between the total number of atoms, $N$:

$$U = \sum_{i=1}^{N} U_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} U_{ij} + \frac{1}{6} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} U_{ijk} + \ldots,$$

(3.1)

where, the first term represents the self energies, the second the pairwise energies, and so on. It is known however, that the contribution from higher order terms progressively diminishes for most systems. Consequently, GULP neglects terms beyond four body interactions. A certain degree of parametrization of the remaining terms is introduced in the force fields to compensate for the omitted terms. The following potentials were included for the calculation of dielectric constants:

1. Coulomb interaction - Coulomb interaction is a dominant term in systems with
ionic materials. It's given by Coulomb's law;

$$U_{ij}^{\text{Coulomb}} = \frac{q_i q_j}{4\pi \varepsilon_0 r_{ij}}$$  \hspace{1cm} (3.2)

where, $q_i, q_j$ are the partial charges on the atoms, $\varepsilon_0$ is the permittivity of the medium, and $r_{ij}$ is the distance between the two atoms. Since the term is given by a conditionally convergent series, it is a complicated term to evaluate for periodic systems. The method of Ewald [43] was employed to yield a convergent series with a well-defined limit.

2. Two-body short-range interactions - The interactions between atoms when they are bonded, or between ions when they are in the immediate coordination shells, are included under these terms. The attractive dispersion term is combined with the exponentially varying repulsive term as follows:

$$U_{ij}^{\text{Buckingham}} = A \exp \left( \frac{-r_{ij}}{\rho} \right) - \frac{C_6}{r_{ij}^6}.$$  \hspace{1cm} (3.3)
For covalently bonded atoms, a harmonic potential is used of the form

\[ U_{ij}^{\text{Harmonic}} = \frac{1}{2} k_2 (r - r_0)^2. \]  

(3.4)

Here, \( k_2 \) is the spring constant, and \( r_0 \) is the expected distance between the covalently-bonded atoms.

3. Three-body interactions - From a covalent perspective, the three-body potential represents the repulsion between bond pairs. It penalizes deviation from the expected angle for the coordination sequence, and it is given as

\[ U_{ijk} = \frac{1}{2} k_2 (\theta - \theta_0)^2 + \frac{1}{6} k_3 (\theta - \theta_0)^3 + \frac{1}{24} k_4 (\theta - \theta_0)^4. \]  

(3.5)

Here, \( \theta_0 \) represents the planar angle between the three atoms, and \( k_2, k_3, \) and \( k_4 \) represent the force constants.

4. Four-body interactions - This term describes energy associated with twisting the
torsional angles, and have the following form:

\[ U_{ijkl} = k_4[1 + m \cos(n\phi - \phi_0)] . \]  

(3.6)

5. Polarizability - The dipole moment of an atom cannot be assumed to be fixed. It changes both in magnitude and direction according to the polarizability of the species. Polarizability is treated in the simulations via the ion-pair shell model [44].

The model separates an ion into two parts, namely a core, which represents the nucleus and inner electrons of the ion and has the atom mass associated with it, and a shell that mimics the valence electrons. The core and shell are Coulombically screened from each other, but coupled through a harmonic spring of force constant \( K_{cs} \). The \textit{in vacuo} polarizability of the ion is given by

\[ \alpha = \frac{q_s^2}{K_{cs}} \]  

(3.7)

where \( q_s \) is the shell charge. The short range forces by convention act on the shell, whereas the Coulomb potential acts on both. The short range forces change the
spring constant between the shell and the core, thus changing the polarizability, which now depends on the environment.

3.3 Calculation of the dielectric constant

Depending on the frequency of the applied electromagnetic field, the actual value of the dielectric constant, \( k \), varies. It is common practice to quote two extreme values of \( k \), namely the static and high frequency dielectric constants. In the static limit both the electronic and nuclear degrees of freedom respond to the applied electric field, thus providing screening. In the high frequency limit, however, only the electrons respond to the perturbation.

The 3 × 3 tensor of static dielectric constant is calculated from \( D_{\alpha\beta} \), the Cartesian second derivative matrix of all particles, and the vector, \( q \), which contains the charges on all particles.

\[
\epsilon_0^{\alpha\beta} = \delta_{\alpha\beta} + \frac{4\pi}{V} (qD_{\alpha\beta}^{-1}q) \tag{3.8}
\]
The high frequency dielectric constant is calculated with the same formula; the only change being that $D_{\alpha\beta}$ now only includes the Cartesian components for all shells present in the model.

The dielectric constant at high frequency has a relatively direct correlation to the shell model as the second derivatives matrix, $D_{\alpha\beta}$, includes only the Cartesian components of the shells that are present within the model. Also, since the dielectric constant tensor depends on the inverse of $D_{\alpha\beta}$, it shares several characteristics of the Hessian matrix, and hence is a sensitive indicator of the goodness of the potential model.

3.3.1 Results

Three different systems were simulated using GULP.

1. ZSM-5: After the application of symmetry operators, the unit cell consisted of 288 atoms (96 silicon and 192 oxygen).

2. ZSM – 5 – TPA$_1$: This system had a total of 340 atoms that included 4 TPA
molecules. The subscript refers to one of the two orientations of the TPA molecule that was found in the Koningsveld paper [1].

3. ZSM – 5 – TPA₂: The second orientation of the TPA molecule was included in this simulation.

Simulations were carried out at constant pressure i.e. the unit cells were allowed to change shape and size. The force field parameters were taken from the force field due to Jackson and Catlow [45]. Cut and shift approach was used in order to make all the potentials calculable. Once the energy was defined, the structure was optimized using the Newton-Raphson procedure. Optimization was achieved in each of the three cases. Once optimization is achieved, a wide range of physical properties could be calculated based on the curvature of the energy surface about the minimum. The results for the dielectric constant are given in the following tables, 3.3, 3.4, and 3.5. The framework structure of ZSM – 5 – TPA₁ is shown in Fig. 3.1.
Table 3.3  ZSM-5: Dielectric tensor constants

<table>
<thead>
<tr>
<th>Static dielectric tensor</th>
<th>High frequency dielectric tensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>3.45630</td>
<td>0.00000</td>
</tr>
<tr>
<td>0.00000</td>
<td>3.40811</td>
</tr>
<tr>
<td>0.00000</td>
<td>0.00002</td>
</tr>
</tbody>
</table>

Table 3.4  ZSM – 5 – TPA₁: Dielectric tensor constants

<table>
<thead>
<tr>
<th>Static dielectric tensor</th>
<th>High frequency dielectric tensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>3.51045</td>
<td>0.00026</td>
</tr>
<tr>
<td>0.00026</td>
<td>3.43768</td>
</tr>
<tr>
<td>-0.00042</td>
<td>-0.00197</td>
</tr>
</tbody>
</table>
Figure 3.1  ZSM – 5 – TPA$_1$ framework structure, looking down the straight channel ($b$-axis).
Table 3.5  ZSM – 5 – TPA₂: Dielectric tensor constants

<table>
<thead>
<tr>
<th>Static dielectric tensor</th>
<th>High frequency dielectric tensor</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>y</td>
</tr>
<tr>
<td>3.70680</td>
<td>-0.00075</td>
</tr>
<tr>
<td>-0.00075</td>
<td>3.44273</td>
</tr>
<tr>
<td>-0.00228</td>
<td>-0.00001</td>
</tr>
</tbody>
</table>

Due to limitations in experimental measurements it is sometimes difficult to make the measurements of the dielectric constant in thin film zeolites [46]. However, it will be interesting to compare the calculated values with the experimental ones as they become available from the researchers at UCR. The important thing to note in these simulations was the effect the organic template molecule (TPA) has on the dielectric constant of ZSM-5. From the tables, the high-frequency dielectric tensors for ZSM-5 and ZSM-5+TPA systems are within 1% of each other. The low-frequency dielectric tensors for the two systems vary up to 7%. Also, the change of orientation of the template seems to have
a pronounced effect on low-frequency dielectric tensors; with values changing up to 5% between the two orientations.

3.4 Calculation of the elastic constants

The following discussion is taken from an article due to Sprik et al. [47].

Calculation of the elastic constants from MD simulations relies on evaluation of strain fluctuations in the $NpH$ ensemble. These fluctuations are used to determine the adiabatic compliances, which in turn yield the elastic constants [48]. During the MD simulation, a system of particles under external stress is simulated. The time dependence of three basic vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ that specify the edges of the MD cell is recorded [48, 49]. The vectors are saved in a $3 \times 3$ matrix $h$. Parrinello-Rahman Lagrangian gives the time evolution of the $3N + 9$ coordinates $\{\vec{r}, \dot{h}\}$ as

$$L = \frac{1}{2} \sum_i m_i \dot{s}_i^T \mathbf{G} \dot{s}_i - V + \frac{1}{2} W \text{Tr}(h^T \dot{h}) + V_d, \quad (3.9)$$

where $\vec{r} = h \vec{s}$, $\mathbf{G} = h^T h$, and $h^T$ is the transpose of $h$. $W$ is the mass associated with
the coordinates $h_{\lambda,\mu}$. $V$ represents the potential energy that determines the interaction between the particles. $V_{el}$ is the elastic energy of the system, and is a function of the applied stress $S$ and the strain $\varepsilon$. Under the conditions of zero external stress, $V_{el}$ is given as

$$V_{el} = \Omega_0 Tr(S \varepsilon),$$

(3.10)

where $\Omega_0 = ||h_0||$ represents the unstrained volume. A closed set of dynamical equations are derived from the Lagrangian (Eq. 1). The strain tensor $\varepsilon$ is expressed in terms of the $3N + 9$ coordinates as

$$\varepsilon(t) = \frac{1}{2}[h_0^{-1}G(t)h_0^{-1} - 1].$$

(3.11)

Here, $G(t)$ represents the instantaneous value of the metric tensor in the strained state, and $h_0$ is the average matrix defining the simulation cell in the reference state. The strain in Eq. 3 describes a homogeneous deformation that fluctuates in time. The dynamics generated by the Lagrangian with the given definition of the elastic energy
conserves the enthalpy $H$. Therefore, time averages over an MD trajectory correspond to the expectation values in the $HpN$ ensemble. The equilibrium fluctuations in this ensemble are related to the adiabatic compliances, $\Gamma_{ijkl}^s$, according to Parrinello and Rahman [48, 49].

$$\langle \Delta \epsilon_{ij} \Delta \epsilon_{kl} \rangle_{S,H,N} = \frac{k_BT}{\Omega_0} \Gamma_{ijkl}^s. \quad (3.12)$$

Eq. 4 enables the determination of the elastic compliances from the time averaged fluctuations of the strain (Eq. 3) in the reference state. Due to the permutation symmetry for the Cartesian indices, the $9 \times 9$ matrix $\Gamma_{ijkl}^s$ is singular. In order to invert the matrix, it is reduced to a non singular (and now non-symmetric) $6 \times 6$ matrix $\Gamma_{p\sigma}^s$. The Voigt convention [50] relates the two matrices as

$$\Gamma_{p\sigma}^s = \Gamma_{ijkl}^s, \quad k = l(\sigma \leq 3) \quad (3.13)$$

$$= 2\Gamma_{ijkl}^s, \quad k \neq l(\sigma > 3). \quad (3.14)$$

The inverse of $\Gamma_{p\sigma}^s$ is also a $6 \times 6$ non-symmetric matrix. The symmetric $9 \times 9$ matrix
of the elastic constants in the Cartesian representation is obtained from this matrix.

3.4.1 MPDyn

One of the approaches in calculation of the elastic constants of solids involves molecular dynamics (MD) under the $NtH$ ensemble [49, 51, 52]. Under this scheme, the simulation cell containing $N$ atoms changes its shape conforming to the external stress $t$ and enthalpy $H$. However, such MD simulations are time consuming [53, 47]. Even under the zero-stress condition, millions of sampling steps are required during the calculation of the elastic constants in order to get satisfactorily converged results. As an alternative, Shinoda and co-workers proposed an algorithm in the $NtT$ ensemble [54]. Non-Hamiltonian equations of motion were used to generate the $NtT$ ensemble. The equations were then solved using the time-reversible RESPA algorithm [53, 47]. Three types of thermostats could be used to control the system temperature, Nosé-Hoover [55], Nosé-Hoover chain [56], and massive Nosé-Hoover chain [56]. Of the three methods, the
method of massive Nosé-Hoover chain was found to be the most efficient [54], and hence used in the current simulations. The algorithm was incorporated in MPDyn, a suite of programs for molecular simulations [57]. Calculation of the elastic constants form the ZSM-5 and ZSM-5+TPA systems were carried out using the MPDyn program.

**Force field**

The potentials of the force field used in the present calculations for the zeolite structure were taken from the force field due to Burchart and co-workers [58]. This force field has been used in simulations to reproduce geometry, heats of formation, and the vibrational frequencies of several zeolites reasonably well [58]. The bonding and non-bonding interactions of the TPA ions, and the interaction of the TPA ions with the oxygen of the zeolite framework were treated using the potential parameters of Oie et al [59]. Partial charges on the TPA atoms were taken from [60]. The interaction of the framework Si atoms and the TPA ions was considered entirely Coulombic. There is no comprehensive
set of parameters used for the molecule/zeolite interactions. However, the parameters
due to Oie et al. have been used in literature to model interaction between ZSM-5 and
TPA ions [60, 61] with success. The force field consists of the following terms:

\[
E_{\text{total}} = K_b(b - b_0)^2 + K_\theta(\theta - \theta_0)^2 + K_\phi(1 + \cos(n\phi - \delta)) + A_{ij}\exp(B_{ij}r_{ij}) - \frac{C_{ij}}{r_{ij}^6} + \frac{q_iq_j}{4\pi\varepsilon r_{ij}}.
\] (3.15)

Here, the first term is a bond-stretching potential with \( K_b \) as the force constant and
\( b, \) and \( b_0 \) as the actual and expected bond lengths respectively. The second term is
the angle-bending term with \( K_\theta \) as the force constant, and \( \theta \) and \( \theta_0 \) as the actual and
expected angle values. The torsion angle term and the non-bonded interactions are in
form similar to those described under for the GULP simulation package.

Simulation details

The Ewald method was used to calculate the Coulombic interactions. The cut-off
distance of 15 Å was used. With the assumed \( Pnma \) symmetry (8 symmetry operators)
each unit cell of ZSM-5+TPA system has 4 molecules of TPA. Since the cut-off distance in a simulation needs to be greater than the half-box length during the simulation, a super cell of $2 \times 2 \times 3$ dimensions with respect to the unit cell of ZSM-5 was generated to obtain a simulation box of $40.04 \, \text{Å} \times 39.78 \, \text{Å} \times 40.149 \, \text{Å}$. Consequently, the ZSM-5 system had 3456 atoms whereas the ZSM-5+TPA system had 5424 atoms in a three-dimensional periodic boundary box. During the MD simulations, variable size time-step was used with the smallest being 2 fs, and fluctuation times of the barostat and thermostats were set at 2 ps and 0.5 ps, respectively. MD simulation for ZSM-5 was run for 2.75 ns. MD simulation for ZSM-5-TPA system was carried out for 1.5 ns. The first 0.5 million steps steps were treated as the equilibration period, while the remaining 1.5 million steps were used for calculating averages. The cell parameters were stored every 10 steps.

3.4.2 Results

The energy conservation plot for the MD simulation of ZSM-5 is shown in Fig. 3.2.
Energy conservation in MD
ZSM-5 (2.75 ns)

Figure 3.2 Energy conservation during MD simulation of ZSM-5.
The MD simulations gave rapid convergence of the cell length. Table 3.6 lists each component of the MD cell length averaged over the simulation period along with the statistical errors. Block averages, with each block consisting of $2.75 \times 10^3$ (ZSM-5), and $1.5 \times 10^3$ (ZSM-5-TPA) MD time steps were used to calculate the statistical errors. The dielectric constant was calculated for the two simulations and are shown in Fig. 3.3 and 3.4.

<table>
<thead>
<tr>
<th>System</th>
<th>$2a$ (Å)</th>
<th>$2b$ (Å)</th>
<th>$3c$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZSM-5</td>
<td>40.50435 ± 0.0098</td>
<td>40.13439 ± 0.0048</td>
<td>39.28555 ± 0.0076</td>
</tr>
<tr>
<td>ZSM-5-TPA</td>
<td>40.54238 ± 0.0460</td>
<td>40.48785 ± 0.0244</td>
<td>41.00933 ± 0.0426</td>
</tr>
</tbody>
</table>
Figure 3.3  Elastic modulus (ZSM-5). Values are block averages with each block consisting of $5 \times 10^4$ steps.
Figure 3.4 Elastic modulus (ZSM-5-TPA). Values are block averages with each block consisting of $2.5 \times 10^4$ steps.
3.4.3 Discussion and Conclusion

Experimental studies to determine mechanical properties of zeolite single crystals have been limited. One early study that investigated MFI single crystals with an in-house micromechanical tester reported a low value of elastic modulus (\(\sim 4\) GPa) [62]. During the test the entire crystal was compressed in the tester. This might have amplified the effects of crystal-defects on the observed elastic modulus values. More recently Yan et al. reported a value of \(\sim 54\) GPa using nano-indentation [46]. The in situ MFI film has been reported to have an elastic modulus between 30-40 GPa as calculated by nano-indentation [63]. The method of nano-indentation does not compress the entire crystal at once. Rather it takes measurements at different locations on the crystal surface, thus reducing the effect of crystal defects on the measured values. As seen from Fig. 3.2 the value from the simulation lies between the reported values for ZSM-5 single crystal. A value of \(\sim 50\) GPa has been reported for the elastic modulus of ZSM-5 single crystal by
Yan and co-workers [46]. The difference could be traced to the use of different force fields (BKS force field [64] for Yan et al.) in the two simulations. Comparison could be made with the experimental value between the ZSM-5-TPA system once they are available. So far attempts to prepare large single crystal without template by calcination have proved to be unsuccessful due to cracking related problems. From the simulation results, it is apparent that inclusion of the template increases the elastic modulus of the system by more than 30%, at the same time keeping the dielectric constant (at low frequencies) nearly unchanged.

These computed values of dielectric constant and elastic modulus for ZSM-5 and ZSM-5-TPA systems indicate that these pure silica zeolites are able to achieve low $k$ values while at the same time maintaining a high mechanical strength as indicated by their elastic modulus (> 6 GPa). This result is encouraging as it paves way for further research in the subject of PSZ films being used as a possible substrate for building integrated circuits.
Chapter 4

Comparison of Monte Carlo methods for simulation of cyclic peptides

4.1 Introduction

The free energy landscapes of biological molecules are complex, in that the conformations characteristic of room or body temperatures are separated by high energy barriers. Standard simulation methods such as Metropolis Monte Carlo (MC) and conventional molecular dynamics (MD) fail to sample these conformations from the correct Boltzmann distribution. The problem is particularly acute in the case of constrained peptides. High temperature [65] or potential scaled [66] molecular dynamics have been shown to cross these barriers, but these methods sample from a distribution that is not the one of interest. Frenkel and Smit [67] and de Pablo [68] developed a configurational bias Monte Carlo (CBMC) method that uses local information to propose new moves. This method has been successful in sampling the complex energy landscapes from the correct
Boltzmann distribution. It has been applied to the study of long chain molecules [69], such as long chain alkanes [70, 71], and to the study of hydrocarbon conformations in zeolite channels [72]. Based on a method for the alkane chains [73], Deem and co-workers developed a generalized concerted rotation scheme to study both linear and cyclic peptides [74]. Especially for constrained molecules, the ability to change local topology up to two to three amino acids per move meant that even without the knowledge of the location of the conformations viz a viz energy barriers, CBMC proved to be considerably more efficient than the Metropolis scheme. In biological molecules effective equilibration of bulky side-chains is yet another challenge. A rebridging scheme [75], inspired by a method for polymers [76, 77, 78] has helped make further gains in terms of efficiency of the simulations.

For more complex cyclic peptides, the CBMC method is combined with the method of parallel tempering [79]. A rigorous Monte Carlo method, parallel tempering was first proposed to study glassy systems, and has since been successfully used to study such
systems as spin glasses [80, 81], self-avoiding random walks [82], lattice QCD [83], linear peptides [84], and crystal structure determination [85].

Monte Carlo methods allow unphysical moves that help systems under study to escape local energy minima. For constrained geometries though these local moves may still fail to achieve equilibrium in a feasible amount of computational time, rendering the simulations unergodic for all practical purposes. Molecular dynamics on the other hand though deterministic, brings about change on a global scale throughout the system, imparting a breathing motion that could prove helpful for escaping the valleys in the energy landscape. Hybrid Monte Carlo method tries to combine the benefits of both MC and MD moves. The method was originally developed by Duane et al. in quantum chromodynamics [86]. Researchers have since used it to study condensed-matter systems [87], dense polymer chains [88, 89], and crystal structures [90].

Depending on the system under study and the goal of a particular simulation, one combination of these methods could prove more useful than another. In this chapter
results from four different combinations of simulation methods in the NVT ensemble on
four different cyclic peptides are compared. Parallel tempering method was used in all
four cases to speed up equilibration. These combinations are as follows:

1. Molecular dynamics (MD)

2. Molecular dynamics + configurational bias-regrowth-rebridging (MD+CBMC+RG+RB)

3. Hybrid Monte Carlo (HMC)

4. Hybrid Monte Carlo + configurational bias-regrowth-rebridging (HMC+CBMC+RG+RB)

One of the aims of the study was to ascertain the consistency of the results from all
the methods for a particular peptide. The methods could then be used to study the more
complex cyclic peptides with disulfide bridges in future. The next section describes these
methods.
4.2 Simulation Methods

An all atom force field, Amber-94 [91] was used to model the peptides. The force field was developed for simulation of nucleic acids and proteins. Water is treated in an implicit manner with a dielectric constant. The energies were matched against those obtained from the TINKER [92] suite of programs for tri-peptides of each amino acid to confirm the veracity of the implementation.

4.2.1 Configurational bias Monte Carlo method

The CBMC method is based on the assumption that at room or body temperatures it is reasonable to consider the intramolecular potential energy to be composed of only torsional and non-bonded terms. The bond lengths and bond angles could be fixed at their equilibrium values, and only the torsional degrees of freedom could be sampled. This approach reduces the total degrees of freedom available to the system from $3N$ to $N_{\text{torsion}}$, where $N$ is the total number of atoms in the system, and $N_{\text{torsion}}$ is the total
number of torsion angles. This coarse-grain restriction in mapping the phase space is easily relaxed during the MD and HMC moves.

With the above assumption, the CBMC move assumes a molecule to be comprised of a set of ‘rigid units.’ A rigid unit consists of a set of atoms that form a rigid body. The relative distance between any pair of atoms within a rigid unit is constant. Adjacent rigid units are connected by a single sigma bond. Rotation is allowed only around the sigma bond.

The following discussion describing the CBMC moves is taken from [74]. There are two types of atoms in a peptide, backbone atoms and side chain atoms. Consequently, there are two types of CBMC moves.

Type I moves change the positions of side chain atoms alone. It begins by identifying the side chain that it then regrows. The $M$ rigid units involved are removed and then added one by one. The unit closest to the backbone is the first to be regrown. Each addition of the rigid unit is accompanied by the following four actions.
1. \( k \) different values of torsion angle \( \phi_{ij} \) that connects rigid unit \( i \) to the previous rigid unit are generated with probability

\[
P_i^{\text{int}}(\phi_{ij}) \propto \exp \left[ -\beta u_i^{\text{int}}(\phi_{ij}) \right]. \tag{4.1}
\]

Here, \( \beta = 1/k_B T \) is the inverse temperature, and \( u_i^{\text{int}}(\phi_{ij}) \) is the part of internal energy. It couples unit \( i \) to the rest of the molecule, excluding units \( i+1 \) to \( M \).

Partitioning of the energy into internal and external parts is arbitrary, and in this case \( u_i^{\text{int}} \) is assumed to be zero.

2. Of the generated values, one is picked with probability

\[
P_i^{\text{ext}}(\phi_{ij}) = \exp\left[ -\beta u_i^{\text{ext}}(\phi_{ij}) \right] / w^{\text{ext}}(i), \tag{4.2}
\]

where

\[
w^{\text{ext}}(i) = \sum_{j=1}^{k} \exp \left[ -\beta u_i^{\text{ext}}(\phi_{ij}) \right]. \tag{4.3}
\]

The term \( u_i^{\text{ext}}(\phi_{ij}) \) is the part of external energy that couples unit \( i \) with the rest of the molecule, excluding units \( i+1 \) to \( M \).
3. The first two steps are repeated until all the $M$ units are regrown.

4. The Rosenbluth weight is calculated as

$$W^{(n)} = \prod_{i=1}^{M} w^{\text{ext}}(i). \quad (4.4)$$

It is used to account for the bias introduced in the move and the attempted move is accepted with a probability

$$\text{acc}(o \rightarrow n) = \min[1, W^{(n)}/W^{(o)}]. \quad (4.5)$$

The term $W^{(o)}$ is the Rosenbluth weight for the reverse move. It is calculated as per steps 2 to 4, but one of the $k$ orientations is the original geometry for each unit.

Type II move changes the positions of backbone atoms, rigidly rotating the attached side chains. It is applied only to a linear peptide case, and is very similar to a type I move.

For any Monte Carlo scheme to properly sample the Boltzmann probability distribution, detailed balance needs to be satisfied. That the CBMC moves described above...
satisfy the detailed balance is proved in [73, 93]. These moves have been proved to be especially effective while sampling cyclic peptides [94].

4.2.2 Analytical Rebridging

The analytical rebridging scheme [75] causes a local conformational change within the molecule, and leaves the rest of the molecule fixed. It is used to equilibrate backbone atoms of a cyclic peptide. Consider a segment of peptide backbone shown in figure 4.1. The angles $\phi_0$ and $\phi_7$ are rotated, causing the rigid units between 0 and 6 to change.
The two rotations break the connectivity of the molecule. The task is to find analytical solutions to the geometrical problem of reconnection of the backbone units. It involves the reduction of twenty linear equations to an $8 \times 8$ determinant equation of one torsional angle. The determinant equation is equivalent to a polynomial of degree sixteen, limiting the maximum number of new geometrical solutions to sixteen. The determinant equation is further reformulated as an eigen value problem and solved using the QR algorithm [95].

In order to incorporate the method in MC simulations, a Jacobian factor necessary to satisfy detailed balance is included. The Jacobian accounts for the correction to the nonuniform distribution generated by rebridging moves. The rebridging move makes a fairly dramatic change in the backbone positions. This makes the move effective in equilibrating complex peptides.
4.2.3 Parallel Tempering

For cyclic peptides, biased rebridging Monte Carlo is not optimally efficient. To overcome the remaining barriers to effective sampling, parallel tempering move is used. In parallel tempering, an ensemble with \( n \) identical systems is considered. Each system is equilibrated at a distinct temperature \( T_i, i = 1, \ldots, n \). In addition to the biased rebridging moves, swapping moves are proposed that exchange the configurations between two systems \( i \) and \( j = i + 1, 1 \leq i < n \). The system with the lowest temperature is the one of interest. The higher temperature systems are added solely to allow the lowest temperature system to escape from local energy minima without explicit knowledge of the barriers. Since there are no energetic interactions between the replicas, for the entire system, the partition function could be written as

\[
Q = \prod_{i=1}^{M} \frac{q_i}{N!} \int d\mathbf{r}_i^N \exp \left[ -\beta_i U(\mathbf{r}_i^N) \right],
\]  

(4.6)
where \( q_i = \prod_{j=1}^{N} (2\pi m_j k_B T_i)^{3/2} \) is the partition function of system \( i \), \( m_j \) is the mass of atom \( j \), \( r_i^N \) specifies the positions of the \( N \) particles in the system \( i \), and \( U \) is the potential energy of the system \( i \). The swapping move between ensembles \( i \) and \( j \) is accepted with the probability

\[
\text{acc}(i \rightarrow j) = \min\{1, \exp[(\beta_i - \beta_j)(U(r_i^N) - U(r_j^N))])\}.
\] (4.7)

Swaps are generally attempted between systems adjacent in the temperature ladder. This method significantly reduces the equilibration time. In order to achieve efficient sampling, the maximum temperature is such that no significant free energy barriers are observed. Also, to improve the acceptance probability of the swapping moves, the energy histograms of adjacent systems are allowed to overlap.

4.2.4 Hybrid Monte Carlo (HMC)

The ensemble of Monte Carlo techniques described so far is still inefficient in sampling the phase space of highly constrained peptides such as those with disulphide bridges. In
order to speed up the proceedings, a HMC move is introduced.

Monte Carlo simulations allow local moves for searching the conformation space. Even with clever moves such as CBMC, updating more than a few particles leads to prohibitively low acceptance ratios. Although the stochastic nature of the method ensures eventual sampling of the entire phase space of a system, for complex proteins it often fails to do so within feasible computer time. The result is poor statistics with large relaxation times and high correlations.

Molecular dynamics simulations propose global moves. The scheme, however, is prone to errors and instabilities due to a finite step size. Also, typical MD simulations are run for $10^3$ to $10^4$ ps. This time scale is many orders of magnitude smaller than the $10^{-3}$ to $10^2$ seconds range over which many interesting phenomena in protein folding are known to take place [96].

Hybrid MC method combines the advantages of both MC and MD methods. It allows for global moves, but at the same time, it is an exact algorithm that does not suffer from
A single HMC move of the whole system consists of the following steps.

1. Initial momenta are generated from the correct equilibrium distribution $P(p_i) \propto \exp(-\beta p_i^2 / 2m_i)$. This is achieved by generating a set of Gaussian random numbers with unit variance and rescaling them appropriately.

2. Employing a time reversible and area preserving algorithm, the system is integrated through phase space for a fixed time $t$ (we carry out the integration for time greater than the first auto-correlation time in the potential energies). The new coordinates are accepted with the following Metropolis probability.

$$P_{\text{acc}} = \min(1, e^{-\beta \Delta H})$$

(4.8)

where $\Delta H = H(q', p') - H(q, p)$. $H$ denotes the Hamiltonian of the system; $q, q'$ denote initial and final coordinates, and $p, p'$ denote initial and final momenta.

Since the momenta are influenced by the interaction forces, the coordinates tend to
evolve in the direction of the decreasing potential energy. However, as the momenta are drawn at random (from a Maxwell-Boltzmann distribution), 'uphill' moves in energy are possible. After every HMC move, the momenta are refreshed irrespective of whether the move is accepted or not. This is essential in order to satisfy the condition of detail balance. The HMC move helps achieve a purely canonical ensemble without compromising the inherent stability and exactness of MC simulations. The step size used during the MD integration has no bearing on the observables. The discretization error produced is accounted for in the acceptance criteria and only affects the acceptance rate.

For the MD part of the move, all $3N$ degrees of freedom are considered. The Newtonian equations of motion are integrated using the velocity verlet algorithm [97].

4.2.5 Simulation details

As mentioned in the introduction, four different combinations of the simulation methods were tried on a test set of four cyclic peptides cyclo(dWdDPdVL), cyclo(dPAAdPAA),
cyclo(dPFASFF), and a protein loop region corresponding to residues 1491-1498 (sequence: TGRGDSPA) of human fibronectin repeat 10 (pdb-entry: 1FNF). In the amino acid sequence, d indicates a residue in the D chirality. Each peptide was simulated for a million steps in all the four methods. During the CBMC+RG+RB method, backbone units and side-chain units were selected with equal probability. The MD was carried out in the NVT ensemble. The Anderson [98] thermostat was implemented to keep the temperature constant. A time step of 1 fs was chosen for both MD as well as HMC moves. Each HMC move was carried out for a modest 100 steps. The temperatures for parallel tempering were chosen such that sufficient overlap was obtained between energy histograms of adjacent systems. The simulation methods were evaluated on their ability to equilibrate different properties of the system, the speed of equilibration, and their ability to generate uncorrelated structures. A comparison was also made between simulations of polyalanine with and without the inclusion of the GBSA model using HMC method.
4.2.6 Results and Discussion

Energy histograms during a parallel tempering run of peptide dPFASFF with MD+RG+RB method are shown in Figure 4.2. For swapping between adjacent replicas to occur with sufficient probability, their energy histograms need to overlap. Table 4.1 lists the acceptance probabilities for swapping moves in this run. These values are representative of all the 16 different runs.

![Energy histograms in parallel tempering](image)

**Figure 4.2** Energy histograms
Table 4.1  Acceptance probability for swapping moves in a parallel tempering simulation.

<table>
<thead>
<tr>
<th>Swap</th>
<th>$P_{\text{acc}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 K ↔ 400 K</td>
<td>0.07700</td>
</tr>
<tr>
<td>400 K ↔ 540 K</td>
<td>0.05155</td>
</tr>
<tr>
<td>540 K ↔ 720 K</td>
<td>0.06253</td>
</tr>
<tr>
<td>720 K ↔ 900 K</td>
<td>0.047144</td>
</tr>
<tr>
<td>900 K ↔ 1170 K</td>
<td>0.09527</td>
</tr>
<tr>
<td>1170 K ↔ 1500 K</td>
<td>0.060121</td>
</tr>
</tbody>
</table>
Figure 4.3 shows the temperature distribution for the lowest temperature (300 K) replica of dPAAdPAA during MD simulation. The Anderson thermostat maintains the temperature distribution around 300 K. The figure demonstrates that the MD method is indeed sampling conformations from the NVT ensemble.

![Temperature distribution during MD simulation.](image)

Figure 4.3 Temperature distribution during MD simulation.

The first test for the performance of the simulation methods is that one should get identical results for the equilibrium properties from each of the methods for a particular peptide. Equilibration of potential energy in all cases is shown in Figure 4.5. For each
peptide, energy histograms from all four different methods are the same.

The energy of the system is an easier property to equilibrate than the radius of gyration, \( R_g \). It is defined as the root mean squared distance of each atom in a peptide from its center of gravity, and it describes the overall spread of the molecule. Figure 4.6 shows equilibration of \( R_g \) in all four peptides with all four methods. Histograms in \( R_g \) for dWdDpVDVL and TGRGDSpA are the same from all four methods. However, those for dPAAAdPAA and dPFASFF differ from each other. For dPFASFF, the histograms from HMC, and HMC+RG+RB methods are the same, whereas, those from MD and MD+RG+RB are different. The reason for this discrepancy could be traced to the acceptance ratios between adjacent replicas during parallel tempering. The low acceptance values, as seen from Table 4.1 indicate a suboptimal choice of temperatures for the two peptides that translates into a lack of equilibration of certain properties.

Same is true for histograms in torsion angle \( \phi \) (\( C' - C_{\alpha} - N - C' \)) of proline residue as shown in 4.7 for these peptides. Histograms for dWdDpVDVL and TGRGDSpA are the
same for all four methods, whereas those for dPAAAdPAA and dPFASFF are different.

In the case of peptide dPAAAdPAA, it is clear that simulations without the RG and RB moves were unable to equilibrate the torsion angle in question. Addition of the CBMC moves helped improve the equilibration.

Figure 4.4 shows the movement of the lowest temperature replica of two peptides, dWdDPdVL and dPFASFF, through the temperature ladder during the course of MD+RG+RB simulations. One can see that in case of dWdDPdVL the replica was able to jump to the higher temperatures much more frequently as compared to the replica for dPFASFF. This explains the relative lack of equilibration of $R_g$ and the proline dihedral angle, $\phi$, in dPFASFF.

At high temperatures though, the system is able to overcome the barriers more easily, and the properties are well equilibrated. Histograms in $R_g$ for dPFSAFF at 1500 K (Fig. 4.8) from all four methods are the same. This is also true for the histograms in proline torsion angle $\phi$ for dPFSAFF as seen from Fig. 4.9. Comparison between histograms in
Figure 4.4 Movement of the lowest temperature replica through the temperature ladder for dWdDPdVL and dPFASFF.
proline dihedral angle $\phi$ (Fig. 4.10) of dPASFF at 300 K and 1500K shows that at 1500 K, the distribution of torsion angles is much flatter, and the molecule is free to cross the barrier at $-100^\circ$. Thus, parallel tempering allows the molecule to escape from local energy minima where it might otherwise have been trapped.

The relative speeds of different methods are shown in Table 4.2.

**Table 4.2** Comparison of simulation results of dWdDPdVL in terms of relative CPU time.

<table>
<thead>
<tr>
<th>Simulation method</th>
<th>Number of steps</th>
<th>Relative CPU time</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMC</td>
<td>$1 \times 10^6$</td>
<td>1.0</td>
</tr>
<tr>
<td>HMC+RG+RB</td>
<td>$1 \times 10^6$</td>
<td>1.5</td>
</tr>
<tr>
<td>MD</td>
<td>$1 \times 10^6$</td>
<td>1.0</td>
</tr>
<tr>
<td>MD+RG+RB</td>
<td>$1 \times 10^6$</td>
<td>1.5</td>
</tr>
</tbody>
</table>

The auto-correlation of an observable in a system shows the similarity between the observable as a function of time separation. For an observable $O$, the autocorrelation
Figure 4.5  Energy histograms at 300 K for four peptides with four simulation methods.
Figure 4.6  Histograms in $R_g$ at 300 K for four peptides with four simulation methods.
Figure 4.7  Histograms in proline dihedral angle, $C'$ - $C_{\alpha}$ - $N$ - $C'$, for four peptides with four simulation methods.
Figure 4.8 \( R_g \) histograms at 1500 K for dPFASFF.
Figure 4.9  Histograms in torsion angle at 1500 K for dPFASFF.
Figure 4.10  Histograms in proline torsion angle in dPFASFF at two different temperatures with HMC+RG+RB method
function is given as

$$C_O(t) = \langle O_{t_0+t} O_{t_0} \rangle - \langle O \rangle^2. \quad (4.9)$$

Here, $O_{t_0}$ and $O_{t_0+t}$ are measurements that are separated by $t$ steps. Figure 4.11 shows the auto-correlations in the $R_g$ values for the four peptides at 300 K. In all four cases, either HMC+RG+RB or MD+RG+RB methods provide the quickest drop in correlations as compared to pure HMC and pure MD methods. Addition of CBMC+RG+RB moves helps sampling of configurations from different energy basins. A pure HMC move almost always seems to be the worst choice in terms of reducing the correlations. Even at high temperatures, when the auto-correlation function quickly decays to zero for other simulation methods, those from HMC method show longer auto-correlation times (Fig 4.12). This, however, in part could due to the lower value of time step (1 fs) during HMC simulation. This value of the time step yields higher acceptance ratios ($>70\%$). This is too high an acceptance rate that makes the overall simulation less efficient by not being
able to sample independent, representative configurations. Using higher values (> 1 fs) of time step during HMC simulation could help improve efficiency of the method. In such cases, the increased error in energy conservation due to discretization of the equations of motion would be accounted for in the acceptance criteria for the step.

A clustering algorithm was implemented to look at the configurations that were saved after every $10^4$ steps. For each run, 100 configurations were thus saved. Table 4.3 shows the number of clusters and configurations in each cluster for the four peptides from each simulation. The probability of sampling a particular structure is directly related to the number of configurations in that cluster. Figure 4.13 shows representative structures from the most probable cluster for dPAAdPA with different simulation methods.
Figure 4.11  Auto-correlation functions in $R_g$ at 300 K for four peptides with four simulation methods.
Table 4.3  Peptide configurations clustered according to RMSD (0.75 Å)

<table>
<thead>
<tr>
<th>Peptide</th>
<th>configurations in cluster 1</th>
<th>configurations in cluster 2</th>
<th>configurations in cluster 3</th>
<th>non-cluster configurations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of configurations</td>
<td>Number of configurations</td>
<td>Number of configurations</td>
<td>Number of configurations</td>
</tr>
<tr>
<td>dPAAdPAA</td>
<td>HMC</td>
<td>HMC+RG+RB</td>
<td>MD</td>
<td>MD+RG+RB</td>
</tr>
<tr>
<td></td>
<td>62</td>
<td>58</td>
<td>70</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>34</td>
<td>40</td>
<td>27</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>dPFASFF</td>
<td>configurations in cluster 1</td>
<td>59</td>
<td>56</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>configurations in cluster 2</td>
<td>38</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>configurations in cluster 3</td>
<td>0</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>non-cluster configurations</td>
<td>3</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>dWdDPdVL</td>
<td>configurations in cluster 1</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>non-cluster configurations</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>TGRGDSPA</td>
<td>configurations in cluster 1</td>
<td>98</td>
<td>99</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>non-cluster configurations</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
**Figure 4.12** Auto-correlations in $R_g$ at 1500 K for dPFSAFF

Implicit solvent model

In computer simulations of biological molecules, solvent environment plays an important role. Structure, dynamics, and energetics are all influenced by the solvent. The most accurate way to account for solvent effects is to use the explicit solvent models. Such models use large number of discrete solvent molecules to simulate the liquid environment [99, 100, 101]. These models, however, do suffer from convergence problems with the required computational times being 100-1000 times more than those involving gas phase
Figure 4.13  Representative structures from most probable clusters of dPAAdPAA from four simulation methods. Clockwise from top left, the methods that generated the structures are, HMC, HMC+RG+RB, MD and MD+RG+RB
molecules [102]. In order to lower the computational cost of including solvent effects, an implicit description of solvation is generally used to replace the explicitly modeled solvent molecules. The implicit description approximates the high dielectric solvent by a continuum electrostatics model. The generalized-Born (GB) model [103] is one such model. This model has been implemented and was tested on alanine dipeptide. The HMC method was used to simulate the system without parallel tempering. The simulation was carried out for a $10^5$ steps. The results from the simulation were compared with HMC simulation where the GB energies were considered only during the acceptance criteria. Both methods satisfy detailed balance. The results comparing energy histograms, $R_g$ histograms, and one $C' - C_\alpha - N - C'$ torsional angle are represented in Figure 4.14. The two sets of simulations provided equivalent results for the potential energy and backbone torsion angle. Here again, $R_g$ proved to be the more difficult property to equilibrate. Also, the simulation using the GB forces was nearly 12 times slower than one that used GB energies only during the acceptance criteria.
Figure 4.14  Histograms in energy, $R_g$, and the $C' - C_\alpha - N - C'$ ($\phi$) torsion angle for alanine dipeptide for two sets of HMC simulations. One model uses forces due to GB energy during the simulation, whereas, the other only uses GB energies in the acceptance criteria.
4.2.7 Conclusion

Four different combinations of simulation methods have been used to equilibrate four different cyclic peptides. The HMC and and the MD methods augmented with RG+RB moves show considerable improvement in sampling independent configurations. Parallel tempering method helps the peptides escape the local energy minima through swapping moves as is clearly seen from the flatness of the histograms at higher temperatures. Of the four peptides, dWdDPdVL was equilibrated the easiest as evidenced from the various histograms and the fact that all the configurations could be collected in a single cluster for all four simulation methods. The peptide dPFASFF proved to be more difficult judging from the longer correlation times and histograms in $R_g$. It was also the only peptide that was found to have three representative structures that were separated from each other by at least 0.75 Å.

There are several variables that could be adjusted depending on the system involved,
for example, more weight to side chain moves than backbone moves in RG+RB for peptides with bulky side chains, the range through which the driver angles are rotated during the analytical rebridging move to cause bigger changes per move etc. Also, choice of temperatures for the parallel tempering run should be optimized for each system. Combination of these simulation techniques along with implementation of GB models would be useful for studying more complex systems such as the cysteine-rich peptides with criss-crossing disulphide-sulfide bridges.
Chapter 5
Probing the role of Cripto in the regulation of E-cadherin with docking methods

5.1 Introduction

Human terato-carcinoma growth factor (TDGF1), or Cripto, is a small membrane-linked molecule that acts as a soluble growth factor when shed into the extracellular environment. Cripto is a membrane-bound co-receptor in tumor growth factor (TGF)-signaling, and it has previously been isolated from a pluripotent cancer cell line, NTERA2, and is now considered an oncogene. Cripto is a member of the epidermal growth factor (EGF)-Cripto/FRL-1/Cryptic (CFC) gene family. The human and murine Cripto genes display 94% homology in the EGF-CFC domain, whereas the sequences of the N- and C-terminal regions are unique [104]. More recently, Cripto expression has been described to be part of a 'pluripotency signature' of several progenitor cells [105, 106, 107].

A research group at M. D. Anderson cancer center at the university of Texas recently
examined TDGF1 expression in human bladder urothelial carcinoma cell lines and found that it was expressed in the cytosol and nucleus of highly invasive bladder cancer cells and that it was associated with no expression of E-cadherin (E-cad) [108]. It was demonstrated that the nuclear form of human Cripto (truncated form) binds histone deacetylase (HDAC)-1 and -2 and Snail with which it cooperates in inducing the loss of E-cad expression and increases the promoter-targeted HDAC activity. Using human tissue samples it was also showed that the nuclear form of Cripto possibly plays an important role in bladder cancer progression. This was the first study to report that Cripto regulates of E-cad expression when in the nucleus.

This chapter describes part of aforementioned study that explored the possibility of specific DNA binding 'signature' to Cripto within the E-cad promoter, and the role played by Snail in DNA-Cripto interactions from a computational point of view. To that end, docking studies were carried out between homologs of Cripto, Snail, and the DNA fragment. The results were compared with experimental findings.
5.1.1 Docking

Molecular docking tries to predict the structure of receptor-ligand complexes. The receptor is normally a protein molecule, and the ligand might be either a small molecule or another protein. Over the years, molecular docking has evolved from a tool to reproduce the crystal structures of known molecular complexes [109] into a tool that is used extensively in rational drug design, for tasks such as optimizing lead compounds through database screening, and pre-screening real or virtual compounds to design combinatorial libraries [110].

The docking algorithm contains two important steps. The first step involves a thorough search of the conformational degrees of freedom of the molecules involved, and the second step assigns an appropriate docking score to rank the conformations thus searched. In general, the molecules are treated as rigid bodies in docking. The search algorithm explores different orientations of the ligand molecule in the experimentally
known receptor active site using translational and rotational degrees of freedom. Ideally, the search algorithm should be able to locate the global energy minimum for the system. The scoring function tries to assess both steric as well as chemical complementarity of the different configurations. It should be realistic enough to assign the lowest score to the experimentally determined complex structure.

Over the years, many research groups have tried to come up with docking algorithms that balance the contradictory requirements of speed and accuracy. In early days, docking program represented both receptor and ligand with explicit atoms [111]. Such representation resulted in numerous complex configurations, all of which needed to be evaluated. The applications of molecular surface calculations [112] in docking algorithm reduced the number of possible complexes significantly, thus increasing the search speed. The next major update came from the application of grids to store receptor properties. With the assumption of rigid receptor, the grid needed to be calculated only once. DOCKER [113] and GRID [114] were two of the early programs that incorporated the grid method in
their algorithms. Flexibility in ligand molecule was the next step in terms of improving
the search for the native conformations. Since the knowledge of ligand orientation that
interacts most favourably with the receptor is often missing, this was an important devel­
opment. These algorithms used one or more of three types of search methods, systematic
[115], stochastic [116], and deterministic [117]. The systematic search methods relied on
combinatorial search of grid values of each degree of freedom. The stochastic methods
relied on random changes in degrees of freedom. These included Monte Carlo methods
[116], and evolutionary algorithms [118]. The deterministic search methods on the other
hand used initial state values to generate the next state to go downhill in energy space
in search of the global minimum. Allowing for receptor flexibility was the next logical
progression. Researchers have showed that methods allowing for receptor flexibility could
improve the success rate of locating the native structures by as much as 25% in certain
cases [119]. Programs that treat both receptor and ligand in a flexible manner are useful
when the structure of either molecule undergoes conformational changes due to induced
Protein-protein docking is a more difficult task compared to protein-small molecule docking. Not only is the number of generated complexes huge, but also prediction of protein interacting site is a difficult task. The scoring function needs to be rather soft because even in near-native configurations, some atoms are likely to clash. Often the scoring takes place in two stages. A geometric filter eliminates those conformers that are unlikely to be complex structures in the first stage. In the second stage, geometrically feasible conformers are scored with a more elaborate energy function. This study uses one such algorithm designed for protein-protein docking called Hex.

5.1.2 Hex

Hex [120] is an interactive protein docking and molecular superposition program, written by Dave Ritchie. The program approaches a docking problem through representation of the steric shape, electrostatic potential, and charge density of each protein as expa-
sions of spherical polar Fourier basis functions. Unlike three-dimensional fast Fourier
transform docking approaches that accelerate translational correlations, Hex favours ro-
tational searches. The original shape-sampling algorithm and the improvements are
described in [121, 120]. This method was able to locate conformations from top 20 solu-
tions in four of the seven target protein-protein complexes in the ‘critical assessment of
predicted interactions’ (CAPRI) blind docking trial [121].

Docking large molecules

Since the radial basis functions fall off rapidly beyond 30 Å from the origin, the
spherical polar approach by itself is incapable of handling large protein molecules. Hex is
able to generate multiple local coordinate systems with which the initial ligand docking
orientations about a large receptor are generated. The docking process is a four step
procedure described as follows:

1. The smaller molecule is oriented to face the receptor.
2. A low resolution spherical harmonic surface is calculated for the receptor. By projecting the surface onto an icosahedral tessellation of the sphere, the surface is discretized. A normal vector is calculated at each triangular facet of the surface, and a sphere of 15 Å radius is centered on each outward pointing normal, tangential to the surface. The step smoothes the receptor surface with spheres.

3. In order to get an even distribution of spheres over the receptor surface, the surface spheres are culled through iterative identification and removal of a sphere that has the most volume overlap with its neighbours. The iteration continues until no sphere with a volume overlap of more than 5 Å³ is left.

4. Each remaining sphere (normal vector) is used to define a local intermolecular axis for docking.
Soft molecular mechanics refinement

In order to reduce the number of false positives, Hex implements a scoring function with Lennard-Jones (12-6) and hydrogen bond (12-10) terms that are constructed from the OPLS force field [122]. The parameters are modified so as to retain the long range nature and minima of the original set, but to reduce the short-range repulsive behaviour. The correlation search is followed by energy minimization of the first five hundred orientations. The final docking score is thus a sum of the soft scoring function and shape-based correlation energies.

\[ E_{\text{total}} = E_{\text{shape}} + E_{\text{OPLS}} \]  

(5.1)

Solution clustering

The procedure to smother the surface of macromolecules tends to over-sample the orientational search space. Hence, all low energy orientations are clustered so as to identify distinct orientations. After ordering the solutions in order of increasing energy,
the clustering algorithm uses the lowest energy orientation as the seed for the first cluster.

The rest of the orientations are searched to find entries where the $C_\alpha$ atoms of the ligand fall within 2 Å rms of the corresponding atoms in the seed member for the current cluster.

The procedure is repeated until all solutions are allocated to a cluster. This process is useful to avoid a good solution being pushed down the order in the final list.
5.1.3 Methods and results

The E-cad promoter was divided into 4 regions namely P1, P2, P3, and P4. Since crystal structures of both Cripto and Snail were unavailable, a homology search was carried out using 3D-PSSM [123], a fold recognition server. For Cripto, a 101 residue chain L of coagulation factor VIIA [124] (amino acid homology 21%), and for Snail, an 83 residue chain C of zinc finger protein [125] (amino acid homology 45%) were selected as suitable homologs. Three dimensional structures of the selected compounds were retrieved from the protein data bank. Heteroatoms and water molecules were removed from the respective pdb files. The DNA fragments P1 and P4 were modeled and energy minimized using standard B-DNA geometry with sibyl-7.1 [126], a computational tool kit for molecular design and analysis. The following three systems were studied using HEX.

1. P1-1DVA
2. P4-1DVA

3. P4-1MEY-1DVA

The sequences for the DNA fragments, the two proteins, and their respective homologs are provided below.
The DNA sequence (double stranded DNA oligos) used for EMSA

P1: GGTAC GGGG GCGGT

P4: GCTCC GGGGC TCACC TGGCT GCAGC CAC

Sequence of protein Cripto:

MDCRK MARFS YSVIW IMAIS KAFEL GLVAG LGHQE FARPS RGYLA FRDDS

IWPQE EPAI RPRSS QRVPP MGIQH SKELN RTCCL NGGTC MLGSF CACPP

SFYGR NCEHD VRKEN CGSVP HDTWL PKKCS LCKCW HGQLR CFPQA FLPGC

DGLVM DEHLV ASRTP ELPPS ARTTT FMLVG ICLSI QSYY

Sequence of chosen homologous protein exosite inhibitor E-76 (PDB code 1DVA; chain L similarity 21%):

ISYSD GDQCA SSPCQ NGGSC KDQLQ SYICF CLPAF EGRNC ETHK DDQLIC

VNENG GCEQY CSDHT GTKRS CRCHE GYSLL ADGVS CTPTV EYPC GKIPI
Sequence of the protein Snail:

MPRSF LVRKP SDPNR KPNYS ELQDS NPEFT FQQPY DQAHL LAAIP PPEIL
NPTAS LPMLI WDSVL APQAQ PIAWA SLRLQ ESPRV AELTS LSDED SGKGS
QPPSP PSPAP SSFSS TSVSS LEAEA YAAFP GLGQV PKQLA QLSEA KDLQA
RKAFN CKYCN KEYLS LGALK MHIRS HTLPC VCGTC GKAFS RPWLL QGHVR
THTGE KPFSC PHCSR AFADR SNLRA HLQTH SDVKK YQCQA CARTF SRMSL
LHKHQ ESGCS GCPR

Sequence of the homologous zinc finger protein (PDB code 1MEY; chain C similarity 45%):

MET GLU LYS PRO TYR LYS CYS PRO GLU CYS GLY LYS SER PHE SER GLN
SER SER ASN LEU GLN LYS HIS GLN ARG THR HIS THR GLY GLU LYS PRO
TYR LYS CYS PRO GLU CYS GLY LYS SER PHE SER GLN SER SER ASP LEU
GLN LYS HIS GLN ARG THR HIS THR GLY GLU LYS PRO TYR LYS CYS PRO
GLU CYS GLY LYS SER PHE SER ARG SER ASP HIS LEU SER ARG HIS GLN

ARG THR HIS GLN ASN LYS LYS
As per convention, the larger molecule (DNA fragment) is referred to as receptor and the smaller one (protein homologs) are referred to as ligands. Owing to the large size of the molecules involved, the macromolecular docking option was used in each case. A sphere smoothing algorithm generated multiple local coordinate systems with which initial ligand (protein) docking orientations were defined about the receptor (DNA). For each target, the algorithm performed an initial low-order shape complementarity scan. The best 10000 orientations were then refined by using high-order shape correlations. The top 500 orientations from each high-order correlation were energy minimized and clustered to get a final list of distinct docking orientations. The top 5 orientations were inspected in each case to identify the DNA residues interacting (< 2.5 Å) with the protein ligand. In order to investigate the possibility of existence of a complex containing Cripto and Snail, docking was carried out between P4 and 1MEY. The orientation where 1MEY was found interacting with the E-box on P4 (2nd lowest energy) was selected as the new receptor for docking 1DVA.
Table 5.1 lists the energies (docking score) and DNA bases that interact with the peptide ligands for the top 5 conformations for 1MEY-P4-1DVA, and 1DVA-P1 systems. DNA bases and residues on ligands within a distance of 2Å of each other were considered to be interacting.

Figure 5.1 shows the lowest energy conformation for the 1MEY-P4-1DVA complex.

5.1.4 Discussion and conclusions

The results from the docking study provide interesting insights in the context of the experimental observations recorded in [108]. Gel-shift assays carried out with Cripto and E-cad promoter have showed that Cripto interacts with P1, but not with the P2 or P3 regions. The P1 region contains a CGGG sequence specific for binding of transcriptional activators of the E-cad promoter [127, 128]. Two of the top five orientations for the 1DVA-P1 system showed CGGG to be a potential interaction area between these two molecules.
Table 5.1  Energies and ligand interacting DNA base pairs in the top 5 orientations for the systems of P4-1MEY-1DVA and P1-1DVA.

<table>
<thead>
<tr>
<th>Orientation rank</th>
<th>Docking score (kJ/mol)</th>
<th>Interacting DNA (P4) base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-578.1</td>
<td>7-10</td>
</tr>
<tr>
<td>2</td>
<td>-559.7</td>
<td>19-20</td>
</tr>
<tr>
<td>3</td>
<td>-558.6</td>
<td>7-9</td>
</tr>
<tr>
<td>4</td>
<td>-552.7</td>
<td>19-20</td>
</tr>
<tr>
<td>5</td>
<td>-537.9</td>
<td>2-4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Orientation rank</th>
<th>Docking score (kJ/mol)</th>
<th>Interacting DNA (P1) base pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-509.8</td>
<td>8-10</td>
</tr>
<tr>
<td>2</td>
<td>-477.2</td>
<td>7-9</td>
</tr>
<tr>
<td>3</td>
<td>-468</td>
<td>8-10</td>
</tr>
<tr>
<td>4</td>
<td>-463</td>
<td>10-13</td>
</tr>
<tr>
<td>5</td>
<td>-461.2</td>
<td>3-7</td>
</tr>
</tbody>
</table>
Figure 5.1  Lowest energy orientation for 1MEY-P4-1DVA system.
The addition of an anti-Cripto antibody induced a 'sub-shift' of the molecular complex that binds the E-cad promoter sequence in the P4 region, suggesting the loss of one or more interacting proteins from the complex. Similarly addition of the anti-Snail antibody induced a similar shift of the same complex, suggesting that Cripto and Snail are parts of the same complex that binds the E-cad promoter sequence P4. The binding of Snail protein to the P4 region was showed to be dependent on the integrity of the so called E-box (CACCTG). Mutations of the type C to T in the E-box abolished the binding between Snail and P4. This was the reason behind selecting the low energy 1MEY-P4 complex as the receptor for the Cripto homologue, 1DVA. In four of the top five orientations for the 1MEY-P4-1DVA system, the 1DVA molecule was found interacting with the GGCT base pairs flanking both ends of the E-box. This suggested that presence of Snail could increase Cripto affinity and that these two proteins could be part of a complex. The predicted interaction between Snail and Cripto has since been confirmed through experiments [108].
To conclude, preliminary docking study performed on homologs of Cripto and Snail proteins and E-cad promoter corroborate experimental findings suggesting that Cripto plays an important role in regulating E-cad expression through some promoter priming mechanism. It is important to remember that the homology between Cripto and 1DVA is on the lower side (< 30%). Also, the docking algorithm for larger molecules are yet in the early stages of development in terms of their ability to exhaustively search the large number of possible orientations. Lack of proper description of the solvent effects is yet another drawback of the method. These factors urge caution in overstating importance of the results thus obtained.
Chapter 6

Computer-assisted vaccine design

6.1 Introduction

Circulating influenza virus uses the ability to change its surface proteins, along with its high transmission rate, to flummox the adaptive immune response of the host. The random accumulation of mutations in the hemagglutinin (HA) and neuraminidase (NA) epitopes, the regions on surface of the viral proteins that are recognized by host antibodies, poses a formidable challenge to the design of an effective annual flu vaccine. Under current practice, the World Health Organization (WHO), and the respective government health agencies rely on historical experience and phylogenetic analysis of HA and NA protein sequences from the circulating human strains to decide the components in the annual influenza vaccine [129]. Every year, the concerned authorities make a projection about the circulating influenza strains for the coming flu season. The flu vaccine currently contains three strains that are as similar as possible to those strains that are predicted
to be the most prominent. At present, the three strains in the vaccine are one H3N2 A, one H1N1 A, and one influenza B component. Historically, the vaccine efficacy has seldom reached the 100% mark. Over the years, it has hovered between 30-60% against influenza-like illnesses. In fact, due to a phenomenon known as the ‘original antigenic sin’ [130, 131, 132], the vaccine efficacy has even been negative at times. If original antigenic sin is operative, the host antibodies that are produced in response to a viral strain tend to suppress creation of new and different antibodies in response to a different viral strain. Whether such a phenomenon takes place or whether the vaccine is effective depends to a great extent on how similar the vaccine component strains are to the circulating viral strains. The methods employed to calculate such an antigenic distance draw heavily from the ferret antisera hemagglutinin inhibition assays. It has been assumed that the antigenic distance thus obtained from ferrets correlates well with the efficacy of the influenza vaccine in humans. However, to our knowledge there is limited evidence in the literature of such correlations.
We introduce a new and effective way of measuring the antigenic distance between different strains. We define a quantity, $p_{\text{epitope}}$, that measures the difference between the dominant epitope of the H3N2 vaccine component and the dominant circulating viral strain. A dominant epitope is one that elicits the most significant response from the adaptive immune system for a particular strain in a particular year [133, 134, 135, 136]. We show that for data spanning last 35 years, $p_{\text{epitope}}$ correlates better with the efficacy studies in humans of the influenza vaccine than do the current measures of antigenic distance, even those derived from ferrets.
6.2 Methods

We have developed a statistical mechanics based theory to model the response of an immune system, free of immunoscenescence, to disease and vaccination. We compare this theory to experimental studies of vaccine efficacy for 18-64 year old subjects over the past 35 years, when the H3N2 subtype of influenza A was the dominant strain. H3N2 is the most common strain and has caused significant morbidity and mortality. As is the norm, we focused on the five epitopes of the hemagglutinin protein. We used the generalized NK model to calculate the affinity constants quantifying the immune response following exposure to an antigen after vaccination. The model takes into account three types of interactions within an antibody. These include interactions within subdomains, interactions between different subdomains, and interaction between an antibody and an antigen [137]. The binding constant is given as $K = \exp(a - b<U>)$, where $U$ is the energy function of an antibody [132], $a = -18.56$, and $b = 1.67$. The values of the
constants are determined after comparing the dynamics of the model with experimental results [132]. In order to capture the antigenic distance between the vaccine strain and the circulating strain, we use $p_{\text{epitope}}$ as an order parameter in our model. It represents the fraction of amino acids that differ between the dominant epitope regions of the two strains.

$$p_{\text{epitope}} = \frac{\text{number of amino acid differences in the dominant epitope}}{\text{total number of amino acids in the dominant epitope}}$$

In order to build the model we needed the identity and sequence of the dominant epitope in both, the candidate vaccine and the circulating strain. This definition of the five epitopes in the H3N2 hemagglutinin protein was taken from reference [138]. Our model assumes the following correlation between the vaccine efficacy, $E$, and the binding constants: $E = \alpha \ln[K_{\text{secondary}}(p_{\text{epitope}})/K_{\text{primary}}]$, where the constant $\alpha$ is selected such that a perfect match between the vaccine and the circulating strain corresponds to
a vaccine efficacy of 45%. This model is in agreement with the historical data [139].

$K_{\text{primary}}$ is the binding constant corresponding to the primary immune response, and $K_{\text{secondary}}$ is the binding constant corresponding to the secondary immune response that follows vaccination. Other than the constant, $\alpha$, no parameter was fitted to data, making the model predictive. For example, the point where vaccine efficacy becomes zero is independent of the value of $\alpha$. 
6.3 Results

Table 6.1 compares vaccine efficacy values from experimental studies [140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153] and predictions from our theory as a function of $p_{\text{epitope}}$. In literature, vaccine efficacy is defined as

\[
\text{Efficiency} = \frac{u - v}{u},
\]

where $u$ and $v$ are the influenza-like illness rates in unvaccinated and vaccinated individuals, respectively. The epidemiological estimates of $u$ and $v$ contain some noise; however, these are the best estimates there are of influenza vaccine efficacy in humans.

Our model demonstrates the effectiveness of using $p_{\text{epitope}}$ as a measure of antigenic drift between the circulating strain and the vaccine strain. Crystallographic data and immunoassays have shown that only the epitope regions of the viral proteins are significantly involved in the recognition of the antigen via the antibodies [154]. Our definition
of $p_{\text{epitope}}$ follows from this observation. When $p_{\text{epitope}}$ value in the dominant epitope is greater than 0.19 according to experimental records or is greater than 0.22 according to our theory, the efficacy of the influenza vaccine drops to the negative territory [Table 6.1 and Fig. 6.1] indicating that while designing the vaccine, this regime needs to be avoided.

As an example, during the 1997/98 influenza season in the Northern hemisphere, when the Sydney/5/97 strain was widespread, the value of $p_{\text{epitope}}$ was 0.238 resulting in an efficacy of -17% [148]. Only one data point falls outside our theoretical predictions, that for the 1989/90 epidemic [155]. During that year there were likely multiple strains circulating, including the strains of influenza B [143, 156].

The WHO uses as a first approximation an alternative definition of the antigenic drift. It considers the sequence difference of the entire hemagglutinin protein of the circulating and viral strains.
Figure 6.1 Vaccine efficacy for influenza-like illness as observed in epidemiological studies as a function $p_{\text{epitope}}$. A linear least squares fit to the data is also shown (long dashed, $R^2=0.81$).
The correlation with the experimentally observed efficacy is not as positive when using the WHO definition for the antigenic distance. This could be seen from Table 6.1 and Fig. 6.2, and it follows from the fact that many domains in the protein are either inaccessible to the human antibodies or simply unrecognizable by them. This results in the vaccine efficacy and the drift in the entire protein sequence being not as correlated with the approximate measure of distance in Eq. 3.

The gold standard measure of antigenic drift is that derived from ferret antisera [140, 157]. Comparison of the vaccine efficacy to this measure of antigenic drift also shows less than ideal success. From Table 6.1 and Fig. 6.3, the hemagglutinin assays in ferret antisera are unable to capture a significant amount of human efficacy information. For example, an antigenic distance of zero according to ferret antisera experiments does not
Figure 6.2  Vaccine efficacy for influenza-like illness as observed in epidemiological studies as a function of $p_{\text{sequence}}$. A linear least squares fit to the data is also shown. (long dashed, $R^2=0.59$). Figure shows the same epidemiological data as in Fig. 6.1. Only the definition of the x-axis is different.
mean that the two strains are identical. As an example, for the 1996/97 season, the ferret antisera derived antigenic distance between the vaccine strain of A/Nanchang/933/95 and the circulating strain of A/Wuhan/359/95 was zero, whereas $p_{eptope}$ value was 0.095. The corresponding vaccine efficacy in the Northern and Southern hemispheres was 28% [150] and 11% [149], respectively. When compared to the average corresponding to a perfect match between the vaccine strain and the circulating strain, which is 45%, these values are clearly much lower.
Figure 6.3  Vaccine efficacy for influenza-like illness as observed in epidemiological studies as a function of antigenic distance $d_1$ and $d_2$ derived from ferret antisera experiments. A linear least squares fit to the data is also shown. $d_1$ (long dashed, $R^2=0.57$) and $d_2$ (short dashed, $R^2=0.43$). Results were averaged when multiple hemagglutinin inhibition (HI) studies had been performed for a given year. These HI binding arrays measure the ability of ferret antisera to block the agglutination of red blood cells by influenza viruses. The epidemiological data is same as in Fig. 6.1. Only the definition of the x-axis is different.
6.4 Discussion

The design of influenza vaccine is always in a race against time. Currently, under the supervision of the WHO and the national health agencies, in the Northern hemisphere, the components of the annual influenza vaccine are determined between February and April. The mass production of the vaccine is then carried out by growing the virus in hen's eggs. After the regulatory tests in mid-July, the vaccine is distributed in September. Data are collected to determine the vaccine efficacy from October. By January a good measure of the effectiveness of the season's vaccine is obtained. Choice of the vaccine strain is contingent upon various biological and manufacturing constraints. For example, among the egg-cultured strains, the availability of high growth strains is an additional criterion. The organizations in charge have to weigh in all these constraints before they decide on the best match for the anticipated circulating strain for the following flu season.

The $p_{\text{epitope}}$ measure of antigenic distance can influence and improve the vaccine design
process in two ways. First, it can help identify the strains to be included in the annual flu vaccine. For every season, given a list of strains and their probabilities of outbreak, the value of $p_{\text{epitope}}$ can help define the weighted distance of a particular vaccine strain from the circulating strains. This procedure will enable us to identify the vaccine strain closest to the strains causing a possible outbreak. Alternately, $p_{\text{epitope}}$ can also help identify the 'like' strains. Various manufacturing constraints in the vaccine production process mean it is often not feasible to grow large quantities of the exact strain that is desired for the annual vaccine. Under such scenario, it is required to choose several strains that are similar to the chosen one. The value of $p_{\text{epitope}}$ can help quantify the 'likenesses' of a given strain to the desired strain. We can extend this approach to other strains of influenza as well. Using information about the epitope regions of other HxNy strains of influenza A or influenza B, we can calculate the value of $p_{\text{epitope}}$ to quantify the antigenic distance for these strains of influenza.

It is clear that the immune system response is neither linear nor monotonic in the
antigenic distance. As a result, the original antigenic sin leading to a negative vaccine efficacy exists only in an intermediate regime of the antigenic distance. If the vaccine falls in this regime, however, it makes the vaccinated individual more susceptible to influenza-like illnesses compared to an unvaccinated individual. Over the past 33 years, this phenomenon seems to have taken place 26% of the time for the relevant H3N2 strains from the epidemiological data (as indicated by 5 of the total 19 data points that are negative in Table 6.1 and Fig. 6.2). The negative points are not necessarily a result of experimental errors; rather they may have their roots in the phenomenon where the immune system is unable to distinguish between the new and the old viral strains. It is desired that the regime corresponding to the original antigenic sin is avoided, not only for the obvious immunological consequences, but also for the negative impact it creates against the acceptance of public health policy. Our theory rightfully captures the underlying physics of immune response, and corroborates the experimental findings. Our theory and $p_{epitope}$ measure of antigenic distance can be used to ensure that vaccines
chosen would not fall in the original antigenic sin region against expected strains.

It appears that \( p_{\text{epitope}} \) is a better measure of the antigenic distance between viral strains than the sequence analysis or the ferret antisera. The population at large may benefit if the health authorities incorporate \( p_{\text{epitope}} \) in the prediction and design of the influenza vaccine in addition to (or even instead of) the current practices. We now present a few examples to show how our theory can assist shaping public health policies. Here we examine the 2004/2005 flu season in the Northern hemisphere. With the aid of \( p_{\text{epitope}} \) we can stipulate a priori the extent of protection a particular vaccine strain provides against the circulating strain for a particular season. The comparison among various candidates would provide us selection criteria for their inclusion in the annual flu vaccine. During the 2003/2004 flu epidemic, A/Fujian/411/2002 strain was predominant. In order to counter that strain, the Advisory Council of the FDA recommended using A/Wyoming/2003 as the H3N2 component of the 2004/2005 vaccine [158]. According to the prevalent measures, this strain and A/Fujian/411/2002 were found to be 'anti-
genically equivalent.' Our calculations, however, yielded a $p_{\text{epitope}}$ value of 0.095 for the pair in question, suggesting that the two strains were not antigenically equivalent, and the predicted efficacy is 20% (Fig. 6.1). The vaccine efficacy that year was roughly 20% [153, 159]. We note that another of the WHO approved candidates for that year was A/Kumamoto/102/02 (ISDN38180) [160], and we found $p_{\text{epitope}}$ value to be zero versus A/Fujian/411/2002. According to our theory, a vaccine constituting this strain would have provided a better protection against the Fujian as compared to the Wyoming strain. The vaccine efficacy that year was a modest 28% [159].

As another example, we consider the 2008/2009 flu season in the Northern hemisphere. The reference vaccine strains for the 2007/2008 season were A/Wisconsin/67/2005 or A/Hiroshima/52/2005 [161]. For the 2008/2009 season, although some isolates were antigenically similar to the Wisconsin strain, the majority of the isolates were distinguishable from it, and more similar to A/Brisbane/10/2007 strain [162]. According to our analysis, the $p_{\text{epitope}}$ value between Wisconsin and Brisbane strains is 0.1579. This
suggests that the people who were vaccinated with the Wisconsin strain will have a low level protection against the Brisbane strain, and will need vaccination with the WHO recommended Brisbane-like virus [162]. Similarly, the $p_{\text{epitope}}$ value for the pair of Brisbane and Hiroshima strain turns out to be 0.095. People vaccinated with Hiroshima strain a year earlier are expected to fare better than those vaccinated with Wisconsin strain; however, the protection level is still low enough to recommend vaccination with a Brisbane-like strain. Thus, $p_{\text{epitope}}$ may help authorities decide when changes to the vaccine are necessary.

6.4.1 Suggestions to improve predictability of vaccine efficacy

At the heart of our approach lies the identification of the dominant hemagglutinin epitope recognized in humans. The identity of the dominant epitope for humans is currently not measured. We define and postulate the dominant epitope for humans for a certain strain and for a certain season as the epitope that undergoes the largest
fractional change in the sequence of amino acids in comparison to the vaccine strain. A measurement of the epitope that is dominant in humans for various circulating strains and vaccines should help improve the predictive power of our approach even more, as we can then use the measured dominant epitope to calculate $p_{\text{epitope}}$ instead of using our postulated dominant epitope. It is important to continue measurement and sequence analysis of the prominent circulating strains in the flu season. Another important piece of the puzzle is the epidemiological study that relates vaccine efficacy to the antigenic drift. We believe that continuous efforts in these areas would enable the health authorities predict the severity of the yearly flu season, design better vaccines for the same, and help manage the health resources for this period.

In more general terms, the results presented here have implications for the fight against diseases that stem from rapidly mutating viruses and that are managed through antibody responses. The $p_{\text{epitope}}$ measure of antigenic distance can predict efficacy for vaccine strains against multiple circulating strains, and it could then be used towards
redesigning of the vaccine in terms of frequency and composition. One such disease that has been threatening to turn into a pandemic of late is the avian influenza or bird flu disease. In the next section, we present a multiple-strain transmission model for avian influenza that would enable one to manage the risk from such an outbreak. We use $p_{epitope}$ to evaluate the effectiveness of a multiple-component vaccine to counter such threat.
6.5 Vaccination strategies and risk management of flu pandemic

Since its first appearance (Hong Kong, 1997), H5N1 avian influenza is known to have spread to various parts of the world, including South-East Asian countries, parts of central and Middle-East Asia, Africa, and Europe. This rapid spread has prompted the World Health Organization (WHO) to suggest that “we are closer to a pandemic than at any time since 1968 [163, 164].” Bird flu has been observed in pigs [165], and occurrences of person-to-person transmission have likely been detected [166, 167, 168]. Efficacy of the vaccine produced against the original Hong Kong strain of H5N1 has been poor against some of the new strains. In addition, the high mutation rates observed in the avian influenza have raised the prospect of simultaneous introduction of multiple strains [164, 169, 170, 171], putting a question mark against the efficacy of a single-component vaccine. Although in the event of appearance of multiple strains, eventual emergence of a single dominant strain is likely; the lack of a priori knowledge of which
strain will be dominant makes it worthwhile to look into the alternative of a multiple-component vaccine. Here, we use an epidemiological model to study the efficacy and cross-protection of a multiple-component bird flu vaccine. We further extend the concept of $p_{\text{epitope}}$ through a stochastic model to manage the risk of a flu pandemic. The strategy is inspired by similar risk management studies in finance, and $p_{\text{epitope}}$ plays the role of the risk variable.

6.5.1 Multiple-strain introduction transmission model

As mentioned above, there is evidence, both theoretical and otherwise that deems introduction of multiple-strains of avian influenza as a distinct possibility. There have been no mathematical models, however, that explore such a scenario and suggest competent vaccine strategies. Our model uses the concepts of hierarchical scale free network, and virus transmission, and viral evolution to take into account possible, and worst case scenarios and to provide answers to the questions posed by the same.
6.5.2 Hierarchical scale free network

We divide the total human population \((6.7 \times 10^9)\) into \((6.7 \times 10^6)\) groups. Each group consists of \(10^3\) people that exhibit similar health status and social behavior. These groups are distributed over \(N_{\text{city}}\) cities. The distribution of cities with \(i\) groups \(N_{\text{city}}(i)\) goes as \(N_{\text{city}}(i) \propto i^{-2.1}\) [172, 173]. The largest and the smallest city areas in our simulation have \(6.2 \times 10^4\) and 400 groups respectively [173]. The total number of cities fluctuates around 4000 [174].

The cities are connected through a network that extends globally and is defined by the network spawned by the airlines. The distribution of cities with \(i\) contacts is given by \(N_{\text{city-contact}}(i) \propto i^{-2.0}\). The network connecting groups within a city is defined by the ground transportation network. In each city, the distribution of groups with \(i\) contacts is given by \(N_{\text{group-contact}}(i) \propto i^{-2.8}\) [175]. Population size differences between cities mean that movement from one city to another leads to a final averaged distribution over cities.
that is different for different contact numbers.

6.5.3 Virus transmission and evolution

In our simulation, the viruses are introduced in groups that are chosen randomly. The virus has a latency period of 2 days, which is followed by the infectious period. During the infectious period, the virus can either be killed by the host immune system, or it can be transmitted between different groups globally. During both these periods, the virus can mutate as well as be killed. The term $R_{\text{mutation}} = 1.6 \times 10^{-5}$/amino acid/day [176] gives the mutation rate, whereas, the virus killing rate $c_{\text{Kill}}$ is a function of the host immune history as well as the identity of the infecting virus. The transmission rate takes into account the seasonal effect [177], and is different for transmission within a city ($F_{\text{fac1}} \times \tau_0 \times (1+0.25 \sin(2 \pi T/360))$, $F_{\text{fac1}}=1.0$) and transmission between cities ($F_{\text{fa}} \times \tau_0 \times (1+0.25)$, $F_{\text{fa}}=7.0$). $F_{\text{fa}}$ is the number of different groups that people traveling on a flight come from, and it is set to 100 for our model. The term $\tau_0$ is the
intrinsic transmission ability of the virus, and equals 0.07 for the virus introduced on day 1. The mutants, once created, can transmit according to probability cProb. For a baseline parameter, cProb equals 0.26.

6.5.4 Vaccination Strategies

As an example, for a very rapid public health policy response, we consider that after 40 days, individuals are vaccinated. We consider single-component as well as multiple-component. In single-component case, the most dominant strain at day 10 is taken as the vaccine strain, whereas, for multiple-component case, the top 10 strains that are most dominant on day 10 are incorporated into the vaccine. Day 10 is taken as the data collection day by public health authorities. We calculated the efficacy and the cumulative attack rates for both cases. Efficacy is defined by equation 2 above. Thus, efficacy represents the percentile reduction in disease occurrences in people who were vaccinated compared to those who were not [178, 179].
The model's ability to predict an influenza pandemic was tested against the existing epidemiological data (Fig. 6.4a). The model successfully predicted the average trend of H3N2 isolates data in FluNet database of WHO from 1995 to 2006 (Fig. 6.4b). We also successfully predicted the fixation rate for dominant and non-dominant epitopes of influenza (Fig. 6.5) versus those measured [177].

6.5.5 WHO FluNet data analysis

We acquired the WHO FluNet data for the period of 1995-2006 [180]. We separately averaged the isolates data for countries from the Northern and the Southern hemisphere over the 1997-2006 seasons [Fig. 6.4a]. For this duration, the annual flu epidemic seems to have several peaks. The highest peak occurs in summer. It is interesting to note the role played by China, with an incidence peak in the summer, which is very different from other Northern hemisphere countries. China probably acts as a reservoir for the flu virus and also as a resource for transmission of the virus from pigs to humans in the
Figure 6.4  Average isolates cases of human H3N2 influenza. (a) Cases reported during last 10 years for various countries around the world. (b) Comparison between the WHO FluNet database and as predicted by our immunological and epidemiological model.
Figure 6.5  Influenza fixation rate for dominant and non-dominant epitopes. Dominant: fixation rate for dominant epitope from simulation, Non-dominant: fixation rate for non-dominant epitopes from simulation; Dominant: fixation rate and error bar for dominant epitopes from sequence analysis [177], Non-dominant: fixation rate and error bar for non-dominant epitope from sequence analysis [177].
summer. Thus, the summer peak in China is likely related to the role played by this virus reservoir. We aligned the week with the highest peak, week\textsubscript{peak}, to week\textsubscript{1} so that we could match the burst times for different years. This results in all data from a week preceding the peak week being shifted to (week\textsubscript{2} - week\textsubscript{peak}), and all data from a week following the peak week being shifted to (week\textsubscript{1} - week\textsubscript{peak}). After rearranging the data in such manner, we compared the results for average isolates from our simulation with those from the WHO FluNet data. A sensitivity analysis was carried out for all the parameters involved in Fig. 6.4a.

6.5.6 Multiple-component vaccine

We evaluated the viability of using a multiple-component vaccine against simultaneous introduction of multiple viral strains. In our model, we considered the possible scenario where two initial viruses, differing by $p_{\text{epitope}}$, cause a pandemic. A $p_{\text{epitope}}$ value of zero indicates that viruses are the same, whereas a $p_{\text{epitope}}$ value of 1 indicates that the viruses
are completely different. The model then predicted the efficacy of both, single and multiple-component vaccines as a function of the $p_{\text{epitope}}$ (Fig. 6.6a). The cumulative attack rates are also predicted (Fig. 6.6b). A multiple-component vaccine was found to exhibit greater efficacy for the case where multiple-strains are introduced. There always exists a lag between viral outbreaks and administration of pertinent vaccine. We considered the effect such lag has on single and multiple-component vaccines. It was found that the lag affects single-component vaccine more than it affects multiple-component vaccine (Fig. 6.7a and 6.7b).

### 6.5.7 Population at risk

In order to manage the risk of possible introduction of multiple strains of avian influenza, we combined together our stochastic model and a bioinformatics study. Such risk management methods are in common use in financial institutions [181]. In economics, Value at Risk (VaR) is used as a metric in risk management studies. The term stands
Figure 6.6  The effect of a vaccine for the initial introduction of two-strains. This indicates the multiple-component vaccine excels single-component vaccine for different vaccination day and vaccination populations. Sc: single-component vaccine. Mc: multiple-component vaccine. V: vaccination population over total population. Vac Day: The day when vaccine is administered. Vaccination populations dependency at Vac Day = 40: (a) Efficacy as a function of $p_{\text{epitope}}$ for single and multiple component vaccines with different vaccination populations. (b) Cumulative attack rate as a function of $p_{\text{epitope}}$. Vaccination day dependence at $V=40\%$. 
Figure 6.7  The effect of a vaccine for the initial introduction of two-strains. This indicates the multiple-component vaccine excels single-component vaccine for different vaccination day and vaccination populations. Sc: single-component vaccine. Mc: multiple-component vaccine. V: vaccination population over total population. Vac Day: The day when vaccine is administered. Vaccination populations dependency at Vac Day = 40: (a) Efficacy as a function of $p_{\text{epitope}}$. (b) Cumulative attack rate as a function of $p_{\text{epitope}}$. 
for the maximum loss over a given period of time with a certain confidence level. We use an analogous term, Population at Risk (PaR) in our study. It defines the maximum percentage of the world’s population that is at risk from viral infection in the event of an influenza pandemic over the period of one year with a confidence level. As an example, a PaR value of 0.1 at the 95% confidence level means that if there are 100 pandemics, the maximum percentage of the global population that would be infected by the circulating viral strains in 95 of these stands at 0.1. The method could be applied for the risk management of infectious diseases in general. In the case of avian influenza, we analyzed all H5N1 strains in the National Center for Biotechnology Information (NCBI) database [182], and calculated the $p_{\text{epitope}}$ for all pairs of the viral strains. The average value turned out to be $p_{\text{epitope}} = 0.118$. We generated two initial viruses using the $p_{\text{epitope}}$ value, and carried out simulations for 2000 yearly pandemics for 3 cases: with no vaccination, vaccination with single-component vaccine, and vaccination with multiple-component vaccine. The PaR was calculated for confidence interval between 0.9 and 0.99. As seen
from Fig. 6.8, the fewer the vaccine components, the more the number of people that get infected in these pandemics. This clearly shows that a multiple-component vaccine not only improves the efficacy of a vaccine, but is also important to manage worse case scenarios.

**Figure 6.8** Population at Risk (PaR) over a year for three cases: without vaccination, single-component vaccination, and multiple-component vaccination.
6.5.8 Summary

We have developed a pandemic model for avian influenza. Virus evolution, inter­
personal transmission, vaccination, and immune history are taken into account. The
model is able to reproduce the observed data for H3N2 isolates. Vaccination against
multiple-components is not a foregone conclusion in case of simultaneous introduction
of multiple strains. The parameter \( p_{\text{epitope}} \), helps decide if use of multiple-component
vaccine is likely to be beneficial. The \( p_{\text{epitope}} \) measure of antigenic distance also helps
decide what strains to incorporate under such scheme. The model shows that a vac­
cine with multiple- components outperforms a single-component vaccine when there are
greater than one circulating strains that are antigenically different from each other. The
model also shows that compared to the traditional single-component vaccine, a multiple-
component vaccine is affected to a lesser degree by the lag between viral outbreak and
vaccine administration. This result has important implications if there is a sudden at-
tack of avian influenza pandemic, and the process of designing, producing, delivering, and administrating a vaccine is delayed significantly. The model proves that the spread of the virus is contingent upon among other things, fraction of the population vaccinated, and the time of administration of the vaccine following the outbreak. Lastly, the model provides a new parameter PaR that foretells how severe a potential epidemic could be.

Combined with multiple-component vaccine, PaR will be valuable to alter actual vaccination strategies, beyond a broad notion of addressing multiple isolates. One can calculate daily virus difference through WHO FluNet, and get an average $p_{epitope}$ value that could be used in our model. The model can yield PaR values for different areas of the World. This will alert the policy makers and general public of the current pandemic risk level, and also help pharmaceutical companies for preparing optimal vaccine components. An improved predictive ability regarding the extent of impending epidemic and appropriate vaccine design would enhance our ability to better manage the precious resources.
Table 6.1 Summary of results. The definitions for $p_{\text{epitope}}$, $p_{\text{sequence}}$ are described previously (fractional change in the immunodominant epitope and fractional change in the whole sequence respectively). The terms $d_1$ and $d_2$ are the two available measures of distance based on ferret anti-serum. Figures below are to provide the analysis of the data presented in the table. The bracketed numbers are the references used. When more than one antisera assay has been performed, the calculated distances are averaged.

<table>
<thead>
<tr>
<th>Year</th>
<th>Vaccine Strain</th>
<th>Circulating Strain</th>
<th>Vaccine Efficacy</th>
<th>Dominant Epitope</th>
<th>$p_{\text{epitope}}$</th>
<th>$p_{\text{sequence}}$</th>
<th>$d_1$</th>
<th>$d_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1971-72</td>
<td>Aichi/2/68 (V01085)</td>
<td>Hongkong/1/68 (AP201874)</td>
<td>7 % [183]</td>
<td>A</td>
<td>0.158</td>
<td>0.033</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1972-73</td>
<td>Aichi/2/68 (V01085)</td>
<td>England/42/72 (AP201875)</td>
<td>15 % [183]</td>
<td>B</td>
<td>0.190</td>
<td>0.055</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1973-74</td>
<td>England/42/72 (ISDNENG722)</td>
<td>PortChalmers/1/73 (AF092062)</td>
<td>11 % [183]</td>
<td>B</td>
<td>0.143</td>
<td>0.018</td>
<td>5 [140]</td>
<td>4 [140]</td>
</tr>
<tr>
<td>1975-76</td>
<td>PortChalmers/1/73 (AF092062)</td>
<td>Victoria/3/75 (ISDNVICT75)</td>
<td>-3 % [140]</td>
<td>B</td>
<td>0.190</td>
<td>0.055</td>
<td>4 [184]</td>
<td>16 [184]</td>
</tr>
<tr>
<td>1984-85</td>
<td>Philippines/2/82 (AF233691)</td>
<td>Mississippi/1/85 (AP008893)</td>
<td>-6 % [142]</td>
<td>B</td>
<td>0.190</td>
<td>0.033</td>
<td>2 [185]</td>
<td>2 [185]</td>
</tr>
<tr>
<td>1985-86</td>
<td>Philippines/2/82 (AF233691)</td>
<td>Mississippi/1/85 (AP008893)</td>
<td>-2 % [145, 186]</td>
<td>B</td>
<td>0.190</td>
<td>0.033</td>
<td>2 [185]</td>
<td>2 [185]</td>
</tr>
<tr>
<td>1987-88</td>
<td>Leningrad/360/86 (AF008903)</td>
<td>Shanghai/11/87 (AP008885)</td>
<td>17 % [143, 145]</td>
<td>B</td>
<td>0.143</td>
<td>0.024</td>
<td>2 [185]</td>
<td>1 [185]</td>
</tr>
<tr>
<td>1989-90</td>
<td>Shanghai/11/87 (AP008885)</td>
<td>England/427/88 (AP204238)</td>
<td>-5 % [143]</td>
<td>A</td>
<td>0.105</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1992-93</td>
<td>Beijing/32/92 (AF008812)</td>
<td>Beijing/32/92 (AF008812)</td>
<td>59 % [146]</td>
<td>0.0</td>
<td>0.0</td>
<td>0 [187]</td>
<td>0 [187]</td>
<td></td>
</tr>
<tr>
<td>1993-94</td>
<td>Beijing/32/92 (AF008812)</td>
<td>Beijing/32/92 (AF008812)</td>
<td>38 % [186]</td>
<td>0.0</td>
<td>0.0</td>
<td>0 [187]</td>
<td>0 [187]</td>
<td></td>
</tr>
<tr>
<td>1994-95</td>
<td>Shanghai/9/93 (Z464417)</td>
<td>Johannesburg/33/94 (AP008774)</td>
<td>25 % [144]</td>
<td>A</td>
<td>0.108</td>
<td>0.021</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1995-96</td>
<td>Johannesburg/33/94 (AP008774)</td>
<td>Johannesburg/33/94 (AP008774)</td>
<td>45 % [147]</td>
<td>0.0</td>
<td>0.0</td>
<td>0 [188, 189]</td>
<td>0 [188, 189]</td>
<td></td>
</tr>
<tr>
<td>1996-97</td>
<td>Nanchang/933/95 (AF008725)</td>
<td>Wuhans/359/95 (AF008722)</td>
<td>28 % [150]</td>
<td>B</td>
<td>0.095</td>
<td>0.006</td>
<td>0 [188, 189]</td>
<td>0 [188, 189]</td>
</tr>
<tr>
<td>1997</td>
<td>Nanchang/933/95 (AF008725)</td>
<td>Wuhans/359/95 (AF008722)</td>
<td>11 % [149]</td>
<td>B</td>
<td>0.095</td>
<td>0.006</td>
<td>0 [188, 189]</td>
<td>0 [188, 189]</td>
</tr>
<tr>
<td>1997-98</td>
<td>Nanchang/933/95 (AF008725)</td>
<td>Sydney/5/97 (AJ311466)</td>
<td>-17 % [146]</td>
<td>B</td>
<td>0.238</td>
<td>0.040</td>
<td>4.5 [189, 190]</td>
<td>27.3 [189, 190]</td>
</tr>
<tr>
<td>1998-99</td>
<td>Sydney/5/97 (AJ311466)</td>
<td>Sydney/5/97 (AJ311466)</td>
<td>34 % [148]</td>
<td>0.0</td>
<td>0.0</td>
<td>0 [189, 191]</td>
<td>0 [189, 191]</td>
<td></td>
</tr>
<tr>
<td>1999-00</td>
<td>Sydney/5/97 (AJ311466)</td>
<td>Sydney/5/97 (AJ311466)</td>
<td>43 % [152]</td>
<td>0.0</td>
<td>0.0</td>
<td>0 [189, 191]</td>
<td>0 [189, 191]</td>
<td></td>
</tr>
<tr>
<td>2001-02</td>
<td>Panama/2007/99 (ISDN1CDA001)</td>
<td>Panama/2007/99 (ISDN1CDA001)</td>
<td>55 % [151]</td>
<td>0.0</td>
<td>0.0</td>
<td>0 [189, 191]</td>
<td>0 [189, 191]</td>
<td></td>
</tr>
<tr>
<td>2003-04</td>
<td>Panama/2007/99 (ISDN1CDA001)</td>
<td>Fujian/411/2002 (ISDN38157)</td>
<td>12 % [153]</td>
<td>B</td>
<td>0.143</td>
<td>0.040</td>
<td>2 [191]</td>
<td>8 [191]</td>
</tr>
<tr>
<td>2004-05</td>
<td>Fujian/411/2002 (ISDN38157)</td>
<td>Wyoming/3/2003 (AY531033)</td>
<td>28 % [159]</td>
<td>B</td>
<td>0.095</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.6 Sensitivity analysis

For network structure, we found that the flight distribution is more sensitive than population distribution and group distribution (Fig. 6.9 a-c). Although, the initial transmission possibility \( r_0 \) and killing rate \( c_{\text{Kill}} \) affect the cumulative attack rate considerably, they seem to have little effect on the epidemic trend (Fig. 6.10 and Fig. 6.11). The immune history capacity seems to have little effect on the epidemic trend as well as the cumulative attack rate (Fig. 6.12).

We also tested the sensitivity to the mutation rate \( R_{\text{mut}} \) and \( c_{\text{Prob}} \). Although both of these factors are found to affect the valley of the epidemic trend curve, they do not affect cumulative rate or the peak (Fig. 6.13).

As the last sensitivity test for Fig. 6.4a, we studied the effect of changing the initial infected group number (IG). IG has a limited effect on cumulative attack rate and almost no effect on the epidemic trend curve (Fig. 6.14). Since fixation rate depends on the
mutation and cProb the most, we did sensitivity tests to these two parameters for Fig. 6.9 (Fig. 6.15). The trend is sensitive to these two factors only in the case of the dominant epitope. For non-dominant epitopes, the curve remains flat for different values of mutation rates and cProb.
Figure 6.9  Sensitivity test for network structure.
Figure 6.10  Sensitivity test for transmission parameters.
Figure 6.11  Sensitivity test for immune killing.
Figure 6.12  Sensitivity test for immune history capacity.
Figure 6.13  Sensitivity test for mutation and diversity parameters.
Figure 6.14 Sensitivity test for initially infected group.
Figure 6.15  Sensitivity test for fixation rate.
References


104. Saloman D. S., Bianco C., Ebert A. D., Khan N. I., M. De Santis, Normanno N.,


126. Sybyl 7.0, Tripos Inc., 1699 South Hanley Rd., St. Louis, Missouri, 63144, USA.


156. Ikonen N., Kinnunen L., Forsten T., and Pyhala R. Recent influenza B viruses in


175. Eubank S., Guclu H., Anil Kumar V. S., Marathe M. V., Srinivasan A., Toroczkai
Z., and Wang N. Modelling disease outbreaks in realistic urban social networks.


182. Influenza virus resource at national center for biotechnology information NCBI.


187. Ellis J. S., Chakravarty P., and Clewly J. P. Genetic and antigenic variation in
the hemagglutinin of recently circulating human influenza A H3N2 viruses in the

188. Information for FDA vaccine advisory panel meeting Atlanta, GA: Centers for
disease control. 1997.

189. Corias M. T., Aguilar J. C., Galiano M., Carlos S., Gregory V., and Lin Y. P.

Rapid molecular analysis of hemagglutinin gene of human influenza A H3N2 viruses

190. Pontoriero A. V., Baumeister E. G., Campos A. M., Savy V. L., Lin Y. P., and


for the vaccines and related biological products advisory committee, CBER, FDA,
WHO collaborating center for surveillance epidemiology and control of influenza.

2003.